CHAPTER 5

ALGORITHMS AND RESULTS

Chapter 5. Algorithms and Results

The research work involves development of new pattern matching algorithms for DNA sequence analysis. The algorithms use Signal Processing approach for optimization.

The three new algorithms designed for addressing the Bioinformatics issues are:

- Data Reduction for Analysis Processing or Transmission of DNA sequences
- Finding identical reads from a DNA sequencing data
- Recognizing Short Tandem Repeats in DNA sequences

The algorithms are developed to deal with the above mentioned issues. Signal processing approach using Wavelet Transforms has been applied for data reduction, and this reduced data is used in the pattern matching algorithms, to optimize time involved in searching. The following sections are describing the purpose, methodology and results in context of each of the three algorithms.

5.1. Algorithm 1: Data Reduction for DNA sequence Analysis Processing and Transmission

5.1.1. Purpose

High throughput sequencing techniques has led to large scale sequencing of genomes, metagenomes and transcriptome. As a result, sequencing data is growing exponentially. This mounting data needs to be efficiently stored, analysed optimally and transmitted correctly.

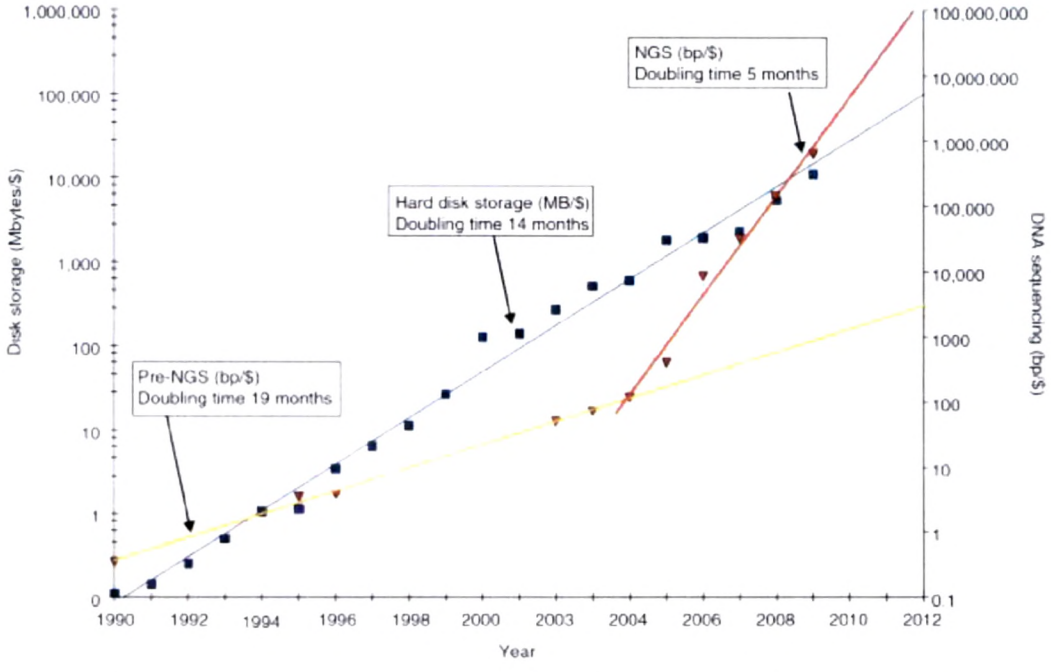


Figure 21. DNA sequencing per dollar is increasing faster than storage capacity per dollar[95]

Applying various techniques for storage or analysis of raw DNA data usually involves string or textual storage or processing, sometimes along with different data structures for string processing. String processing algorithms are space and time intensive and usually involve linear or exponential computational complexity. Hence performing some kind of reduction on these large data sets would be advantageous in reducing the total size of data. The reduction techniques should be lossless in nature, without altering any type of behavioural or structural properties of an organism.

The existing compression or reduction tools applied to DNA sequences are primarily based on the classic Lempel-Ziv algorithm[96], Burrows-

---

95 http://ai.stanford.edu/~serafim/CS374_2011/presentations/lecture6.pdf - Source: Stein 2010

Wheeler transform[97] (bwt), Huffman Coding[98], FM-Index[99] and Wavelet Trees[100]. These algorithms or tools for DNA compression take statistical or dictionary based approaches. Statistical based algorithms need prior knowledge of the data in concern and hence are data dependent. "formatdb"[101] and GRS[102] uses Huffman coding, but deals efficiently only with minor symbols[103]. The Huffman code is an optimal prefix code in the case where exact symbol probabilities are known in advance and are integral powers of 1/2 [104] [105]. The algorithms based on Arithmetic coders are slow in decompression process[106]. Arithmetic Coders[107] and Huffman Coding use probability based compression which is difficult to predict, as DNA sequences comprises of nucleotide bases belonging to an alphabet size of 4, and hence each nucleotide base has almost similar

---

96 Ziv J, Lempel A. A universal algorithm for sequential data compression, IEEE Transaction of Information Theory 1977; 23(3):337-343.
97 Burrows M, Wheeler D: A block sorting lossless data compression algorithm. Tech. Rep. 124, Digital Equipment Corporation 1994.
98 Huffman D.A. A Method for the Construction of Minimum-Redundancy Codes, Proceedings of the I.R.E., September 1952, pp 1098-1102. Huffman's original article
99 Paolo Ferragina, Giovanni Manzini, Veli M¨akinen, and Gonzalo Navarro, "An Alphabet-Friendly FM-index? " In International Symposium on String Processing and Information Retrieval (SPIRE), (2004) pages 150-160
100 O˘guzhan M, Vitter J S, Xu B, Efficient Maximal Repeat Finding Using the Burrows-Wheeler Transform and Wavelet Tree, IEEE/ACM Transactions On Computational Biology And Bioinformatics.
101 http://www.ncbi.nlm.nih.gov
102 Congmao Wang1 and Dabing Zhang, A novel compression tool for efficient storage of genome resequencing data
103 Hisahiko Sato1 Takashi Yoshioka, Akihiko Konagaya3 Tetsuro Toyoda,DNA Data Compression in the Post Genome Era, Genome Informatics 12: 512-514 (2001)
104 Ana Balevic, Lars Rockstroh, Marek Wroblewski, and Sven Simon, Using Arithmetic Coding for Reduction of Resulting Simulation Data Size on Massively Parallel GPGPUs
105 Sayood, K., Lossless Compression Handbook. Academic Press (2003)
106 Hisahiko Sato1 Takashi Yoshioka, Akihiko Konagaya3 Tetsuro Toyoda,DNA Data Compression in the Post Genome Era, Genome Informatics 12: 512-514 (2001)
107 Introduction to Arithmetic Coding - Theory and Practice, Amir Said. Academic Press, April 21, 2004,
Chapter in Lossless Compression Handbook by Khalid Sayood, Approved by External Publicaations

probability[108]. The other algorithms applied in tools like GSR[109] or BioCompress[110] take into account characteristics of DNA like reverse complement or point mutation or characteristic structures of sequences like Single Nucleotide Polymorphisms (SNP) or repeat regions[111]. But SNP map may not be available as SNP may not be the common bi-allele[112], Index of repetitiveness may vary in different genomes[113], or organisms have very little variation and hence may become difficult to compress. If, a *priori* knowledge of the characteristics of the given DNA sequence is not available, then statistical approach becomes irrelevant and hence, the problem of data compression becomes difficult[114]. Reference sequence is used for data compression of DNA[115]. Another algorithm used in software tool 'coil'[116] uses Levenstein distances and encoding trees for compressing entire database of DNA data, but takes into account an initial DNA sequence to compare length-k substrings amongst all the other DNA sequences in database. This again is data

---

108 Toshiko Matsumoto Kunihiko Sadakane, Hiroshi Imai, Biological Sequence Compression Algorithms, Genome Informatics 11: 43-52 (2000) 43

109 Chen, X., Kwong, S., and Li, M., A compression algorithm for DNA sequences and its applications in genome comparison, Genome Informatics , 10:52-61, 1999

110 Grumbach, S. and Tahi, F., A new challenge for compression algorithms: genetic sequences, Information Processing and Management, 30:875-886, 1994

111 Kenny Daily, Paul Rigor, Scott Christley, Xiaohui Xie, Pierre Baldi, Data structures and compression algorithms for high-throughput sequencing technologies, BMC Bioinformatics 2010, 11:514

112 Samantha Woodward, A Critical Analysis of DNA Data Compression Methods BIOC 218, Winter 2011-2012

113 Bernhard Haubold andThomas Wiehe, How repetitive are genomes?, BMC Bioinformatics 2006, 7:541 doi : 10.1186/1471-2105-7-541

114 Ziv J, Lempel A. A universal algorithm for sequential data compression, IEEE Transaction of Information Theory 1977; 23(3):337-343

115 Markus Hsi-Yang Fritz, Rasko Leinonen, Guy Cochrane, and Ewan Birney, Efficient storage of high throughput DNA sequencing data using reference-based compression, Genome Research, 21:734-740_2011 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/11

116 W Timothy J White and Michael D Hendy, Compressing DNA sequence databases with coil, BMC Bioinformatics 2008, 9:242 doi:10.1186/1471-2105-9-242

dependent and will work efficiently only if all the sequences are nearly similar. For widely varied sequences in the database, the algorithm may prove inefficient[117]. Techniques implemented on hardware based acceleration device, have also been used for data compression to accelerate DNA sequence alignment[118], but these techniques need special purpose devices involving costs.

The research work, presents an *ab initio* method of data transforms which can be applied for optimal analysis. The algorithm emphasizes on the use of Wavelet Transforms particularly the *Haar Wavelets* for data reduction of genomic data, principally the DNA sequences. After single level of Wavelet Transformation, there are always N elements, same as the original input sequence. But the fact is that we do not need all N elements when we want to compare two sequences or need to perform sequence analysis. For local alignment, fine comparison is needed and this can be achieved through detailed co-efficient C_d, which is generated after performing *Haar* Wavelet Transforms. For global alignment of two DNA sequences, the coarse components i.e. approximate co-efficient C_a are enough for comparison. So, in all N/2 elements are sufficient for actual use. Thus, the algorithm proposes to reduce the number of elements used for further analysis, with an *apriori* approach. The content and positional information is preserved even after Wavelet

117 W Timothy J White and Michael D Hendy,Compressing DNA sequence databases with coil, BMC Bioinformatics 2008, 9:242 doi:10.1186/1471-2105-9-242

118 Al Junid, S. A. M. Tahir, N.M. Haron, M.A. Abd Majid, Z., Idros, M.F. Osman, F.N., Development and Implementation of Novel Data Compression Technique for Accelerate DNA Sequence Alignment Based on Smith-Waterman Algorithm, IJSSST, Vol. 11, No. 3 ISSN: 34 1473-804x online, 1473-8031

transform is performed on data which is used for the analysis[119] [120]. Since, N/2 elements are sufficient to represent the original sequence, for analysis purpose, one can work on this reduced data. Thus, the time complexity of Wavelet transforms is *O(log n)*, n being the length of the input sequence. Also, use of *Haar* Wavelets is proved to be memory efficient.

This concept of data reduction of DNA sequences using *Haar* Wavelets have been proven to be efficient in design and development of the two other algorithms, which are part of research work and are discussed in later part of this chapter.

5.1.2. Methodology

Algorithm:

1) Read the Fasta File that contains DNA sequences.
2) For each DNA sequence, assign binary values to nucleotide bases, thus converting nucleotide sequence into single indicator sequence.
3) Combine four binary values and put it as a single byte. Thus compressing the sequence one-fourth of the original sequence i.e. reduces the sequence by 2^2.
4) Perform four levels Haar Wavelet Transform on Numerical Sequence, which in turn reduces the sequence by 2^4.
5) Perform Steps 2 to 4 for all sequences in the Fasta File.

---

119 Mamta C. Padole, B. S. Parekh, D. P. Patel, Signal Processing Approach for Recognizing Identical Reads From DNA Sequencing of Bacillus Strains, IOSR Journal of Computer Engineering, Mar-Apr 2013, pp 19-24

120 Mamta C. Padole, Recognizing Short Tandem Repeat Regions in Genomic Sequences Using Wavelet Transforms, unpublished

6) Validate the total reduction, which is 2^6 times of the original sequence.

7) At each level of transform, if one removes repeating sequences, it would give further reduction of data.

The suggested algorithm for creating a compressed form of a given sequence is applied as follows:

Read the Fasta file which contains the sequence $S = \{s_1, s_2, \ldots s_{i\ldots}s_n\}$ where each $s_i \in \sum = \{A,C,G,T\}$,

Where,

$\sum$ is an alphabet set for DNA sequences with alphabet size = 4, and

i = 1, n where n is the length of S.

Convert the nucleotide sequence S into its numerical representation X. The single indicator sequence using 2-bit binary representation of {A, C, G, T} as {00, 01, 10, 11} is used. This binary representation of the nucleotide bases is than combined into a group of four. Each of these four binary representations of the four nucleotide bases is than stored in a single byte. Thus, only one byte is used to represent four consecutive nucleotide bases. The use of binary representation in a single byte reduces the computational overhead to 25% compared, to the other representation of nucleotide sequences like dipole moments[121] or EIIP

[121] J. K. Meher, M. R. Panigrahi, G. N. Dash, P. K. Meher, "Wavelet Based Lossless DNA Sequence Compression for Faster Detection of Eukaryotic Protein Coding Regions ," *I.J. Image, Graphics and Signal Processing 2012, 7, 47-53*

values[122]. Only numerical representations can be applied for wavelet transformations.

Thus, if a given nucleotide sequence is:

$S = \{s_1, s_2, ...s_i\}$ = GTGCAGGATGCTGCAA

Where,

i = [1, n], n being length of nucleotide sequence

Then, its numerical representation i.e. a Digital signal can be given as:

$X = \{x_1, x_2, ... x_j\}$ = 185, 40, 231, 144,

Where,

j = [1, m], m being length of numerical representation of binary values, as a single byte for four nucleotides.

$\therefore$ m = 1... n/4

Perform first-level decomposition of this digital signal X, using Haar Wavelet Transform. This would reduce the signal into one-half of the original numerical representation of the nucleotide sequences. Thus 50% compaction is achieved by single level wavelet transform.

The 1st level Haar Wavelet Transform for the above digital signal X is:

$$Ca = \frac{225}{\sqrt{2}}, \quad \frac{375}{\sqrt{2}}$$

[122] A. S. Nair and S. P. Sreenathan, "A Coding Measure Scheme Employing Electron-Ion Interaction Pseudopotential (EIIP)," Bioinformation, Vol. 1, No. 6, 2006, pp. 197-202

$$Cd = \frac{145}{\sqrt{2}}, \qquad \frac{87}{\sqrt{2}}$$

Further decomposition of a signal, to 2nd level, 3rd level and 4th level would further reduce the size of the signal by two, at each level of transform.

