International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 07 | July -2017 www.irjet.net p-ISSN: 2395-0072

Design & Implementation of a DNA Compression Algorithm

Manju Rani [1], Pawan Kumar Mishra [2]

[1]*M.Tech Scholar, Uttarakhand Technical University, Dehradun, Uttarakhand, India*
[2]*Assistant Professor, Uttarakhand Technical University, Dehradun, Uttarakhand, India*

Abstract - *If we talk about the DNA arrangements, we understand that it oversees only four pictures addressing four nucleotide bases {A, C, T, G}. these four pictures could have been shown as {00, 01, 10, 11} independently, where we can watch that every nucleotide base having 8 bit is made to include 2 bits, when encoded in the already said parallel shape. This could have been a champion among the most capable encoding designs, if and only if there were the same pictures in the plan, other than A, G, T and C base characters. Here, however the encoding ought to be conceivable, yet essential issue will occur in the midst of decompression as the twofold code of the unanticipated picture like N or S will organize with the matched code of A, G, T and C. An another sort of figuring used for DNA weight is Differential Direct (2D) Coding Algorithm, which can vanquish this issue by isolating between the base characters and the astounding pictures. The 2D coding computation uses the social occasion of three characters (triplets), being supplanted by some other character [28].*

Key Words: DNA Compression, Nucleotide Sequence Compression, Look Up Table, Compression Algorithm, Lossless Compression

1.INTRODUCTION

Compression is a technique to reduce the size of some data by lowering the number of bits used in its formation. In other words, we can say compression means reduction of size of data by changing it to a format which requires fewer bits than the original format [1].

DNA is a contraction for deoxyribonucleic corrosive, which conveys genetic data. A large portion of the DNA's found in people are same which in gathered in cell core, with the exception of some, which are found in cell's mitochondria. The previous one is known as atomic DNA, while the last one is mt DNA. There are four distinct sorts of nucleotides found in DNA, contrasting just in the nitrogenous base. They are: A, G, T, C. there are about 3 billion bases of human DNA, out of which 99% are same in all people [9]. These bases of these groupings are essential characterizing parts of organisms.
In DNA bonds up with T and C bonds up with G, forming base pairs. The twofold helix structure of a DNA contains nucleotides (mix of phosphate particle, sugar atom and base combine). Twofold helix structure resembles a step, with base framing its rungs and sugar and phosphate atoms shaping its vertical sidepieces. DNA has a unique property of multiplying or replicating.

With the increase in size of the databases containing the nucleotide arrangements, which are utilized as a part of seeking applications to find successions homologous to an inquiry grouping, the need of pressure methods have happened. It is important to store information minimally with the goal that it can be exchanged effortlessly. Furthermore, groupings can be gotten to autonomously. Moreover, the circle costs are regularly bottleneck in seeking additionally [10, 11].

Compression rate is the measurement of the reduction in size of the original file. There are four main methods of measuring the compression rate . The first one known as Bit per Byte or bpb refers to the replacement of one byte (particularly the collection of 8 bits) by less than 8 bits. It can be formulated as follows: (compressed length / original length) * 8. If a file of 800 bytes has been compressed to a file of 200 bytes, the compression will be - (200/800)*8=2 bpb. The second method is measuring of compression in terms of percentage, which can be formulated as (compressed length / original length) * 100. If a file of 800 bytes has been compressed to a file of 200 bytes, the compression will be - (200/800)*100= 25% of the original file. Third method can be representation in ratio form, which is (original size: compressed size). This is a general representation technique and is widely used. But it has low precision. i.e (4: 1) or (3:1) Bit per Char is another technique. It is same like bpb in some cases only and it cannot be used to compress binary files.

2. Literature survey

Chen,X. et al. (2000) characterized Differential Direct Coding (2D) likewise isolates the grouping into components of length three . It suggests that pressure methodologies must suit vast informational collections, comprise of various arrangements and helper information. The arrangement of expected images for the 2D show are {A, T, G, C, and U}, which expels the weight of express assertion of succession sort like DNA or RNA.

Adjeroh,D. et al. (2002) depicted the examine disconnected word reference arranged ways to deal with DNA succession pressure, in view of the Burrows-Wheeler Transform (BWT).

The dominance of short rehashing designs is a critical wonder in natural groupings.

Cherniavski,N. and Lander,R. (2004) proposed strategy for pressure of nucleotide grouping information. While present day equipment can give immense measures of economical stockpiling for organic databases, the pressure of nucleotide succession information is still of foremost significance keeping in mind the end goal to encourage quick pursuit and recovery operations through a diminishment in plate movement.

