IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-8, Issue-1, Jan 2020 E-ISSN: 2347-2693

Detection of Longest Common Sub Sequence in Normal DNA and Dengue Virus Affected Human DNA using Self Organizing Map

G. Tamilpavai[1*], C. Vishnuppriya[2]

[1,2]Dept. of Computer Science and Engineering, Government College of Engineering, Tirunelveli, Tamil Nadu, India

**Corresponding Author: tamilpavai@gcetly.ac.in, Tel.: +91-9442523888*

DOI: https://doi.org/10.26438/ijcse/v8i1.17 | Available online at: www.ijcseonline.org

Accepted: 01/Jan/2020, Published: 31/Jan/2020

***Abstract*—** Bioinformatics is an active research area which combines biological matter as well as computer science research. Detection of disease causing human Deoxyribo Nucleic Acid (DNA) sequence analysis is one of the major application areas under bioinformatics. Among the severe diseases, the number of Dengue cases and deaths are raised in Tamil Nadu. Identification of sequence motifs involved in Dengue virus is essential for early prediction and saving human life. It includes wide ranges of steps for disease diagnosing. The scope of this proposed work is to provide the longest common subsequence which present in a normal and Dengue virus affected human DNA sequence. The human DNA sequences are collected from National Center for Biotechnology Information (NCBI) database. Human DNA sequence is separated as k-mer using k-mer separation rule. From that, the separated k-mers are clustered using Self Organizing Map (SOM) algorithm. In which mean, median and standard deviation are used as features for clustering k-mers. Then obtained k-mers clusters are given to the Longest Common Subsequence (LCSS) algorithm to find common subsequence with higher length, which presents in every k-mers clusters. Time consumption for identification of LCSS is compared for both normal and Dengue virus affected DNA.

***Keywords*—**Bioinformatics, K-mers, Longest Common Sub Sequence (LCSS), String pattern matching algorithms.

I. INTRODUCTION

Bioinformatics combines biology, computer science, mathematics and statistics to analyze and interpret biological data. The recent advent of bioinformatics role in human health related application includes genome annotation, DNA sequence analysis, protein strands prediction, drug discovery, etc [1]. DNA carries the genetic information of an organism which consists of four nucleotide bases are Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) [2]. Figure 1 shows the structure of DNA.

Generally biological data are very large in size. So, it requires computational algorithms to perform the analysis on DNA sequences [3], genomic sequences, etc [4]. DNA sequence analysis is a process of determining the exact order of nucleotides within a DNA molecule. Changes of nucleotides order in the DNA sequence is called as mutation. A short repeated pattern of nucleotide bases which presents in human DNA sequence is called motif [5], [6]. If nucleotide order is changed in motif then it is called as “mutated motifs”. Biologists suggest that the diseases can be categorized from DNA according to the number of occurrence of mutated motif [7].

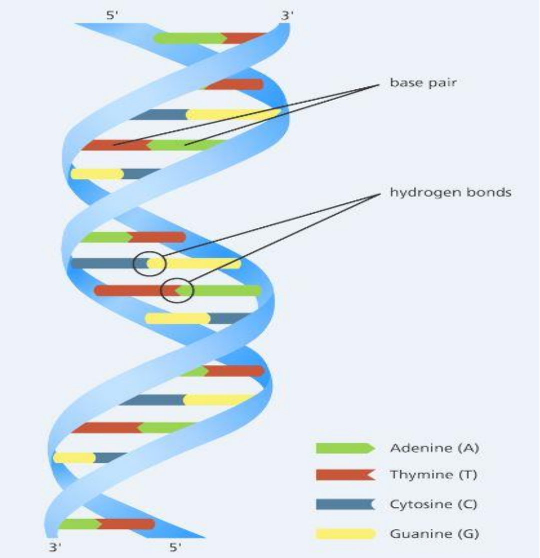


Figure 1: DNA structure

In protein the disease prediction is done using the single residue mutation, protein-protein interaction data, domain-domain interaction data, etc [8], [9], [10]. Evolutionary relationships of different living organisms are analyzed using Phylogenetic tree [11].

DNA sequence comparison is useful to identify the presence of abnormality in human DNA sequence. Calculation of DNA sequence similarity can be done based on the number of string matches in DNA sequence and number of characters matches between different DNA sequences [3].

Various similarity measuring algorithms are following the idea of LCSS. However, the new methodologies are developed in data exploration tools, still the time complexity is a major issue for researchers [3]. LCSS identification is one of the steps of disease causing pattern detection process. Hence in this proposed work, it aims to detect the LCSS in normal and Dengue affected human DNA sequences with lesser time consumption.

