Lossless compression of repetitive and non repetitive DNA sequences using A Novel first Art methodology

V.Hari prasad
Research Scholar in CSE
Jawaharlal Nehru Technological University (JNTUK),
Andhra Pradesh
hariprasadvemulapati@gmail.com

Dr.P.V.Kumar
Professor of CSE
Osmania University
Hyderabad, Andhra Pradesh
Pvkumar58@gmail.com

Abstract— Excess accumulation of living organisms in a daily routine Genbank sizes are growing exponentially. Now, a challenge to the real world is that accessing and transmitting t the genetic data through network is a cumbersome task. Hence, a feasible solution is compression!.State of the art existing compression algorithms may work on repetitiveness of the sequences and the results are not bountiful in a comparative study. Here a new methodology is proposed for better compression gain.

Keywords— *encoding; decoding; bio compress; Huff bit compress; dnabit compress; LSBD compression*

I. INTRODUCTION

Bio informatics is one of the emerging fields in computer science includes processing and maintenance of biological databases. This is the one of the active area of research which will more helpful in different areas of specialty, including (but in no means limited to) statistics, computer science, physics, biochemistry, genetics, molecular biology and mathematics Computational Biology is the mathematical and algorithmic study of bio informatics allied areas like DNA computing, protein docking and visualization protein information etc.Bio informatics and computational biology are two multidisciplinary fields typically refers to the field concerned with the collection and storage of biological information, where as computational biology refers to the aspect of developing algorithms and statistical models necessary to analyze biological data through the aid of computers.

II. MOTIVATION

Life is strongly associated with organization and structure **[1]**.With the completion of 1000 genomes project, the project is estimated to generate about 8.2 billion bases per day, with the total sequence to exceed 6 trillion Nucleotide bases. The DNA molecule is made up of a concatenation of four different kinds of nucleotides namely: Adenine, Thymine, cytosine and Guanine (A,T,C,G).Today, more and more DNA sequences are available, due to the excessive surge of genomes storage databases size is two or three times bigger annually. Thus, it becomes very hard to download and process the data in intra and internetworking systems. To maintain it compression is came into the existence .compression can performed in two ways either Loss or Loss- less. Lossy compression is applicable for images because if we remove unnecessary pixels also image doesn't violates its property. But sequences like DNA and RNA encoded information in textual format. So Lossy compression is not advisable to compress such sequences. Text compression is always Loss-less because we have to retain its original property after decoding.

Universal compression algorithms are fails to compress genetic sequences due to specificity of 'text'. Some standard algorithms are worked on it and achieved negative compression rates. General purpose compression algorithms do not perform well with biological sequences. Giancarlo *et al*. **[2]** have provided a review of compression algorithms designed for biological sequences. Finding the characteristics and comparing Genomes is a major task (Koonin 1999**[3]**; Wooley 1999**[4]**). In mathematical point of view, compression implies understanding and comprehension (Li and Vitanyi 1998) **[5]**. Compression is a great tool for Genome comparison and for studying various properties of Genomes. DNA sequences, which encode life should be compressible. It is well known that DNA sequences in higher eukaryotes contain many tandem repeats, and essentials genes (like rRNAs) have many copies. It is also proved that genes duplicate themselves sometimes for evolutionary purposes. All these facts conclude that DNA sequences should be compressible. The compression of DNA sequences is not an easy task. (Grumback and Tahi 1994**[6]**, Rivals *et al*. 1995 **[7]**; Chen *et al*.

2000 **[8]**) DNA sequences consists of only four nucleotides bases {a,c,g,t}. Two bits are enough to store each base. The standard compression software's such as "compress", "gzip", "bzip2", "winzip" expanded the DNA genome file more than compressing it.

