International Journal of Theoretical & Applied Sciences, 9(2): 269-273(2017)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

DNA Compression Algorithm –using Inter Vs Intra Chromosomal Repeats

Kakoli Banerjee* and Dr. D. V. Rai**
**Ph.D. Research Scholar, Shobhit University, Gangoh, (Uttar Pradesh), INDIA*
***Guide, Vice Chancellor, Shobhit University, Gangoh, (Uttar Pradesh), INDIA*

(Corresponding author: Kakoli Banerjee)
(Received 02 October, 2017 accepted 12 November, 2017)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: The human genome in its base format occupies almost thirty terabyte of storage space. Computational resources are limited. Not only storage, transmission capabilities and run time memory is also limited. Data compression is a challenge when the input is exponentially increasing genetic data. It is important to conserve the integrity of genetic data while compressing it Standard compression algorithms fail to compress genetic data. Hence, we require special algorithms for compressing genetic databases. In this paper, we show a comparative study between the compression algorithm using Inter chromosomal repeats and Intra chromosomal repeats.

***Keywords*:** DNA, Genome, Sequence, Compression, Chromosome.

I. INTRODUCTION

Biological sequences are like blue prints of all living organisms. Special biological sequences like DNA are functional description of cells that are the basic building unit of each organism. DNA sequences are made of four chemical bases namely – Adenine (A), Thymine (T), Guanine (G) and Cytosine(C) [15]. Due to latest enhancement in technology, the DNA sequences of most of the organisms are known. The human genome is almost known, and occupies almost 30 terabyte of space in its base format. It consists of 3 billion such bases. The special databases like GenBank, which stores genetic information shared by researchers all over the world, doubles itself every 35 months [15]. Storage is limited and so is the capacity of transmission channels. With these limited resources, handling of such exponentially growing data is a challenge. Hence we require compression [15]. What is Compression – a big question to think about – is it just reducing the size data. No. Compression is much more than that. Compression is "Modeling + Coding". Modeling is where we find different type of methods to find redundancy in data and coding is where we replace these redundancies by some type of references. Hence for handling such big volume data compression is must [15].

Longer the repeat, more compression is achieved [15]. When we consider the inter chromosomal comparison, we are able to find very long repeats. This study will present a comparison between inter and intra chromosomal repeats and show how compression improves when we consider inter chromosomal repeats along with intra chromosomal repeats.

II. LITERATURE SURVEY

DNA compression was initiated by Grumbach and Thai and it was known as Bio Compress and its second version known as BioCompress- 2 [1,2]. Ziv- Lempel compression techniques was the base for both the Algorithms. BioCompress-2 came with the added feature of searching exact repeats. C fact was the next Algorithm, which uses a suffix tree to exploit the lengthiest exact repeat [3]. GenCompress-1 and GenCompress-2 came next. GenCompress- 1 uses the technique of hamming distance or substitution only for repeats [4]. GenCompress-2 uses insertion, deletion and substitution for encoding the repeats. In 2011 DNABIT was introduced. It was also a two-phase algorithm. [5].

In 2012 CTW+LZ came which was a combination of context tree weighting + LZ77 [6]. The algorithm used the context tree weight for encoding long approximate repeats whereas LZ77 was used to encode short repeats.

III. INTRA CHROMOSOMAL REPEATS

Intra chromosomal repeat are those pieces of identical sequences that are present in one particular chromosome of a genome. As in the study we are considering only exact repeats, hence they need to be identical in nature. Any two subsequences can be called a pair of intra chromosomal repeat, if they are at least in a group of 2 nucleotides and are exactly same.