The paper is organized as follows. Section 2 deals about related works. Section 3 deals about methodology. Section 4 deals about experimental results. Section 5 deals about conclusion and future work.

II. RELATED WORK

This section discusses about the literatures related to the computational algorithms used in string processing and DNA sequence analysis.

S.Rajesh, S.Pramitha, Dr.L.S.S.Reddy [7], have proposed a method to detect the unusual pattern in DNA data using Knuth-Morris-Pratt (KMP) algorithm. The main objective of their work is to find out the start point and end point of the given sequence and the number of repetitions of a sequence. They have shown that KMP algorithm provides the minimum number of string matching comparisons and low time complexity.

Sumedha S.Gunewardena [12], has implemented the k-mer analysis algorithms for whole genome sequences. Author proposed, (i) A linear time algorithm for short k-mer analysis of the whole genome sequences (ii) An optimal time algorithm for k-mer analysis (GK-MER-COUNT) and (ii) A heuristic algorithm for large k-mer analysis (PGK-MER-COUNT). Author tested the proposed algorithms on various genome sequences namely human, mouse, 681 bacteria and 50 archaea.

Teuvo Kohonen, Panu Somervuo [13], have proposed the supervised and unsupervised learning methods for Self Organizing Maps (SOM) of symbol strings. For performing the multi-speaker word recognition experiment, they have used the 9x9 SOM with the data set of 20 speakers. The experiments were repeated four times with the different combination of training set and testing set. Finally, they have concluded that mean and median can be used for any dataset which containing the members were related by distance function. Marghny Mohamed, Abeer A. Al-Mehdhar, Mohamed Bamatraf, Moheb R.Girgis [14], have used the enhanced Self Organizing Map method for the classification of DNA sequence.

Izzat Alsmadi, Maryam Nuser [3], have evaluated the Longest Common Substring (LCS) and Longest Common Subsequence (LCSS) algorithm using different types of codes implementations for DNA sequence comparison. LCS was defined as longest common string which contains the consecutive characters and LCSS was defined as longest common subsequence in which characters need not be contiguous, but characters should be same order in forward direction. They have described the seven pseudo codes namely LCS1, LCS2, LCS3, LCS4, LCS5, LCS6 and LCS7 for LCS algorithm and six pseudo codes namely LCSS1, LCSS2, LCSS3, LCSS4, LCSS5 and LCSS6 for LCSS algorithm. For Longest Common Substring algorithm, pseudo codes from LCS1 to LCS5 were implemented using loops concepts. Then LCS6 and LCS7 were implemented using dynamic programming method. For Longest Common Subsequence algorithm, pseudo code LCSS1 was implemented by recursion. Authors noted that the LCSS2 was implemented by the Wiki books concept. LCSS3 was implemented using the similar concept of LCSS2 and Dynamic programming with back tracking method. LCSS4 and LCSS6 were developed using two-dimensional arrays and both uses two nested loops for back tracking, LCSS5 was implemented by dynamic programming method that also uses back tracking process. They have used 60 (randomly selected genome sequences) DNA sequence datasets those taken from National Center for Biotechnology Information (NCBI), for comparing the accuracy and performance of the described pseudo codes. DNA dataset sequence length includes 100, 500 and 1,000. Finally, they have concluded that evaluating same DNA sequences on different algorithms have shown different results.

Dr.S.A.M.Rizvi, Pankaj Agarwal [15], have presented the algorithm for finding the Longest Common Subsequence from two DNA or protein sequences. Authors used bucket based concepts for implementing the algorithm. Several authors Xuyu Xiang, Dafang Zhang, Jiaohua Qin [16], and Coasts S. Iliopoulos, M. Sohel Rahman [17], performs LCSS analysis in various ways using dynamic programming method.

The following observations are made from the literature survey. (i) For LCSS identification, dynamic programming method is suitable when compared to other looping concepts. (ii) Available every K-mer analysis algorithm has its own advantages and disadvantages, time and space complexity are dependent on the size of 'k'. (iii) Mean and median is the strongly proven features for SOM clustering.

Hence, in this proposed work, a linear time based short k-mer analysis principle is used for k-mer separation. Mean, median and standard deviation are used for clustering. Dynamic programming based LCSS identification method is used. Performance of LCSS identification in human DNA data set is measured using the elapsed time of LCSS identification.

III. METHODOLOGY

The steps involved in the proposed work are k-mer separation, feature extraction, k-mer clustering and LCSS identification. Figure 2 depicts the architecture diagram of the proposed system.