Hoebeke,M., Chiapello,H., Gibrat,J.F. et al. (2005) characterized in this paper The right translation of any natural investigation depends in a fundamental path on the exactness and consistency of the current explanation databases. Such databases are omnipresent and utilized by all life researchers in many trials. Be that as it may, it is outstanding that such databases are fragmented and numerous comments may likewise be inaccurate. In this paper we portray a system that can be utilized to break down the semantic substance of such explanation databases David Salomon (with contributions by Giovanni Motta and David Bryant) (2006) examined in this paper about the Compression Maximizes the capacity limit of Cassandra hubs by diminishing the volume of information on plate and circle I/O, especially for read-commanded workloads. Cassandra rapidly finds the area of lines in the SS Table file and decompresses the significant column lumps.

Hall N (May 2007). Characterized the DNA sequencing incorporates a few techniques and innovations that are utilized for deciding the request of the nucleotide bases—adenine, guanine, cytosine, and thymine—in a particle of DNA. Information of DNA arrangements has turned out to be fundamental for essential natural research, other research branches using DNA sequencing, and in various connected fields, for example, analytic, biotechnology, measurable science and organic systematics.

Korodi,G. and Tabus,I. (2008) depicted in this paper DNA promoter groupings a novel transformative calculation for administrative theme revelation in DNA promoter arrangements.

Haiminen,N. et al. (2009) portrayed in this paper of Textual information pressure, and the related procedures originating from data hypothesis, are frequently seen as being of enthusiasm for information correspondence and capacity.
Gregory Vey (2012) proposed in this paper strategy for pressure of genomic information. Matter unveils a framework and a strategy for pressure of genomic information. In one encapsulation, the strategy for pressure of genomic information incorporates getting adjusted genomic information from genomic information construct in any event to a limited extent in light of middle person information recognized from the genomic information.

K.N. Mishra, (2014) the creator is characterized in this paper huge measures of modest stockpiling for organic databases, While current equipment can give huge measures of cheap stockpiling for natural databases, the pressure of Biological groupings is still of vital significance with a specific end goal to encourage quick pursuit and recovery operations through a diminishment in circle movement.

R.K.Bharti (2016) the creator is characterized in this paper huge measures of modest stockpiling for organic databases, While current equipment can give huge measures of cheap stockpiling for natural databases, the pressure of Biological groupings is still of vital significance with a specific end goal to encourage quick pursuit and recovery operations through a diminishment in circle movement.

3. Existing Methodology

The model and coding used by existing methodology is show below-

3.1 Model : There exist two type of key progressions which are DNA and mRNA courses of action having base characters {A, C, G, T} and {A, C, G, U} independently. The 2D coding computation works subsequent to taking a union of both of these sets to avoid the outside attestation of the sort of collection i.e. it is a DNA gathering or a mRNA progression and base characters are {A, C, G, T, U} independently. This computation uses the ASCII characters of the range 0 to 127, to address the base characters, unexpected character. The non-printable ASCII characters going from - 1 to - 127 address the triplets for goes over. Whatever is left of the picture passing on - 128th ASCII regard is used to store the dark character. This model contains signify 125 remarkable blends of triplets and in this way uses - 1 to - 125 ASCII characters to address these triplets. Out of every one of these mixes, some are never utilized as they abuse the nucleotide base subsets of DNA and mRNA like {UUT, TTU, UTU…}[33].

3.2 Coding : Here, a marked byte is used to address the data. In this checked byte, the last seven bits are used to address the data and the remaining first MSB is used as a sign piece. If this bit is 1 (negative), this suggests the accompanying seven bits will address triplet or darken character and if this bit is 0 (positive),it will speak to the nucleotide bases. These are represented in the table below:[30]

Table-1: Coding Data Model for Existing Methodology

Type	Description	Range	Compressible
Auxiliary	ASCII	0 to 127	No
Sequence	Triplet	-1 to -125	Yes
Unknown	?	-128	No

4. Shortcoming of existing algorithm

As discussed, the algorithm stores all the possible combinations of {A, G, T, C, U}, though some of the combinations are not acceptable and are never used, therefore leaving some of the non-printable ASCII characters unused.

It only concentrates over the triplets, even though other combinations may yield better compression ratio.
Boundary preservation property increases the output file size as the base characters forming triplets are copied to the output stream as they are.

