Identification of Disease Causing DNA Motifs in Human DNA Using Clustering Approach

G. Tamilpavai, C. Vishnuppriya

***Abstract*—**Studying DNA (deoxyribonucleic acid) sequence is useful in biological processes and it is applied in the fields such as diagnostic and forensic research. DNA is the hereditary information in human and almost all other organisms. It is passed to their generations. Earlier stage detection of defective DNA sequence may lead to many developments in the field of Bioinformatics. Nowadays various tedious techniques are used to identify defective DNA. The proposed work is to analyze and identify the cancer-causing DNA motif in a given sequence. Initially the human DNA sequence is separated as k-mers using k-mer separation rule. The separated k-mers are clustered using Self Organizing Map (SOM). Using Levenshtein distance measure, cancer associated DNA motif is identified from the k-mer clusters. Experimental results of this work indicate the presence or absence of cancer causing DNA motif. If the cancer associated DNA motif is found in DNA, it is declared as the cancer disease causing DNA sequence. Otherwise the input human DNA is declared as normal sequence. Finally, elapsed time is calculated for finding the presence of cancer causing DNA motif using clustering formation. It is compared with normal process of finding cancer causing DNA motif. Locating cancer associated motif is easier in cluster formation process than the other one. The proposed work will be an initiative aid for finding genetic disease related research.

***Keywords*—**Bioinformatics, cancer motif, DNA, k-mers, Levenshtein distance, SOM.

I. Introduction

DNA is the hereditary information in human and almost all other organisms. Nearly every cell in a human body has the same DNA. Most DNA is located in the cell nucleus. DNA is represented by four chemical bases called adenine (A), guanine (G), cytosine (C), and thymine (T). In genreral 3 billion bases are there in human DNA. The sequence, of these bases determines the information available for building and maintaining an organism. In DNA, A and C are paired with T and G Respectively. There pairs are called as base pair. Each base is handy with sugar phosphate Combination of base, sugar, and phosphate are called as nucleotide. Nucleotides are settled in two long strands that form a spiral called a double helix.

A DNA motif is defined as a nucleic acid sequence pattern that has some biological significance. Normally, the pattern is fairly short (5 to 20 base pairs (bp) long) and is known to recur in different genes or several times within a gene [1]. Finding motifs in genomic DNA sequence is one of the most important and challenging problems in both bioinformatics and computer science. Motifs are short, recurring patterns in DNA sequence having biological function [22].

Many ongoing developments are in this field. Experimental approaches for finding DNA motifs, e.g., ChIP-chip [2], ChIP-seq, and micro-array technology [3], are still laborious, time consuming, and expensive. Last two decades, more number of computational approaches have been used for DNA motif analysis [4]. Over these years, many motif search algorithms and web-based tools have been developed based on computational intelligence systems and data mining. Several tools are used for computational processing yields not adequate performance. String representation and matrix representation are the two popular methods for motif representation [5], [6]. In order to prevent the consequence of disease, it is necessary to find the presence of disease causing sequence in the human DNA at early stage. Hence it is required to find an accurate detection of sequence match in DNA. Thus, effective motif discovery remains challenging in spite of the good number of endeavors over the previous years, which requires the investigation of potential outcomes for enhanced advancements.

The paper is organized as follows; Section II explains the related works. Section III describes the methodology used for identifying the disease-causing DNA motifs from DNA sequence. Experimental results are discussed in section IV. Section V deals about the conclusion and future enhancements of the proposed work.

II. Related Works

To process the DNA sequence and discover the DNA motif from the massive DNA data set [13], a web service called Argo Compute Unified Device Architecture was designed. Clustering using genetic algorithm (GA) [7] helps to find regulatory motifs in DNA sequence. GA with three kinds of operations such as addition, deletion, mutation is considered for motif finding [10]. Other algorithms such as swarm intelligence based Gravitational Search algorithm (GSA) and Artificial Bee Colony algorithm also helpful in DNA motif discovery [11]. Moreover, disease specific and healthy control specific motifs can also be discoverable one using computational algorithm [12]. Mutation in DNA motif can be identified [15] using scalar and vector scoring representation methods. A(C/G)AA(C/G)(A/T) is a motif in association with hotspots in various human cancers such as blood cancer, breast cancer, lung cancer, kidney cancer, stomach cancer, liver cancer, large

G. Tamilpavai and C. Vishnuppriya are with Computer Science and Engineering Department, Government College of Engineering, Tirunelveli, Tamil Nadu, India, 627007 (e-mail: tamilpavai@gcetly.ac.in, c.vishnuppriya@gmail.com).

intestine cancer, etc. [21]. Real DNA dataset uses for analysis require some kind of DNA owner's privacy. A private DNA motif finding algorithm was proposed [8] to protect the privacy.