Most of the Existing software tools worked well for English text compression (Bell *et al*. 1990**[9]**) but not for DNA Genomes. There are many text compression algorithms available having quite a good compression ratio. But they have not been proved well for compressing DNA sequences as the algorithm does not incorporate the characteristics of DNA sequences even though DNA sequences can be represented in simple text form

III. BASIC KNOWLEDGE OF GENOME DATA

3.1 DNA Characteristics

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),as shown in Table 3.1.These nucleotide literals are arranged in a double helix format tied with two hydrogen bonds. The pair of nucleotides (A,T) and (C,G)are arranged as opposite pair in the DNA structure, as shown in Figure 3.2.Due to its opposite bonding, if one strand needs to be encoded while another strand can be easily decoded.

Figure 3.1.DNA chain with complement pairs A<->T and C<->G

Bases	Nucleotides	Complement
Adenine	**A**	**T**
Cytosine	**C**	**G**
Gunine	**G**	**C**
Thymine	**T**	**A**

Tab1e2.2. Four types of nucleotides, Adenine (A), Guanine (G), Thymine (T) and Cytosine (C), and their complements.

DNA sequences are not random in nature; in fact it will contain long term repetitions in which sub sequences are similar to each other. The long term repetitions may contain approximate repeats and complementary palindromes. Based on these similarities different compression algorithms are exploited in the literature.

2.1.1 Approximate Repeats

Normally DNA repeats include exact matches and approximate matches. An exact match may found if two sub sequences consists of identical nucleotides along the whole subsequence. An approximate match may found in the sub sequence by performing different operations are illustrated in Figure 3.2

Figure3.2 Approximate Repeats. The sequence is a part of a DNA

A-Substitution	B-Addition	C-Deletion
AGTACGGTACGA	AGTACGGTACGA	AGTACGGTACGA
1:AGTACG :6	1:AGTA-G :6	1:AGTACG :6
7:AGTAGG :12	7:AGTAGG :12	7:AGTA-G :12

In Figure 3.2(A), the DNA sequence is “AGTACGGTACGA". The first six nucleotides are identical to the remaining six nucleotides, if the 5th base of first sub sequence “C” is substituted by “G”, the second sub sequence can be obtained from the first one. In a similar manner in Figure3.2 (B) an addition can be performed by inserting “G” in the first sub sequence to obtain the second one and in Figure3.2(C) deletion can be performed in the first one to obtain the second one i.e. by deleting the 5th base “C” from first sub sequence. These operations can be performed to make the sub sequence identical to some other sequence.

2.1.2 Reverse complements

The reverse complement is also referred to in the literature as complementary palindromes, reverse repeats and inverted complemented repeat. A Reverse Repeat is said to be complementary palindrome if each nucleotide in the first sub sequence is replaced by its complement in the second sub sequence. In Figure2.4, “AGTACG CGTACT” is a sub-part of DNA sequence, the sub sequence “CGTACT” formed from the last six bases, the complement of the six bases is “GCATGA” and its reverse ordering is “AGTACG”, which is the same as the first six bases of original DNA sequence.

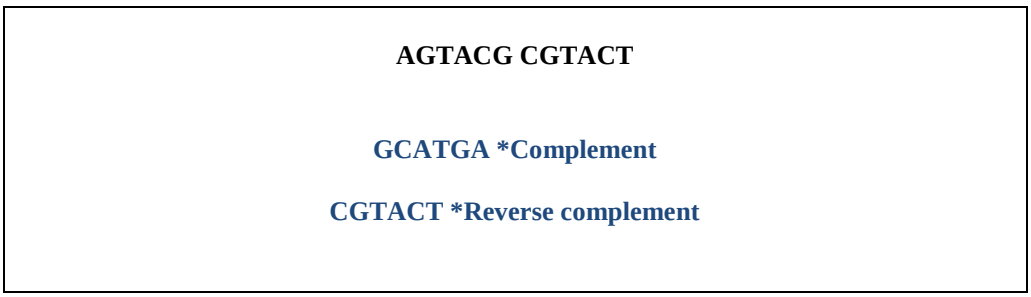

Figure2.4 An example of a Complementary palindrome which is a part of DNA sequence

A. Work flow of the paper

This paper is organized as follows. Section 4 describes general compression algorithms. Section 5 describes cognate subsisting algorithms to compress genome data. Section 6 describes proposed algorithms analysis how it is more preponderant one than subsisting techniques. Section 7 describes comparative study on a sample sequence. Section 8 is concluding with future work.

