BIOLOGICAL SEQUENCE COMPRESSION BASED ON COMPLEMENTARY PALINDROME USING VARIABLE LENGTH LOOK UP TABLE (LUT)

Rajendra Kumar Bharti[1] & R. K. Singh[2]
[1]Research Scholar UTU, Dehradun, Assistant Professor, B.T. Kumaon Institute of Technology, Dwarahat, Almora, Uttarakhand, INDIA
[2]Professor, B.T. Kumaon Institute of Technology, Dwarahat, Almora, Uttarakhand, INDIA
Email: [1]raj05_kumar@yahoo.co.in, [2]rksinghkec12@rediffmail.com

ABSTRACT
As the technology progress, the DNA sequencing increases the size of genome database. Data storage capacity has become an appreciable proportion of total capacity in the creation and analysis of DNA sequence database. The rate of increase in DNA sequencing is significantly outstripping the rate of increase in storage capacity. Efficient storage removes redundancy from the data being stored in the DNA molecule. Data compression remove redundancy used in data i.e. DNA molecule. In this paper we present a compression algorithm based on the properties complementary palindrome of a biological sequences. Our algorithm works on both repetitive and non repetitive biological sequences. This algorithm achieves the best compression ratio for biological sequences for larger genome.
It is very useful in sequence comparisons and multiple sequence alignment analysis. The simplicity and flexibility of our algorithm could make it valuable tool for biological sequence compression in clinical research.

Keywords: *Genome Database; redundancy; complementary palindrome; DNA compress; LZ 77; Gen compress; LUT.*

1. BACK GROUND

Life is strongly related with organization and structure. With the completion of 1000 genomes project, the project is estimated to generate about 802 billion bases per day, with the total sequence to exceed 6 trillion nucleotide bases. The DNA molecule has four nucleotide bases: A(Adenine), T(Thymine), G(Guanine), and C(cytosine). General purpose compression algorithms do not perform well with biological sequences. Giancarlo et al. have provided a review of compression algorithms designed for biological sequences. Finding the characteristics and comparing Genomes is a major task . 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.DNA sequences consists of only four nucleotides bases {a,c,g,t}. Two bits are enough to store each base. The standard compression softwares 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 but not for DNA Genomes. Increasing genome sequence data of organisms lead DNA database size two or three times bigger annually. Thus it becomes very hard to download and maintain data in a local system. Other algorithms specifically designed for DNA sequences compression did not manage to achieve average compressionrate below 1.7 bits/base. Algorithms for Compressing DNA sequences, such as Ziv-Lempel compression algorithms . Biocompress ,Gencompress and DNAcompress compress DNA sequences rate below 1.7 bits/base. Algorithms for Compressing DNA sequences, such as Ziv-Lempel compression algorithms . Biocompress ,Gencompress and DNAcompress compress DNA sequences. Hence we present a new compression algorithm which is based on properties complementary palindromes, whose compression rate is below 1.7 bits per base.

A region of sequence, that when it’s been read left to right it is complementary to the sequence that been read right to left (A match T, and C match G).
For example, the DNA sequence =ACCTAGGT
palindrome Sequence =TGGATCCA

2. PROPOSED ALGORITHM

An Algorithm consists of two phases. Fist, we shell search for all palindromes in a specific length, a range of allowed mismatches and a range of allowed gaps. Searching for palindromes done by checking all the possible places in the sequence (in order to be correct and not to miss even one palindrome). For each place in the sequence

we will check all possible sizes of gap, checking whether the pal' has allowed number of mismatches. The "heart" of the algorithm compare the first letter to the last letter, the second letter to the letter second form the end, etc. (A matches T and C matches G). If we found a palindrome that correlated with our demands, we will print it to the output.

In second phase we apply the LUT base variable length compression algorithm.

Phase 1: A double strand DNA locus whose 5'-to-3' sequence is identical on each DNA strand. The sequence is the same when one strand is read left to right and the other strand is read right to the left. In other words, a region of sequence, that when it's been read left to right it is complementary to the sequence that been read right to left (A match T, and C match G). Approximate Palindrome contain a certain number of mismatches and allow gap. "palindrome fingerprints" -Each DNA sequence has it's unique number, sizes of palindromes, and location in sequence. We find the palindrome sequence as:

Input

1) Sequence, genome of different organisms, text file in a FASTA format .
2) Length of palindrome (one side).
3) Maximum gap between repeated regions.
4) Number of mismatches allowed.

Output

All the palindromes within a specified length range and also a range of mismatch.

The Algorithm :

1.Search for the palindrome within a sub sequence, in the size of MaxSize.
2. Each iteration incrementing the size of palindrome, until MaxSize is reached.
3. Shift left of the sequence .

Phase 2:

In second phase we apply the LUT base variable length compression algorithm.