A. K-mer separation

The length of human DNA sequence (strings) is very large in size. It cannot be processed as it is. This is because search space of human DNA sequence will increase in further processing. So, the human DNA sequence is separated into k-mer (i.e. separating a lengthy human DNA sequence into substrings of the length k over alphabets {A, C, G, T}) using k-mer separation principle [12]. k-mer separation is done by eqn. (1),

$$\text{Number of k\_mer} = \mathrm{M} - \mathrm{K} + 1 \qquad (1)$$

where M is the length of human DNA sequence and K is the size of k-mer where 1≥K≤12. In this proposed system, human DNA k-mer size is considered as 7, which is randomly assumed.

B. K-mer clustering

Clustering is the process of grouping more similar and dissimilar things into individual groups. For grouping relevant human DNA k-mer pattern, three features are extracted from human DNA k-mer namely mean, median and standard deviation. ASCII values of characters are used for feature extraction of separated k-mer. Based on these three input feature vectors, every human DNA k-mer are grouped into clusters using Self Organizing Map.

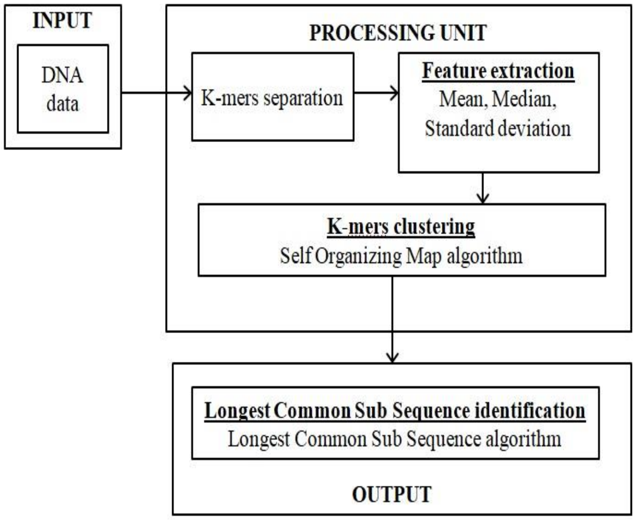

INPUT
DNA data
PROCESSING UNIT
K-mers separation
Feature extraction
Mean, Median, Standard deviation
K-mers clustering
Self Organizing Map algorithm
Longest Common Sub Sequence identification
Longest Common Sub Sequence algorithm
OUTPUT

Figure 2: Proposed architecture diagram

1) Self Organizing Map

Self Organizing Map is one type of Artificial Neural Network (ANN) [14]. It follows unsupervised learning [13], [14]. It maps high dimensional data onto a low dimensional grid, like hexagonal or rectangular two dimensional grids [14]. Based on distance measures SOM can also be performed well for strings too, not only restricted to numerical data [13]. Four steps of SOM algorithm are (a) initialization (b) activation (c) updating and (d) continuation [14].

a) *Initialization*: Randomly values are chosen for initial weight vectors W_j and a small positive value is assigned to the learning rate parameter α.

b) *Activation:* Input vector X is applied to activate the SOM network. Using the minimum Euclidean distance measure, the Best Matching Unit (BMU) neuron $\mathrm{X_i}$ at iteration p is determined. It is given by eqn. (2),

$$\mathrm{E} = \min_j \left\| \mathrm{X} - \mathrm{W_j(p)} \right\| = \sqrt{\textstyle\sum_{i=1}^{n} \left[\mathrm{X_i} - \mathrm{W_{ij}\,(p)} \right]^2} \qquad (2)$$

where $\mathrm{X_i}$ is the input vector and i=1,2,…n, where n is the number of neurons in the input layer, where $\mathrm{W_{ij}}$ (p) is the weight repairing at iteration p and i=1,2,…n where n is the number of neurons in the input layer and j=1,2,…m where m is the number of neurons in the SOM layer.

c) *Updating:* Weight update equation is applied. It is given by eqn.(3),

$$\mathrm{W_{ij}\,(p+1) = W(p) + \Theta(p)\alpha(p)(X(p) - W_{ij}\,(p))} \qquad (3)$$

Where $\mathrm{W_{ij}}$ (p) is the weight repairing at iteration p and i=1,2,…n where n is the number of neurons in the input layer and j=1,2,…m where m is the number of neurons in the SOM layer, where Θ is the distance from the BMU i.e. neighborhood function.

d) *Continuation*: Until no changes stage occurs in the feature map, repeat from step (b).

C. Longest Common Sub Sequence identification

Obtained clusters of DNA k-mer are used as input for LCSS algorithm, to find out the LCSS. Number of LCSS identified, for one cluster is given by eqn. (4)

$$\text{Number of LCSS} = \mathrm{n} * (\mathrm{n} - 1)/2 \qquad (4)$$

where n is the number of DNA k-mer in a cluster.