IV. GENERAL COMPRESSION ALGORITHMS

State of the art general compression algorithms are worked based on the properties mentioned below.

- Finding the candidate reiterate segments.
- Considering approximate reiterates.
- Encoding of the reiterate segments.
- Encoding of the non-reiterate segments.

V. RELATED EXISTING ALGORITHMS

We can encode every base of DNA by two bits. Compression method mainly categorized into two ways one statistical and other is substitution. In statistical method longer stream are replaced by shorter code and other is dictionary based mechanism. The existing algorithm based on two bits encoding schemes like A (00), C(01), G(10) nd T(11). HUFFBIT[13],GENBIT[14], and DNABIT[15] algorithms are evaluated in Best,Avg and Worst case analysis based on fragments repetitions in the sequences .Suppose in the given sequence more fragments are repeated they achieve Best case if not worst case In this connection our existed techniques achieve 2.25 bits/Bases. But Sequences like AT-rich DNA, which constitutes a distinct fraction of the cellular DNA of the archaebacterium Methanococcus voltae, consist of non-repetitive sequences .So existed algorithms may run in worst case and we will achieve less compression rates. Our proposed algorithms DNASP (DNA Sequence pack) are well suitable for non repetitive DNA sequences and we achieve more compression rates than existing algorithms.

VI. PROPOSED METHODOLOGY AND ITS ANALYSIS

The sample DNA sequence (FASTA format) can be applied as an input to the proposed new utility dReaM.The proposed new implement will work on the principle of Divide and conquer strategy. The DNA sequence is sub divided into L/16

fragments (fragment contains four bases in any combination (A, C, G, T)).In the stage1 process, group the quadrupled fragments into F_{pt} and $S_{pt.}$ The indexes of these two can be stored separately and then group F_{pt} and S_{pt} into C_{st}. Now the C_{st} can contain L/32 fragment bases. To get the total number of encoded bits sub sequent sub-clusters grouped into R_{cv}. Here after, measure the storage space needed to encode R_{sv} (in terms of bytes).

L=Length of the DNA sequence F_{pt}=First sequence pattern

S_{pt}=Second Sequence Pattern F_{pv}=First pattern Value

S_{pv}=Second pattern Value C_{st}=Cluster Set

C_{stv}=Cluster Set Value Rcv=Regional Cluster Value

T_{neb}=Total number of encoded bits Zr=Compression Ratio

F_{pt} and S_{pt} can be measured as follows.

$$F_{pt}=f_{p1}+f_{p2}+f_{p3}+\ldots..+f_{pn}$$

$$S_{pt}=S_{p1}+S_{p2}+S_{p3}+\ldots+S_{pn}$$

Here, F_{pt} and S_{pt}(contains quadrupled fragment bases) can be stored in two separate array indexes. Here, the array index will represent binary equivalent information. Now measure the binary equivalent numeric of Fpt and Spt i.e. FpV and Spv and then group them into C_{st}. The values F_{pv}, S_{pV}, C_{stv} and R_{cv} can be measured as follows.

$$F_{pv} = \sum_{t=0}^{L/16} F_{pt}$$

$$S_{pv} = \sum_{t=0}^{L/16} S_{pt}$$

The Cluster Set C_{st} can be measured as follows.

$$C_{St} = Cs_1 + Cs_2 + Cs_3 + ... + Cs_n$$

$$Cs_1 = (\mathrm{F}_{pv} + S_{pv}) / 2$$

$$Cst_v = \sum_{s=0}^{m} Cst$$

$$R_{cv} = Csv1 + Csv_2 + Csv_3 + ... + Csv_n$$

$$Rc_v = \sum_{v=0}^{m} Cst_v$$

Here, Rcv will represent memory storage. The proposed model is java based implement so that to store every numeric array index will require four bytes of storage.

