IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 10, Issue 1 (Mar. - Apr. 2013), PP 19-24
www.iosrjournals.org

Signal Processing Approach for Recognizing Identical Reads From DNA Sequencing of Bacillus Strains

Mamta C. Padole[*1], B. S. Parekh[2], D.P. Patel[3]
[1]*Department of Computer Science and Engineering, The Maharaja Sayajirao University of Baroda, India.*
[2]*Department of Computer Science and Engineering, The Maharaja Sayajirao University of Baroda, India.*
[3]*Department of Applied Mathematics, The Maharaja Sayajirao University of Baroda, India.*

Abstract *: DNA sequencing generates a large number of reads of lengths varying from 100bp to 1000bp, when sequenced using different methods of sequencing. These reads are further assembled to form contigs which are useful in annotation. The library generation using different amplification technique is involved in DNA sequencing process, which generates several identical reads, which are redundant, resulting in degraded quality of sequencing, besides also causing longer time for assembly. Existing computationally complex algorithms use string processing. The paper discusses the signal processing approach with application of Wavelet Transforms, designed to find exact and near exact identical reads. The string processing approach for pattern matching in search of similar patterns is computationally very expensive because the order of complexity of String comparisions is exponential in nature. Whereas Wavelet Transforms translates the sequence in co-efficients which are half of the length of the original sequence. On applying Wavelet Transforms repeatedly on the sequence, the sequence get transformed to half the length of the sequence used for transformations. Thus the order of complexity reduces to O(log n), which is much efficient compared to string processing.*

Keywords – *Haar wavelets, identical reads, pattern recognition, signal processing, wavelets,*

I. INTRODUCTION

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 [1], but has limitations in terms of throughput and cost of large genome sequencing[2]. The other methods are sequencing-by-hybridization (SBH), nanopore-sequencing and sequencing-by-synthesis [3]. "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 [2] sequences 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 [4] introduced by these identical reads, especially in the cases of analyzing quantification and transcriptome profiling sequence data.

In this paper, we present an *a priori method*, for recognizing identical reads, which does not require any mapping reference for recognizing identical read, nor does it need to compare any string pattern as an input parameter for comparison [4] neither does it use clustering on basis of seeds [6]. The paper discusses 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.

II. METHODS

The suggested algorithm for recognizing identical reads from the set of *DNA sequenced* reads is applied as in Fig. – 1 and the explanation following thereafter.

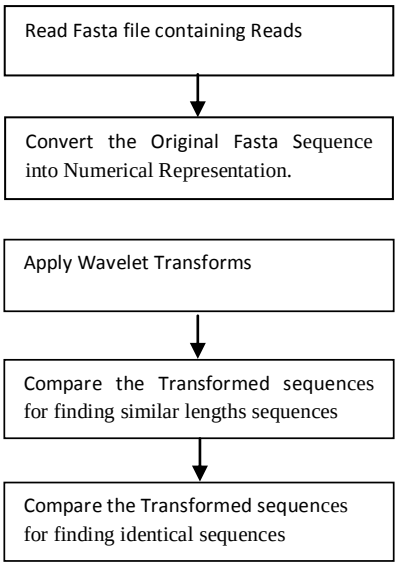

Fig. 1. Steps to Identify the Identical Reads

Read the fasta file which contains the sequence $S_i = \{s_1, s_2, \ldots s_n\}$ where $s_i \in \sum = \{ A,C,G,T \}$, 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. EIIP values for A= 0.1260, C = 0.1340, G = 0.0806, T = 0.1335 [25]. 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 [25]. Only numerical representations can be applied for Wavelet transformations.

The next step is to perform multi-level 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 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 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. Haar transforms are also memory efficient, as computations are performed in place.

2.1 Wavelet Transforms

A wavelet transform is a transformation of a signal or data into time-scale domain on a basis of wavelet functions [7][8]. The wavelet transform representations enable exploring the hidden information about the signal. Two co-efficient vectors [9] are generated, the approximate and the detail co-efficient vectors, after Wavelet transform is performed of the original signal.

When a signal *x* is passed through low pass filters (scaling functions) and high pass filters (wavelet functions) simultaneously, it is defined as performing the discrete wavelet transform (DWT) which along with down-sampling, generates co-efficients with half the length of the original input to each filter.