1) Longest Common Sub Sequence algorithm

LCSS algorithm is used to find out the similarity between two sequences. LCSS is the longest sub-sequence obtained from two different sequences. Obtained sequence should contain the characters in same order but need not be contiguous [3]. In the proposed system, LCSS is performed using dynamic programming method. It follows the three cases are (i) if either sequence is empty (ii) if characters match (iii) if characters do not match. They are shown in eqn.(5) to eqn.(7) respectively,

$$s[i, j] = 0, \quad \text{if } i = 0 \text{ or } j = 0 \qquad (5)$$

$$s[i, j] = s[i-1, j-1] + 1, \quad \text{if } i, j > 0 \text{ and } x_i = y_j \qquad (6)$$

$$s[i, j] = \max(s[i, j-1], s[i-1, j]), \text{if } i, j > 0 \text{ and } x_i \neq y_j (7)$$

where i and j is the length of first and second sequence respectively, where i=0,1,…, m and j=0,1,…, n, where s[i, j] represents the length of subsequence in the dynamic programming table, where x_i is the i^{th} character of first sequence, y_j is the j^{th} character of second sequence.

After finding the length of subsequence, the actual LCSS is extracted using back tracking process from the dynamic programming table which is already constructed to find the length of subsequence. Three cases of back tracking is shown in eqn.(8) to eqn.(10)

$$b[i, j] = 3, \text{if } x_i = y_j$$

$$(\text{i.e. character present in LCSS, if } s[i, j] = s[i-1, j-1] + 1)$$

(8)

$$b[i, j] = 2, \text{if } x_i \neq y_j$$

$$(\text{i.e. character } y_j \text{ not in LCSS, if } s[i, j] = s[i, j-1]) \quad (9)$$

$$b[i, j] = 1, \text{if } x_i \neq y_j$$

$$(\text{i.e. character } x_i \text{ not in LCSS, if } s[i, j] = s[i-1, j] \quad (10)$$

Where b[i, j] represents the back tracking case of subsequence in the dynamic programming table.

IV. EXPERIMENTAL RESULTS

In the proposed system, MATLAB 2013b tool is used for k-mer separation, k-mer feature extraction for clustering and LCSS identification. Orange 2.7 tool is used for k-mer clustering.

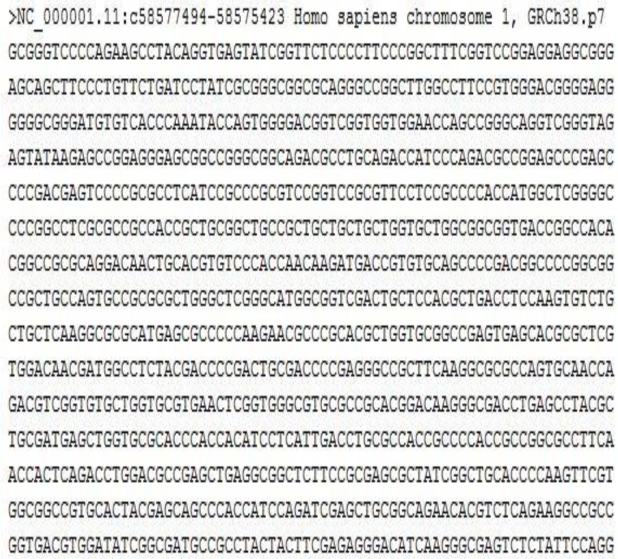
A. *Data set*

FASTA format of human DNA sequences are collected from National Center for Biotechnology Information (NCBI) database. Collected data set consists of 12 normal human DNA data and 4 types of Dengue virus affected human DNA data. Table 1 shows the details of data set.

Table 1. Details of data set

Accession number of data in NCBI	**Length of data (base pairs i.e. bp)**	**Category of data**
NC_000001.11	2072	Normal human
NC_000002.12	14866	Normal human
NC_000003.12	20571	Normal human
NC_000004.12	206053	Normal human
NC_000005.10	185501	Normal human
NC_000006.12	81390	Normal human
NC_000007.14	78524	Normal human
NC_000008.11	1841	Normal human
NC_000009.12	27321	Normal human
NC_000010.11	108493	Normal human
NC_000011.10	116962	Normal human
NC_000012.12	24663	Normal human
NC_001477.1	10735	Dengue virus 1 affected human
NC_001474.2	10723	Dengue virus 2 affected human
NC_001475.2	10707	Dengue virus 3 affected human
NC_002640.1	10649	Dengue virus 4 affected human

