ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 21, August 2017

A NOVAL APPROACH TO COMPRESS DNA REPETATIVE SEQUENCES IN BIO-INFORMATICS

N. DEEPAK[1], K. Satyanarayana[2], P. Bharath Siva Varma[3]
[1]Assistant Professor, Department of CSE, Sir C R Reddy College of Engineering, Eluru, Andhra Pradesh, INDIA.
[2]Assistant Professor, Department of IT, Sir C R Reddy College of Engineering, Eluru, Andhra Pradesh, INDIA.
[3]Assistant Professor, Department of CSE, S R K R Engineering College, Bhimavaram, Andhra Pradesh, INDIA.
(Corresponding Author E-Mail: nedunurideepak@gmail.com)

ABSTRACT

Compressing Deoxyribonucleic acid (DNA) sequences is a very important task as on a daily basis thousands of gigabytes of sequences of nucleotides and amino acids gets archived in Genbank. Storing giant Genomes in a laptop computer within the compressed type is an economical means that of victimization Deoxyribonucleic acid sequences for biological functions. Today, a lot of and a lot of Deoxyribonucleic acid sequences are getting out there. The knowledge concerning Deoxyribonucleic acid sequences square measure hold on in biology databases. Need for Compression arises as a result of close to forty five billion bases in four corer living organisms within the GenBank database (http://www.ncbi.nlm.nih.gov/Genbank/) better compression may additionally reveal some inherent biological structures, abet in phylogenic tree reconstruction. To forecast the future needs the simulation results of different living organism's records to be hold on. For a four-letter alphabet in DNA (Adenine(A), Cytosine(C), Guanine(G) and Thymine(T)), an average description length of 2 bits per base is that the max length requiredto encode DNA. An attempt is made to review the previous art of compression techniques and its merits and de merits. Based on the comparative study of existing algorithms a new method proposed for DNA compression without depending on statistics of sequence set.

KEY WORDS: DNA, GenBank, Phylogenetic Tree, Genomes.

1. INTRODUCTION

The amount of human genomic data available in public data bases is increasing exponentially [1] in day to day life. As a result, manipulation of genomic data (DNA or protein) also becomes more difficult. DNA sequences contain four types of bases like adenine (A), cytosine (C), Guanine (G) and Thymine (T) .Currently DNA sequences are usually stored in text using 8 bits per character.

Without compression two bits required to encode each base by information theory [2].Even in existing general compression tool like gzip are used, it has been found that more than two bits required to encode each base and some of them are extended to negative rates [3].Thus, it became an essential need to compress DNA sequences by developing specific compression algorithms.

1.1 DNA

DNA contains the heredity information can be passed from one progeny to another which be used in the development and functioning of living organisms. DNA consists of two long polymers of backbone units referred to as nucleotides, with backbones product of sugars and phosphate teams joined by organic compound bonds (Fig.1). These 2 strands run in opposite directions to every alternative and area unit so anti-parallel. Connected to every sugar is one in every of four kinds of molecules referred to as nucleobases (informally, bases). It is that the sequence of those four bases on the backbone that encodes data. In each cell, DNA is organized into long structures called chromosomes, for ex the human genome contains 23 chromosome pairs. In DNA replication chromosomes are duplicated before cell division.

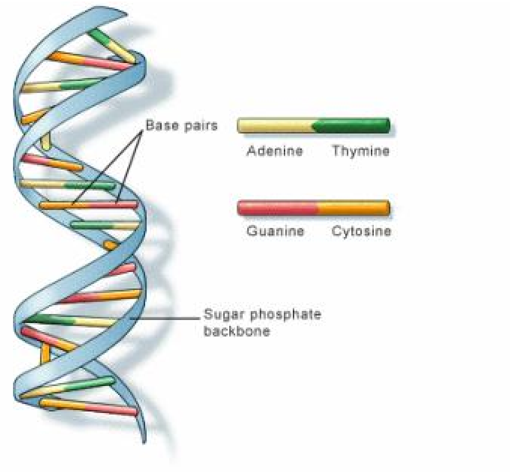


Figure 1: Helical Structure of DNA

1.2 CHARACTERISTICS OF DNA SEQUENCES

DNA (Deoxyribonucleic acid) contains genetic information carried from one generation to next generation.DNA fragments consisting of four nucleotides: Adenine, Cytosine, Thymine and Guanine (A, C,G and T) [4],as shown in Table 1. The pair of nucleotides (A, T) and (C, G)are arranged as opposite pair in the DNA structure, as shown in Figure 2. Due to its opposite bonding, if one strand needs to be encoded while another strand can be easily decoded.