To handle large size of data sets k-mer subsequence analysis is essential in computational algorithm [17]. It will reduce processing time and occupy less space in DNA sequence processing. Clustering based DNA motif discovery methods extracts clusters of same length k-mer sequences. SOM is a powerful clustering algorithm and fuzzy SOM was used for DNA motif discovery [9]. SOM is a very good structural classifier for finding novel motifs [16]. To improve the performance of SOM, noise filtering from the input DNA sequence is useful [14]. Distance of strings is calculated using Levenshtein method and feature distance method [18]. Similarity between two strings can be measured by the distance between them. Number of transformations required to transform one string to another string is termed as Levenshtein distance. A transformation represents the insertion, deletion and substitution of characters in a string [20]. Classification of DNA sequences are done using the algorithm of enhanced SOM and Eugene's Hom, MEME is used to predict the system efficiency [19].

Observations made from the above discussed literature are (i) Motif discovery from a DNA sequence is a challenging task. (ii) Clustering and artificial intelligence algorithms are suitable for motif discovering process (iii) without losing DNA information, computationof k-mer separation (sub sequencing of DNA data) is required (iv) SOM performs well for clustering of DNA sequence in motifsfinding. (v) Distance metrics is essential to calculate the similarity between strings.

III. Methodology

The proposed method contains three steps as follows *A.* K-mer separation *B.* Clustering k-mers*C.*Identification of cancer motif. The diagrammatic representation of the proposed method is shown in Fig. 1.

A. K-mer Separation

Human DNA sequences are collected from National Center for Biotechnology Information (NCBI). DNA is separated as k-mers using k-mer separation rule. Due to the large size (total number of characters presents in the DNA sequence) of the DNA sequence, it is very difficult to process the total sequence. Hence k-mer separation is performed. K-mers are substrings of the DNA sequences over the alphabets {A, C, G, T}. Initially the k-mer length (k) is fixed according to the length of the disease associated motifs. Total numbers of k-mers [17] are obtained from (1),

$$Total\ \text{numbers}\ of\ k_{mers} obtained = N - K + 1 \qquad (1)$$

where N is the length of human DNA sequence and K is the size of k-mer & 1≤ K ≤12. In this work, the size of the reference pattern of cancer associated human DNA motif is six. So, the size of k-mer is fixed as six for processing.