In this section experimental results and analysis are discussed for one data of normal human DNA i.e. NC_000001.11 and one data of Dengue virus affected human DNA i.e. NC_001477.1. Portion of FASTA format of human DNA data is shown in Figure 3.



```
>NC_000001.11:c58577494-58575423 Homo sapiens chromosome 1, GRCh38.p7
GCGGGTCCCCAGAAGCCTACAGGTGAGTATCGGTTCTCCCCTTCCCGGCTTTCGGTCCGGAGGAGGCGGG
AGCAGCTTCCCTGTTCTGATCCTATCGCGGGCGGCGCAGGGCCGGCTTGGCCTTCCGTGGGACGGGGAGG
GGGGCGGGATGTGTCACCCAAATACCAGTGGGGACGGTCGGTGGTGGAACCAGCCGGGCAGGTCGGGTAG
AGTATAAGAGCCGGAGGGAGCGGCCGGGCGGCAGACGCCTGCAGACCATCCCAGACGCCGGAGCCCGAGC
CCCGACGAGTCCCCGCGCCTCATCCGCCCGCGTCCGGTCCGCGTTCCTCCGCCCCACCATGGCTCGGGGC
CCCGGCCTCGCGCCGCCACCGCTGCGGCTGCCGCTGCTGCTGCTGGTGCTGGCGGCGGTGACCGGCCACA
CGGCCGCGCAGGACAACTGCACGTGTCCCACCAACAAGATGACCGTGTGCAGCCCCGACGGCCCCGGCGG
CCGCTGCCAGTGCCGCGCGCTGGGCTCGGGCATGGCGGTCGACTGCTCCACGCTGACCTCCAAGTGTCTG
CTGCTCAAGGCGCGCATGAGCGCCCCCAAGAACGCCCGCACGCTGGTGCGGCCGAGTGAGCACGCGCTCG
TGGACAACGATGGCCTCTACGACCCCGACTGCGACCCCGAGGGCCGCTTCAAGGCGCGCCAGTGCAACCA
GACGTCGGTGTGCTGGTGCGTGAACTCGGTGGGCGTGCGCCGCACGGACAAGGGCGACCTGAGCCTACGC
TGCGATGAGCTGGTGCGCACCCACCACATCCTCATTGACCTGCGCCACCGCCCCACCGCCGGCGCCTTCA
ACCACTCAGACCTGGACGCCGAGCTGAGGCGGCTCTTCCGCGAGCGCTATCGGCTGCACCCCAAGTTCGT
GGCGGCCGTGCACTACGAGCAGCCCACCATCCAGATCGAGCTGCGGCAGAACACGTCTCAGAAGGCCGCC
GGTGACGTGGATATCGGCGATGCCGCCTACTACTTCGAGAGGGACATCAAGGGCGAGTCTCTATTCCAGG
```

Figure 3: FASTA format of human DNA data

B. *K-mer separation*

Collected FASTA format of human DNA data is given as input to the k-mer separation. In human DNA, same pattern of k-mers are presented more than one time. Number of occurrence of k-mers is counted and the pattern of k-mer is taken one time for further processing. Separated k-mers (k-mer size=7) count for normal human DNA data and Dengue virus 1 affected human DNA data is 1,857 and 6,773 respectively. Due to the large count of separated k-mers, some samples of separated k-mers for normal human DNA data and Dengue virus 1 affected human DNA data are shown in Table 2.

Table 2. Separated k-mers of size 7, for normal human DNA data and Dengue virus 1 affected human DNA data

Separated K-mers for normal human DNA data	
DNA k-mer	**Number of occurrence of k-mer**
CCCACCA	4
CTGGTGC	4
AGGGCCG	3
CACCGCC	3
AAAATGT	2
Separated K-mers for Dengue virus 1 affected human DNA data	

AGGAAAA	12
CATGGAA	12
ATGGAAC	11
AAAGAAA	9
AAAAATG	8

C. *K-mer clustering*

Separated k-mers are used in k-mer clustering. Characters in each k-mer are converted into numerical value using the ASCII value and then features are extracted. Calculated features like mean, median and standard deviation of some samples of separated k-mers, for normal human DNA data and Dengue virus 1 affected human DNA data are shown in Table 3.