Tab1e 1: Fourtypes ofN ucleotides, Adenine(A), Guanine(G),Thymine(T)and Cytosine(C),andtheircomplements.

Bases	Nucleotides	Complement
Adenine	A	T
Cytosine	C	G

Gunine	G	C
Thymine	T	A

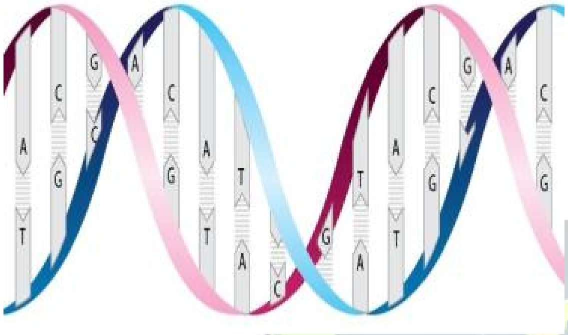


Figure 2: DNA chain with complement pairs A<->T and C<->G

1.3 APPLICATIONS OF BIO INFORMATICS

The following section describes some of the available Bio informatics applications.

- **3-D protein prediction:** The 3-D structure of protein [49] determines a detailed analysis and regular patterns of a macroscopic structure called "secondary structure" .Therefore, by the prediction of the secondary structure of a protein helps researchers to predict 3-D structure and its function.

- **Phylogenetics:** The study of phylogeny is called phylogenetics, which is the study of evolutionary development and history of species. The outcome of Phylogenetics is evolutionary trees which explain the evolutionary distance between the set of sequences, if the sequence are from different organisms then evolutionary relationships can be found among the organisms.

2. LITERATURE REVIEW

2.1 INTRODUCTION TO DATA COMPRESSION

The world has a changed a lot from an era where a computer is occupied a room to the present where super computers can conveniently carried personally. This transformation as a technological growth; rather, it should be termed as technological explosion. This transformation has certainly occurred as part of wonderful contributions by many eminent personalities the world over. In this context, the period in history which marked the advent of data compression has got a remarkable role to play in this aspect. The massive world of internet is extensively using data compression techniques in innumerable ways, without which the dreams of web technology booms would never have been possible.

2.2 SIMILARITIES OF SUB SEQUENCES AMONG CHROMOSOMES

State of the art DNA compression methods are based on searching repetitions, better compression ratios can be achieved by similar sub sequences along the current DNA sequence. Similarities of sub sequences could exist among different species that are close in terms of evolutionary distance [4] or different chromosome species of one species [6]. Finally, similarities of sub sequences among chromosomes are classified into two ways: one is self reference and other is cross reference similarities. Self and cross reference similarities are verified on a sample data set s.cerevisiae.

2.3 EXISTING COMPRESSION METHODS FOR DNA SEQUENCES

Many classical algorithms are fails to compress DNA sequences due to the encoded specialty of text [7].DNA compression includes many technologies, metrics and computational algorithms, presents the following list.

- Probabilistic Markov models[8]
- Burrows Wheeler Transform[9]
- Greedy algorithms[10]
- Dynamic programming approaches
- Based on the Normalized maximum Likelihood[11]

2.4 ABOUT S.CEREVISIAE DATA SET

The S.cerevisiae is a huge data set consists of 16 chromosome sequences which can be downloaded freely from the Web ftp://ftp.ncbi.nih.gov/genomes/. In S.cerevisiae, the sixteen chromosome sequences are denoted as chromosome I to chromosome XVI shown in

Table 2: The longest and shortest chromosome

Chr	Reference	No.ofbases
I	GI:50593113	143157
II	GI:50593115	605184
III	GI:42759850	217332
IV	GI:50593138	1129605
V	GI:7276232	391086
VI	GI:42742172	183702
VII	GI:50593213	784707
VIII	GI:50882583	402792
IX	GI:6322016	310041
X	GI:42742252	557103
XI	GI:50593424	478620
XII	GI:42742286	784695
XIII	GI:44829554	693291
XIV	GI:50593505	576585
XV	GI:42742309	785568
XVI	GI:50593503	687666
		Avg:545696

