International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 07, July 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

# DNA SEQUENCE COMPRESSION BASED ON SUBSTITUTION TECHNIQUE

**Syed Mahamud Hossein[1], Pradeep Kumar Das Mohapatra[2], Debashis De[3]**

*[1]Research Scholar , [2]Department of Microbiology*

*[1,2] Vidyasagar University ,Midnapur-721102, West Bengal, (India)*

*[3]Department of Computer Science and Engineering,*

*Maulana Abul Kalam Azad University of Technology, BF-142, Sector-I, Kolkata,West Bengal, (India)*

## ABSTRACT

*The volume of DNA sequence are gradually increase twice or more in a year and protection of information in digital form is becoming more important. During storage as well as in transmission, the genome information is being exposed to unauthorized entities unless otherwise adequate security measures are built around the information system. Encryption of large DNA sequence takes time before they can be transmitted, causing considerable delay in successive transmission of information in real-time. In order to minimize the storage space, efficient encryption algorithms are needed. This algorithm is based on the combinations of repeat & genetic palindrome / reverse & genetic palindrome substring substitution and creates online Library file acting as a Look Up Table. The repeat & genetic palindrome / reverse & genetic palindrome substring is replaced by corresponding ASCII character. This substring length depends on user. It can provide the data security, by using ASCII code and on line Library file acting as a signature. This algorithm reduced the storage space and transmission time over internet.*

*To improve the compression rate, the output string is again compressed by WINRAR compressor. This proposed method can approach a compression rate of 1.933940 bit/base.*

***Keywords: DNA sequence, Compression, Repeat, Reverse, Genetic Palindrome, compressor and Security.***

## I. INTRODUCTION

The DNA sequence only consist of 4 nucleotide bases {a, c, g & t}, 8 bits are enough to store each base. However, if one applies standard compression software, they all expand the file with more than 8 bits per base [1], because the regularities in DNA sequences are much subtler. There is also strong biological evidence that supports this claim: it is well known that DNA sequences, especially in higher eukaryotes, contain many approximate repeats [2]; it is also well known that many essential genes have many copies. All this evidence gives more concrete support that the DNA sequences should be reasonably compressible. It is well recognized that the compression of DNA sequences is a very difficult task [3]. However, searching for all exact repeat & genetic palindrome / reverse & genetic palindrome in a very long DNA sequence is not a trivial task. This algorithm take a long time in order to find approximate repeat & genetic palindrome / reverse & genetic palindrome that are optimal for compression. Simultaneously achieving high speed and best compression ratio

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

remains to be a challenging task. Proposed algorithm consists of two phases i) find all exact repeats & genetic palindromes / reverses & genetic palindromes and ii) encode exact repeat & genetic palindrome / reverse & genetic palindrome regions and non-repeat regions. We will discuss details of the algorithm, provide experimental results and compare the results with the one most effective compression algorithm [4]. The string (DNA sequence) is divided into no. of substring [5] for implementing this technique. This String pattern matching [6] method is scanned left-to-right using character shift rule[7-8]. We assume that the string i,e the DNA sequence of 4 nos. of characters from a finite characters set $\sum$ . The pattern of DNA are not exact, we may not remember the exact repeat & genetic palindrome / reverse & genetic palindrome in all place in the long string. Using this technique we can compress DNA sequences to reduce their size and I/O overhead considerably and reduce the time to search patterns, communication time to sender to receiver end directly in compressed DNA sequence with information security. Various compression methods have been extensively studied in the last decade from both theoretical and practical points of view [8-17]. We propose a compression method for DNA sequences best on search method to repeat-genetic palindrome / reverse-genetic palindrome pattern directly in the DNA sequences. The search method is our compressed pattern matching algorithm, which is based on the exact repeat-genetic palindrome / reverse-genetic palindrome subsequences match.

The output of OUR programme, again compressed by WINRAR compressor for reducing the compression rate, which is open source software[18].

