Secured Encrypted DNA Compression Through Extended-Huffman-Codec:A Novel Methodology

Dr.V.Hari prasad
hariprasadvemulapati@gmail.com, drvhp60@gmail.com

***Abstract*—**State of the art the existing techniques of DNA sequences worked on standalone sequences and very few techniques were explained in connected with data base support but the results are not bountiful. In a daily routine different organisms are developed, for their archival infrastructure will require substantially larger .So to maintain Qos over computer network is a cumbersome task. Due to excess growth of living organisms in a daily routine , the maintenance of database administration becoming problem. Hence, a feasible solution is compression!. State of the art existing compression algorithms may work on repetitiveness and non-repetitiveness. In case of repeated codons the compression ratio and compression gain is feasible but whereas in case of non repetitive Codons the result of compression Gain and ratio is far ahead. So in this work the new methodology multistage E(Extended)- H-Codec (Bi-stage) has been proposed to maintain consistent results even with non repetitive sequences. The proposed methodology work in two phase's .In first stage phylogenetic tree codec can be constructed and at a later stretch extended to H-Codec methodology

Keywords— compression; encoding; decoding; bio compress; Huffbit compress; dnabit compress; LSBD compression.

1. INTRODUCTION

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

2. BASIC KNOWLEDGE OF GENOME DATA

2.1 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 2.1.

Tab1e2.1. Four types of nucleotides, Adenine (A), Guanine (G), Thymine (T) and Cytosine (C), and their complements

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

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.

3. GENERAL COMPRESSION ALGORITHMS

State of the art general purpose of compression algorithms

throws a light on the following methods

- Finding approximate repeats and non-repeats.
- Identifying reverse complements and reverse complements
- Identifying palindrome sequences
- Self Vs cross reference similarities

4. RELATED EXISTING ALGORITHMS AND PROPOSED METHODOLOGY

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 .State of the performance analysis of existing algorithms are found to be better in case of repetitive codons whereas the same techniques may run in worst case if the sequence is non repetitive codons. The proposed NDCP (Non-repetitive codlings of DNA compression will find out the remedy for so far state-art algorithm. Some sequences like AT-Rich Volate contains very rare frequent repeats. The existing techniques were applied on this the compression gain and compression Ratio's may run in worst cases. This will overcome in the proposed H-codec methodology. This methodology is purely developed based on mathematical analysis.

A. Idea behind the algorithm

The proposed Utility will work in two stages. In first stage the Input sequence converted into binary Huffman Tree and the output of first stage codec will be the Input for the second stage of H-codec.

Compression ratio is calculated encoded bits per Bases.

Compression Ratio = Encoded Bits/Bases.

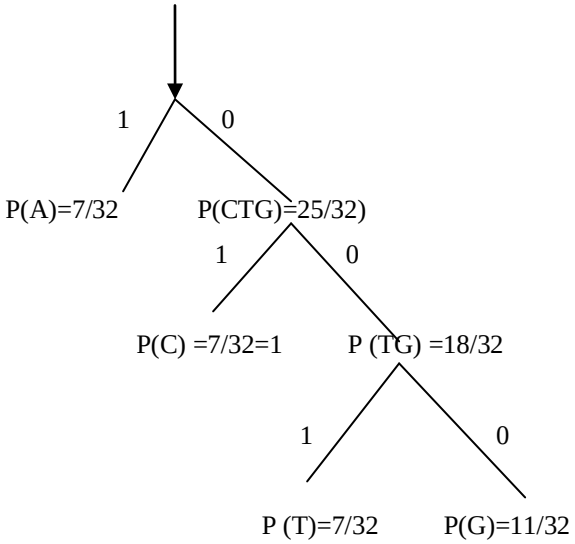
B. Plan of work

First stage:-

Let us consider a sample sequence and construct the Huffman tree for minized codecs.Here the sequence is divided into equal fragments(each fragement contains four bases)

Sequence:-

ACGT GTAC GACT TGAC
TACG AGCT GATC GGGC (32BASES)



Now the entropy of Huffman codec is noted as follows

P(A)=1 P(T)=001

P C)=01 P(G)=000

So the above sequence can be encoded in 84 bits and compression raio is nearer to 2.5bpb still it is compressible. The above sequence contains non repetitive fragments so the existing compression algorithms like Huffbit[14] can compress in worst case i.e 2.5bpb.

In our proposed NDCP methodology the first stage o/p will be the input for the second stage for better compression Ratio.

Second stage:

Our algorithm is array based implementation dynamic technique and the compression rate vary with the length of the sequence. NDCP method will work as follows. Re substitute the Huffman Codec bits in the above sequence and prepare a stream of bit pattern as input for the second stage.