The decomposition of a signal is possible until there is only one element remaining or the resolution level is $2^k \approx m$ where k is the level of decomposition and m is the number of elements in the signal.

Thus, if the length of a signal is 8, then for such signal three levels of decomposition are possible, because $2^3 = 8$.

The wavelet transforms are also proved to provide lossless compression, when no elements of transforms are discarded. Hence, the transformed signal would contain same information as of the original signal, but in reduced form.

Thus, applying repeated wavelet transforms reduces the signal size and hence requires less storage space compared to storing the original nucleotide sequence. This reduced size would also reduce the time involved in processing the transformed data.

Moreover, searching on the transformed signal is equivalent to searching on a reduced data set of the original sequence and hence is more efficient. If the correct mapping/transformation is applied, the Parseval Theorem implies that frequency distance FD is less than or equal to edit distance ED (Distance preserving transformation)[123].

---

[123] Aghili, Alireza A, Agrawal, D El Abbadi, A, "Sequence Similarity Search Using Discrete Fourier and Wavelet Transformation Techniques," Postprints, UC Santa Barbara, 2005
http://www.escholarship.org/uc/item/9w7094b9

With every transform, we are discarding half the values as per Nyquist's rule[124] and hence optimizing the search, with time complexity of *O(log n),* excluding other overheads.

With the Haar Wavelet Transforms, the approximate co-efficient C_a i.e. running average represents the trend; the other sub-signal, C_d i.e. a running difference, represents fluctuations[125]. Hence, Haar wavelets can be used to transform the signals for data analysis too, besides applying it for data reduction or compaction.

Thus, use of this algorithm does reduction at two levels. Initially converting four nucleotide bases into a single byte through binary representation compresses the data to one-fourth. Further, applying wavelet transforms four times reduces the data to $1/2^4$ times i.e. one-sixteenth of the numerical representation of the original nucleotide sequence. Thus, the proposed algorithm applies reduction twice i.e.

The length C_σ of the compressed signal C becomes,

$$C_\sigma = X_\sigma * 1/2^4$$

$$C_\sigma = X_\sigma /16$$

Where,

X_σ is the length of the Digital Signal X, which is Numerical form of original DNA sequence S

---

[124] D. Lee Fugal, Conceptual Wavelets in Digital Signal Processing, Space and Signals Technologies, 2006

[125] I. Daubechies, "Orthonormal Bases of Compactly Supported Wavelets,"Comm. Pure Appl. Math, Vol 41, 1988, pp. 906-966

But, $X_\sigma = S_\sigma * ¼$

Where,

S_σ is the length of the original DNA sequence S

$\therefore C_\sigma = (S_\sigma /4)/16$

$\therefore C_\sigma = (S_\sigma /64)$

$\therefore C_\sigma = (S_\sigma * 1/64)$

So, 64 times compaction of data can be achieved using the proposed algorithm.

Storing this compact data would result in storage efficiency. It will be much optimal for any type of analysis too.

Overall, the complexity of the proposed algorithm applied for reduction would be linear.

Complexity of proposed algorithm for reduction = Complexity of converting nucleotide sequence to numerical representation + Complexity of Wavelet transform

$= O(n) + O(\log n)$

$= O(n)$.

5.1.3. Result And Discussion

Example-1: Example of Synthetic Nucleotide sequence, before and after applying the proposed algorithm. The sequence can be perfectly reconstructed as demonstrated in Fig. 1.

If, in a given FASTA file containing several DNA sequences, if some one Original Sequence (Total 120 characters) S_i is given as:

TTTGGGGCGTGCAGGATGCTGCAAAACGTTATTTCGGCAAGAATGCGAGCCAACT
GGATGCCAGTGAAGGTGCTATTTAGGGATGTTACGTAACCCGACGTATTATAATA
CCGGCCGTGA

Numerical Representation through Binary Representation of {A, C, G,T} as {00, 01, 10, 11} respectively and storing four consecutive characters in binary form as a single byte.

The following are the decimal values of the converted binary form: (Total 30 elements i.e. 4 times reduction of S_i) i.e. Digital Signal X_j is :

254, 169, 185, 40, 231, 144, 6, 243, 246, 144, 131, 152, 148, 30, 142, 82, 224, 174, 115, 242, 163, 188, 108, 21, 134, 207, 48, 197, 165, 184

After applying Fourth Level Wavelet Transform, Data appears (Total 2 elements i.e. 2^4 times thus 30/16 ~ 30/15 = 2 elements) as C_k:
49.2500000000001, -12.2500000000001

The graphical representation of the numerical form of original sequence i.e. the digital signal, transformed signals at 4-levels of decompositions and the reconstructed sequence, is given in Figure 22.
The X axis represents the Time or Space Localization and the Y axis represents the Scale or Frequency Localization. Figure 22 represents the output of the proposed algorithm for a sample sequence as explained in Example 1. The four level decompositions clearly display reducing length trend, each time being the half of the previous signal,

using multi-level Haar Wavelet Transform. This is visible in the graph with label "level2 decomposition" is having half the length on time-axis of "level1 decomposition" graph. Similarly if you observe the time-axis of graph "level4 decomposition" its length is only 2, which is 1/15 ~ $1/16^{th}$ of the "level1 decomposition"

Moreover the signal can be perfectly reconstructed using multi-level Inverse Haar Wavelet Transform. The graph labeled "reconstructed seqn" is exactly same as the graph labelled as "original seqn". The graph labeled "reconstructed seqn" clearly represents that the signal can be perfectly reconstructed, as it is exactly same as the graph labeled as "originalseqn".

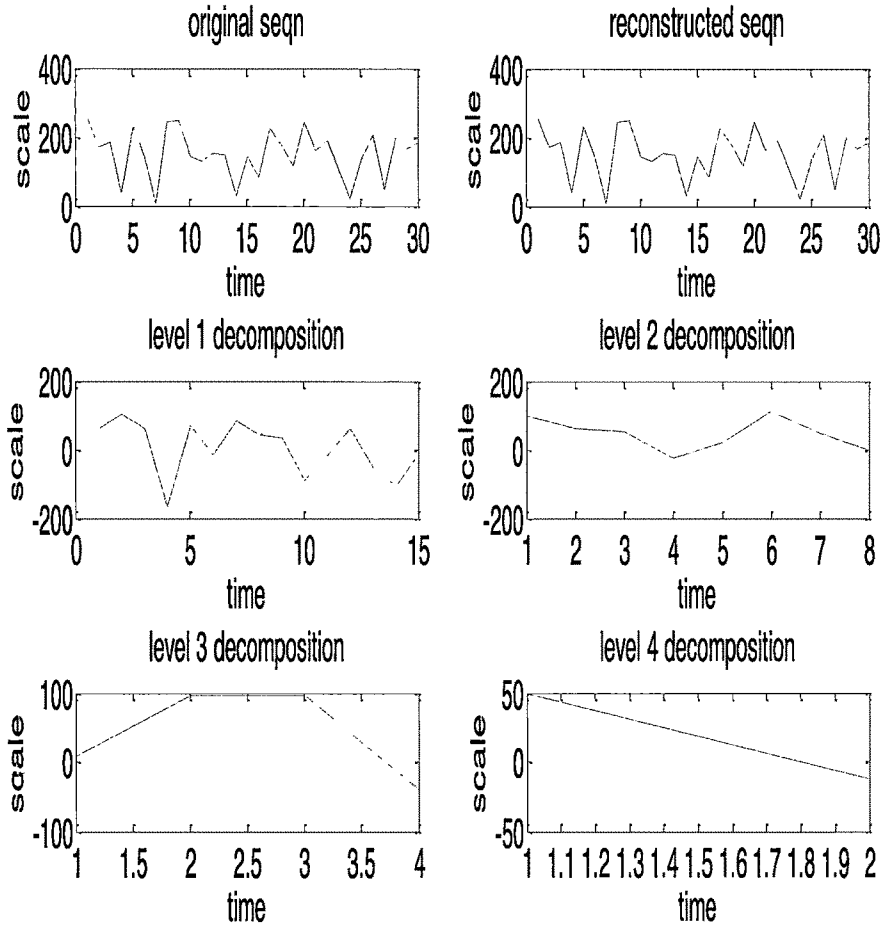


Figure 22. Visual representation of four level decompositions displaying reducing length trend, of a sequence as explained in Example 1.

Example-2: SRA Read of Metagenomics sequence having Accession Number >gnl|SRA|SRR000675.1.2 EXHS9OF01EH7NX.2 is used to apply the proposed algorithm (As shown in Figure 23).

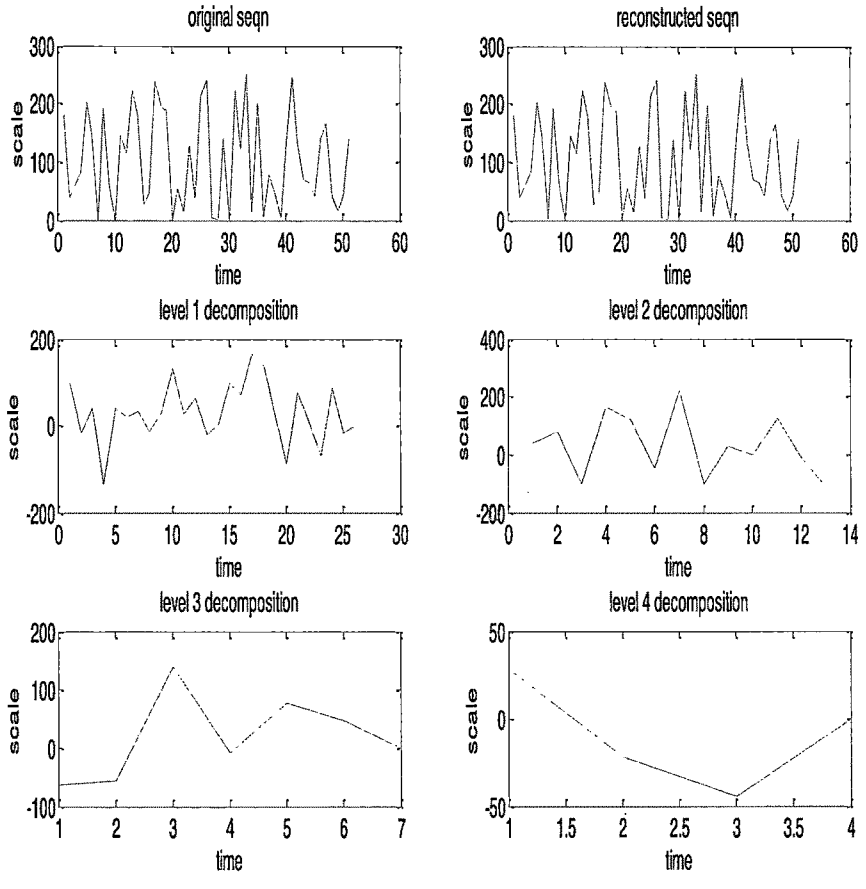


Figure 23. Graph representing Data Reduction upto nearly 64 times for metagenomic sequence with Accession No. SRR00675.1.2

Example-3: Contig 46 of E. Coli K-12 strain having Accession Number gi|305690406|gb|AEFE01000046.1 is used to apply the proposed algorithm. (See Figure 24)

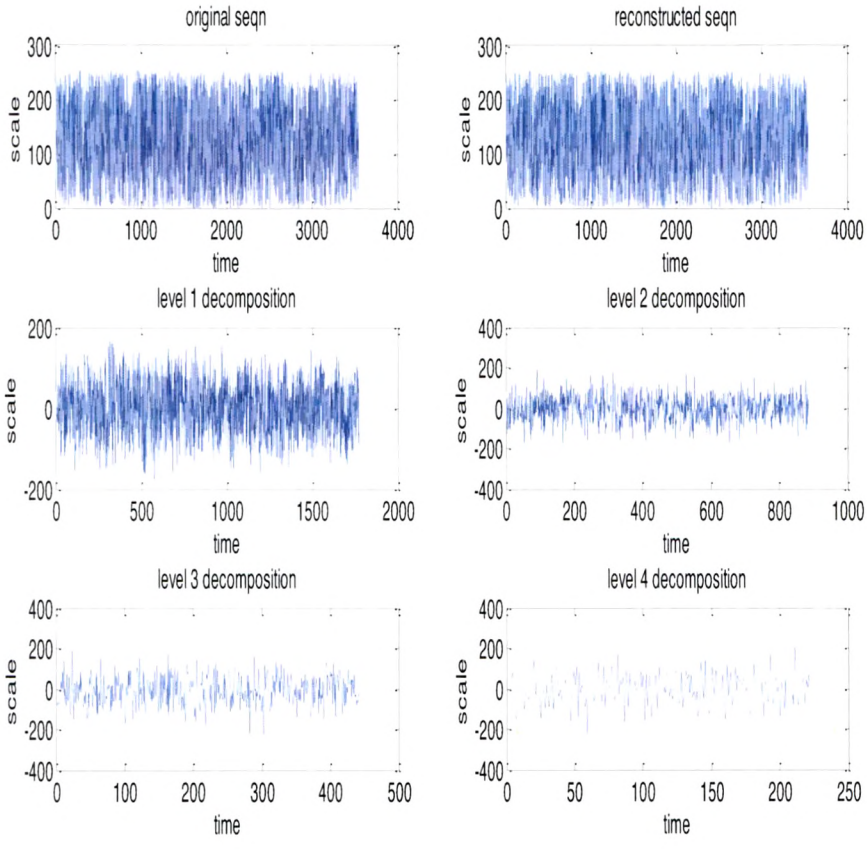


Figure 24. Graph representing Data Reduction upto nearly 64 times for Contig 46 of E.Coli K12 strain

Example-4: Chromosome 7 of Homo Sapiens having Accession Number gi|224514930|ref|NT_007741.14 is used to apply the proposed algorithm. (See Figure 24)