CCACA**CCA**CAC**CCA**CACAC**CCA**CACA**CCA**CA**CCA**CACA**CCA**CA**CCA**CAC**CCA**CACACACACATCCTAACACTAC
CCTAACACAGCCCTAATCTAACCCTGG**CCA**ACCTGTCTCTCAACTTACCCT**CCA**TTACCCTGCCT**CCA**CTCGTTACC
CTGTC**CCA**TTCAA**CCA**TA**CCA**CTCCGAA**CCACCA**T**CCA**TCCCTCTACTTACTA**CCA**CTCAC**CCA**CCGTTACCCT**CC**
AATTAC**CCA**TAT**CCA**AC**CCA**CTG**CCA**CTTACCCTA**CCA**TTACCCTA**CCA**T**CCACCA**TGACCTACTCA**CCA**TACTG
TTCTTCTAC**CCACCA**TATTGAAACGCTAACAAATGATCGTAAATAACACACACGTGCTTACCCTA**CCA**CTTTATA**C**
CACCACCACATG**CCA**TACTCACCCTCACTTGTATACTGATTTTACGTACGCACACGGATGCTACAGTATATA**CCA**T
CTCAAACTTACCCTACTCTCAGATT**CCA**CTTCACT**CCA**TGGC**CCA**TCTCTCACTGAATCAGTA**CCA**AATGCACTCA
CATCATTATGCACGGCACTTGCCTCAGCGGTCTATACCCTGTG**CCA**TTTAC**CCA**TAACGC**CCA**TCATTAT**CCA**CAT
TTTGATATCTATATCTCATTCGGCGGTC**CCA**AATATTGTATAACTGCCCTTAATACATACGTTATA**CCA**CTTTTGCA
CCATATACTTA**CCA**CT**CCA**TTTATATACACTTATGTCAATATTACAGAAAAATCC**CCA**CAAAAATCACCTAAACAT
AAAAATATTCTACTTTTCAACAATAATACATAAACATATTGGCTTGTGGTAGCAACACTATCATGGTATCACTAAC
GTAAAAGTTCCT

Fig. 1. Figure representing 52 Intra chromosomal Repeats of Subsequence "CCA" in first 847 Bases of *Saccharomyces cerevisiae* S288C chromosome I.

Figure number explains the occurrence of intra chromosomal repeat. The size of the repeat can vary from 2 bases to thousands of bases. The Figure Number 1 shows the occurrence of 52 repeats of Subsequence "CCA" in first 847 Bases of *Saccharomyces cerevisiae* S288C chromosome I.

IV. INTER CHROMOSOMAL REPEAT

In case of Intra Chromosomal Repeats we search for Repeats within one single Chromosome. Where as in case of Intra Chromosomal Repeats the algorithm try to find out the similarities not only within the chromosome itself but also in other chromosomes of the same genome. Because the compression algorithm developed in this study replaces repeats with ASCII codes, hence compression ratio is directly proportional to the size and number of the repeat. In other words, bigger the size and number of the repeat more compression can be achieved. As this algorithm compressors one genome at a time, hence there is a possibility that bigger repeat are found if two different chromosomes of the same genome are compared. Inter chromosomal repeat should have a size length of at least two nucleotides, should be exactly same and can belong to two different chromosomes of the same genome. Figure Number 4 shows that by comparing two different chromosomes the number of repeats increases.

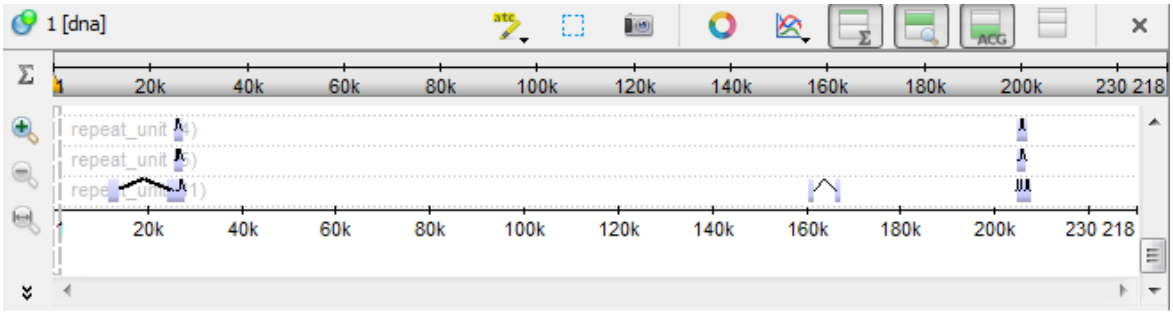

Fig. 2. Figure showing the output of UGENE while trying to find out the number of repeats in *Saccharomyces cerevisiae* S288C chromosome I (Having size at least 100bpb).