3. METHODS

3.1 DATA EXTRACTION

The open source data sets S.cerevisiae and S.pombe can be used as an input data for the dReaM algorithm.
The Saccharomyces genome database (SGD) of Yeast (bacteria) directory contains sequence files of the sixteen nuclear chromosomes and the mitochondrial order of S. cerevisiae strain S288C, in FASTA format. Individual chromosome sequence files are updated once there's an update to the systematic reference sequence stored in SGD.In S.Cerevisiae contains 16 chromosome sequences names as ChrI to ChrXVI.The longest and the shortest chromosomes are ChrIV and ChrI according to their sizes. [5] proposed a method in which the minimization is per-formed in a sequential manner by the fusion move algorithm that uses the QPBO min-cut algorithm. Multi-shape GCs are proven to be more beneficial than single-shape GCs. Hence, the segmentation methods are validated by calculating statistical measures. The false positive (FP) is reduced and sensitivity and specificity improved by multiple MTANN.

3.2 THE dReaM ALGORITHM

In earlier, differencing has its origins in each longest common subsequence (LCS) algorithms [12] and also the string-to-string correction drawback. Miller and Myers [13] established the constraints of LCS after they made a brand new file compare program that executes at fourfold the speed of the diff program whereas manufacturing considerably smaller variations.

The edit distance [14] proved to be an improved metric for the distinction of files and techniques supported this technique increased the utility and speed of file differencing.

Compression –based distance measures (CBMs) that rely on probabilistic mismatching [15] with neglect the locations of variations, aren't distinct enough among completely different categories. As an example, 3 sequences construct table listing all Variance (Variance matrix) because the following.

P1=Pattern 1: ACGTGTAC

P2=Pattern 2: CGATGCAT

P3=Pattern 3: TCGAACGT

Table 3: Variance Matrix(R* -Repeats-self similarity-Vice-Versa)

S.no	P1	P2	P3
P1	-	*5*	*6*
P2	*R**	-	*6*
P3	*R**	*R**	-

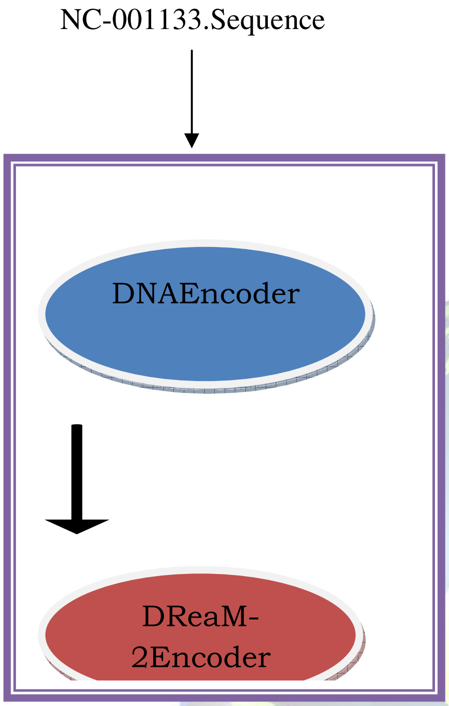
The proposed algorithm dReaM (DNA repetitive and non repetitive encoding analysis model) developing of distinction algorithmic rule particularly for DNA information sets.

The basis of our algorithmic rule is that the economical storage of knowledge that's identical or close to a dead ringer for input “reference” sequences. The key feature of our algorithmic rule is that some sequences similar to the reference have minimal impact on storage despite their length or depth of sequencing coverage. To boot, we tend to use the reference ordering strictly as a compression framework, and don't need any biological correctness for the reference. Our strategy is to store information set from a spread of “similar” sequences, e.g., those exploitation identical individual or species. These sequences then function as secondary compression framework upon that we are able to offer economical storage. This can be similar however not similar to the task of finding the distinction sequences and their locations from whole information set.

3.3 The DNACoder

The first stage of dReaM is DNACoder which is one of the important algorithmic components of the proposed compression algorithm. The main function of this component is to identify the similarities among DNA sequences and then encode them to achieve better compression ratios. The DNACoder have two sub components Encoder and Decoder for compression and Decompression processes.

The compression process as shown in Figure 4.1.This process will include DNACoder and dReaM-2 Encoder. The sequence such as NC-001133(shown in Figure4.2) is applied as input into the DNACoder and the output is bypassed to dReaM-2 encoder to make compressed file. Obviously the compressed one is having lesser in size than that of the original one.