The total number of encoded bits to encode the DNA sequence can be measured as follows.

$$Tn_{eb} = \sum_{v=0}^{L} (\mathrm{Rc}\, v)$$

Finally, the compression ration can be calculated as follows

$$Zr = (Tneb / \mathrm{L})$$

Let us take the sample sequence1 (which is the part of S.cerevisiae)

4.1 Analysis of the proposed methodology

Sequence1:-

ACGT GCGC GATC GCCT GCTA GGCG TACG TCGC AGGC GATC GATG TGCT AGAT CAGA TGAC TCAG TGCA CGAT CGAG TGCA GCCT GACT TACG CGAT .

The above sequence contains 96 bases which is scattered into 3 cluster sets, 6 patterns and 1 Regional Set as per earlier explanation. The values F_{pv}, S_{pV}, C_{stv} and R_{cv} can be measured as follows.

$$C_{st}=C_{s1}+C_{s2}+C_{s3}$$

Now the first sub cluster will store L/32 bases and its binary equivalent numeric value can be stored in separate array index. To represent this array index 4 bytes will require in java based system and then sub sequent cluster values can be grouped into Regional Cluster as per the explanation.

So Total number of Encoded bits to encode the above sequence will be

$$T_{neb} = C_{st} = (\mathrm{Cs}_1 + Cs_2 + Cs_3)$$
$$Tneb = (4+4+4=12)Bytes(96bits)$$

Finally, the Compression Ratio (Zr) can be calculated as

Zr=Total number of Encoded bits / Total Number of Bases

$$Z_r = T_{neb} / L$$
$$=96 /96=1.002bpb$$

Hence, 1.002 bits will require encoding each base from the above analysis. The compression Ratio around of 87% and storage space reduced from one byte to one bit nearly.

Huffbit, GenBit and Dna compress =204 bits (2.428)
Genbit Compress (Tool based) = 202 bits (2.404)
DNASC Compress = 128 bits (1.523)
Splinted Binary compression (SBC) =96bits (1.142)
A novel based approach =96 bits(1.012)

4.1.1 The dReaM- Encoding Algorithm

- INPS : Input Sequence
- O/P: Out Put(Encoded)
- Encoding Procedure Begin
- Begin
- Divide the input sequence into equivalent fragment bases, where each fragment contains four bases(Adenine-A,Cytosine-C,Guanine-G and Thymine-T)
- All the possibilities can be generated for non repetitive sequences
- Group quadrupled fragments into patters(Divide and Conquer)
- Substitute the binary digits (0 and 1) for DNA fragments
- Adenine=00, Cytosine=01, Guanine=10 and Thymine=11
- Substitute all the combinations of 00, 01, 10 And 11 in non repetitive DNA sequence.

- Store the Encoded value for every pattern in any data structure and calculate the number of bytes required for it.
- Return the number of storage bytes to encode all the patterns, called as total number of encoded bits
- Repeat the process of substitution of binary digits till the end of the input sequence i.e. INPS.
- Move the total number of encoded bit sequence to the Output file.
- End

4.1.2 The dReaM- Decoding Algorithm

Due to the lossless property of DNA, the decoding is carried out in the reverse process of encoding

- INPS:- Input Sequence
- O/p:-Output Sequence(Decoded)
- Decoding Procedure Begin
- Begin
- All the possible combinations of (A,C,G and T) can be generated.
- Read the data from the Output file and assign the base for binary digits in a reverse order like 00=Adenine (A), 01=Cytosine(C), 10=Guanine (G) and 11=Thymine (T).
- Group all the bases into different patterns.
- Group the entire patterns into Decoded sequence in the reverse order.
- Repeat the above step until the retention of input sequence.
- Finally. Transfer Decoded bits to the input sequence for validity.
- End

VII. CONCLUSIONS AND FUTURE WORK

The proposed new implement can encode every base in 1.002 bpb which will save in and around of 8 bits per base. so finally concluding that the proposed new implement is a first art methodology and which can be extended to any tool based mechanism.