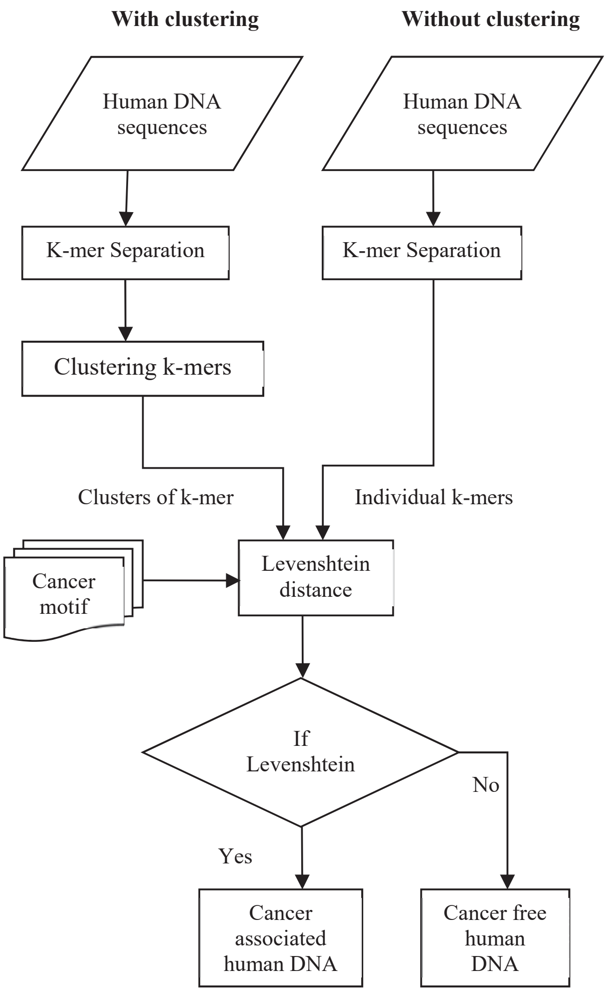


Fig. 1 Flow diagram of proposed methodology

B. Clustering K-mers

An amount of the similarity between two patterns drawn from clustering procedure. Mean and median feature values are extracted from human DNA for grouping k-mers. Based on these two input features, k-mer is grouped into clusters using SOM.

1. Self-Organizing Map

SOM is a kindof Artificial Neural Network [19]. It follows unsupervised learning [18], [19]. It maps high dimensional data into low dimensional grid, like hexagonal or rectangular two-dimensional grids [19]. Based on distance measures SOM can performed well for strings and numerical data [18]. SOM is implemented in four steps such as (a) initialization (b) activation (c) updating and (d) continuation [19].

(a) Initialization

Random 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 utilized for activating SOM network. Using minimum Euclidean distance measure, the Best Matching Unit (BMU) neuron X_i at iteration p is determined. It is given by (2):

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

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

(c) Updating

Weight update is done using (3):

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

where $W_{ij}(p)$ is the weight repairing at iteration p,$i = 1,2, \dots n$,here n is the number of neurons in the input layer, $j = 1,2, \dots m$ 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 change occurs in the feature map, repeat step (b) and step (c).

C. Identification of Cancer Motifs

Human DNA sequence containing a short length of nucleotides pattern called motif. Reference pattern of cancer associated human DNA motifs [21] are used, for identifying the presence of cancer motifs in the human DNA k-mers cluster. TABLE I shows the cancer associated human DNA motifs (length of motif is 6). Levenshtein distance measure is used for the identification of cancer motifs.

TABLE I
CANCER ASSOCIATED HUMAN DNA MOTIFS

Motif No.	Cancer associated motifs
1	ACAACA
2	ACAACT
3	ACAAGA
4	ACAAGT
5	AGAACA
6	AGAACT
7	AGAAGA
8	AGAAGT

1. Levenshtein Distance Measure

Similarity between two strings (source, target) is measured by Levenshtein distance [20]. It is also called as edit distance. In this work, Levenshtein distance is measured using dynamic programming. K-mer string is considered as source and cancer associated motif string is considered as target. The distance obtained from this measure refers to the number of insertions, deletions or substitutions are required to transform the k-mer string to cancer associated motif string.

If the obtained distance is zero (i.e. all the characters of k-mer string is exactly matches with the characters of cancer associated motif string), then it represents the human DNA sequence having the cancer associated motifs. While the obtained distance is other than zero (i.e. 1, 2, 3 . . . , k where k is the length of k-mer string), then it considered human DNA sequence having no such cancer associated motifs.

Dynamic programming has two cases (i) Match occurrence of source and target characters (ii) Mismatch occurrence of source and target characters. These two cases are shown in (4) and (5) respectively:

$$d[n,m] = d[n-1,m-1] + 0,$$
$$if\ n,m > 0\ and\ s_n = t_m (4)$$
$$d[n,m] = 1 + min\begin{pmatrix} d[n,m-1], d[n-1,m], \\ d[n-1,m-1] \end{pmatrix},$$
$$if\ n,m > 0\ and\ s_n \neq t_m \qquad (5)$$

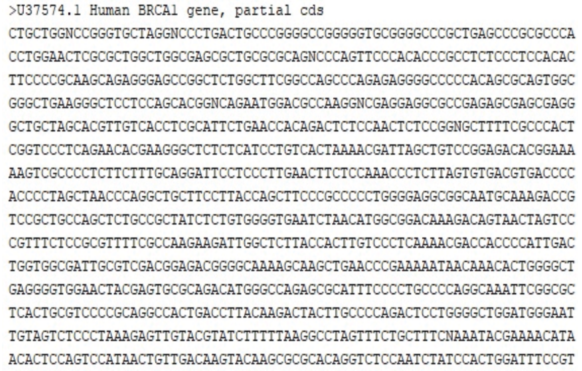
where n and m is the length of k-mer and cancer associated motif respectively, $n = 0,1,2, \dots k$ and $m = 0,1,2, \dots k.d[n,m]$ represents the distance value between k-mer and cancer associated motif. Where s_n is the n^{th} character of k-mer (source string), t_m is the m^{th} character of cancer associated motif (target string).