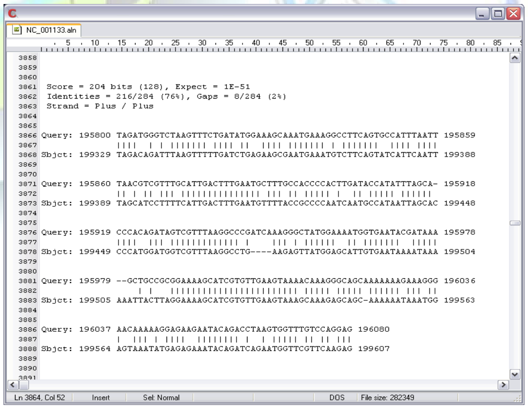


The first sub component of DNACoder is DNAEncoder and it will consist of the following steps.

- Retrieval of similar subsequences among different DNA sequences.
- Similar subsequences can be sorted out according to their priority.
- Removal of redundancy from different DNA sequences.
- Sort out the non-overlapping subsequences according to their index position.
- Residual sequences after removal of redundancy, bypassed to dReam-2 Coder for further compression.

In the first step mentioned above, similar subsequence can be extracted by using pattern Hunter. Similar subsequences may contain approximate repeats and reverse complements leads to redundancy. In addition to that, similarities may found in self and cross chromosomal sequences. The results of two similar chromosome sequences can be stored in variance table which will contains two files i.e. (self-similar.aln and cross-reference.aln) shows variation point (which shows variations) among chromosomes.

In every .aln file, info concerning similar subsequences together with the scores, the direction, the beginning and also the ending positions of query and subject Sequences are recorded. An example record is shown in Figure 3 The query and the subject sequences see the similar subsequences within the current sequence in the case of self-referencing (i.e., self-cross-reference.aln). They see the reference sequence and the current sequence severally within the case of cross-referencing (i.e., cross-reference.aln). The scores relate to the similarity between the query and also the subject sequences. High score indicates that they're just like one another. The direction are often either 'plus' or 'minus'.



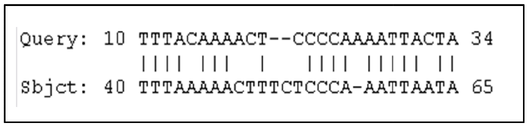
```
 Score = 204 bits (128), Expect = 1E-51
 Identities = 216/284 (76%), Gaps = 8/284 (2%)
 Strand = Plus / Plus


Query: 195800 TAGATGGGTCTAAGTTTCTGATATGGAAAGCAAATGAAAGGCCTTCAGTGCCATTTAATT 195859
              ||||  | | ||||||| |||| ||  |||| ||||||| | |||||||  |||| ||||
Sbjct: 199329 TAGACAGATTTAAGTTTTTGATCTGAGAAGCGAATGAAATGTCTTCAGTATCATTCAATT 199388


Query: 195860 TAACGTCGTTTGCATTGACTTTGAATGCTTTGCCACCCCACTTGATACCATATTTAGCA- 195918
              || | || ||| ||||||||||||||| ||| || ||||| |  || ||||| ||||||
Sbjct: 199389 TAGCATCCTTTTCATTGACTTTGAATGTTTTACCGCCCCAATCAATGCCATAATTAGCAC 199448


Query: 195919 CCCACAGATAGTCGTTTAAGGCCCGATCAAAGGGCTATGGAAAATGGTGAATACGATAAA 195978
              ||||  ||| ||||||||||||| |    ||| | ||||||  || |||||||  |||||
Sbjct: 199449 CCCATGGATGGTCGTTTAAGGCCTG----AAGAGTTATGGAGCATTGTGAATAAAATAAA 199504


Query: 195979 --GCTGCCGCGGAAAAGCATCGTGTTGAAGTAAAACAAAGGGCAGCAAAAAAAGAAAGGG 196036
                  | |   |||||||||||||||||||||||| ||||| ||||| |||||| ||| ||
Sbjct: 199505 AAATTACTTAGGAAAAGCATCGTGTTGAAGTAAAGCAAAGAGCAGC-AAAAAATAAATGG 199563


Query: 196037 AACAAAAAGGAGAAGAATACAGACCTAAGTGGTTTGTCCAGGAG 196080
              |  ||| | ||||  |||||||| |  | ||||| || || |||
Sbjct: 199564 AGTAAATATGAGAGAAATACAGATCAGAATGGTTCGTTCAAGAG 199607
```

Figure 3: NC-001133.aln file

The 'plus' direction suggests that an approximate repeat, that the sequence ought to be browse in ascending order.