It only uses the ASCII symbols till the date whose character set is not very large.[35]

5. Proposed Method

Whenever we need to pack some organic succession, we know that at once just a DNA or mRNA grouping will be compacted. Henceforth if the instance of arrangement of triplets is considered, with mix of four images {A, C, G, T} or {A, C, G, U} for DNA or mRNA separately, we will experience most extreme 64 mixes. These 64 triplet blends can be taken care of by 64 non-printable ASCII characters, though there exist add up to 127 non-printable ASCII characters.
Along these lines, the rest of the 63 characters can be utilized to store some different mixes of size more than 3, which can yield a superior pressure.

We partition the look-into table into two areas here: settled length LUT and variable length LUT. The past one will constantly exist there containing 64 mixes of triplets, however the variable length LUT can be of variable size, containing the blend bases of the size which is in different of 3. This will yield to the proper utilization of the available non-printable ASCII characters.[31]

5.1 Model: We consider the ASCII characters between the extents 0 to 127 for the Auxiliary images. The other range between - 1 to - 127 is isolated keeping in mind the end goal to change settled size LUT and variable size LUT. The remaining - 128th character is utilized for encoding the unknown character.

Table-2: Coding Data Model for Proposed Methodology

Type of Data	Description	Range	Look-Up Table
Auxiliary Symbol	ASCII	0 to 127	
Triplet	Group of three DNA bases	-1 to -64	Fix LUT
Multiple of Triplet	Group of DNS bases in multiple of 3	-65 to -127	Variable Length LUT
Unknown	?	-128	

6. Proposed Algorithm

Our proposed algorithm is given below-
Initialize: String s, st and t as NULL
Step 1: Read first 3 unprocessed bases into string t. If t is not equal to NULL then go along to step 2.
Else process the last one or two DNA bases by step 5.
Step 2: If t has all non N bases then go to step 3.
Else if t has N characters then go to step 4 otherwise go to step 5.
Step 3: If string st found in the LUT then s=st
Else write the ASCII code for s that is mapped in LUT from -1 to -127 and Add st to the LUT table for future reference.
s=t;
Step 4: count (c) total number of appearing in successive Ns and write all such Ns with "Nc" into destination file. After this jump to step 6. If number of successive Ns appears more than one time repeat the step 4.
Step 5: write non-N bases whose number is less than three into destination file directly. After that, jump to step 6.
Step 6: Return to step 1 and repeat the process until EOF is reached.

6.1 Coding: The coding data model of proposed algorithm is show in table 3.

Table-3: Coding Data Model for Proposed Methodology

Step	Input Sequence	Triplet (t)	Multiple of Triplet (st)	Look-Up Table		Encoded Sequence (s)
				Status of st	Entry	
1	TCTGCTTCTGCTNNNGC					
2	TCT	TCT	TCT	Found	TCT = # GCT = +	
3	GCT	GCT	TCTGCT	Not Found	Add with TCTGCT=$	#
4	TCT	TCT	GCTTCT	Not Found	Add with GCTTCT=@	#+
5	GCT	GCT	TCTGCT	Found	TCTGCT=$	#+
6	NNN					#+$N3
7	GC	GC	GC	<3 Char		#+$N3GC

7. Results and Analysis

We can contrast both the calculations and the assistance of indicated preview. In this depiction "2D coding with settled length LUT", calculation is compacting a source record of 2048 Bytes into target document of 684 Bytes and aggregate pressure accomplished is 67%. Be that as it may, with the assistance of "2D Coding with variable length LUT" calculation, the source record of 2048 Bytes can be packed into 655 Bytes and aggregate pressure accomplished is 69%. Consequently, this demonstrates a significant change more than "2D Coding with settled length LUT". Our proposed calculation can give better pressure as number of rehashes increments for group of 6, 9, 12… characters.

The table 4 shows that on an ordinary direct 2D Coding estimation packs the DNA groupings of 64338.9 Bytes into

21446.9 Bytes while our proposed figuring 2D Coding using variable length LUT packs 64338.9 Bytes into 20914 Bytes.

Table-4: Result Analysis

S.N	Type of Sequence	Original size of sequence before compression (Bytes)	Size of Sequence after Compression	
			Using Existing Algorithm (Bytes)	Using Proposed Algorithm (Bytes)
1	ATATSGS	9647	3217	3101
2	ATEF1A23	6022	2008	1957
3	ATRDNAF	10014	3338	3276
4	ATRDNAI	5287	1763	1734
5	CHMPXX	15180	5060	4874
6	CHNTXX	155844	51948	50540
7	HEHCMVCG	229354	76452	74736
8	HUMDYSTROP	105265	35089	34347
9	HUMHDABCD	58864	19622	19201
10	VACCG	47912	15972	15374
AVERAGE		**64338.9**	**21446.9**	**20914**

The compression result can be analyzed using graph shown in figure 1.