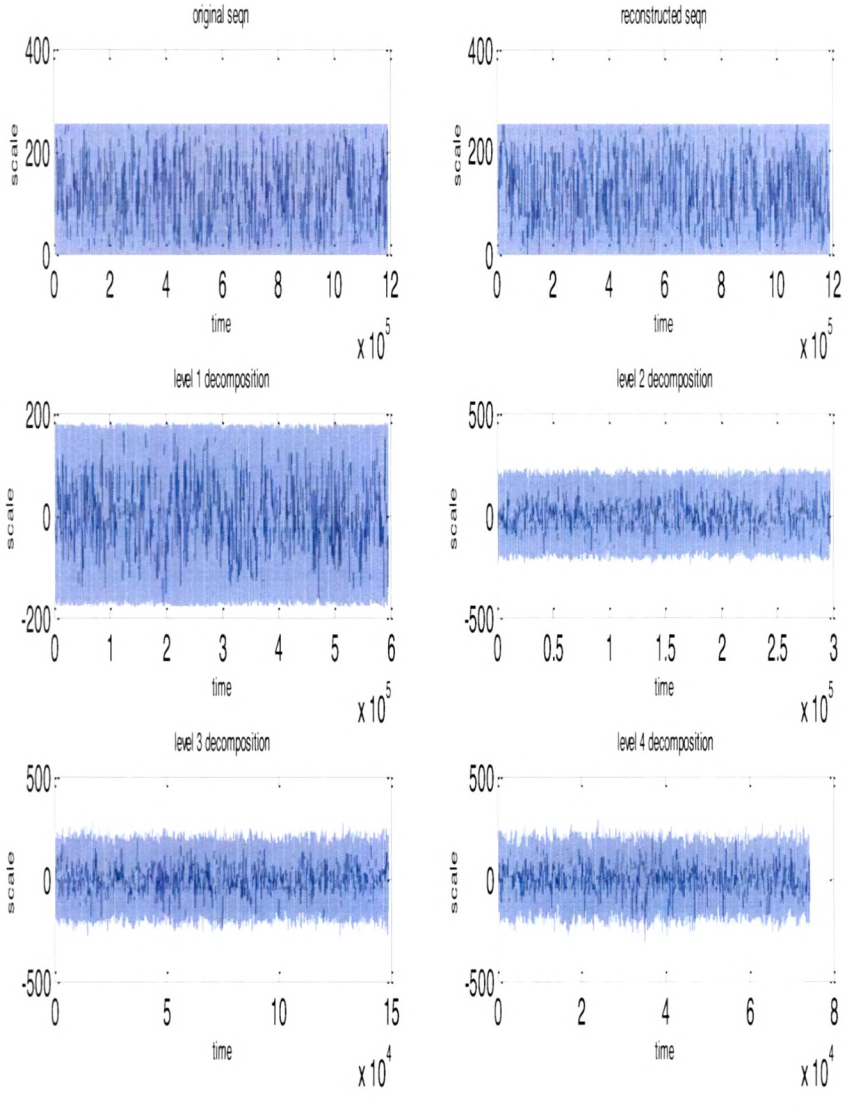


Figure 25. Graph representing Data Reduction upto nearly 64 times for Chromosome 26 of Homo Sapiens

Example-5: Chromosome 3 of of Caenorhabditis Elegans having Accession Number gi|392976480|ref|NC_003281.9| is used to apply the proposed algorithm. (See Figure 26)

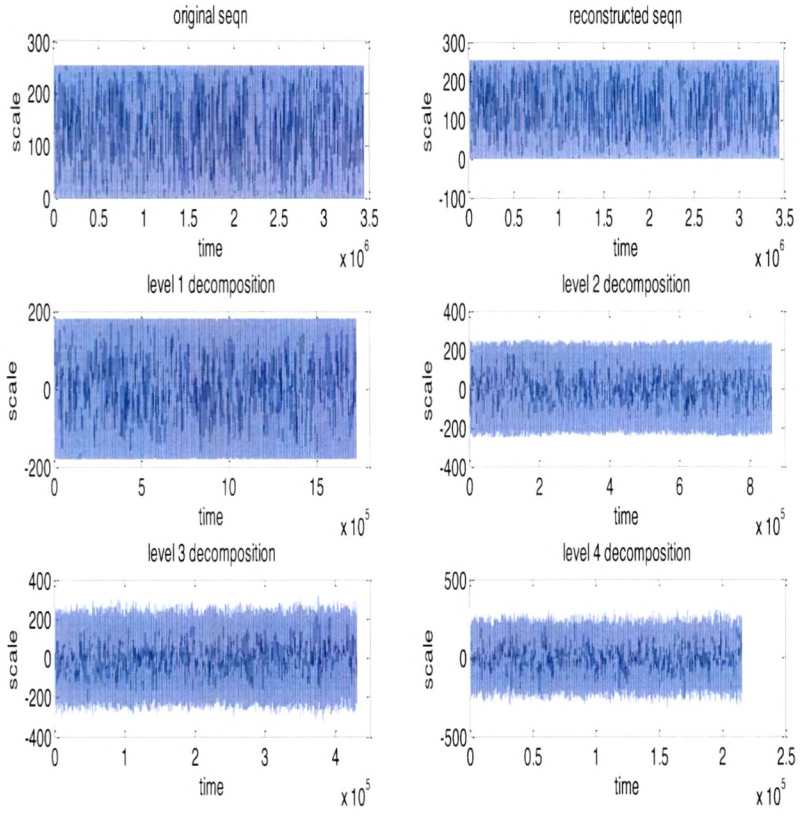


Figure 26. Graph representing Data Reduction upto nearly 64 times for Chromosome 3 of C. Elegans

Thus, any type of Nucleotide sequence may be used for reduction through application of the proposed algorithm. The sequence can be perfectly reconstructed, and can be compressed to almost 64 times, as demonstrated in Figure 27.

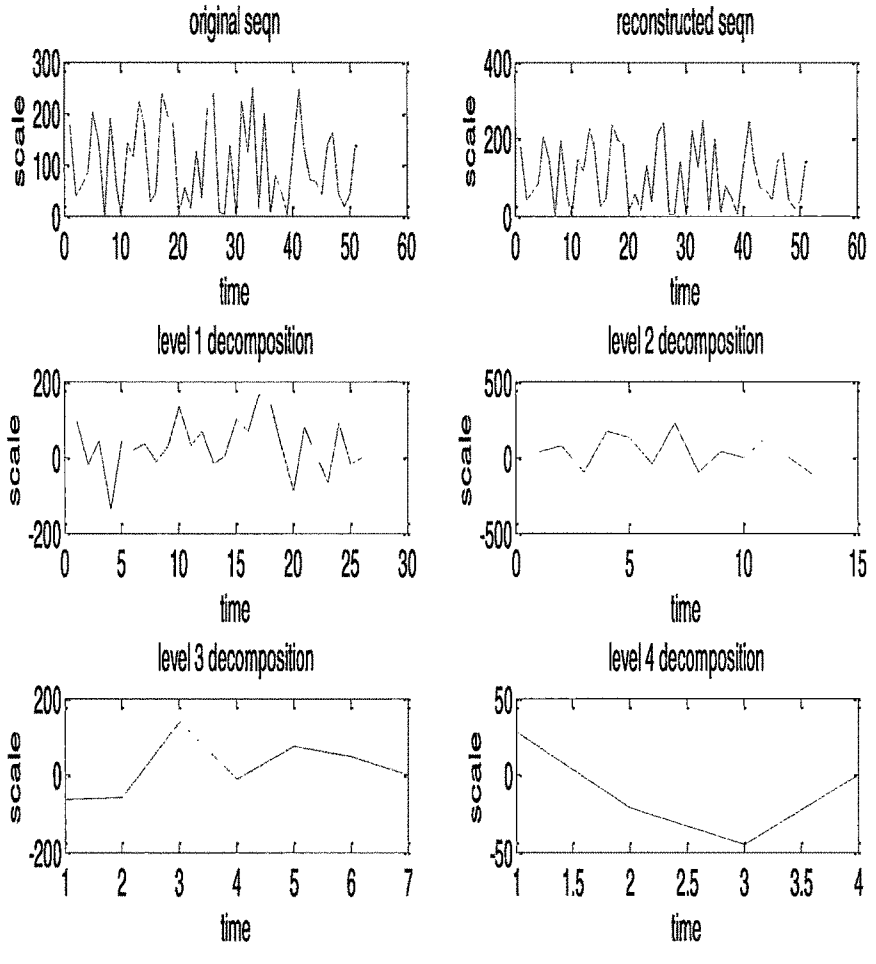


Figure 27. Visual representation of compressed form metagenomics sequence having Accession Number >gnl|SRA|SRR000675.1.2 EXHS9OF01EH7NX.2, upto 64 times reduction, using proposed algorithm applying multi-level Haar Wavelet Transform

Table 7, clearly indicates that the algorithm can be applied for nearly 64 times of reduction of DNA sequences.

Table 7 and Table 8, and the all the Graphs presented for the examples considered for this algorithm's discussion, clearly supports the efficiency of the proposed algorithm because it has been applied for variety of data like reads, contigs as well as large chromosomes. The cases considered for reduction, show nearly 64 times of reduction ratio which is worth for a big data of genomics.

The computational time was calculated for an experiment conducted on regular laptop with Intel Core2 Duo CPU, T6500 @ 2.10 GHz x 2.10 GHz processor and 32-bit Windows 7 Ultimate Operating System with 4.00GB RAM using MATLAB R2009a version 7.0.8.

The computing environment did not consider any special arrangement; instead the experiments were performed, while all routine services including Oracle Database service and Apache Tomcat Application Server was running on an average personal machine.

This type of environment used for computing, simply proves that special purpose computer or clustered computing environment is not essential for executing this algorithm. It can be smoothly executed on a simple desktop machine too. The computational time can be optimized to a great extent, if special purpose computers are used instead of shared resources.

Thus, the proposed algorithm can be applied on any form of DNA or genomic data.

Table 7. Represents the Reduction Ratio of Original Sequence and Compressed Sequence after binary representation and after four levels of Wavelet Transforms, as proposed in the algorithm

Accession Number	*Description of Sequence*	*Original Nucleotide Sequence Size (In base pairs)*	*Size in Decimal values of Binary Form of Nucleotide (In Number of elements)*	*Size 4th level Wavelet Transformed Sequence (In Number of elements)*	*Reduction Ratio*
(a)	*(b)*	*(c) = S_i*	*(d) = X_j*	*(e) = C_k*	*(f) = (c)/(e)*
-	Sample Read	120	30	2	60.5
gnl|SRA|SRR000675.1.2 EXHS9OF01EH7NX.2	Metagenomic Read	208	51	4	52
gi|305690406|gb|AEFE01000046.1|	Contig 46 of E.Coli K12 strain	14516	3540	222	65.387
gi|224514930|ref|NT_007741.14| Homo sapiens chromosome 7 genomic contig, GRCh37.p5 Primary Assembly	Chromosome 7 of Homo Sapiens	47580 40/ 67973 59	11895 07/ 16993 40	74345	63.999

gi|392976480|ref|NC_003281.9|	Chromosome 3 of Caenorhabditis Elegans	13980611	3445924	215371	64.914
gi|194719403|ref|NC_007327.3|NC_007327	Chromosome 26 of Bos Taurus	51750744	12937686	808606	63.999

Table 8. Represents Computational time for the proposed algorithm of Data Reduction using Wavelet Transforms

Accession Number	*Description of Sequence*	*Time to compute Wavelet Transforms (in secs)*	*Time to execute the entire algorithm, including plotting of graphs (in secs)*
(a)	*(b)*	*(c)*	*(d)*
-	Sample Read	0.0273	0.2004
gnl|SRA|SRR000675.1.2 EXHS9OF01EH7NX.2	Metagenomic Read	0.0043	0.1004
gi|305690406|gb|AEFE01000046.1|	Contig 46 of E.Coli K12 strain	0.0063	0.1120

gi|224514930|ref|NT_007741.14	Chromosome 7 Genomic Contig of Homo Sapiens	0.9362	2.1011
gi|392976480|ref|NC_003281.9|	Chromosome 3 of Caenorhabditis Elegans	2.7350	5.8772
gi|194719403|ref|NC_007327.3|NC_007327	Chromosome 26 of Bos Taurus	10.8639	23.0478

The algorithm used for reduction of DNA sequence data has proved to be very effective in further analysis of DNA sequences. The application of this data reduction algorithm alongwith few of its variations, has been presented in the following two algorithms, which are used to identify duplicate reads from pyrosequencing data and to recognize the Short Tandem Repeat regions in the given DNA sequences.

The purpose of applying this reduction algorithm is to further apply it to optimize the time involved in transmission of data, when applying distributed computing for the DNA sequence analysis. Hence, these three algorithms are partially the work towards Distributed Computing. Reduced data is used to optimize the time for DNA sequence analysis done using Distributed Computing.

5.2. Algorithm 2: Recognizing Identical Reads

5.2.1. Purpose

DNA sequencing is the method of identifying the arrangement and order of nucleotides in a DNA sequence. The conventional widely used method of sequencing, the Sanger sequencing, implemented chain termination with di-deoxynucleotides[126], but has limitations in terms of throughput and cost of large genome sequencing[127]. The other methods are sequencing-by-hybridization (SBH), nanopore-sequencing and sequencing-by-synthesis[128]. "Sequencing-by-synthesis" involves taking a single strand of the DNA to be sequenced and then synthesizing its complementary strand enzymatically. The process of DNA sequencing requires the library generation as one of the steps, which enable amplification of DNA[129] material which are available for sequencing. This process of amplification has a possibility of biased amplification of a DNA template causing large number of reads of same segment of DNA generated multiple times and thus causing large number of identical reads. It is suggested that special attention should be paid to potential biases[130] introduced by these identical reads, especially in the cases of analyzing quantification and transcriptome profiling sequence data.

As a part of the research work, an *a priori method* has been developed, for recognizing identical reads, which does not require any mapping

126 Sanger, F.et al. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. 1977, 74: 5463-5467
127 Ronaghi, M. Pyrosequencing Sheds Light on DNA Sequencing Genome Res. 2001, 11: 3-11
128 Metzker, M. L. et al. Emerging Technologies in DNA Sequencing. Genome Res. 2005, 15: 1767-1776
129 Ronaghi, M. Pyrosequencing Sheds Light on DNA Sequencing Genome Res. 2001, 11: 3-11
130 Dong, H. et al. Artificial duplicate reads in sequencing data of 454 Genome Sequencer FLX System Acta Biochim Biophys Sin 2011, 43: Issue6: 496-500

reference for recognizing identical read, nor does it need to compare any string pattern as an input parameter for comparison[131] neither does it use clustering on basis of seeds[132]. The research work here presents the use of a heuristic approach of signal processing as a recognition criterion, for detecting identical reads from DNA sequencing reads, including exact and near exact identical reads. This paper emphasizes on the use of efficient Wavelet Transforms particularly the *Haar Wavelets* for identifying these identical reads. The time complexity of Wavelet transforms is *O(log n)*, n being the length of the transformed sequence.