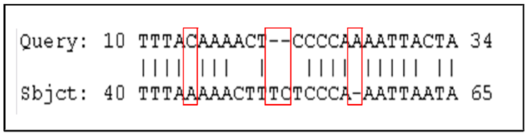
The 'minus' direction suggests that a reverse complement repeat, therefore the sequence ought to be browse in reverse direction. The beginning and ending position of the subsequence marks its location within the original sequence. A listing is made for storing information concerning repetitive records with score over a threshold. The edge is set in order that solely vital similar subsequences square measure thought of by our projected compression algorithm.

In the third step, every repetitive record within the combined list from the second step is examined. Specifically, overlapping similar subsequences from two records are cut. If the topic sequences in 2 repetitive records are overlapped with one another in position, the overlapping half are going to be unbroken within the record with the next score and be removed within the different record with a lower score. The principle behind is to stay a protracted repetitive length instead of a brief one. If the length of the cut repetitive record is a smaller amount than the edge once removing the overlapping part, the repetitive record are going to be off from the list. After removing all the overlapping elements within the similar subsequences, DNACoder saves the variations between the 2 similar subsequences. This step is crucial as approximate repeats, instead of precise repeats, are thought of in DNACoder.



```
Query: 10 TTTACAAAACT--CCCCAAAATTACTA 34
          |||| |||   |   |||| ||||| ||
Sbjct: 40 TTTAAAAACTTTCTCCCA-AATTAATA 65
```

Consider an easy example shown higher than. The query subsequence from the tenth base to the thirty fourth base and also the subject subsequence from the fortieth base to the sixty fifth base are found to be an approximate Repeat. This instance includes all the operations - substitution, deletion and insertion. For example, the base "C" within the fifth position of the query subsequence is replaced by "A" within the corresponding position of the subject subsequence. There square measure 2 a lot of bases extra within the subject subsequence in between the twelve and also the 13 bases of the question subsequence. The ordinal base "A" within the query subsequence is deleted in matching to the subject subsequence.



```
Query: 10 TTTACAAAACT--CCCCAAAATTACTA 34
          |||| |||   |   |||| ||||| ||
Sbjct: 40 TTTAAAAACTTTCTCCCA-AATTAATA 65
```

4. RESULTS

4.1 SIMULATION RESULTS ON S.CEREVISIAE-SGD

The Saccharomyces Cerevisiae Genome Database (S.cerevisiae-SGD) can be considered in this experiment to assess the effectiveness of the compression algorithm. The S.Cerevisiae is an open source repository can be downloadable from the web:- ftp://ftp.ncbi.nlm.nih.gov.in/genomes/

Different representations of SGD can be shown in Figure 4 and Figure 5.