## II. METHODS

### 2.1. File type and substring creating

Consider all DNA sequences are in text format. The Sub sequences of repeat-genetic palindrome / reverse-genetic palindrome are generated into words [19]

### 2.2. Encoding procedure

Consider a finite sequence P over the DNA alphabet {a, c, g & t}. An exact repeat-genetic palindrome/reverse-genetic palindrome is a substring in R that can be transformed from another substring in S with edit operation (repeat-genetic palindrome/reverse-genetic palindrome, insertion). We only encode those exact repeat-genetic palindrome/reverse-genetic palindrome that provide profits on overall compression.

This method of compression is as below

a). Run the program and output all exact repeat-genetic palindrome/reverse-genetic palindrome into a list S in the order of descending score;

b). Extract a repeat-genetic palindrome/reverse-genetic palindrome (R*)* with highest score from list S, then replace all R by corresponding ASCII code into another repeat-genetic palindrome/reverse-genetic palindrome list O and place R in library file.

c). Process each repeat & genetic palindrome / reverse & genetic palindrome in S so that there's no overlap with the extracted repeat-genetic palindrome/reverse-genetic palindrome R.

d). Goto step 2 if the highest score of repeat-genetic palindrome/reverse-genetic palindrome in S is still higher than a pre-defined threshold; otherwise exit.

### 2.3.Decoding procedure

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Decoding time, first require on line Library file, which was created at the time of encoding the input file. On the particular value, the encoded input string is decoded and produce the output original file.

## III. PROCEDURE EVALUATION

### 3.1: Accuracy

It is not tolerable that any mistake exists either in compression or in decompression because change in single nucleotide would result in huge change of phenotype.

### 3.2: Efficiency

This algorithm can compress original file from substring length (R) into 1 character for any DNA segment, and destination file uses less ASCII character to represent successive DNA bases than source file.

### 3.3: Space Occupation

Our algorithm read characters from source file and writes them immediately into destination file. It costs very small memory space to store only a few characters.

## IV. EXPERIMENTAL RESULT

This technique test on benchmark data used in [20] and testing purpose we use two sets of data. The compression rate[21], which is defined as (|O|/| *I*|), where |*I*| is number of bases in the input DNA sequence and |O| is the length (number of bits) of the output sequence. The improvement[22] over WINRAR, which is defined as (( Rate_of_WINRAR– Ratio_of_OUR) / Rate_of_WINRAR*100). The compression rate and improvement are presented in table-I.

**Table-I**

| Data set | Sequence Name | Base pair/ File size | Repeat-Genetic palindrome Technique | | | | Reverse-Genetic palindrome Technique | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Using Repeat-Genetic palindrome Technique | Using WINRAR Software | Using Repeat-Genetic palindrome Technique +WINRAR S/w | Improvement | Using Repeat-Genetic palindrome Technique | Using WINRAR software | Using Repeat-Genetic palindrome Technique +WINRAR S/w | Improvement |
| | | | Compression rate ( bits /base) | Compression rate ( bits /base) | Compression rate ( bits /base) | | Compression rate ( bits /base) | Compression rate ( bits /base) | Compression rate ( bits /base) | |
| Data set-I | MTPACGA | 100314 | 3.608190 | 2.347588 | 2.220866 | 7.22% | 3.618398 | 2.347588 | 2.231074 | 7.17% |
| | MPOMTCG | 186608 | 3.598816 | 2.390165 | 2.297865 | | 3.596073 | 2.390165 | 2.295121 | |
| | CHNTXX | 155844 | 3.602166 | 2.383203 | 2.303123 | | 3.605400 | 2.383203 | 2.306357 | |
| | CHMPXX | 121024 | 3.561475 | 2.329058 | 2.194209 | | 3.561475 | 2.329058 | 2.194209 | |
| | HUMGHCSA | 66495 | 3.635401 | 2.339784 | 1.491721 | | 3.635401 | 2.339784 | 1.491721 | |
| | HUMHBB | 73308 | 3.601244 | 2.332951 | 2.226660 | | 3.601244 | 2.332951 | 2.226660 | |
| | HUMHDABCD | 58864 | 3.564555 | 2.294781 | 2.172737 | | 3.564555 | 2.294781 | 2.172737 | |
| | HUMDYSTROP | 38770 | 3.620118 | 2.383492 | 2.315604 | | 3.620118 | 2.383492 | 2.315604 | |
| | HUMHPRTB | 56737 | 3.616828 | 2.324127 | 2.212876 | | 3.616828 | 2.324127 | 2.212876 | |