IV. Experimental Results

In this proposed system experiments for k-mer separation, feature extraction for SOM clustering and identification of cancer associated motifs has been done using Matlab 2013b tool. Orange 2.7 tool is used for clustering k-mers. FASTA format of cancer associated human DNA sequences are collected from NCBI. Details of dataset are shown in TABLE II. Experiments are done for the dataset mentioned in TABLE II, in which experimental results of Human BRCA1 gene (breast cancer) and stomach cancer therapeutic agent-target gene are discussed in this section. Due to the large size (number of characters) of data, portion of human DNA data in FASTA format is shown in Fig. 2.

TABLE II
CANCER ASSOCIATED HUMAN DNA MOTIFS

S.No	NCBI data accession number	Human DNA sequence length (Total number of characters in DNA data)	Name of the data in NCBI
1	U37574.1	3798	Human BRCA1 gene (breast cancer)
2	DQ115319.1	383	Homo sapiens breast and ovarian cancer BRCA2 gene
3	Y08757.1	1482	Homo sapiens breast and ovarian cancer BRCA1 gene
4	AF348515.1	2145	Homo sapiens breast cancer BRCA2 gene
5	BC047121.2	1314	Homo sapiens cDNA clone
6	LV501046.1	1983	Stomach cancer therapeutic agent -target gene
7	DI158975.1	1967	Gene therapeutic agent – large intestine cancer, bladder, lung cancer



```
>U37574.1 Human BRCA1 gene, partial cds
CTGCTGGNCCGGGTGCTAGGNCCCTGACTGCCCGGGGCCGGGGGTGCGGGGCCCGCTGAGCCCGCGCCCA
CCTGGAACTCGCGCTGGCTGGCGAGCGCTGCGCGCAGNCCCAGTTCCCACACCCGCCTCTCCCTCCACAC
TTCCCCGCAAGCAGAGGGAGCCGGCTCTGGCTTCGGCCAGCCCAGAGAGGGGCCCCCACAGCGCAGTGGC
GGGCTGAAGGGCTCCTCCAGCACGGNCAGAATGGACGCCAAGGNCGAGGAGGCGCCGAGAGCGAGCGAGG
GCTGCTAGCACGTTGTCACCTCGCATTCTGAACCACAGACTCTCCAACTCTCCGGNGCTTTTCGCCCACT
CGGTCCCTCAGAACACGAAGGGCTCTCTCATCCTGTCACTAAAACGATTAGCTGTCCGGAGACACGGAAA
AAGTCGCCCCTCTTCTTTGCAGGATTCCTCCCTTGAACTTCTCCAAACCCTCTTAGTGTGACGTGACCCC
ACCCCTAGCTAACCCAGGCTGCTTCCTTACCAGCTTCCCGCCCCCTGGGGAGGCGGCAATGCAAAGACCG
TCCGCTGCCAGCTCTGCCGCTATCTCTGTGGGGTGAATCTAACATGGCGGACAAAGACAGTAACTAGTCC
CGTTTCTCCGCGTTTTCGCCAAGAAGATTGGCTCTTACCACTTGTCCCTCAAAACGACCACCCCATTGAC
TGGTGGCGATTGCGTCGACGGAGACGGGGCAAAAGCAAGCTGAACCCGAAAAATAACAAACACTGGGGCT
GAGGGGTGGAACTACGAGTGCGCAGACATGGGCCAGAGCGCATTTCCCCTGCCCCAGGCAAATTCGGCGC
TCACTGCGTCCCCGCAGGCCACTGACCTTACAAGACTACTTGCCCCAGACTCCTGGGGCTGGATGGGAAT
TGTAGTCTCCCTAAAGAGTTGTACGTATCTTTTTAAGGCCTAGTTTCTGCTTTCNAAATACGAAAACATA
ACACTCCAGTCCATAACTGTTGACAAGTACAAGCGCGCACAGGTCTCCAATCTATCCACTGGATTTCCGT
```

Fig. 2 Human DNA data (FASTA format)

A. K-merSeparation

Human DNA sequences are separated as k-mers, according to (1). Based on the size of cancer associated motifs, k size is taken as 6. Repeated pattern of k-mers are considered one time for further process. Obtained total number of k-mers are 2267, 1256 for breast cancer data and stomach cancer data respectively. Table III shows some sample of separated k-mers for breast cancer data and stomach cancer data.