Table 3. Features of k-mer – Mean, Median and standard deviation

Features for Normal human DNA data			
DNA k-mer	**Mean**	**Median**	**Standard deviation**
CCCACCA	66.4285	67	0.9759
CTGGTGC	73.5714	71	7.3452
AGGGCCG	69	71	2.5819
CACCGCC	67.2857	67	1.7994
AAAATGT	71.2857	65	8.9575
Features for Dengue virus 1 affected human DNA data			
DNA k-mer	**Mean**	**Median**	**Standard deviation**
AGGAAAA	66.7142	65	2.9277
CATGGAA	69.7142	67	6.8487
ATGGAAC	69.7142	67	6.8487
AAAGAAA	65.8571	65	2.2677
AAAAATG	68.5714	65	7.1614

Calculated features of all separated k-mers are given to SOM algorithm. Based on those feature values, SOM forms the k-mer clusters. In this work, the size of 8x8 SOM mapping topology is constructed (topology size is randomly assigned), that creates totally 64 nodes (i.e. 64 k-mer clusters). Sample of first layer of SOM node position (i.e. (0,0) cluster) and number of instances for normal human DNA data and Dengue virus 1 affected human DNA data are shown in Table 4. Sample of K-mer cluster details, for normal human DNA data and Dengue virus 1 affected human DNA data is shown in Table 5. Due to large number of instances in every cluster, only 5 instances of first cluster i.e. (0,0) are shown in Table 5.

Table 4. Sample of first layer of SOM node position and number of instances for normal human DNA data and Dengue virus 1 affected human DNA data

Node (cluster) position in SOM	**Number of instances in a node (Number of k-mers in cluster for Normal human DNA data)**	**Number of instances in a node (Number of k-mers in cluster for Dengue virus 1 affected human DNA data)**
(0,0)	48	201
(0,1)	6	168
(0,2)	17	123
(0,3)	20	160
(0,4)	26	0
(0,5)	15	127
(0,6)	30	177
(0,7)	11	0

D. *Longest Common Sub Sequence identification*

Sample of obtained LCSS and its length for kmer clusters of normal human DNA data and Dengue virus 1 affected human DNA data are shown in Table 6. In Table 6, LCSS for nodes (0, 0) is discussed for both normal and Dengue virus 1 affected human DNA data.

Table 5. Sample of K-mer cluster details for Normal human DNA data and Dengue virus 1 affected human DNA data

K-mer cluster details for Normal human DNA data				
Node (cluster) position in SOM	**DNA k-mer**	**Mean**	**Median**	**Standard deviation**
(0,0)	CCTCATT	74.0000	67	9.3808
	TCCCCTT	74.2857	67	9.0868
	AATTCCT	73.7142	67	9.6559
	ACTACTT	73.7142	67	9.6559
	ATCCATT	73.7142	67	9.6559
K-mer cluster details for Dengue virus 1 affected human DNA data				
(0,0)	GTGTGGT	76.5714	71	6.9487
	ATGGTGT	75.7142	71	8.0356
	GGTGTTG	76.5714	71	6.9487
	GTGTGTG	76.5714	71	6.9487
	TGTGGTA	75.7142	71	8.0356

Table 6. LCSS and its length for k-mer cluster of Normal human DNA data and Dengue virus 1 affected human DNA data

LCSS for node or cluster at (0,0) for normal human DNA data			
First Sequence (First K-mer)	**Second Sequence (Second K-mer)**	**Length of Longest Common Sub sequence**	**Longest common sub Sequence**
CCTCATT	TCCCCTT	5	CCCTT
TCCCCTT	AATTCCT	4	TCCT
AATTCCT	ACTACTT	4	ATCT
ACTACTT	ATCCATT	5	ATCTT
LCSS for node or cluster at (0,0) for Dengue virus 1 affected human DNA data			
GTGTGGT	ATGGTGT	5	TGTGT
ATGGTGT	GGTGTTG	5	GGTGT
GGTGTTG	GTGTGTG	6	GGTGTG
GTGTGTG	TGTGGTA	5	TGTGG

Time taken for LCSS identification is calculated for all data in dataset and those values are tabulated in Table 7. Time consumption for LCSS identification of all data is shown as a graphical representation in Figure 4. Based on the length of base pairs time consumption is varry. DNA data which contains large number of base pairs consumes more time than the less number of base pairs.