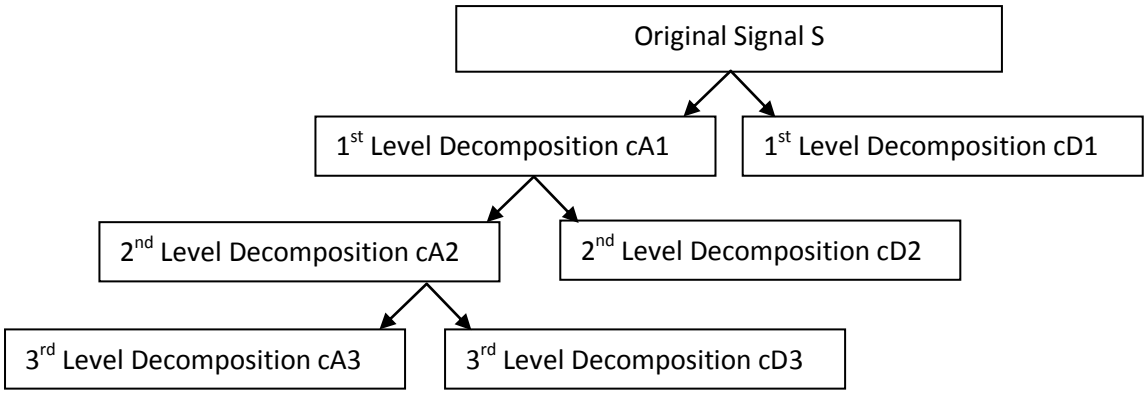

Fig. 2. The Decomposition Phase in Discrete Wavelet Transforms. After each level of transform and down-sampling, half the length of co-efficient are generated at each pass.

Wavelet Transform W_T can also be represented as in (1),

$$W_T = X.W, \quad where\ W = [\varphi(x); \psi(x)\,] \qquad (1)$$

As in (Equation 1.),

$\varphi(x)$ is called scaling function to find the approximate co-efficients and

$\psi(x)$ is called wavelet function to find the detail co-efficient

Wavelet decomposition can be applied to Haar wavelets are related to a mathematical operation called the Haar transform. All the wavelet transforms refer to Haar Wavelets as its prototype. [15]. any sequential data, including strings, where, in case of strings, the position of a character in string represents the time series data. Wavelets are tools used to study regularity and to conduct local studies [27]. The zero moments [24] of the function are related to the regularity of scaling function & wavelets [14].

2.1.1. Haar Wavelets

Haar wavelets are conceptually simple, fast and memory efficient [17], [18], can be computed without a temporary buffer, are exactly reversible, can be perfectly reconstructed and are defined to be orthonormal [16].

Applications of Haar wavelets are dimensionality reduction [11], approximate querying of database [12], image processing [10], selectivity estimation tasks, digital network synthesis [19], binary logic design [20] [21] [22].

The Haar transform decomposes a discrete signal *x* into two sub-signals of half its length. One sub-signal is a running average or trend (C_a) as in Table-1; the other sub-signal is a running difference or fluctuation (C_d) as in Table-1. [23].

1.1.2. Computation of Haar Wavelet Transform of Time Series Data

Consider a one-dimensional data vector X containing the N = 8 data values X = [8,8,0,8,12,20,16,16]

Table 1. Representation Of Computations Of Haar Wavelet Transform

Transformation Level or Decomposition Level (n)	**Resolution or Granularity (Order k)**	**Length of signal (L)**	**Averages / Approximate Co-efficients (Ca) Ca = (xi + xi+1)/2**	**Differences / Detail Co-efficients (Cd) Cd = (xi - xi+1)/2**
Original signal	3	8	**[**8,8,0,8,12,20,16,16**]**	-
1	2	4	[8,4,16,16]	[0, -4, -4, 0]
2	1	2	[6, 16]	[2,0]
3	0	1	[11]	[-5]

Haar wavelet transform, are computed by iteratively performing pair-wise averaging and differencing [13].

The values are first averaged together pair-wise to get a new "lower-resolution" representation of the data with the following average values [8, 4, 16, 16]. To restore the original values of the data array, additional detail coefficients must be stored to capture the information lost due to this averaging. In Haar wavelets, these detail coefficients are simply the differences of the second value and the first value, of the pair from the computed pair-wise average, divided by 2, i.e., [8-8, 4-8, 16−20,16−16] = [0, −4, −4, 0].