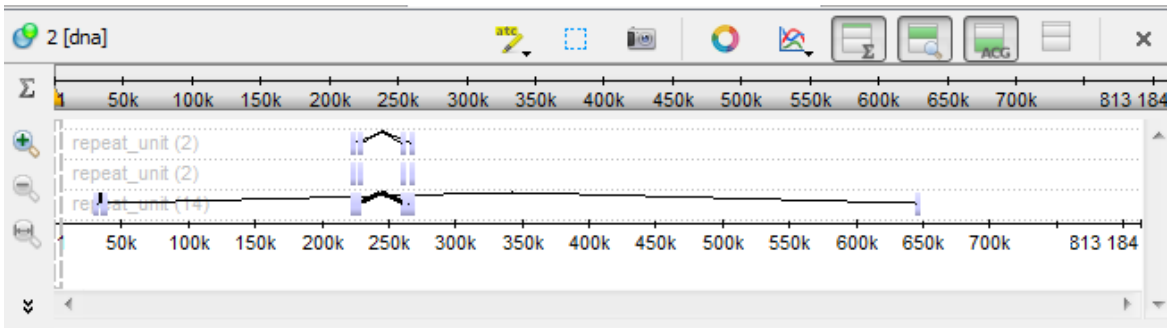

Fig. 3. Figure showing the output of UGENE while trying to find out the number of repeats in *Saccharomyces cerevisiae* S288C chromosome II (Having size at least 100bpb).

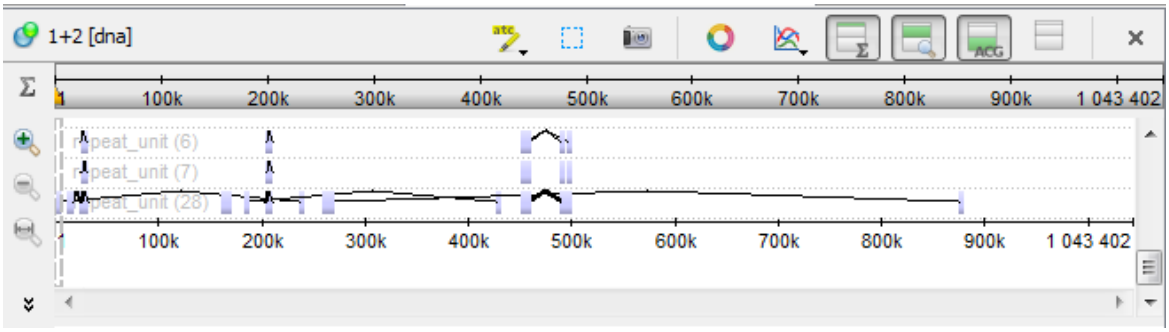

Fig. 4. Figure showing the output of UGENE while trying to find out the number of repeats in *Saccharomyces cerevisiae* S288C chromosome I + chromosome II (Having size at least 100bpb).

This simple example explains inter chromosomal repeats and also shows the advantage of using inter chromosomal repeats along with intra chromosomal repeats. Table number 1 and figure 5 shows the advantage of using inter chromosomal repeats than using only intra chromosomal repeats.

From the above example it is clear that if we only try to find intra chromosomal repeats in the sequence 1, the number of repeats we get is 58 with minimum size of 100bpb. Intra chromosomal repeats in the sequence 2, the number of repeats we get is 36 with minimum size of 100bpb. But when we combine both the sequences and try to find the repeats, the number of repeats found is 94 in number. This study shows how compression can improve if we use inter chromosomal repeats instead of only intra chromosomal repeats.

Table 1: Comparison of Inter and Intra Chromosomal Repeats.