5.2.2. Methodology

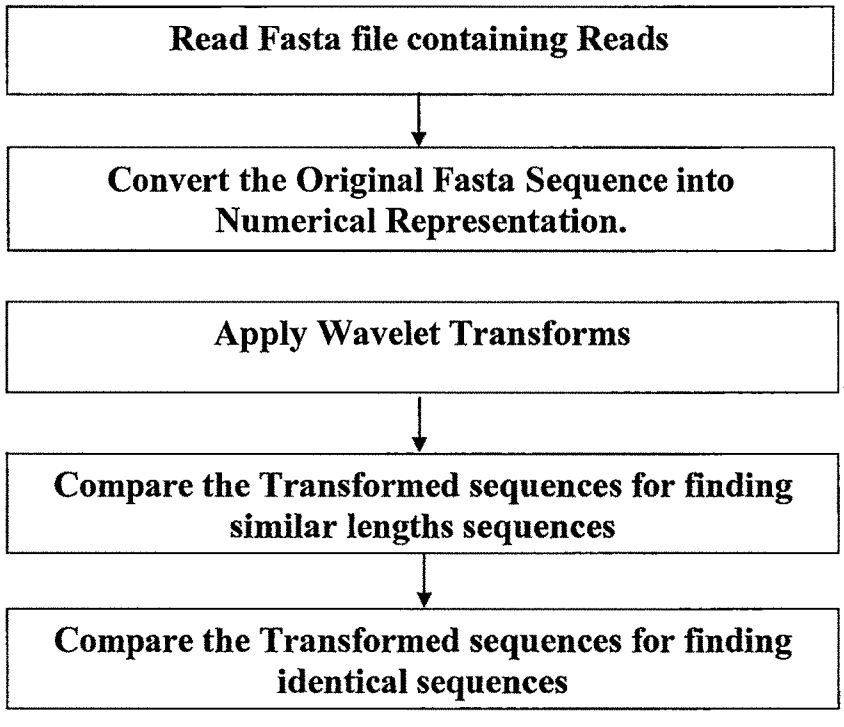


Figure 28. Steps to Identify the Identical Reads

131 Dong, H. et al. Artificial duplicate reads in sequencing data of 454 Genome Sequencer FLX System Acta Biochim Biophys Sin 2011, 43: Issue6: 496-500

132 Gomez-Alvarez V. et al. Systematic artifacts in metagenomes from complex microbial communities. ISME J, 2009, 3 : 1314-1317

Algorithm 2(a) – Algorithm to Recognize Identical Reads

The suggested algorithm for recognizing identical reads from the set of DNA sequenced reads is as given in

Figure 28 alongwith the explanation thereafter.

Steps to Recognize Identical Reads from Pyrosequenced data:

4) Read a FASTA file containing pyrosequenced reads
5) For Each Read, convert the DNA sequence form of Read into a Numerical Representation i,e, convert it into Digital Signal form, so that it can be considered for Signal Processing
6) For Each Digital Signal of the Read, Perform Multi-level Discrete Wavelet Transform
7) Compare a Wavelet Transformed Signal with another Wavelet Transformed Signal for same sequence length.
8) If the Lengths of two transformed signals are same then compare them for similarity.

In a given Fasta file which contains the several sequences S_i's where each $S_i = \{s_1, s_2, \ldots s_n\}$ where $s_i \in \Sigma = \{ A,C,G,T \}$, i = 1, n and n is the length of S_i. Convert the nucleotide sequence S_i into its numerical representation X_i. The single indicator sequence using Electron-ion interaction pseudo potentials - EIIP property of nucleotides, is used for numerical representation.

Table 9. EIIP values for Single Indicator of Sequence Representation

Nucleotide Base	*EIIP value*
A	0.1260
C	0.1340
G	0.0806
T	0.1335

The use of EIIP values for single indicator sequence representation reduces the computational overhead by 75% compared to the conventional four-base binary sequence representation of nucleotide sequence[133]. Only numerical representations can be applied for Wavelet transformations. The next step is to perform multi-level Haar Wavelet transforms on the numerical representation of the sequences. We performed four-level Haar Wavelet transforms on the sequences. The Haar Wavelet Transform applied up to fourth level, reduces the length of the original sequence to one-eighth. This reduced length of transformed sequences can be efficiently used for comparison. Compare the length of transformed sequences, to check whether the sequences are comparable. If the lengths of the transformed sequences are same, then the element by element equality of the two transformed sequences for finding the identical reads is performed. Thus, data-reduction without loss of information using Wavelet transform is applied to recognize identical reads. If a single element of a four-level Haar Wavelet Transform is found to be equal, it means, eight nucleotide bases in a

133 A. S. Nair and S. P. Sreenathan, "A Coding Measure Scheme Employing Electron-Ion Interaction Pseudopotential (EIIP)," Bioinformation, Vol. 1, No. 6, 2006, pp. 197-202

given read are found to be similar. Thus it is much efficient to perform a single comparision on signal processed data, instead of eight comparisions while implementing string processing. Since, the computational complexity of Haar Wavelet is O(log n), it is much faster than any other string processing based methods of finding identical reads. Here we do not consider the other overheads like reading the FASTA file etc. Haar transforms are also memory efficient, as computations are performed in place.

5.2.3. Result and Discussion

The algorithm defined is tested on the Short Reads Archive (SRA) data. The SRA data is downloaded from NCBI site ftp://ftp.ncbi.nlm.nih.gov/sra/ containing short reads. The sequenced reads are for various strains of Bacillus. Table 10 and Table 11 show the results of Identical Reads recognized using Wavelet Transforms. The tables represent the output in terms of reads as well as nucleotide base pairs.

Table 10. Result Showing Number and Percentage of Identical Reads Recognized Using the Wavelet Transforms based Algorithm from various Strains of Bacillus

SRA Accession No.	*Total No. of Reads*	*Total No. of Copies of Identical Reads*	*Total Percentage of Identical Reads (%)*	*Total No. Unique Reads amongst Identical Reads*	*Total No. of Redundant Reads*	*Total Percentage of Redundant Reads %*
A	*b*	*C*	*d*	*e*	*(c-e)*	*((c-e) * 100) / b*
SRR149222	2182	249	11.4115	95	154	7.0577
SRR065619	3404	410	12.0447	156	254	7.462
SRR153778	3670	417	11.3624	158	259	7.0572

SRR393844	10932	681	6.2294	254	427	3.906
SRR052290	74076	12634	17.055	4291	8343	11.263

Table 11. Time Taken to find the Identical reads Using Wavelet Transforms based algorithm

SRA Accession No.	*Total No. of Reads*	*Total No. of Copies of Identical Reads*	*Time Taken for Recognizing Identical Reads*
a	*b*	*c*	
SRR149222	2182	249	9.4601 secs.
SRR065619	3404	410	15.3080 secs.
SRR153778	3670	417	16.6342 secs.
SRR393844	10932	681	82.5258 secs
SRR393839	12563	58	98.9385 secs.
SRR052290	74076	12634	1933.6 secs.

Table 12. The subset of records of the Result generated by Matlab program using Wavelet Transforms Algorithm, which represents the Read Nos. of all Identical Reads found in the Bacillus with SRA Reference Id. SRR149222

Read No.	*1st Identical Read No.*	*2nd Identical Read No.*	*3rd Identical Read No.*	*4th Identical Read No.*	*5th Identical Read No.*	*6th Identical Read No.*	*7th Identical Read No.*	*8th Identical Read No.*	*Total No. of Identical Reads*
300	543	867	909	1327					5
313	514	538	624	1192					5
317	359								2
325	494								2
327	479	931	1258	1607					5
328	433	628	748	832	974	1173	1312	1378	9
329	434	629	749	833	1174	1313			7

330	933	1063							3
347	616								2
351	353								2
356	772	1866							3

Table 13. Count of Redundant Reads and the No. of Redundant Base Pairs, for the subset of records of Reads found in SRR149222

1st Read No. whose Identical Reads are found	*Sequence Length in Base Pairs*	*Total No. of Copies of Identical Reads*	*Total No. of Redundant Copies of Reads, after preserving 1 copy of Identical Reads*	*Total No. of Redundant Base Pairs, after preserving 1 copy of sequence of Identical Reads*
(a)	*(b)*	*(c)*	*(d) = (c) - 1*	*(e) = (d) * (b)*
300	236	5	4	944
313	259	5	4	1036
317	218	2	1	218
325	191	2	1	191
327	417	5	4	1668
328	138	9	8	1104
329	54	7	6	324
330	227	3	2	454
347	144	2	1	144
351	117	2	1	117
356	77	3	2	154

From the Table 12 and Table 13, it is observed that there are several copies of identical reads found from DNA sequenced data. If the entire result is stored, without verification, than lot of redundant data may be

preserved unnecessarily, occupying lot of disk space, at the same time causing increased processing time during annotation due to irrelevant data.

In a FASTA file containing DNA sequenced data of Bacillus with Accession Number SRR149222, the total number of redundant reads are 154 and total number of redundant bases are 27536 resulting in wastage of storage space up to 7.0577% in terms of reads (Refer Table 10) and 3.8738% in terms of bases.

Also, it is interesting to know that the SRA sequence with SRA Accession Number SRR393844 contained the Read No. 62 with length 6, whose total number of identical copies were 52 [Fig. 3.].

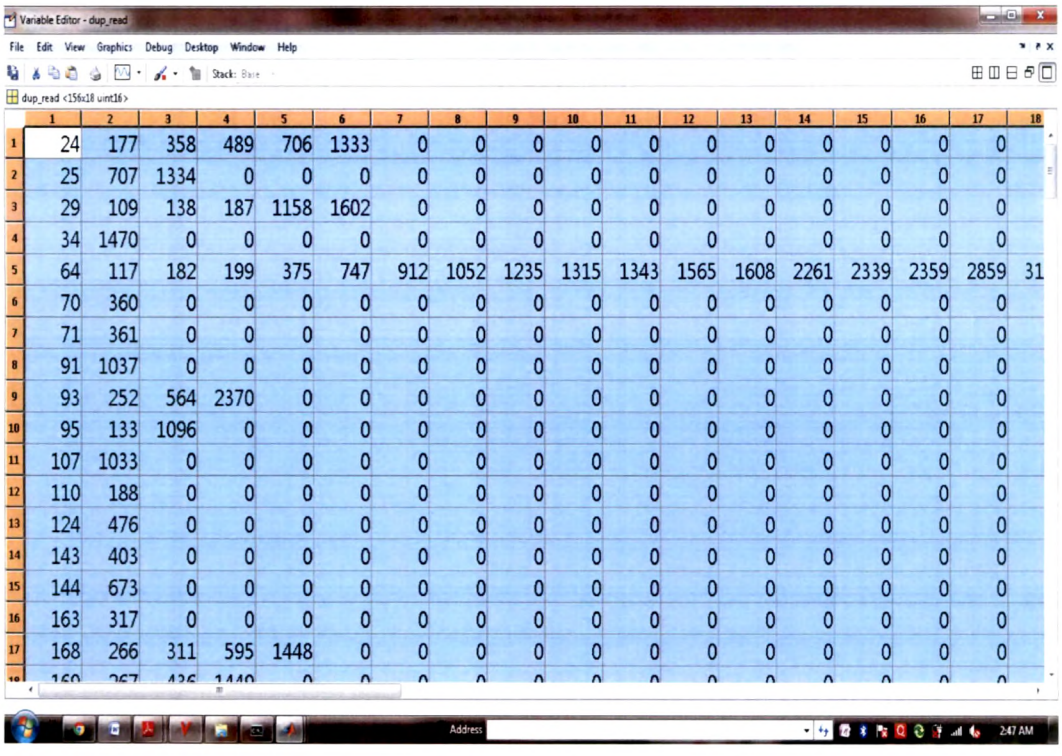
62	64	207	209	821	847	896	1003	1071	1387
1434	1497	1832	2036	2104	2572	2579	2774	2924	3005
3120	3266	3487	3546	3877	4214	4378	5695	5899	5925
6016	6116	6765	7098	7274	7438	7471	7723	7751	7906
8207	8527	8559	8577	9136	9304	10013	10171	10282	10414	10542	10905

Figure 29. List of Read Numbers of Identical Reads (Starting Read Number is 62)

So, this category of reads with irrelevant length and large number of copies can cause increased processing time during further analysis of these reads. Thus Wavelet Transform based algorithm presented here

helps in removing this type of identical and insignificant reads from the generated output of pyrosequencing process.

The proof the the algorithm whose code is developed in Matlab using Wavelet Toolbox, is also presented in the Figure 30 , Figure 31 , Figure 32.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	24	177	358	489	706	1333	0	0	0	0	0	0	0	0	0	0	0	
2	25	707	1334	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	29	109	138	187	1158	1602	0	0	0	0	0	0	0	0	0	0	0	
4	34	1470	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	64	117	182	199	375	747	912	1052	1235	1315	1343	1565	1608	2261	2339	2359	2859	31
6	70	360	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	71	361	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	91	1037	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	93	252	564	2370	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	95	133	1096	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	107	1033	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
12	110	188	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	124	476	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	143	403	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	144	673	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	163	317	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	168	266	311	595	1448	0	0	0	0	0	0	0	0	0	0	0	0	

Figure 30. Matlab Output of Sequence Numbers which are Duplicate Reads for a sequencing data with Accession Number SRR065619

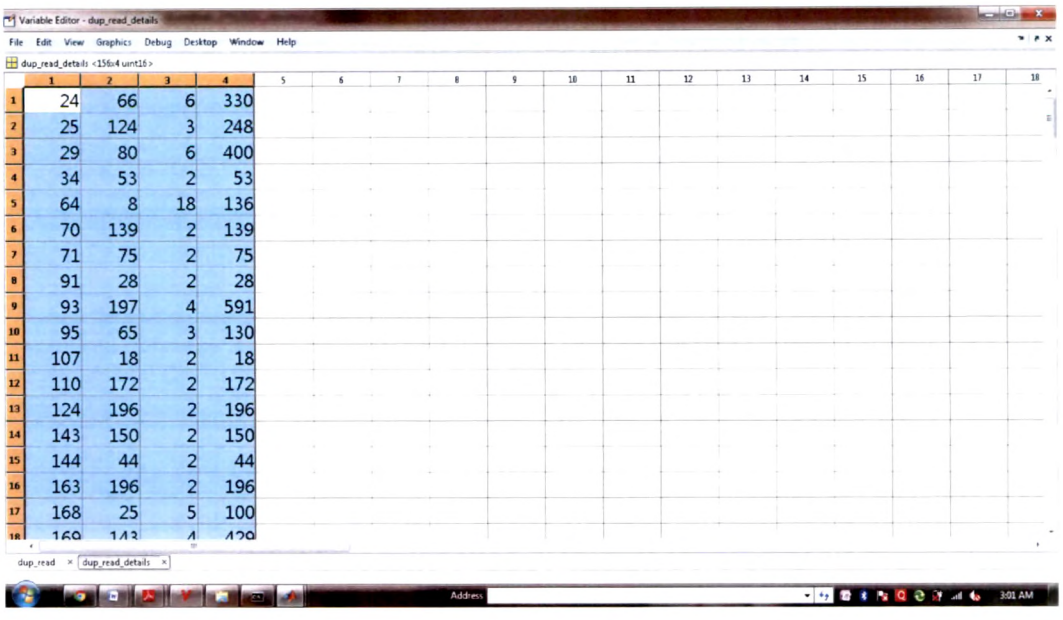


Figure 31. Statistical output generated by Matlab Program for finding Duplicate Reads.