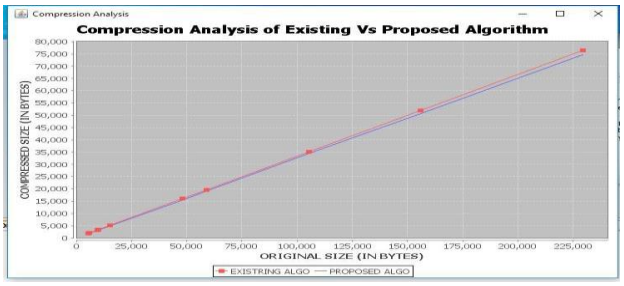

Fig-1: Result Analysis Using Graph

8. Conclusion

The objective of this recommendation is to develop a count that has high weight extent to other existing DNA Sequence Compression computations. This computation moreover uses less measure of memory when appeared differently in relation to interchange counts and is definitely not hard to realize. The proposed estimation packs Nucleotide progressions like DNA and furthermore RNA. Each and every other figuring use simply substitute properties of plans, for instance, reiterated and non-repeated. In case the gathering is pressed using proposed estimation it will be less requesting to make course of action examination between compacted progressions. It will in like manner be less difficult to make multi progression course of action. High weight extent moreover prescribes an especially dismal game plan. With the help of settled length LUT we can pack the DNA gathering up to 1/3 of its one of a kind size. Regardless, now we can achieve higher weight by using variable length LUT. The proposed calculation can give much better pressure as the more drawn out arrangements are found frequently in big sequences.

References

[1] D. A. Huffman, "A method for the construction of minimum-redundancy codes," Proc. IRE, vol. 40, pp. 1098-1101, 1952.

[2] Wilkins M.H.F., A.R. Stokes A.R. & Wilson, H.R. (1953)."Molecular Structure of Deoxypentose Nucleic Acids" (PDF).Nature 171 (4356): 738-740. Bibcode 1953Natur.171..738W .doi:10.1038/171738a0 . PMID 13054693 .

[3] J. Ziv, "Coding of sources with unknown statistics-Part I; Probability of encoding error," IEEE Trans. Inform. Theory, vol. IT-18, pp. 384-394, May 1972.

[4] A. Lempel, S. Even, and M. Cohn, "An algorithm for optimal prefix parsing of a noiseless and memoryless channel," IEEE Trans. Inform. Theory, vol. IT-19, pp. 208-214, March 1973.

[5] A. Lempel and J. Ziv, "On the complexity of finite sequences," IEEE Trans. Inform. Theory, vol. IT-22, pp. 75-81, Jan. 1976.

[6] Jacob Ziv and Abraham Lempel; Compression of Individual Sequences Via Variable-Rate Coding , IEEE Transactions on Information Theory, September 1978.

[7] Leslie AG, Arnott S, Chandrasekaran R, Ratliff RL (1980). "Polymorphism of DNA double helices". J. Mol. Biol. 143 (1): 49-72. doi:10.1016/0022-2836(80) 90124-2 . PMID 7441761.

[8] Saenger, Wolfram (1984). Principles of Nucleic Acid Structure. New York: Springer-Verlag. ISBN 0-387-90762-9.

[9] R. Curnow and T. Kirkwood, "Statistical analysis of deoxyribonucleic acid sequence data-a review," J. Royal Statistical Soc., vol. 152, pp. 199-220, 1989.

[10] T.C. Bell, J.G. Cleary and I.H. Witten, Text compression (1990).

[11] A. Moffat, Implementing the PPM data compression scheme , IEEE Transactions on Communications, Vol. 38 (11), pp. 1917-1921, November 1990.

[12] Grümbach,S. and Tahi,F. (1994) A new challenge for compression algorithms: genetic sequences. Inform. Process. Manage., 30, 875-886.

[13] Rivals,E., Dauchet,M., Delahaye,J.P. et al. (1996) Compression and genetic sequence analysis. Biochimie, 78, 315-322.

[14] Gusfield, Dan. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, 15 January 1997.

[15] Jeffrey L Hansen, Alexander M Long, Steve C Schultz (1997). "Structure of the RNA-dependent RNA polymerase of poliovirus".Structure 5 (8): 1109-22. doi:10.1016/S0969-2126(97)00261-X .PMID 9309225 .