Sequence Name	Min. BPB Size	No of Repeats	Type of Repeat
Saccharomyces cerevisiae S288C chromosome I	100	58	Intra
Saccharomyces cerevisiae S288C chromosome II	100	36	Intra
Saccharomyces cerevisiae S288C chromosome I+II	100	49	Inter

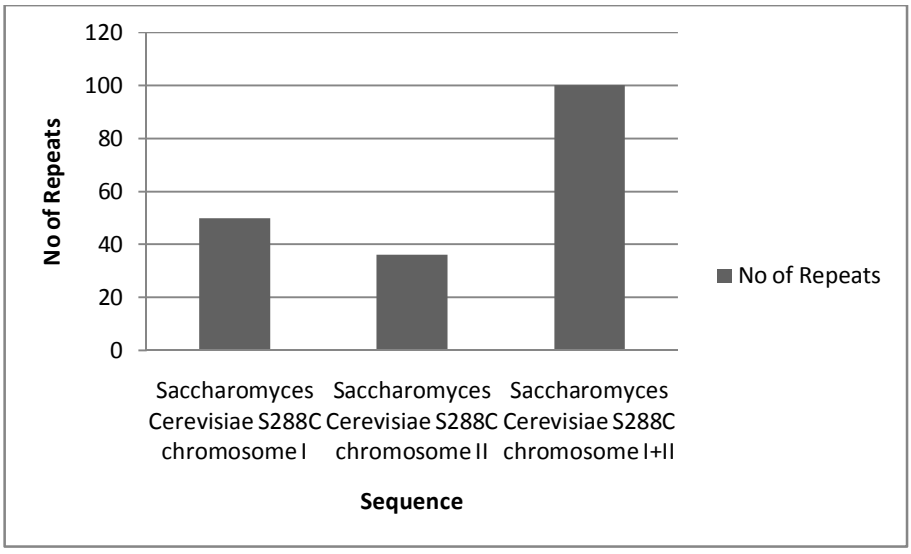

Fig. 5. Comparison of Inter and Intra Chromosomal Repeats.

RESULTS

From the procedure discussed above, the difference between inter and intra chromosomal repeats can be easily understood. The graph depicts the difference between the sum of Repeats in Intra Chromosomal Search and Inter Chromosomal Search.

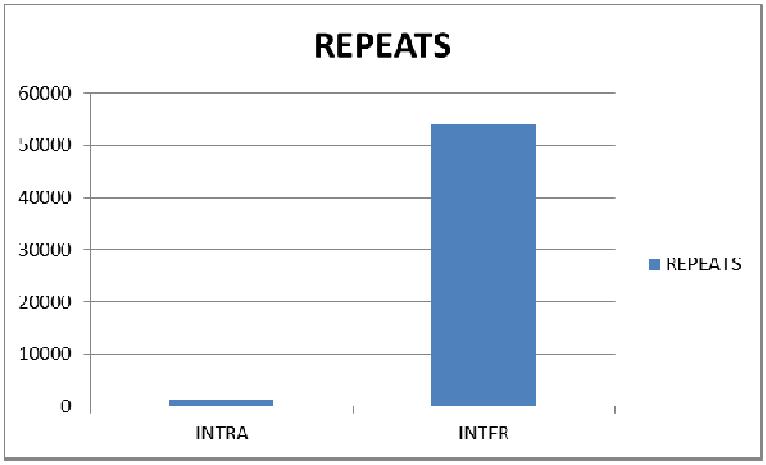

Fig. 6. Shows the difference between the sum of Repeats in Intra Chromosomal Search and Inter Chromosomal Search.

The current study is not only focusing on development of a novel DNA compression Algorithm, but also aims at improving the compression ratio of already existing algorithms. As already concluded – "More number of repeats better is the compression". Table Number shows the huge difference between Intra and Inter Chromosomal Repeats in number. This study is not only end up developing a better compression algorithm but will also change the approach of already existing repeat based DNA compression algorithms, hence forth improve the compression ratio of all existing algorithms.