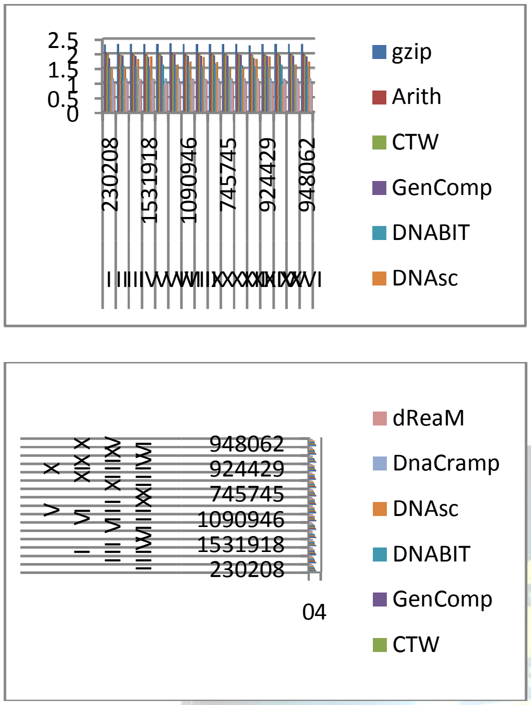


Figure 4 and 5: Bar Chart(s) Representation of SGD

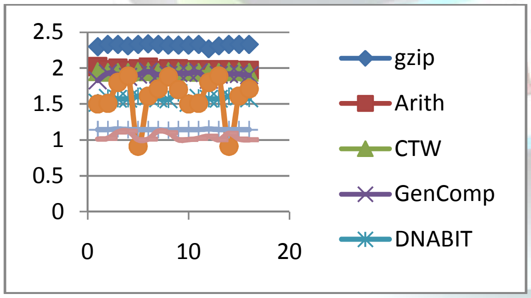


Figure 6: Area Chart Representation of SGD

4.2 Simulation Results on S.Pombe-PBD

In this experiment, the tendency to think about testing our proposed algorithm dReaM on another real dataset. The 3 chromosome sequences of schizosaccharomyces pombe (S. pombe) are tested. S. pombe may be a species of yeast. It's usually used as a model organism in molecular and cell biology. Note that the common length of S. pombe is 4200k, which is considerably longer than that of S. cerevisiae. Thus, the long length would increase the compression time significantly because it would take for much longer time to go looking for the repetitive records within the sequences.

5. CONCLUSION

The simulation study has been performed between the chromosome sequences of Saccharomyces Cerevisiae Genome Database (S.cerevisiae-SGD) and schizosaccharomyces pombe (S. pombe).In this thesis, for the first time an attempt is made to recognize self and cross sequence similarities and make it use in different real time applications. A detailed analysis has been performed in between SGD and PBD to find out the variations in between the sequences, length and locations by constructing variance table. From this analysis it is observed that the cross similarities have much impact than that of self similarities in SGD and PBD. To till date most of the methods consider self references only but the present study shows to attain higher compression ratios it would be more advantages if cross similarities can be considered.

6. REFERENCES

[1] B.A.Gata, "DatabaseSimilaritySearchingUsingBLASTandFastA",Australasian Biotechnology,vol.5,pp.282-290,1995.

[2] U.OzkanNalbantoglu, J.DavidRussellandKhalidSayood," Data CompressionConceptsandAlgorith

msandTheirApplicationstoBioinformatics", OPEN ACCESS,Entropy,vol.12,pp.34-52;2010.

[3] T.Matsumoto,K.Sadakane,H. ImaiandT.Okazaki,"CanGeneral-Purpose Compression Schemes ReallyCompress DNASequences?", Currents in Computational MolecularBiology,pp.76-77,2000.

[4] OpenBioinformaticsFoundation:BOSC",Officialwebsite.Open Bioinformatics Foundation,Retrieved10 May2011.

[5] Christo Ananth, G.Gayathri, M.Majitha Barvin, N.Juki Parsana, M.Parvin Banu, "Image Segmentation by Multi-shape GC-OAAM", American Journal of Sustainable Cities and Society (AJSCS), Vol. 1, Issue 3, January 2014, pp 274-280

[6] Hari Prasad,P.V.Kumar Efficient compression through CBT,vol.4,no.2,2102 CIIT proceedings.

[7] Hari Prasad , Kumar , A new revisited DNACRAMP tool based approach of chopping DNA repetitive and non repetitive sequences of genomes in vol.9,issue 6,November 2012 IJCSI proceedings.

[8] P.J.S.G.Ferreira,A. J.R.Neves,V.Afreixo,andA.J.Pinho,"ExploringThree-BasePeriodicityforDNACompressionandModeling,"presentedatICASSP 2006,vol.5,Toulouse,France,2006

[9] D.Adjeroh,Y.Zhang,A.Mukherjee,M.Powell,andT.Bell,"DNASequence CompressionUsingtheBurrows-Wheeler Transform,"presentedatIEEE ComputerSocietyBioinformaticsConference(CSB'02),2002.

[10] A.ApostolicoandS.Leonardi,"Compression ofbiologicalsequencesbygreedy off-linetextualsubstitution,"presentedat DataCompressionConference,pp.143- 152,2000.

[11] G.KorodiandI.Tabus,"Anefficientnormalized maximum likelihoodalgorithm forDNAsequenceco

[12] Stomg J.A.Storer,andT.G.Szymanski,"Datacompressionviatextualsubstitution,"J. ACM,29,928–951, 1982.

[13] R. WAGNER, andM.FISCHER, "The string-to-string correction problem", JournaloftheACM21,1,pp.168–173,Jan.1973.

[14] M. Li,JH. Badger,X.Chen,S.Kwong,P.Kearney,HY.Zhang,"Aninformation-basedsequencedistanceanditsapplication towholemitochondrial genome phylogeny",Bioinformatics,17:149-154,2001.

[15] Stomg J.A.Storer,andT.G.Szymanski,"Datacompressionviatextualsubstitution,"J. ACM,29,928–951, 1982.