Table 7. Time consumption for LCSS identification for Normal human DNA data and Dengue virus affected human DNA data

Time consumption for LCSS identification - Normal human DNA data			
Accession number of data in NCBI	**Length of data (base pairs i.e. bp)**	**Time (in secs.)**	**Time (in mins.)**
NC_000001.11	2072	16.6161933	0.276936555
NC_000002.12	14866	230.5511	3.8425
NC_000003.12	20571	306.4553	5.107588498
NC_000004.12	206053	732.5936	12.20989
NC_000005.10	185501	705.837	11.76395

NC_000006.12	81390	520.0093	8.666822
NC_000007.14	78524	747.8475	12.46412
NC_000008.11	1841	11.08633	0.184772
NC_000009.12	27321	314.4697	5.241162
NC_000010.11	108493	628.5561	10.47593
NC_000011.10	116962	619.0872	10.31812
NC_000012.12	24663	296.0182	4.933636
Time consumption for LCSS identification – Dengue virus affected human DNA data			
NC_001477.1	10735	152.5332	2.542219
NC_001474.2	10723	136.5997	2.276661
NC_001475.2	10707	157.8308	2.630514
NC_002640.1	10649	184.4124	3.073541

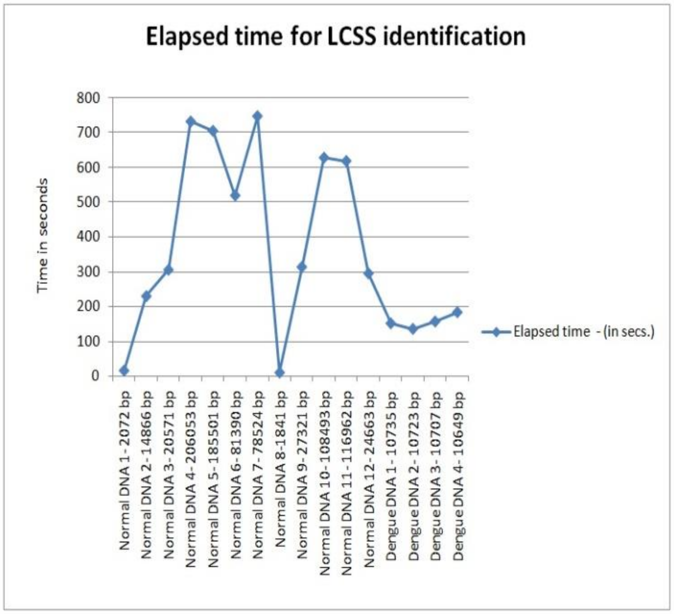


Figure 4: Time consumption of LCSS identification for normal and Dengue virus affected human DNA sequences

V. CONCLUSION AND FUTURE SCOPE

The proposed work is developed for detecting the LCSS from human DNA sequences. Search space of the whole DNA sequence gets reduced by the k-mer size which is used in k-mer separation principle. Hence, k-mer separation seems to be effective for further processing of human DNA sequences. SOM algorithm is used for k-mer clustering and it gives the transparent grouping results for k-mer clusters with mean, median and standard deviation. Dynamic programming method of LCSS algorithm is suitable for detecting the LCSS from the human DNA sequences. Experimental outcomes of this proposed work produce the possible number of LCSS in normal and Dengue virus affected human DNA data. From the analysis of time consumption for LCSS identification, it is concluded that the larger length of DNA sequences takes little more time than the lesser length of DNA sequences. Due to the non continuous pattern in the identified LCSS, it may miss some biological meaning in DNA sequences. So the proposed work still requires improvement to overcome this limitation.

In future work, the research can be focused on detecting Longest Common Substring (LCS) from the human DNA sequence to ascertain the strong biological nature.

ACKNOWLEDGMENT

This work is one of the outcome of the below mentioned funded project. The authors would like to thank, “Science and Engineering Research Board (Statutory Body Established through an Act of Parliament: SERB ACT 2008), Department of Science and Technology, Government of India”, for providing the financial support - grant number EEQ/2016/000255.

REFERENCES

[1] Vinayak Majki, Sudip Paul and Rachna Jain, “Bioinformatics for Healthcare Applictions”, IEEE Conference, pp.**2014-207, 2019**.

[2] Terasa K.Attwood, David J.Parry-Smith and Phukan, Introduction to bioinformatics, Noida(U.P), India: Pearson India Education Services Pvt. Ltd, pp.**221**, **2008**.

[3] Izzat Alsmadi and Maryam Nuser, “String Matching Evaluation Methods for DNA Comparison”, *International Journal of Advanced Science and Technology”*, Vol.**47**, pp.**13-32**, **2012**.

[4] Sasikala S, Ratha Jeyalakshmi T, “Extensive Review on Computational Predictions of Genomic Regulatory Sequences”, *International Journal of Computer Sciences and Engineering*, Vol.**07**, Issue.**08**, pp.**91-94**, **2019**.

[5] Amit U Sinha and Raj Bhatnagar, “Efficient and Scalable Motif Discovery using Graph-based Search”, IEEE symposium on Computational Intelligence in Bioinformatics and Computational Biology, pp.**197-204**, **2007**.