There is no information loss in this process; it is simple to reconstruct the eight values of the original data array from the lower-resolution array containing the four averages and the four detail coefficients.

Recursively applying the above pair-wise averaging and differencing process on the lower-resolution array containing the averages, gives the full transform, which can be explained as follows :

The Haar wavelet transform WT of the original signal X is the single coefficient representing the overall average of the data values followed by the detail coefficients [Table 1.] in the order of increasing resolution, i.e., WT = [11, -5, 2, 0, 0, -4, -4, 0]
Each entry is called a wavelet coefficient.

III. Results

The algorithm defined is tested on the Short Reads Archive (SRA) data. The SRA data downloaded from NCBI site ftp://ftp.ncbi.nlm.nih.gov/sra/ containing short reads. The sequenced reads are for various strains of Bacillus. Table 2. and Table 3. 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 2. 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 3. 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 4. 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 5. 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 4 & 5**, 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.

From DNA sequenced data of Bacillus with SRR149222 reference id, 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 [**TABLE 2**] and 3.8738% in terms of bases.

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

Fig. 3. 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. The Wavelet Transform based algorithm defined in this paper also helps in removing this type of identical and insignificant reads from the generated sequencing output.

IV. DISCUSSION

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. 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. 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.

V. CONCLUSION

The results reflect that the Wavelet Transforms can be applied to identify Duplicate Reads from DNA sequencing reads. It is also helpful in improving the efficiency in applying search for identical reads and is memory efficient.

REFERENCES

[1] Sanger, F.*et al*. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* 1977, **74:** 5463–5467.
[2] Ronaghi, M. Pyrosequencing Sheds Light on DNA Sequencing *Genome Res.* 2001, 11: 3-11
[3] Metzker, M. L. *et al*. Emerging Technologies in DNA Sequencing. *Genome Res.* 2005, 15: 1767-1776
[4] 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
[6] Gomez-Alvarez V. *et al*. Systematic artifacts in metagenomes from complex microbial communities. *ISME J*, 2009, 3 : 1314–1317
[7] Meher,J. K. et al. 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
[8] Daubechies, I. Ten Lectures On Wavelets, 1992
[9] Aggarwal, C. C. On the Use of Wavelet Decomposition for String Classification, *Springer - Data Mining and Knowledge Discovery*, 2005, 10, 117–139
[10] Castleman, K.R. Digital Image Processing, (Englewood Cliffs: Prentice-Hall, 1996)
[11] Keogh E. *et al.* Dimensionality reduction for fast similarity search in large time series databases, *J Knowledge Information Systems*, 2001, 3:263–286
[12] Garofalakis, M. Discrete Wavelet Transform and Wavelet Synopses, Springer Science+Business Media, LLC *Encyclopedia of Database Systems*, 2009, 10.1007/978-0-387-39940-9_539
[13] Sacharidis, D. Constructing Optimal Wavelet Synopses , *Proceedings of the 2006 International Conference on Current Trends in Database Technology* EBDT, 2006, 10.1007/11896548_10 Pg 97-104
[14] Burrus, C.S. *et al.* Introduction to Wavelets and Wavelet Transforms (*Prentice Hall*, 1998)
[15] Mohamed M. I. *et al.* Comparison between Haar and Daubechies Wavelet Transformtions on FPGA Technology, *Proceedings of World Academy Of Science, Engineering And Technology*, 2007, Volume 20 ISSN 1307-6884
[16] Moharir, P.S. Pattern recognition transforms, New York: (Wiley 1992)
[17] Radomir S. *et al.* The Haar wavelet transform: its status and achievements, Elsevier Science Ltd. *Computers and Electrical Engineering*, 2003 29 25–44
[18] Ruiz, G. *et al.* Switch-level fault detection and diagnosis environment for MOS digital circuits using spectral techniques, *IEE Proc Part E*, 1992, 139(4):293–307
[19] Hurst, S.L. The Haar transform in digital network synthesis, *Int Symp Multiple-valued Logic*, Proc, 1981, 11th. p. 10–8.
[20] Falkowski, B.J. Mutual relations between arithmetic and Harr functions, *Proceedings of IEEE Int Symp Circ Syst*, ISCAS, 1998, vol. V. p. 138–41.
[21] Falkowski, B.J. *et al.* Efficient algorithm for forward and inverse transformations between Haar spectrum and binary decision diagrams, Int Phoenix Conf Comput Commun,, 1994, p. 497–503.