International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 07, July 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | VACCG | 191737 | 3.597803 | 2.309622 | 2.209151 | | 3.597803 | 2.309622 | 2.209151 | |
| | HEHCMVCG | 229354 | 3.585514 | 2.375541 | 2.301629 | | 3.585514 | 2.375541 | 2.301629 | |
| | Average | | 3.599283 | 2.346392 | 2.176949 | | 3.604722 | 2.346392 | 2.177922 | |
| Data set-II | atatsgs | 9647 | 3.629729 | 2.241525 | 2.035866 | 8.62% | 3.682802 | 2.241525 | 2.088939 | 9.2% |
| | atef1a23 | 6022 | 3.608103 | 2.130853 | 1.843905 | | 3.489870 | 2.130853 | 1.725672 | |
| | atrdnaf | 10014 | 3.571000 | 2.242460 | 2.123427 | | 3.610145 | 2.242460 | 2.162572 | |
| | atrdnai | 5287 | 3.611878 | 1.839984 | 1.711367 | | 3.552865 | 1.839984 | 1.652354 | |
| | celk07e12 | 58949 | 3.582622 | 2.050314 | 1.885562 | | 3.597007 | 2.050314 | 1.899947 | |
| | hsg6pdgen | 52173 | 3.620723 | 2.298200 | 2.163417 | | 3.660744 | 2.298200 | 2.203438 | |
| | mmzp3g | 10833 | 3.623742 | 2.303332 | 2.100987 | | 3.529216 | 2.303332 | 2.006461 | |
| | xlxfg512 | 19338 | 3.581756 | 1.933602 | 1.705657 | | 3.608232 | 1.933602 | 1.732133 | |
| | Average | | 3.603694 | 2.130034 | 1.946273 | | 3.591360 | 2.130034 | 1.933940 | |