Figure 30 and Figure 31 are comparable with the Output of the Matlab Program.

Figure 30 shows the sequence numbers, whose reads are identical to each other. From the Figure 30 we can make out that Reads with sequence number 24, 177, 358, 489, 706, 1333 are identical to each other.

Figure 31 gives the statistical details like the starting read number i.e. the first read number i.e. 24, identified in the above set of duplicate reads, the length of read is 66 bases, total number of reads which are identical to each other are 6 reads and the fourth column refers to the total number of bases which are redundant and need not be stored or processed, because we can do with storing only one read hence 66 * (6-1) = 330 bases.

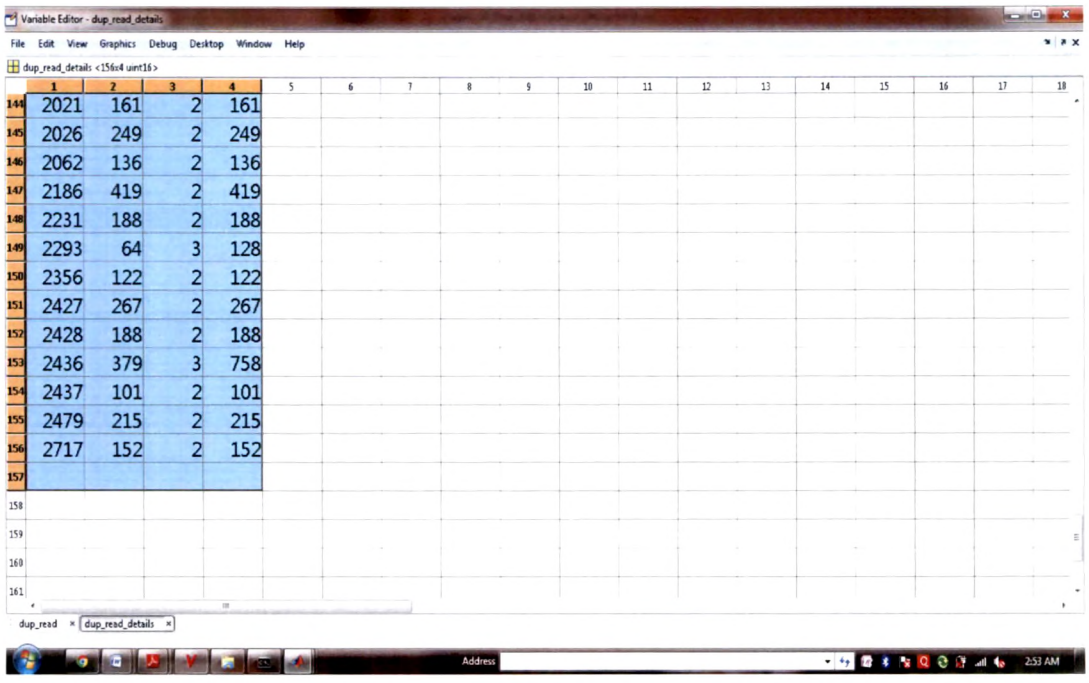


Figure 32. Matlab Program output depicting the Total Number of Unique Duplicate Reads

If you compare Figure 32 and col. (e) of Table 10, you can observe that the Matlab Program creates a more user friendly output, which reflects total number of unique duplicate reads are 156 reads, which can be read from the index of the last row containing the data.

Algorithm 2 (b) : The Improvise Algorithm

A slight improvisation of the above algorithm is also done.

The sequence of steps taken in the improvised algorithm and the results of processing on metagenomics data is given in the tables following the diagram.

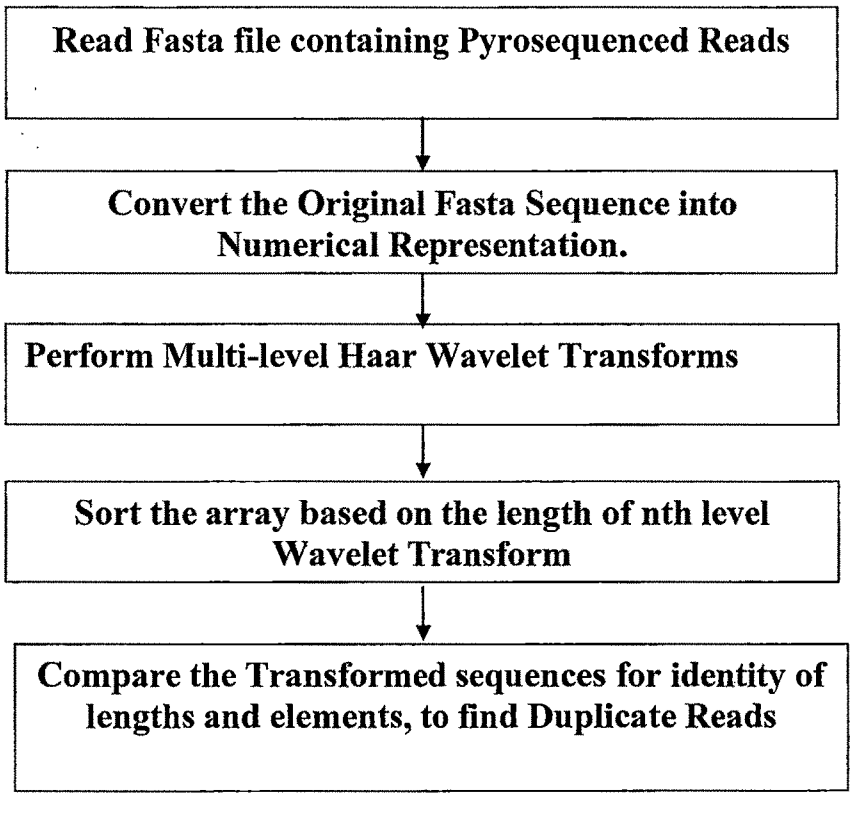


Figure 33. Improvised Algorithm for Recognizing Identical Reads

Algorithm:

Read Fasta File containing Pyrosequenced Reads.

Convert each Fasta sequence to Signal i.e. Numerical Representation, using EIIP property of nucleotides.

Perform four levels Haar Wavelet Transform on Numerical Sequence.

Perform Steps 2 to 3 for all sequences in the Fasta File.

Sort the array of detailed co-efficient of fourth level Wavelet Transforms, based on the signal length of detailed co-efficient. Check for the similarity of lengths of the transformed sequences.

If the lengths are equal, perform element by element comparision of detailed coefficient of transformed sequences, pair wise, to identify duplicate reads.

Perform steps 6 to 7 for all pairs of similar length transformed sequences to identify all the sets of artificial duplicate reads.

Thus, as the algorithm explains, the transformed sequences are first sorted based on length, so that comparision of sequences for identical copy is done only for similar length sequences and not with all the sets of sequences, given in the FASTA file.

The Results are displayed in the following Tables:

Table 14. Result Representing Number and Percentage of Artificial Duplicate Reads with 100% identities, Recognized Using the Wavelet Transforms based Algorithm

SRA Accession No.	***Total No. of Reads***	***Total No. of Copies of Duplicate Reads***	***Total Percent-age of Duplicate Reads***	***Total No. Unique Reads within Duplicate Reads***	***Total No. of Redundant Reads***	***Total Percentage of Redundant Reads***
a	b	c	d	e	(c-e)	((c-e) * 100) / b
SRR000907	5133	124	02.4157	61	63	01.2274
SRR001669	41649	3362	08.072	1486	1876	04.504
SRR001670	55292	12431	22.482	4831	7600	13.746
SRR000675	142545	5210	03.655	2588	2622	01.8394
SRR077225	25698	127	00.494	63	64	00.249
SRR000906	200557	12127	06.0467	5739	6388	03.1851
SRR001663	369811	64164	17.3505	25434	38730	10.4729
SRR065619	3404	410	12.045	156	254	07.462
SRR052290	74076	12634	17.055	4291	8343	11.263

Table 15. Representing details of Redundant Reads (after retaining one copy each) in terms of Number and Percentage of Base Pairs

SRA Accession No.	*Total No. of Reads*	*Total No. of Bases*	*Total No. of Redundant Reads*	*Total No. Redundant Bases*	*Total Percentage of Redundant Bases*
A	b	c	d	e	(e*100) / d
SRR000907	5133	583662	63	6889	1.1803
SRR001669	41649	4422480	1876	194779	4.4043
SRR001670	55292	5856963	7600	789129	13.47335
SRR000675	142545	30932262	2622	423206	1.3682
SRR077225	25698	13166319	64	26939	0.2046
SRR000906	200557	22735927	6388	723900	3.1839
SRR001663	369811	39152427	38730	4187877	10.6963
SRR065619	3404	1081048	254	37937	3.5093
SRR052290	74076	22356978	8343	1211079	5.4170

Table 16. Representing byte conservation, using proposed algorithm

SRA Accession No.	*Total No. of Reads*	*Total No. of Duplicate Reads*	*Total No. Unique Reads within Duplicate Reads*	*Total No. of Redundant Reads*	*Total No. Redundant Bases*	*Total Bytes which can be conserved by removing Duplicate Reads*
a	b	c	D	e=c-d	f= (e) * Length of each Redundant Read	g

SRR000907	5133	124	61	63	6889	6889
SRR001669	41649	3362	1486	1876	194779	194779
SRR001670	55292	12431	4831	7600	789129	789129
SRR000675	142545	5210	2588	2622	423206	423206
SRR077225	25698	127	63	64	26939	26939
SRR000906	200557	12127	5739	6388	723900	723900
SRR001663	369811	64164	25434	38730	4187877	4187877
SRR065619	3404	410	156	254	37937	37937
SRR052290	74076	12634	4291	8343	1211079	1211079

Table 17. Representing the SRA Read No., with highest number of duplicate reads in a given SRA Read, along with Length and Maximum number of copies (SRA No. of read as specified in column (d))

SRA Accession No.	***Total No. of Reads***	***Total No. of Duplicate Reads***	***SRA Read No. of Read with Max No. of Copies (Refers to the 1st occurrence of the duplicate read)***	***Length of a Read with Max No. of Copies (bp)***	***Max No. of Copies of a given Read (units)***
a	b	c	d	e	f
SRR000907	5133	124	SRR000907.865.2	106	3
SRR001669	41649	3362	SRR001669.29.2	98	18
SRR001670	55292	12431	SRR001670.7715.2	103	19
SRR000675	142545	5210	SRR000675.3409.2	143	5
SRR077225	25698	127	SRR077225.1404.3	436	3
SRR000906	200557	12127	SRR000906.3468.2	95	7
SRR001663	369811	64164	SRR001663.17896.2	124	48
SRR065619	3404	410	SRR065619.64.3	8	18
SRR052290	74076	12634	SRR052290.3.3	5	394

Table 18. Representing Computational Time for Identifying Artificial Duplicate Reads with 100% identities

SRA Accession No.	*Total No. of Reads*	*Total No. of Duplicate Reads*	*Computational Time (in secs)*
a	B	c	d
SRR000907	5133	124	34.6411
SRR001669	41649	3362	1440.2
SRR001670	55292	12431	2900
SRR000675	142545	5210	7815.9
SRR077225	25698	127	453.8735
SRR000906	200557	12127	32447
SRR001663	369811	64164	112910
SRR065619	3404	410	15.0854
SRR052290	74076	12634	1933.6

Table 19. Result displaying Computational Time for Identifying Artificial Duplicate Reads with 100% identities Recognized Using the Wavelet Transforms based Algorithm used for Data Reduction, prior to comparing the sequences for finding exact or near exact match

SRA Accession No.	*Total No. of Reads*	*Total No. of Duplicate Reads*	*Time Taken To Perform Four Level Wavelet Transforms (in secs)*	*Time Taken To Recognize Duplicate Reads (in secs)*	*Total Computational Time (in secs) inclusive of generating other statistical information*
a	b	c	D	e	f
SRR000907	5133	124	17.5457	17.0194	34.6411
SRR001669	41649	3362	181.9955	1257.5	1440.2
SRR001670	55292	12431	263.3025	2899.5	3163.4769
SRR000675	142545	5210	1016.9	6797	7815.9
SRR077225	25698	127	111.2499	300.9164	453.8735

SRR000906	200557	12127	1681.6	30761	32447
SRR001663	369811	64164	4679.3	108220	112910
SRR065619	3404	410	11.956	3.8123	15.9854
SRR052290	74076	12634	412.2294	1517.8	1933.6

Thus, Signal Processing Approach can be used to compare the two reads generated from DNA sequencing, for verifying their resemblance. Using Wavelet Transforms we can reduce the data for comparision to one-eighth size of the original sequence, when transformed to three levels. This data reduction using Wavelet Transforms optimizes the computational complexity to logarithmic order and hence provides improved algorithm for recognizing identical reads amongst the DNA sequenced data. Here other overheads like reading a file or time involved in encoding the DNA sequence to form its Digital Signal, is not considered. Once the similar sequences are identified, it is not necessary to store the entire sequence, instead can store only the references to the strings for further annotation. This also optimizes the space requirement for storage of reads in the database. Also Wavelet Transforms are performed in place and hence memory requirement is reduced. Thus the proposed algorithm optimizes both space and time complexity involved in recognizing identical reads from DNA sequenced data. Further, if it is possible to apply distributed computing on this algorithm, improvement in processing time is possible, particularly when data is very large.

5.3. Algorithm 3: Recognizing Short Tandem Repeat Regions

5.3.1. Purpose

DNA sequences exhibit ubiquitously distributed patterns of repeat regions, known as Short Tandem Repeats (STRs). STRs are also known as Short Sequence Repeats (SSRs) or microsatellites are contiguously placed, multiple and approximate copies of pattern of nucleotides, in DNA sequences[134]. They are nucleotide sequences in DNA of 1–6 bp unit length, distributed randomly in eukaryotic and prokaryotic genomes and are highly polymorphic[135].

Short tandem repeats (STRs) has an impact in genetic mapping, population genetic analysis, DNA forensics and phylogenetics[136]. Mutational dynamics of microsatellites play a role in human genetic disorders[137] and may have significant roles in the regulation of gene expression[138] [139]. The STRs are genetic markers playing various regulatory and evolutionary roles and are responsible for causing several human diseases like Huntington's disease, myotonic dystrophy, spinal

---

134 Benson G., "Tandem repeats finder: a program to analyze DNA sequences," PMC, PubMed Nucleic Acids Research, Vol. 27. No. 2 1999; 27:573-80

135 Angelika Merkel, Neil Gemmell, "Detecting short tandem repeats from genome data: opening the software black box," Briefings in Bioinformatics Advance Access, July 10, 2008

136 Goldstein DB, Schlotterer C ,"Microsatellites: Evolution and Applications," Oxford: Oxford University Press,, 1999

137 Pearson CE, Edamura KN, Cleary JD, "Repeat instability: mechanisms of dynamic mutations," Nat Rev Genet, 2005; 6: 729-42

138 Kashi Y, King DG,"Simple sequence repeats as advantageous mutators in evolution," Trends in Genetics, 2006;22: 253-9

139 Moxon ER, Wills C., "DNA microsatellites: agents of evolution? " Sci Am, 1999; 280:94-9

and bulbar muscular atrophy, Friedreich's ataxia[140]. Approximately 10% of human genome comprises of STRs[141].