References

[1] *E Schrodinger. Cambridge University Press: Cambridge, UK, 1944.[PMID: 15985324]*
[2] *R Giancarlo et al. A synopsis Bioinformatics 25:1575 (2009) [PMID:19251772]*
[3] *EV Koonin. Bioinformatics 15: 265 (1999)*
[4] *JC Wooley. J.Comput.Biol 6: 459 (1999) [PMID: 10582579]*
[5] *CH Bennett et al. IEEE Trans.Inform.Theory 44: 4 (1998)*
[6] *S Grumbach & F Tahi. Journal of Information Processing and Management 30(6): 875 (1994)*
[7] *E Rivals et al. A guaranteed compression scheme for repetitive DNA sequences. LIFL, Lille I University, technical report IT-285 (1995)*
[8] *X Chen et al. A compression algorithm for DNA sequences and its applications in Genome comparison. In Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, Tokyo, Japan,*

April 8-11, 2000. [PMID: 11072342]
[9] TC Bell et al. Newyork:Prentice Hall (1990)
[10] J Ziv & A Lempel. IEEE Trans. Inf. Theory 23: 337 (1977)
[11] A Grumbach & F Tahi. In Proceedings of the IEEE Data
[12] X Chen et al. In Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, Tokyo, Japan, April 8-11, 2000.
*[13] X Chen et al. Bioinformatics **18**: 1696 (2002) [PMID: 12490460]*
[14] An Efficient Horizontal and Vertical Method for Online DNA Sequence Compression in IJCA proceedings 2010 vol.3,Issue 1 June 2010.
[15] Allam AppaRao.In proceedings of the Bio medical Informatics Journal [2011].DNABIT compress-compression of DNA sequences
[16] Loss less segment based compression in IEEE confernece proceedings in ICECT-2011 kanyakumari,India.
[17] Srinivasa K G,Jagadish M, Venugopal K R and L M Patnaik "Efficient compression of non repetitive DNA sequances using Dynamic programming " pages 569-574 IEEE 2006
[18] National Center for Biotechnology Information, Entrez Nucleotide Query, http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=n s.
[19] Allam AppaRao in proceedings of the JATIT journal computational biology and Bio informatics:[2011].Huffbit compression of DNA sequances
[20] Allam AppaRao in proceedings of the JATIT journal of computational Biology[2011],Genbit compress fro DNA sequances.

V Hari Prasad , B.Tech CSE from JNTU University,Anantapur,M.Tech CSE from JNTUCEH,HYD and pursuing research in CSE at JNTU KAKINADA, A.P as a Research scholar in CSE .He has 10 years of teaching experience in various Engineering colleges. Presently He is working as Lecturer in computer Engg in Govt Polytechnic proddatur A.P. He is a Life Member of MISTE and Member of IEEE.He presented papers at International & National conferences on various domains. His interested areas are Bio Informatics, Databases, and Artificial Intelligence.

Dr.P.V Kumar ,Professor of CSE in Osmania University Hyd,Completed M.Tech CSE from Osmania university and PhD (CSE) welding from Osmania university. He has 30 years of Teaching & R&D experience. Many students are working under him for PhD .He has to his credits around 56 papers in various fields of Engineering, Indian and international journals, National and International conferences, He worked as Chairman BOS in OUCE and conducted various staff development programs and workshops. He is Life Member of MISTE, Life Member of CSI..His interested area is temporal databases, Bio Informatics, Data mining and Artificial Intelligence.