**Table shown the compression rate using repeat & genetic palindrome / reverse & genetic palindrome technique**

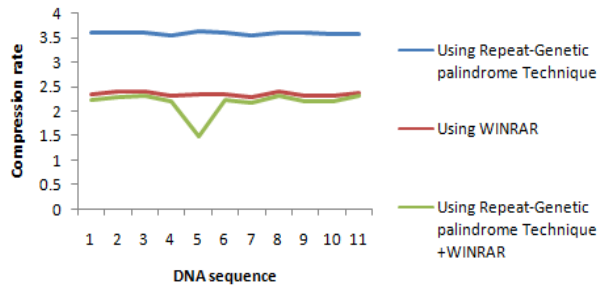


**Fig. 1 : shown the compression rate w.r.t. Repeat-Genetic Palindrome for data set-I**

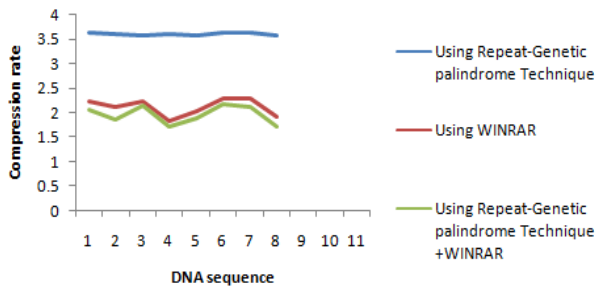


**Fig. 2 : shown the compression rate w.r.t. Repeat-Genetic Palindrome for data set-II**

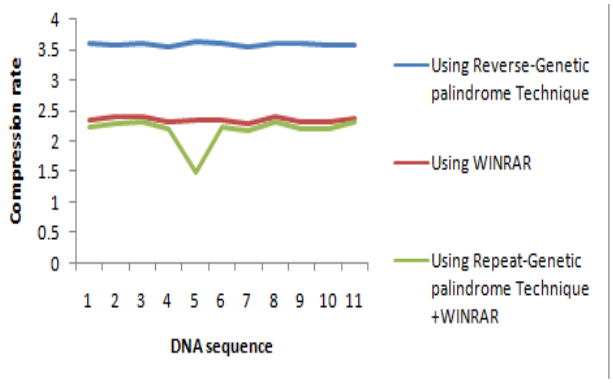


**Fig. 3 : shown the compression rate w.r.t. Reverse-Genetic Palindrome for data set-I**

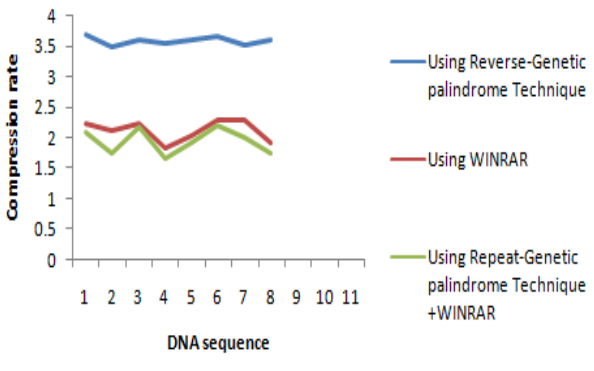


**Fig. 4 : shown the compression rate w.r.t. Reverse-Genetic Palindrome for data set-II**

## V. RESULT DISCUSSION

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

From these experiments, we conclude that internal repeat & genetic palindrome / reverse & genetic palindrome matching patter are same in all type of sources, presented in fig. 1 to fig.4 and plays a key role in finding similarities or regularities in DNA sequences. Output file contain ASCII character with unmatch a,t,g and c, so, it can provide information security which is very important for data protection over transmission point of view. Our algorithm is very useful in database storing. We can keep sequences as record in database instead of maintaining them as file.

**Availability:** The executable file will be available upon request for academics.

## VI. CONCLUSION

We discuss a DNA sequences compression algorithm whose key idea is internal repeat & genetic palindrome / reverse & genetic palindrome. This compression algorithm gives a good model for compressing DNA sequences that reveals the true characteristics of DNA sequences. This method is able to detect more regularities in DNA sequences, such as mutation and crossover, and achieve the best compression results by using this observation. This program achieves a little higher compression rate than that of existing DNA compression algorithms but this algorithm provides the better information security.

Important observation are :

a) Repeat & genetic palindrome / Reverse & genetic palindrome substring length vary from 2 to 5 and no match found in case the substring length becoming six or more.
b) The substring length is three of highly repeated than substring length of four and five, i,e substring length of three is highly compressible over substring length of four and five.

## VII. ACKNOWLEDGEMENT

The authors are grateful to all our colleagues for their interest and constructive criticism of this study.

## REFERENCES

[1] Bell, T.C., Cleary, J.G., and Witten, I.H., Text Compression, Prentice Hall, 1990.
[2] Nour S. Bakr1, Amr A. Sharawi, 'DNA Lossless Compression Algorithms: Review ', American Journal of Bioinformatics Research, 2013 pp 72-81
[3] P Roy, D Dey, S Sinha, D De, Reversible OR Logic Gate Design Using DNA ,Proceedings of Seventh International Conference on Bioinspired Computing, Advances in Intelligent Systems and Computing, Springer, 2013
[4] Syed Mahamud Hossein et al., Lookup Table based Genome Compression Technique-unpublished
[5] Syed Mahamud Hossein et al., DNA Compression Algorithm based on $R^2$ Techniques, Journal of Bioinformatics and Intelligent Control, Vol. 1, pp 1-6,2013
[6] Sun Wu and Udi Manber, 'Agrep-Fast Approximate Pattern-Matching Tools', Usenix-Winter'92
[7] A. Amir, G. Benson and M. Farach, "Let sleeping files lie: Pattern matching in Z-compressed file", Journal of Computer and System Sciences, 52, pp. 299-307, 1996.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