Table 2: Comparison of Intra Vs Inter Chromosomal Repeats of *Saccharomyces cerevisiae* S288C, which clearly shows the difference between the two different type of repeats.

Repeats	Intra Chromosomal Repeats	Inter Chromosomal Repeats
I	58	2712
II	36	2944
III	48	2706
IV	352	7034
V	62	3464
VI	2	2002
VII	88	3692
VIII	32	2742
IX	62	2538
X	50	3248
XI	12	1560
XII	198	5686
XIII	80	3564
XIV	18	2806
XV	66	3292
XVI	94	4062
SUM	1258	54052

REFERENCES

[1]. S. Grumbach and F. Tahi, (1993). "Compression of DNA sequences", *[Proceedings] DCC `93: Data Compression Conference*.

[2]. S. Grumbach and F. Tahi, (1994). "A new challenge for compression algorithms: Genetic sequences", *Information Processing & Management*, Vol. **30**, no. 6, pp. 875-886.

[3]. É. Rivals, M. Dauchet, J. Delahaye and O. Delgrange, (1996). "Compression and genetic sequence analysis", *Biochimie*, Vol. **78**, no. 5, pp. 315-322.

[4]. X. Chen, S. Kwong and M. Li, (2000). "A compression algorithm for DNA sequences and its applications in genome comparison", *Proceedings of the fourth annual international conference on Computational molecular biology - RECOMB '00*, 2000.

[5]. P. Rajarajeswari and A. Apparao, (2011). "DNABIT Compress-Genome Compression Algorithm", *Bioinformation*, Vol. **5**, no. 8, pp. 350-360.

[6]. Cong Li, Zhenzhou Ji and Fei Gu, (2012). "Efficient parallel design for BWT-based DNA sequences data multi-compression algorithm", *International Conference on Automatic Control and Artificial Intelligence (ACAI 2012)*, 2012.

[7]. W. Kinsner, (2010). "Towards cognitive analysis of DNA", *9th IEEE International Conference on Cognitive Informatics (ICCI'10)*, 2010.

[8]. Z. Zhu, Y. Zhang, Z. Ji, S. He and X. Yang, (2013). "High-throughput DNA sequence data compression", *Briefings in Bioinformatics*, Vol. **16**, no. 1, pp. 1-15.

[9]. P. Eric, G. Gopalakrishnan and M. Karunakaran, (2016). "An Optimal Seed Based Compression Algorithm for DNA Sequences", *Advances in Bioinformatics*, pp. 1-7.

[10]. T. Soliman, T. Gharib, A. Alian and M. Sharkawy, (2009). "A Lossless Compression Algorithm for DNA sequences", *International Journal of Bioinformatics Research and Applications*, Vol. **5**, no. 6, p. 593.

[11]. S. Hossein and S. Roy, (2013). "A Compression & Encryption Algorithm on DNA Sequences Using Dynamic Look up Table and Modified Huffman Techniques", *International Journal of Information Technology and Computer Science*, Vol. **5**, no. 10, pp. 39-61.

[12]. Xin Chen, S. Kwong and Ming Li, (2001). "A compression algorithm for DNA sequences", *IEEE Engineering in Medicine and Biology Magazine*, Vol. **20**, no. 4, pp. 61-66.

[13]. S. Hossein, P. Mohapatra and D. De, (2015). "A Compression Algorithm for DNA Sequences Based on R2G Techniques with Security", *Trends in Bioinformatics*, Vol. **8**, no. 3, pp. 93-98, 2015.

[14]. S. Roy, (2012). "An Efficient Biological Sequence Compression Technique Using LUT and Repeat in the Sequence", *IOSR Journal of Computer Engineering*, Vol. **6**, no. 1, pp. 42-50.

[15]. Kakoli Banerjee, Dr. R. Prasad, (2013). “Exact Repeat Based Intra Chromosomal DNA Compression Algorithm”, *GLIMPSES – An International Journal of Multidisciplinary Research*, ISSN: 2250-0561, Vol. **1**, no. 3, pp 259-270.