3. RESULT

Table 1: Comparaison between Different Previous Existing Biological Compression Techniques & LZ 77 Compression (Universel) Techniques

Type of Sequences	Original Size(bits) before compression	Size of the sequences after applying various compression algorithm			
		DNA Compress	Gen Compress	Fixed LUT	Univ.(LZ 77)
Gallus β globin	752	272	360	256	568
Goat alanine β globin	732	256	352	248	516
Human β globin	752	272	360	256	608
Lemur β globin	760	280	376	264	592
Mouse β globin	776	280	376	264	608
Opossum β hemoglobin β- M gene	760	272	376	264	600
Rabbit β globin	736	264	352	256	560
Rat β globin	752	272	360	256	600
Avg	752.5	271	364	258	581.5

Table 2: Comparison between Different Biological Sequence Compression Techniques with variable LUT

Type of Sequences	Original Size(bits) before compression	Size of the sequences after applying various compression algorithm			
		DNACompress	GenCompress	Fixed LUT	Variable LUT
Gallus β globin	752	272	360	256	248
Goat alanine β globin	732	256	352	248	232
Human β globin	752	272	360	256	248
Lemur β globin	760	280	376	264	256
Mouse β globin	776	280	376	264	256
Opossum β hemoglobin β- M gene	760	272	376	264	256
Rabbit β globin	736	264	352	256	248
Rat β globin	752	272	360	256	248
Avg	752.5	271	364	258	249

Table 3: Biological Sequence Compression Based on Complementary palindrome Using Variable length LUT

Type of Sequences	Original Size(bits) before compression	Size of the sequences after applying compression algorithm	Using Complementary Palindrome & Variable length LUT
Gallus β globin	752	376	56
Goat alanine β globin	732	392	88
Human β globin	752	368	56
Lemur β globin	760	344	64
Mouse β globin	776	376	56
Opossum β hemoglobin β- M gene	760	312	64
Rabbit β globin	736	344	64
Rat β globin	752	376	56
Avg	752.5	361	63

4. CONCLUSION AND DISCUSSION

There are various universal compression algorithm like LZ77 but they are suitable for compression of Biological sequences as shown so various specialized Biological sequence compression algorithm were developed like DNA Compress, Gen Compress, Fixed length LUT and Variable length LUT. As the result shows that variable LUT gives better result so we choose it for proposed compression algorithm.

Thus proposed algorithm has high compression ratio to other exiting Biological Sequence Compression. This algorithm also uses less memory compared to the other algorithm and easy to implement.

The proposed algorithm is compress Biological sequences which complementary in nature. All other algorithm only uses other properties of sequences such as repeated and non repeated. If the sequence is compressed using proposed algorithm it will be easier to make sequence analysis between compressed sequences. It will also be easier to make multi sequence alignment. High compression ratio also suggests a highly repetitive sequence.

5. REFERENCES

[1]. Ateet Mehta , 2010, et al., " DNA Compression using Hash Based Data Structure", IJIT&KM, Vol2 No.2, pp. 383-386.

[2]. B.A., 2005, " Genetics: A comceptual approach." Freeman, PP 311.

[3]. Choi Ping Paula Wu, 2008, et al., " Cross chromosomal similarity for DNA sequence compression", Bioinformatics 2(9): 412-416.

[4]. Gregory Vey,2009, "Differential direct coding: a compression algorithm for nucleotide sequence data", Database, doi: 10.1093/database/bap013.

[5]. J. Ziv and A.1977, et al., "A universal algorithm for sequential data compression," IEEE Transactions on Information Theory, vol. IT-23.

[6]. K.N. Mishra,2010, " An efficient Horizontal and Vertical Method for Online DNA sequence Compression", IJCA(0975-8887), Vol3, PP 39-45.

[7]. P. raja Rajeswari, 2010, et al., " GENBIT Compress- Algorithm for repetitive and non repetitive DNA sequences", JTAIT, PP 25-29.

[8]. Pavol Hanus, 2010, et al., "Compression of whole Genome Alignments", IEEE Transactions of Information Theory, vol.56, No.2Doi: 10.1109/TIT.2009.2037052.

[9]. R.K.Bharti,2011, et al., "Biological sequence Compression Based on Cross chromosomal properties Using variable length LUT", CSC Journal, Vol 4 Issue 6.

[10]. R.K.Bharti,2011, et al, "Biological sequence Compression Based on properties unique and repeated repeats Using variable length LUT" CiiT journal, Vol 3 Issue, 4.

[11]. R. Curnow, 1989, et al. "Statistical analysis of deoxyribonucleic acid sequence data-a review," J Royal Statistical Soc., vol. 152, pp. 199-220.

[12]. Sheng Bao, 2005, et al. "A DNA Sequence Compression Algorithm Based on LUT and LZ77".

[13]. U. Ghoshdastider,2005, et al., "GenomeCompress: A Novel Algorithm for DNA Compression", ISSN 0973-6824.

[14]. Xin Chen, 2002, et al.," DNA Compress: fast and effective DNA sequence Compression" BIOINFORMATICS APPLICATIONS NOTE, Vol. 18 no. 12, Pages 1696–1698.

[15]. X. Chen, 2002, et al., "Dnacompress:fast and effective dna sequence compression," Bioinformatics, vol. 18.