[8] A. Amir, and G. Benson, "Effcient two-dimensional compressed matching", In Proc. CC'92, pp. 279-288,2002.

[9] E. S. de Moura, G. Navarro, N. Ziviani and R. Baeza- Yates, "Direct pattern matching on compressed text", In Proc. 5th International Symp. on String Processing and Information Retrieval, IEEE Computer Society, pp. 90-95, 1998.

[10] M. Farach and M. Thorup , "String-matching in Lempel-Ziv compressed strings", Algorithmica, 20, 388-404, 1998.

[11] T. Kida, M. Takeda, A. Shinohara, M. Miyazaki, and S.Arikawa, "Multiple pattern matching in LZW compressed text. In Proc. Data Compression Conference (DCC'98), IEEE Computer Society, pp. 103-112, 1998.

[12] U. Manber, "A text compression scheme that allows fast searching directly in the compressed file", In Proc. 5th Ann. Symp. on Combinatorial Pattern Matching, Springer-Verlag,pp. 113-124, 1994.

[13] M. Miyazaki, S. Fukamachi, M. Takeda, and T. Shinohara, "Speeding up the pattern matching machine for compressed texts" Transactions of Information Processing, Society of Japan, 39, 2638-2648, 1998.

[14] G. Navarro and M. Raffinot, "A general practical approach to pattern matching over Ziv-Lempel compressed text", In Proc. 10th Ann. Symp. on Combinatorial Pattern Matching, Springer-Verlag, pp. 14-36, 1999.

[15] G. Navarro and J. Tarhio, "Boyer-Moore string matching over Ziv-Lempel compressed text", In Proc. 11th Ann. Symp. on Combinatorial Pattern Matching, Springer- Verlag, pp. 166-180, 2000.

[16] Y. Shibata, T. Kida, S. Fukamachi, M. Takeda, A. Shinohara, T. Shinohara, and S. Arikawa, "Speeding up pattern matching by text compression", In Proc. 4th Italian Conference on Algorithms and Complexity, Springer-Verlag, pp. 306-315, 2000.

[17] Y. Shibata, T. Matsumoto, A. Takeda, T. Shinohara and S. Arikawa, "A Boyer-Moore type algorithm for compressed pattern matching", In Proc. 11th Ann. Symp. On Combinatorial Pattern Matching, Springer-Verlag, pp. 181- 194, 2000.

[18] www.win-rar.com/download.html

[19] S.F. Altschul, W.Gish, W. Miller, E.W. Myers, and D.J. Lipman, 1990, A basic local alignment search tool, J.Mol. Biol. 215 : 403-410

[20] S. Grumbach and F. Tahi, "A new challenge for compression algorithms: Genetic sequences," J. Inform. Process. Manage., vol. 30, no. 6, pp. 875-866, 1994.

[21] Xin Chen, San Kwong and Mine Li, "A Compression Algorithm for DNA Sequences Using Approximate Matching for Better Compression Ratio to Reveal the True Characteristics of DNA", IEEE Engineering in Medicine and Biology,pp 61-66,July/August 2001.

[22]Syed Mahamud Hossein, A.Mukherjee, A.K.Samanta, & P.K.Maity, "A Compression Algorithm for DNA Sequences based on genetic palindrome sequences and its applications in Genome comparison with Information security". International multi-conference on Intelligent Systems, Sustainable, New and Renewable Energy Technology& Nanotechnology February 18 - 20, 2011.