STRs are the repetitive portions in DNA sequence that may exist as homopolymers of a single nucleotide type like AAA, CCCC, GGGGG or TTTT, or may exist in small or big number of multi-mers. These multi-mers may be identical units i.e. homogeneous repeats, mixed units i.e. heterogeneous repeats, or degenerate repeat sequence motifs[142] as shown in Table 20. Short tandemly repeated sequences occur in few to thousands of copies, distributed all throughout the genome in several eukaryotes[143] [144].

Table 20. Terms and Notations used to describe STRs / Microsatellites

Generic Term/ Biological Term for Type of Repeat	*Length of Repeat Unit (basepairs)*	*Repeat Sequence*	*Annotation (Includes Repeat Unit and its Frequency)*
Homopolymeric or Monomer/ Perfect	15	5'-TTTTTTTTTTTTTTT-3'	5'-(T)15-3'

[140] Benson G., "Tandem repeats finder: a program to analyze DNA sequences," *PMC, PubMed Nucleic Acids Research*, Vol. 27. No. 2 1999; 27:573-80

[141] Benson G., "Tandem repeats finder: a program to analyze DNA sequences," *PMC, PubMed Nucleic Acids Research*, Vol. 27. No. 2 1999; 27:573-80

[142] Alex Van Belkum, Stewart Scherer, Loek Van Alphen, And Henri Verbrugh , "Short-Sequence DNA Repeats in Prokaryotic Genomes," *American Society for Microbiology MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS*, June 1998, p. 275-293 Vol. 62, No. 2

[143] Jeffreys, A. J., V. Wilson, and S. L. Thein, "Hypervariable "minisatellite" regions in human DNA," *Nature*, 1985 314:67-73

[144] Nakamura, Y., M. Leppert, P. O'Connell, R. Wolff, T. Holm, M. Culver, C. Martin, E. Fujimoto, M. Hoff, E. Kumlin, and R. White, " Variable number of tandem repeat (VNTR) markers for human gene mapping ," *Science*, 1987 235:1616-1622

Multimeric, Dimer/ Perfect	8	5'-ATATATATATATATAT-3'	5'-(AT)8-3'
Multimeric,Trimer / Perfect	8	5'-ATTATTATTATTATTATTATT-3'	5'-(ATT)8-3'
Multimeric,Tetramer/ Perfect	6	5'-GATCGATCGATCGATCGATCGATC-3'	5'-(GATC) 6-3'
Multimeric,Pentamer/ Perfect	6	5'-ATGCCATGCCATGCCATGCCATGCCATGCC -3'	5'-(ATGCC)6-3'
Multimeric,Hexamer/ Perfect	4	5'-AAAATTAAAATTAAAATTAAAATT-3'	5'-(AAAATT)8-3'
Multimeric, Heterogeneous/ Imperfect	Variant	5'- GCC GCC GCC GATC GATC AT AT AT AT -3'	5'-(GCC)3(GATC)2 (AT)4-3'
Multimeric, Heterogeneous/ Imperfect and Interrupt	Variant	5'- GCC GCC GCC T GATC GATC GC AT AT AT AT -3'	5'-(GCC)3 T (GATC)2 GC (AT)4-3'

Several laboratory analytical and bioinformatics tools have been developed, for studying microsatellite evolution[145] [146] since 1996, when the first eukaryotic genome, yeast *Saccharomyces Cerevisiae* was

[145] Thiel T, Michalek W, Varshney RK, "Exploiting EST databases for the development and characterization of gene derived SSR-markers in barley (Hordeum vulgare L.)," *TheorAppl Genet*, 2003; 106:411-22

[146] Smit AFA, Green P, "RepeatMasker,"Available at: http://www.repeatmasker.org, 1996

sequenced[147]. The *in-vitro* methods involve expensive probe hybridization. The *in-silico* or computational approach[148] methods include study of regional distribution bias or putative association[149] with genomic features. These *in-silico* investigations are widely used instead of expensive *in-vitro* method [150] [151] [152].

Widely utilized tools for detecting microsatellites or tandem repeats are *MISA*[153], *REputer*[154], *Sputnik*[155], *TRF*[156], *RepeatMasker*[157]. These tools primarily use regular-expression, Hamming distance, recursive match and penalty scores, Heuristic-with alignment procedure, k-mer with suffix trees[158] and k-tuples, string-comparison based algorithms. These

---

[147] Goffeau A, Barrell BG, Bussey H, "Life with 6000 genes, "*Science*, 1996;274:546-67

[148] Surya Saha and Susan Bridges and Zenaida V. Magbanua and Daniel G. Peterson, "Computational Approaches and Tools Used in Identification of Dispersed Repetitive DNA Sequences," *Springer Science+Business Media, LLC*, 2008

[149] Li YC, Korol AB, Fahima T, "Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review," *Mol Ecol*, 2002; 11:2453-65

[150] Alex Van Belkum, Stewart Scherer, Loek Van Alphen, And Henri Verbrugh , "Short-Sequence DNA Repeats in Prokaryotic Genomes," *American Society for Microbiology MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS*, June 1998, p. 275-293 Vol. 62, No. 2

[151] Morgante M, Hanafey M, Powell W, "Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes," *Nat Genet*, 2002; 30:194-200

[152] Lim S, Notley-McRobb L, Lim M, "A comparison of the nature and abundance of microsatellites in 14 fungal genomes," *Fungal Genet Biol*, 2004; 41:1025-36

[153] Thiel T, Michalek W, Varshney RK, "Exploiting EST databases for the development and characterization of gene derived SSR-markers in barley (Hordeum vulgare L.)," *TheorAppl Genet*, 2003; 106:411-22

[154] Stefan Kurtz, Jomuna V. Choudhuri, Enno Ohlebusch, "REPuter: the manifold applications of repeat analysis on a genomic scale," *Nucleic Acids Research*, 2001 Vol 29 No. 22 pg 4633 - 4642

[155] Chris Abajian, "Sputnik - DNA microsatellite repeat search utility," *http://www.espressosoftware.com/sputnik/*

[156] Benson G., "Tandem repeats finder: a program to analyze DNA sequences," *PMC, PubMed Nucleic Acids Research*, Vol. 27. No. 2 1999; 27:573-80

[157] Smit AFA, Green P, "RepeatMasker,"Available at: http://www.repeatmasker.org, 1996

[158] Surya Saha and Susan Bridges and Zenaida V. Magbanua and Daniel G. Peterson, "Computational Approaches and Tools Used in Identification of Dispersed Repetitive DNA Sequences," *Springer Science+Business Media, LLC*, 2008

string-based and other approaches need pattern, pattern size[159], reference sequence[160] as input parameters. Knowledge about pattern, pattern size, copy number, mutational history, for tandem repeats is limited, due to inability to easily detect them in genomic sequences. Amongst others, the algorithm based on computing alignment scores[161] has a time complexity of $O[n^2 \text{ polylog}(n)]$, *n* being length of the sequence. This exponential time complexity may not be useful for very large sequences.

As a part of research work, an *ab initio method* is developed for finding short tandem repeats, which does not require pattern or pattern size as an input parameter. The signal processing concept are applied as a recognition criterion, for detecting perfect, imperfect or interrupted microsatellite regions as well as simultaneously provides the visual representation of these repeat regions. The study emphasizes on the use of Wavelet Transforms particularly the *Haar Wavelets* for recognizing the regions of short tandem repeats. The time complexity of Wavelet transforms is *O(log n)*, n being the length of the transformed sequence. Using the proposed algorithm, the Graphical Representaion in Matlab, of the Short Tandem Repeat regions is also possible, which is usually not available in the previously discussed tools for identifying STRs.

5.3.2. Methodology

[159] Thiel T, Michalek W, Varshney RK, "Exploiting EST databases for the development and characterization of gene derived SSR-markers in barley (Hordeum vulgare L.)," *TheorAppl Genet*, 2003; 106:411-22

[160] Smit AFA, Green P, "RepeatMasker,"Available at: http://www.repeatmasker.org, 1996

[161] Schmidt JP ," All Highest Scoring Paths in Weighted Grid Graphs and Their Application to Finding All Approximate Repeats in Strings,"*Siam J. Comput*,1998, Vol. 27 Issue 4, 972-992

The suggested algorithm for identifying the short tandem repeat in a given sequence is applied as follows:

Read the Fasta file which contains the several sequences $S_i \in S$, where a single sequence can be represented as Si = {s1,s2,...sn} where si $\in \sum$ = { A,C,G,T }, i = 1, n and n is the length of S_i.

Convert the nucleotide sequence Si into its numerical representation Xi. The single indicator sequence using dipole moments[162] property of nucleotides is used as numerical representation.

Table 21. Representation of Dipole Moments value for Nucleotide Bases

Nucleotide Bases	*Dipole Moments Values*
A	0.4629
G	6.488
C	3.943
T	1.052

The use of dipole-moment property which is a single indicator for nucleotide base, reduces the computational overhead by 75% compared to the conventional four-base binary sequence representation of nucleotide sequence[163]. Only numerical representations can be applied for transformations.

[162] J. K. Meher, M. R. Panigrahi, G. N. Dash, P. K. Meher, "Wavelet Based Lossless DNA Sequence Compression for Faster Detection of Eukaryotic Protein Coding Regions ," *I.J. Image, Graphics and Signal Processing 2012, 7, 47-53*

[163] A. S. Nair and S. P. Sreenathan, "A Coding Measure Scheme Employing Electron-Ion Interaction Pseudopotential (EIIP)," *Bioinformation*, Vol. 1, No. 6, 2006, pp. 197-202

Perform first-level decomposition of this sequence, using Haar Wavelet Transform.

Identify the zero values in detailed coefficients of these transforms. The detail co-efficients of Wavelet Transforms represent the fluctuation. Thus, the occurrences of zero values in first-level decomposition represents that there is no fluctuation in the given part of the signal i.e. all consecutive values are similar. This signifies the existence of monomer of single nucleotides, in this region of the sequence.

Repeat steps 3 through 4 on each decomposed sequence. Further decomposition of a sequence, to 2nd level, 3rd level and 4th level would identify STRs of di-nucleotides, tri-nucleotides and tetra-nucleotides, respectively.

Searching for zero values in transformed signal is searching on a reduced data set and hence is more efficient. If the correct mapping/transformation is applied, the Parseval Theorem implies that frequency distance FD is less than or equal to edit distance ED (Distance preserving transformation)[164].

With every transform, we are discarding half the values as per Nyquist's rule[165] and hence optimizing the search, with time complexity of $O(\log n)$.

Since, the Haar wavelet transforms work on averaging and differencing. The pair-wise consecutive same values when calculated for differencing, will always result in zeroes. The set of consecutive zeros, in the detail-

---

[164] Aghili, Alireza A, Agrawal, D El Abbadi, A, "Sequence Similarity Search Using Discrete Fourier and Wavelet Transformation Techniques," *Postprints, UC Santa Barbara*, 2005 http://www.escholarship.org/uc/item/9w7094b9

[165] J. K. Meher, M. R. Panigrahi, G. N. Dash, P. K. Meher, "Wavelet Based Lossless DNA Sequence Compression for Faster Detection of Eukaryotic Protein Coding Regions ," *I.J. Image, Graphics and Signal Processing 2012, 7, 47-53*

coefficients, in any order of wavelet transform, is a proof that the repeat exists in that region.

The decomposition of a discrete signal using Haar wavelet transform generates two sub-signals. One of these sub-signals, computing a running average represents the trend; the other sub-signal, having a running difference, represents fluctuation[166].

We have used in our method, this signal processing concept to identify the fluctuations, through the second half sub-signal, the running difference. If the signal processing concept can be used to identify fluctuating region, it precisely can also be applied to reflect the non-fluctuating region. We therefore applied the Haar Wavelet Transforms to identify "the non-fluctuating regions", which biologically represents "the repeat regions" or STRs.

If the sequence contains the homopolymer of mono-nucleotide, the region in sequence containing Short tandem repeat, would have its detailed-coefficients, i.e. differencing values set to zeros, in its first level of transform. This subset of zeros is actually a repeat region. Identifying the start and end position of this subset, is sufficient to identify the region of STR.

The repeat region in the sequence, which contains di-nucleotide or more as its repeat pattern, can be identified by recursively doing multi-level discrete wavelet transform (DWT). The multi-level DWT of the second order or more can be performed, to identify the subset containing zeros. For di-nucleotides, perform second-level discrete wavelet transform and further levels of transform for tri-nucleotide and tetra-nucleotide homopolymers respectively. Since, the maximum level of transform

[166] I. Daubechies, "Orthonormal Bases of Compactly Supported Wavelets,"*Comm. Pure Appl. Math*, Vol 41, 1988, pp. 906-966

performed for the given sequence is four, the maximum of tetra-nucleotide repeats can be identified, using this method currently.

Thus, our method can be applied for finding homopolymeric regions with perfect repeats as well as imperfect repeats. The method can also recognize the heterogeneous regions. Both homopolymeric regions as well as heterogeneous regions with interrupts can also be identified. This is clearly indicated by spikes or non-zero regions in the graphical representation (Figure 34 (d) and (e)).