[16] Chen, X., Kwong, S., Li, M.: A compression Algorithm for DNA sequences and its applications in genome comparison. The 10th workshop on Genome Informtics(GIW'99), pp 51-61, Tokyo, Japan, 1999.

[17] Chen,X. et al. (2000) A compression algorithm for DNA sequences and its applications in genome comparison. In RECOMB 00: Proceedings of the 4th Annual International Conference on Computational Molecular Biology. ACM, New York, pp. 107-117.

[18] Higgs PG (2000). "RNA secondary structure: physical and computational aspects". Quarterly Reviews of Biophysics 33: 199-253. doi:10.1017/S0033583500003620 . PMID 11191843 .

[19] Apostolico A. and Lonardi S.: Compression of Biological Sequences by Greedy Offline Textual Substitution. In proc. Data Compression Conference, IEEE Computer Society Press, pp 143-152, 2000.

[20] Apostolico A. and Lonardi S.: Compression of Biological Sequences by Greedy Offline Textual Substitution. In proc. Data Compression Conference, IEEE Computer Society Press, pp 143-152, 2000.

[21] Bernaola-Galván,P. et al. (2000) Finding borders between coding and noncoding DNA regions by an entropic segmentation method. Phys. Rev. Lett., 85, 1342-1345.

[22] Albà M (2001). "Replicative DNA polymerases". Genome Biol2 (1): REVIEWS 3002. PMID 11178285 .

[23] Adjeroh,D. et al. (2002) DNA sequence compression using the Burrows-Wheeler transform. In Proceedings of the IEEE Computer Society Conference on Bioinformatics. IEEE Computer Society, pp. 303-313.

[24] David J. C. MacKay. Information Theory, Inference, and Learning Algorithms Cambridge: Cambridge University Press, 2003. ISBN 0-521-64298-1.

[25] Cherniavski,N. and Lander,R. (2004) Grammar-based compression of DNA sequences. Computer Science & Engineering Technical Report. University of Washington, 2007-05-02, pp. 1-21.

[26] Hoebeke,M., Chiapello,H., Gibrat,J.F. et al. (2005) Annotation and databases: status and prospects. In Lesk,A.M. (ed), Database Annotation in Molecular Biology, John Wiley & Sons, West Sussex, England, pp. pp. 1-21.

[27] Data Compression: The Complete Reference. 4th Edition. David Salomon (with contributions by Giovanni Motta and David Bryant). Published by Springer (Dec 2006). ISBN 1-84628-602-6.

[28] Hall N (May 2007). "Advanced sequencing technologies and their wider impact in microbiology". J. Exp. Biol. 210 (Pt 9): 1518-25.doi:10.1242/jeb.001370 . PMID 17449817.

[29] Korodi,G. and Tabus,I. (2008) Compression of annotated nucleotide sequences. IEEE/ACM Trans. Comput. Biol. Bioinform., 4, 447-457.

[30] Haiminen,N. et al. (2009) Comparing segmentations by applying randomization techniques. BMC Bioinformatics, 7, 171.

[31] Menconi,G., Benci,V. and Buiatti,M. (2010) Data compression and genomes: a two dimensional life domain map. J. Theoret. Biol., 253, 281-288.

[32] David Salomon, Giovanni Motta, (with contributions by David Bryant), Handbook of Data Compression, 5th edition, Springer, 2011, ISBN 1848829027, pp. 16-18.

[33] Gregory Vey,2012, "Differential direct coding: a compression algorithm for nucleotide sequence data", Database, doi: 10.1093/database/bap013

[34] K.N. Mishra, 2014, “An efficient Horizontal and Vertical Method for Online DNA sequence Compression", IJCA(0975-8887), Vol3, PP 39-45.

[35] R.K.Bharti,2016, et al., "Biological sequence Compression Based on Cross chromosomal properties Using variable length LUT", CSC Journal, Vol 4 Issue 6.

[36] Govind Prasad Arya and R.K. Bharti, " A Compression Algorithm for Nucleotide Data Based on Differential Direct Coding and Variable Length Lookup Table (LUT)," IJCSIT, vol. 3, pp. 4411-4416, 2012.

BIOGRAPHIES

Manju Rani is pursuing M.Tech from Uttarakhand Technical University, Dehradun. Her research area includes image processing, compression and operating system.

Pawan Kumar Mishra is assistant professor in the department of computer science and engineering of Uttarakhand Technical University, Dehradun. He has published various research papers in national and international journals. He has attended a number of conferences and workshops also.