$N = length\ of\ the\ bit\ stream(84bits)$
$Ncfb = Noncoded\ first\ partitioned\ block(16bits)$
$Ncsb = Noncoded\ second\ partitioned\ block(16bits)$
$FAb = FirstAvg\ block$
$Cab = Cummulative\ avg\ block$
$Cabv = Cummulative\ avg\ block\ value$
$Esb = Encoded\ segmented\ bits$
$Cr = Compression\ ratio$

Ncfb and Ncsb is calculated as follows .The numeric equivalent value of binary substitution.

$$Ncfb_v = \sum_{b=0}^{p} (Ncf_b)$$

$$Ncsb_v = \sum_{b=0}^{p} (Ncs_b)$$

Here the avg of first and second can be calculated as follows.

$$Fab = \sum_{b=0}^{p} (Ncfbv + Ncsbv / 2)$$

Now the cumulative value for Cab and Cabv calculated as follows.

Cab=Fab1+Fab2+……..+ n

$$Cab_v = \sum_{b=0}^{n} (Cab)$$

Here n is the length of the given sequence and total number of encoded bits to represent the genome sequence is equivalent to E_{sb}

Compression ratio can be computed as follows

$$Cr = E_{sb} / N$$

C. Analysys

Cabv=Fab1+Fab2+Fab3

Esb=2+2+2=6bytes=48bits

Finally, we can calculate the compression ratio in terms of bits per bases.

Cr= 48/32=1.5bpb (bitsperbase)

Compression and Decompression algorithms for DNA sequence is as follows.

D. . Encoding Algorithm

INP: input String
OPS: Encoded String

PROCEDURE ENCODE
Begin

- Group INS into equivalent fragments as four bases
- Generate all possible combinations of DNA and it will contain non- repetitive and repetitive
- Every Ncfb and Ncsb contains n/16 fragment bases
- Assign binary bits of Huffman codec of first phase
- Calculate Ncfbv for every Ncfb in INP till eof INP. Ncfbv represents binary equivalent numeric of Ncfb.
- Calculate Ncsbv for every Ncsb in INP till eof INP. Ncsbv represents binary equivalent numeric of Ncsb.
- Calculate Cab for every Avb till eof of INP
- Calculate E_{sb} for every S_b till eof INP
- Repeat the steps 4 and 5 until the length of the INP
- Transfer the sequence Esb to the output string i.e. OPS String.

End.

E. Decoding Algorithm

INP: input String
OPS: Decoded String

PROCEDURE DECODE

Begin

- Generate all possible combinations of (A,C,G,T)
- Read the binary data of each sub partition from OPS and assign the two bits by equivalent Base s of Huffman codec) and then store it in an array till eof
- Repeat step 2 until eof INS is reached and calculate Dsb and Db in the reverse process..
- Transfer the sequence D_{sb} to the input String i.e. INP

End

5. EXAMPLE AND COMPARISON

Let us consider the sequence.

Sequence1:

ACGT GTAC GACT TGAC
TACG AGCT GATC GGGC (32BASES)

Sequence length (no of bases) = 32.
Bytes required to store in a text file = 32 Bytes.
Bytes required in ASCII representation=28bytes

The above sequence doesn't contain tandem repeats so existing algorithms like Huff bit compress, Genbit Compress and Dnabit compress may run on worst case and require more bits to encode the sequence.

Huffbit, GenBit and Dna compress =90 bits (2.428)
Genbit Compress (Tool based) = 94 bits (2.404)
Extended-H-Codec

TOOL BASED =48 Bits (1.5bpb)

We compared of our technique with existing techniques and found to be the first-one technique .Proposed technique can be applicable for and non repetitive DNA.

6. CONCLUSION AND FUTURE WORK

By using of our algorithm we can encode every base by 1.498bits .By applying of ours we are saving nearer of 6 bytes to encode the given sequence, compression may vary with size of the sequence. In addition to that existing techniques uses dynamic programming to compress the sequence which is complex in implementation and time consuming. Our technique is implemented without dynamic programming approach, so it is simple and fast. The simplicity of this will reduce the complexity in processing and definitely it will be the invaluable tool in Bio informatics era. Our algorithm can be extended to any tool based approach.

ACKNOWLEDGMENT

I would like thank my better half G.Lakshmi Prasanna who bared me with utmost compassion and patience by giving 360^{o} support and my replica new born baby boy given me additional boost in extending variant ideas in Bio Informatics era.

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

Dr.V Hari Prasad , B.Tech CSE from JNTU University,Anantapur,M.Tech CSE from JNTUCEH,HYD and welded his PhD in CSE from JNTU KAKINADA, A.P .He has 13 years of teaching experience in various Engineering colleges. Presently He is working in Department of Technical Education 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.