5.3.3. Result and Discussion

Example-1 : Examples of Short Tandem Repeats (Synthetic Data) :

>FYRUZ7J01B3YZA rank=0001782 x=748.0 y=836.5 length=152
GTACTATGTA**TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT**GTGGCGTGT
CTGTAGCGTGTGGCTTGGCTGTAGTATGTATGTGGCAGTGTCTGTAGTGTGTGGCGTGTCT
GTCGCATGTAGCCGTGTCTGTCAT

>FYRUZ7J01A1PTO rank=0003965 x=312.0 y=1898.5 length=175
CAT**ACACACACACACACACACACACACACACACACACACACACACACACACACACACACAC**
ACACACACACACACACACACACACACACACACACACACACACACACACACACACACACAC
ACACACACACACACACACACACACACACACACACACACACACACACACTACTAC

>FYRUZ7J01A1PT1 rank=0003965 x=399.0 y=2198.5 length=167
GTACTATGTATGTGGCGTGTCTGTAGCGTGTGGCT**TCATCATCATCATCATCATCATCATCA**
TCATCATCATCATCATCATCATCATCATCATCATCATCATCATCATCATCATCATCATCATCA
TCATCATCATTTTGTACTATGTATTGTACTactatcgATGTA

>FYRUZ7J01A1PT2 rank=0003965 x=442.0 y=2898.5 length=248
CATA**CACACACA**TA**TACATACATACATACATACATACATACATACATACATACATACATACA**
TACATACATACATACATACATACACACACACACACACACACACACACACA**CTACTA**CACACA
CACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACA
CACACACACACACACACACACACACACACACACACACACACACACACACACACACACACTACTA

>FYRUZ7J01A1PT3 rank=0003965 x=467.0 y=3908.5 length=248
ACGTCACACAACGTTA*ACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTAC*
*GTACGTACGTACGTACGTACGT***CACACACACACACACACACACACACACACTACTACACAC**
ACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACAC
ACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACT
ACTA

>FYRUZ7J01A1PT4 rank=0003965 x=490.0 y=4208.5 length=240
GTACTATGTAGACGTCAGTCTGTAGCGTGTGGCTTGGCTGTAGTATGTactatcg**ATGTAGTA**
CTATGTAGACGTCAGTCTGTAGCGTGTGGCTTGGCTGTAGTATGTactatcgATGTAGTACTA
TGTAGACGTCAGTCTGTAGCGTGTGGCTTGGCTGTAGTATGTactatcg**ATGTAGTACTATGT**
AGACGTCAGTCTGTAGCGTGTGGCTTGGCTGTAGTATGTactatcgATGTA

The Figure 34 shows the visual representation of the numerical form of original sequence, 4-levels of decompositions and the reconstructed sequence, given for 6 different synthetic DNA sequences containing STR of homopolymers, as in Example-1. The (a), (b), (c) and (d) represent homopolymer consisting of mono, di, tri and tetra nucleotides respectively. The (e) has heterogeneous and interrupted, more than one repeat region. The (f) has no repeat region of homopolymers up to 4 bases. The flat line at 0 frequencies represents the repeat region. The X axis represents the Time or Space Localization and the Y axis represents the Scale or Frequency Localization. Figure 35 and Figure 37 show the zoomed in graphical representation of sequences in Example 1(a) and (b) for clear description of the positions of the Repeat region as shown in form of horizontal line with Y-axis value zero.

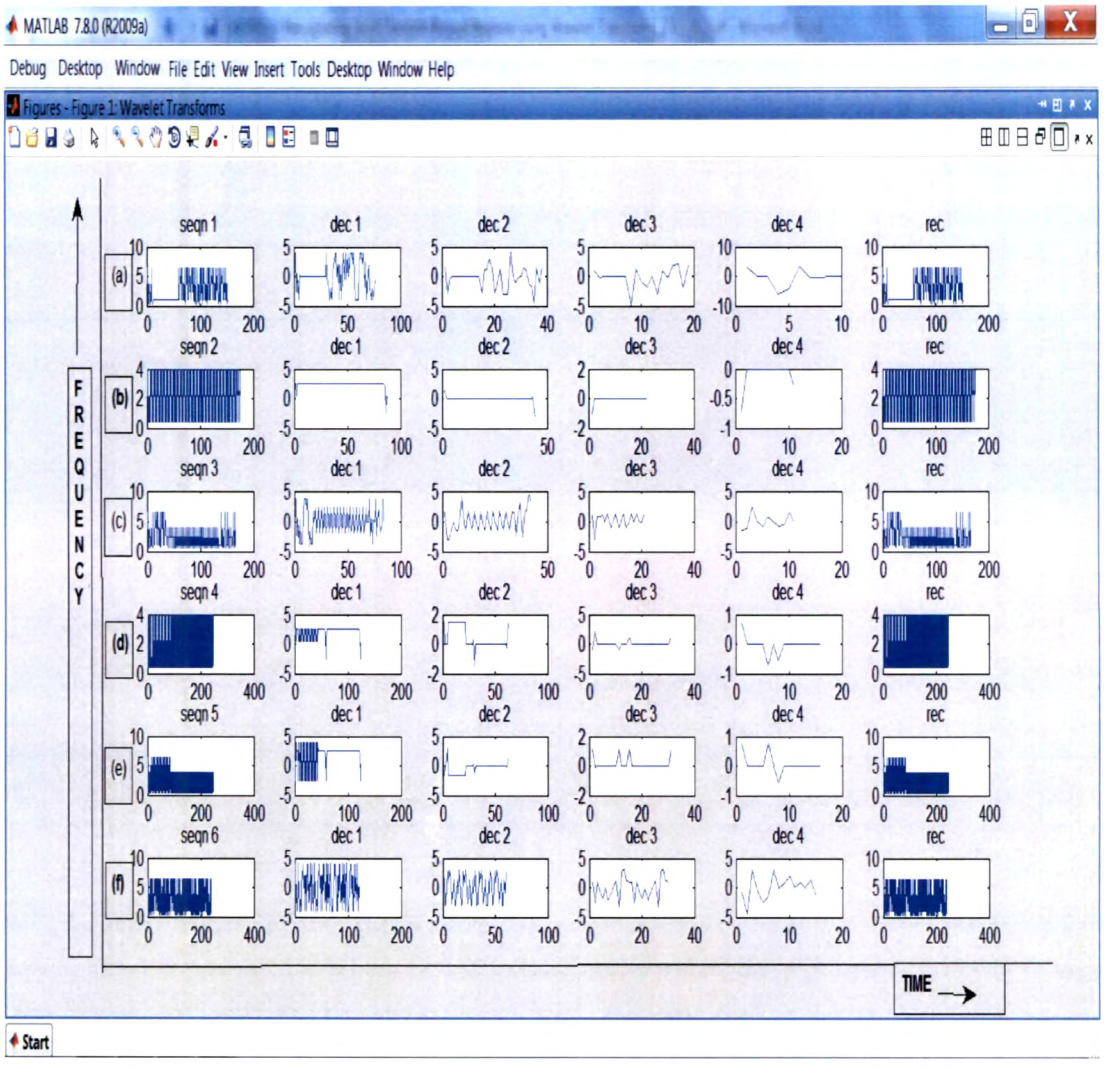


Figure 34. Visual representation of Short Tandem Repeat Region in sequences as in Example 1, using multi-level Haar Wavelet Transform.

Example1 : Signal (a)

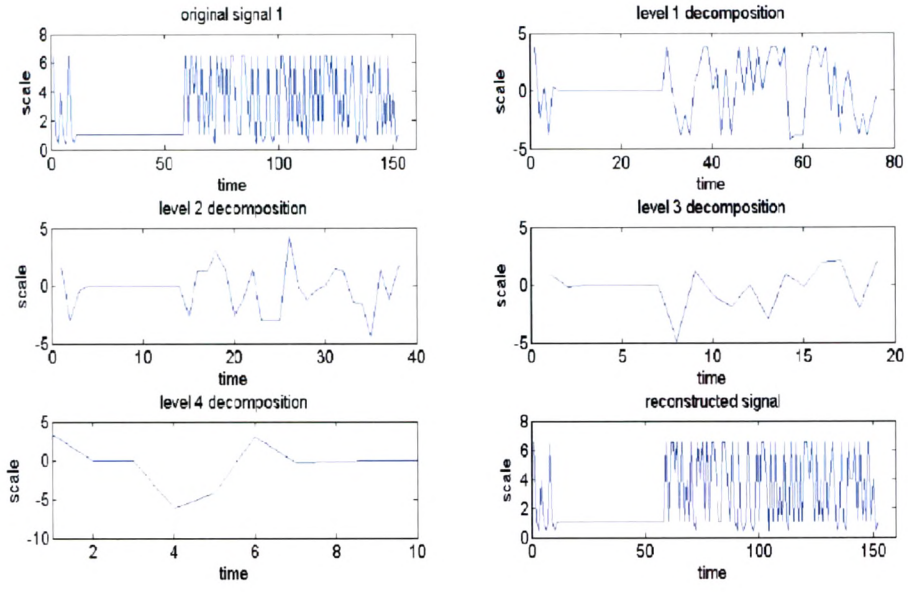


Figure 35. Detailed Graphical representation of Short Tandem Repeat Region in sequences as in Example 1(a), using multi-level Haar Wavelet Transform.

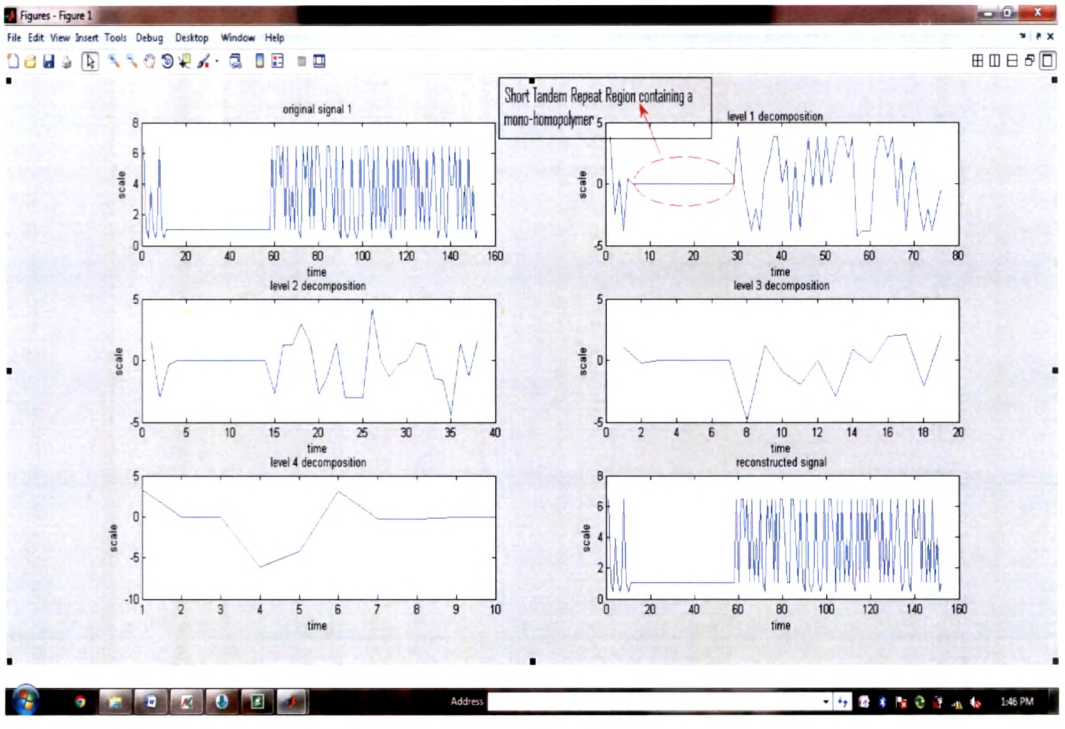


Figure 36. Graphical representation of Short Tandem Repeat Region. RED coloured ellipse defines the STR region containing mono-homopolymer

Example1 : Signal (b)

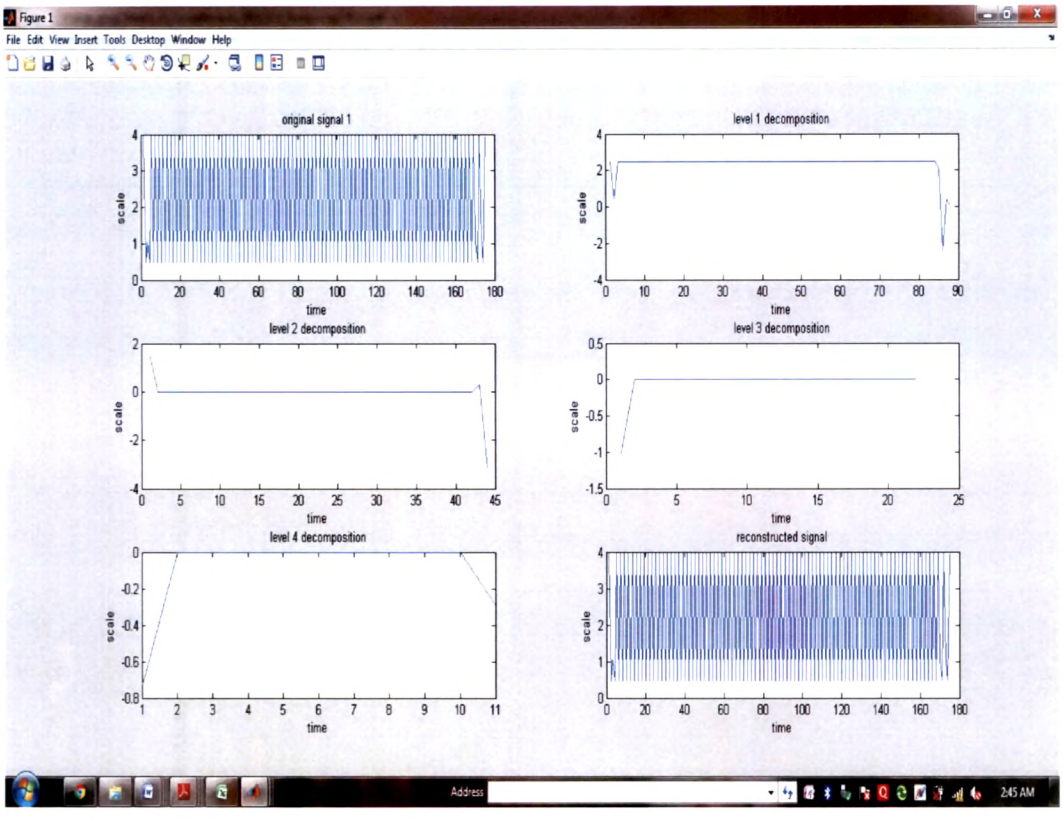


Figure 37. Detailed Graphical representation of Short Tandem Repeat Region in sequences as in Example 1(b), using multi-level Haar Wavelet Transform.