[6] Khumukcham Robindro, Ashoke Das, “Effectiveness of Ssaha Algorithm for Searching Motif in Large Databases of DNA Sequences” , *International Journal of Scientific Research in Computer Science and Engineering*, Vol.**5**, Issue.**4**, pp.**79-87**, **2017**.

[7] S.Rajesh, S.Prathima and Dr.L.S.S.Reddy, “Unusual Pattern Detection in DNA Database using KMP Algorithm”, *International Journal of Computer Applications (0975-8887)*, Vol.**1**, Issue.**22**, pp.**1-5**, **2010**.

[8] Benjamin Schuster-Bockler and Alex Bateman, “Protein interactions in human genetic diseases”, Genome Biology, Vol.**9**, Issue **1**, Article R9, pp.**R9.1-R9.12**, **2008**.

[9] Chein-Hung Huang, Huai Shun Peng and KA-Lok Ng, “Prediction of Cancer Proteins by Integrating Protein Interaction, Domain frequency and Domain Interaction Data using Machine Learning Algorithms”, BioMed Research International, Vol.**2015**,pp.**1-10**, **2015**.

[10] Lei Yang, Xudong Zhao and Xianglong Tang, “Predicting Disease-Related Proteins Based on Clique Backbone in Protein-Protein Interaction Network”, *International Journal of Biological Sciences*, Vol.**10**, Issue.**7**, pp.**677-688**, **2014**.

[11] Pankaj Bhanbri, O.P. Gupta, “Phylogenetic Tree Construction for Distance based Methods”, *International Journal of Scientific Research in Computer Science and Engineering*, Vol.**5**, Issue.**3**, pp.**142-149**, **2017**.

[12] Sumedha S.Gunawardena, “Optimum-time, Optimum-space, Algorithms for k-mer Analysis of Whole Genome Sequences”, *Journal of Bioinformatics and Comparative Genomics*, Vol.**1**, pp.**1-12**, **2014**.

[13] Teuvo Kohonen and Panu Somervuo, “Self-organizing maps of symbol strings”, Elsevier, Neurocomputing 21, pp.**19-30**, **1998**.

[14] Marghny Mohamed, Abeer A. Al-Mehdhar, Mohamed Bamatraf and Moheb R.Girgis, "Enhanced Self-Organizing Map Neural Network for DNA Sequence Classification", *Intelligent Information Management*, Vol.**5**, pp.**25-33**, **2013**.

[15] Dr.S.A.M.Rizvi and Pankaj Agarwal, "A New Bucket-Based Algorithm for Finding LCS from two given Molecular Sequences", IEEE, Third International Conference on Information Technology: New Generations, **2006**.

[16] Xuyu Xiang, Dafang Zhang and Jiaohua Qin, "A New Algorithm for the Longest Common Subsequence Problem", IEEE, International Conference on Computational Intelligence and Security Workshops, pp.**112-115**, **2007**.

[17] Coasts S. Iliopoulos and M. Sohel Rahman, "Algorithms for Computing Variants of the Longest Common Subsequence Problem", Elsevier – Theoretical Computer Science, pp.**255-267**, **2008**.

Authors Profile

Dr.G.Tamilpavai, she completed her B.E in Computer Science and Engineering from Thiagarajar College of Engineering, Madurai, Tamil Nadu, India. She did her P.G in Government College of Engineering, Tirunelveli, Tamil Nadu, India. She Completed her Ph.D. at Anna University, Chennai, Tamil Nadu, India. Her area of interest includes medical image processing, remote sensing, bio informatics and operating systems. She is working as Associate Professor (CAS) and Head in Department of Computer Science and Engineering at Government College of Engineering, Tirunelveli. She has 20 years of teaching experience. She is recognized guide in Anna University, Chennai, Tamil Nadu, India. She has 16 publications in international journals especially in biomedical image processing and bio informatics. She has published many papers in National and International conferences. She has life membership in ISTE, IE and BMESI. She received fund from SERB, Department of Science and Technology, Government of India for the project entitled "Detecting Defective DNA motifs to find genetic disease sequence in a Human DNA using SOM". SERB sanctioned Rs.14,78,000 for the project (project duration 2017- 2020).

C.Vishnuppriya, she completed her B.Tech. degree in Information Technology in 2014 and M.E. degree in Computer Science and Engineering in 2016 from Anna University Regional Campus- Tirunelveli Region, Tirunelveli, Tamil Nadu, India. Her research interests include image processing and bioinformatics. She has one publication in national journal for Siddha medicine related image processing and 3 publications in international journal for bio informatics. She has published 4 papers in international conferences. She is working as Senior Research Fellow in Department of Computer Science and Engineering at Government College of Engineering, Tirunelveli, Tamil Nadu, India.