Example – 2:

>FYRUZ7J01A1MA2 rank=0003965 x=312.0 y=1898.5 length=181
CATACACACACATATACATACATACATACATACATACATACATACATACATACATACATACAT
ACATACATACATACATACATACACACACACACACACACACACACACACACACACTACTACACACAC
ACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACAC
ACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACTACTA

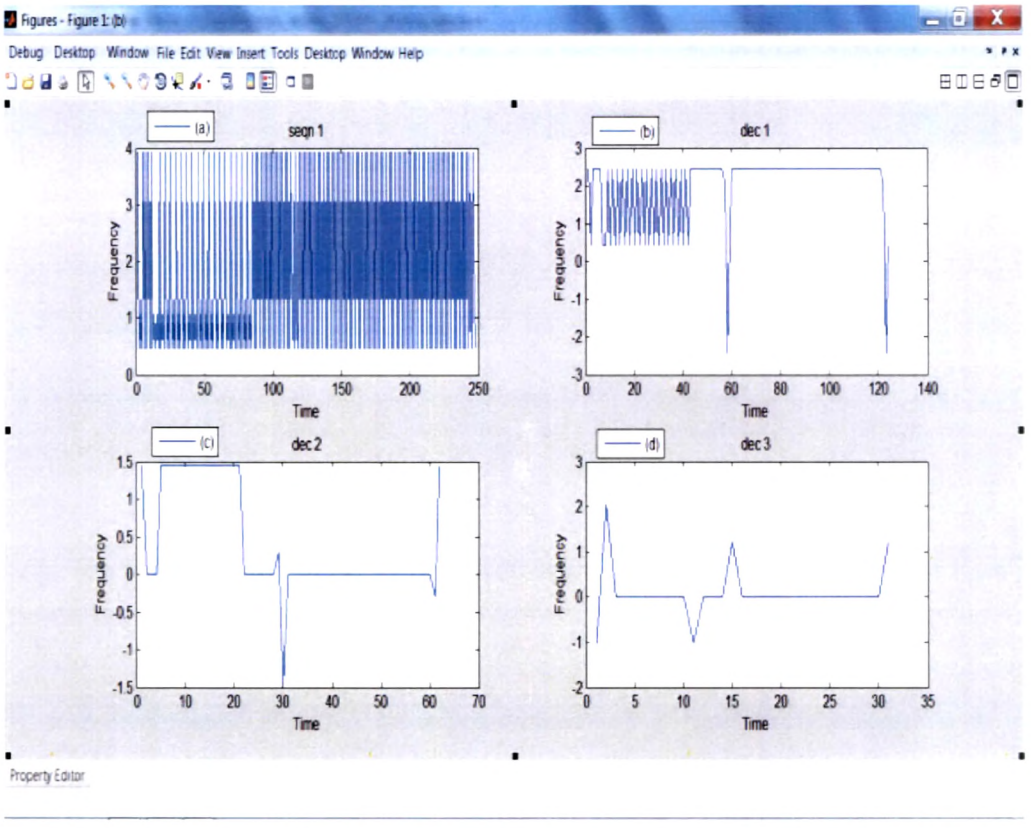


Figure 38. Representing the Discrete Wavelet Transform of the sequence in Example-2 with four repeat regions.

The graphical representation as in Figure 38, of the sequence in Example-2 has four repeat regions which are heterogeneous and interrupted. This is clearly represented in second level decomposition (c) in form of flat horizontal lines around zero frequency at positions between 2 to 4, 5 to 21(regularity, but not zero), 22 to 28 and 31 to 60.

The X axis represents the Time or Space Localization and the Y axis represents the Scale or Frequency Localization

The transform positions seen in the graph has to be multiplied by 2^j, where j is the level of transform that one is referring at zero frequency, to get the original position of STR in the original nucleotide sequence.

Example-3:

The E-Coli strain K-12 substrain MG1655star Contigs Numbered 40, 41, 43, 46 of GenBank Accession No. AEFE01000040.1, are displayed as the graphical representation in

Figure 39

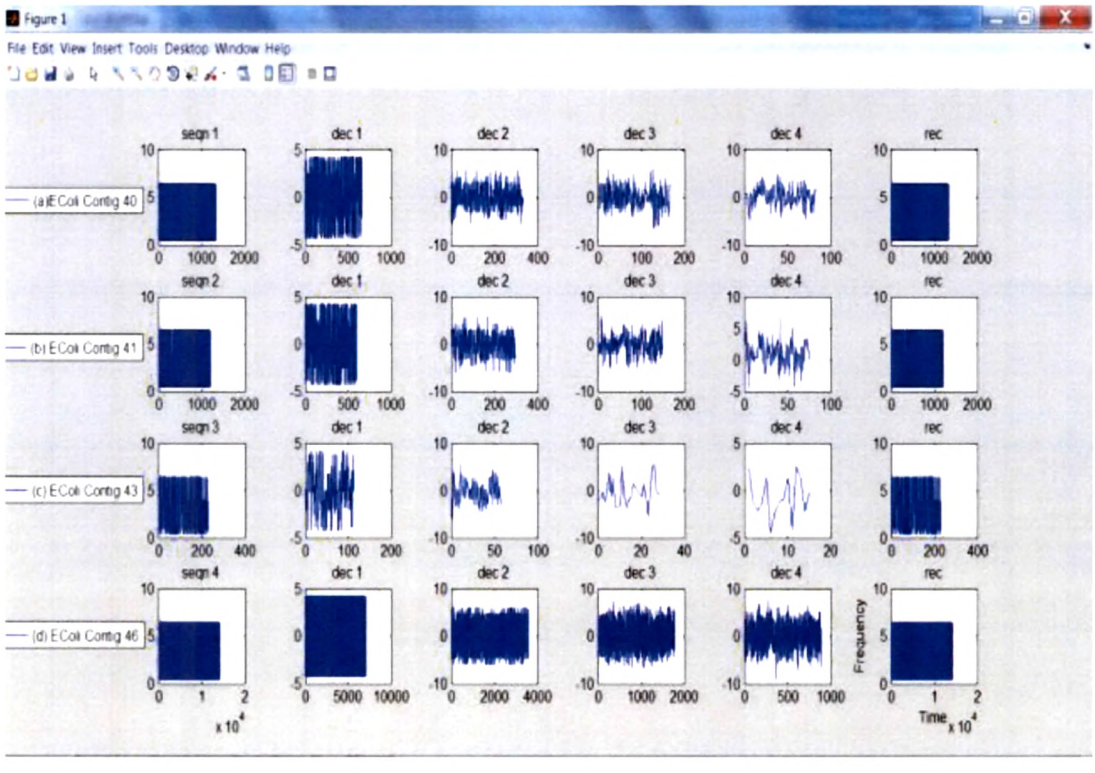


Figure 39. The Graphical Representation of output of Wavelet-based technique for searching STR regions in E-Coli Contigs No. 40, 41, 43, 46 GenBank Accession No. AEFE01000040.1, AEFE01000041.1, AEFE01000043.1 and AEFE01000046.1 respectively

The graphical representation as in

Figure 39, demonstrate that there are no STR regions in the mentioned contigs. The result is comparable with the results of widely used MISA software, executed with default parameters, as displayed in Figure 40.

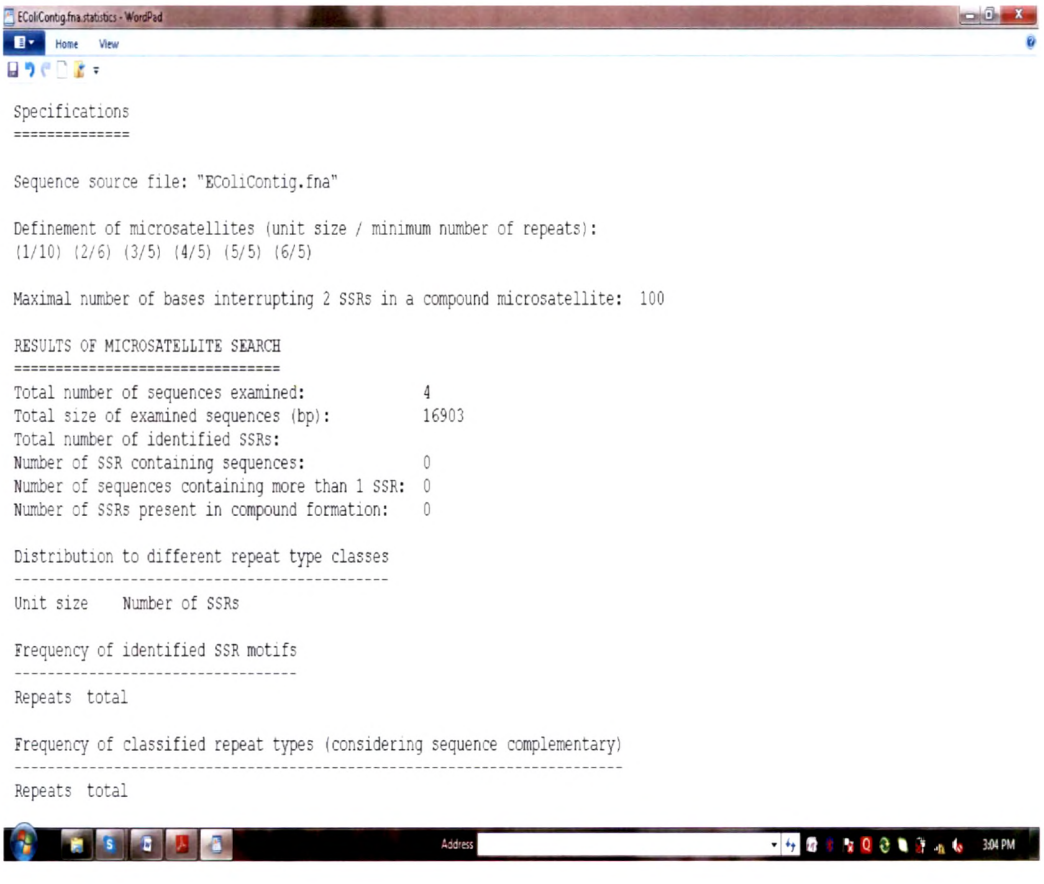


Figure 40. The output from MISA software, repeats found with default parameters.

Example 4:

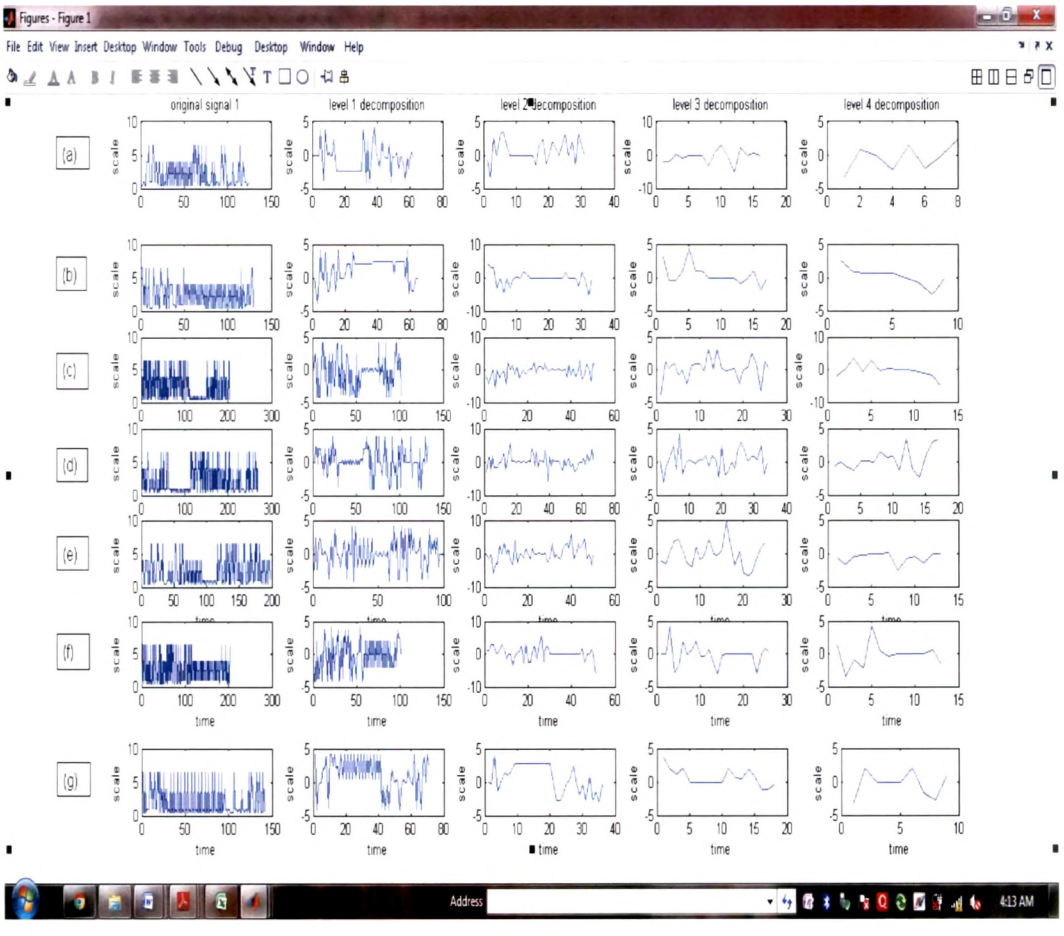


Figure 41. Graphical Representation of Short Tandem Repeat Regions from sequence extracted from the the human genome RefSeq database (release 28).

Figure 41, shows the graphical representation of Short Tandem Repeat Regions in Human Genome RefSeq database (release 28). The output of the algorithm discussed in this work for detecting Short Tandem Repeat Regions confirms with the output as discussed in Research article Sequence determinants of human microsatellite variability[167] . The graph of sample examples is presented here. But, the complete data can be proved using the proposed algorithm. The samples considered here are as given in Table 22 .

Table 22. List of Sample data taken from Human Genome RefSeq (release 8) for representing Short Tandem Repeat Regions

chromosome	marshfieldSetNumber	locusName
1	10	D1S468
4	10	D4S408
1	10_13	D1S1589
5	10	D5S2488
6	10_52	D6S1006
1	10	D1S1677
8	52	D8S1130

Thus it can be concluded that discrete wavelet transform, particularly Haar wavelets, can be used to identify the short tandem repeat regions in a given sequence, by identifying the subsets of detailed co-efficient with values zero. Since, the short tandem repeat regions are an important regulatory mechanism and play a vital role in genetic analysis,

[167] Trevor J Pemberton, Conner Sandefur, Mattias Jakobsson and Noah A Rosenberg, Sequence determinants of human microsatellite variability, BMC Genomics 2009, 10:612 doi:10.1186/1471-2164-10-612

identifying this region is a very crucial. The wavelet transforms have a time complexity of *O(log n),* and hence identifying these repeat regions using Haar wavelet transforms is independent of the length of the original sequence. Thus wavelet based approach is much efficient than other classic string or regex based algorithms, with exponential time complexity. Also identifying STRs using Haar wavelets does not need to specify any fixed pattern to be searched for (which is very difficult to predict) and hence it is not pattern dependent as required to be specified *a priori* while applying other well-known algorithms.

This method of Haar Wavelet Transform can also identify any number of regions, which may be ubiquitously distributed along the nucleotide sequence. Also, single transform can identify the spread of different types of repeat regions, irrespective of the number of occurrences or positions across the nucleotide sequence, which may not be possible using string or regex based algorithm.

Thus, using Haar Wavelet Transforms is an efficient way to identify Short Tandem Repeat regions in variable length nucleotide sequences, in linear time, without requirement of prior pattern specification, or number of occurrences or awareness of the positions.