TABLE III
SAMPLE OF SEPARATED K-MERS (K=6) FOR BREAST CANCER DATA AND STOMACH CANCER DATA

S.No.	K-mer: breast cancer data	K-mer: stomach cancer data
1	AAAACA	ACCATC
2	ACAAAA	ATCCCG
3	CTCTCC	CCATCC
4	CCAGCT	CACCAT
5	GCTCTG	GCTGCG
6	GAGGCT	GGCTTC
7	TGAGAG	TGCACC
8	TACATA	TTCTGG
9	ACAGAT	ACATCG
10	CCGCAA	CATGAT

B. Clustering K-mers

Separated k-mers are clustered based on the features (mean and median) calculated from k-mers. ASCII values of characters are used for calculating mean and median. Table IV shows some samples mean and median of separated k-mers for breast cancer and stomach cancer data.

TABLE IV
SAMPLE OF SEPARATED K-MER FEATURES- MEAN AND MEDIAN

S.No.	Breast cancer data			Stomach cancer data		
	K-mer	Mean	Median	K-mer	Mean	Median
1	AAAACA	65.3333	65	ACCATC	69.1666	67
2	ACAAAA	65.3333	65	ATCCCG	70.1666	67
3	CTCTCC	72.6666	67	CCATCC	69.5	67
4	CCAGCT	70.1666	67	CACCAT	69.1666	67
5	GCTCTG	74	71	GCTGCG	71.8333	71
6	GAGGCT	71.5	71	GGCTTC	74	71
7	TGAGAG	71.1666	71	TGCACC	70.1666	67
8	TACATA	71.1666	66	TTCTGG	76.8333	77.5
9	ACAGAT	69.5	66	ACATCG	69.8333	67
10	CCGCAA	67	67	CATGAT	72.6666	69

SOM constructs a map with calculated features using 8*8 mapping topology. The map size is fixed based on the default size used in MATLAB 2013b tool. Therefore 64 nodes are created (i.e. 64 clusters).Those 64 cluster nodes are represented as (0,0), (0,1), (0,2), (0,3), ….. , (0,7), (1,0), (1,1), (1,2), …. , (1,7), (2,0), (2,1), (2,2), … ,(2,7), (3,0), (3,1), (3,2), …. ,(3,7), (4,0), (4,1), (4,2), …. , (4,7), (5,0), (5,1), (5,2), …. , (5,7), (6,0), (6,1), (6,2), …. , (6,7), (7,0), (7,1), (7,2), ….. , (7,7). Due to more number of clusters, here two k-mer cluster details of breast cancer data and stomach cancer data are shown in Tables V and VI respectively. In Table V, nodes (0,4) and (5,1) of breast cancer data and in Table VI, nodes (0,1) and (7,3) of stomach cancer data are discussed.

TABLE V
SAMPLE OF K-MER CLUSTER DETAILS OF BREAST CANCER DATA

Node (Cluster) position in SOM	Number of instances (k-mers) in cluster	k-mer	Mean	Median
(0,4)	27	GGAGGC	69.33334	71
		GGCAGG	69.33334	71
		GGGGCA	69.33334	71
		GAGGCG	69.33334	71
		GGCGGA	69.33334	71
		GGGCAG	69.33334	71
		ACGGGG	69.33334	71
		AGGGGC	69.33334	71
		CGAGGG	69.33334	71
		GACGGG	69.33334	71
		GAGGGC	69.33334	71
		AGGCGG	69.33334	71
		AGGGCG	69.33334	71
		GCGAGG	69.33334	71
		GGGGAC	69.33334	71
		CAGGGG	69.33334	71
		CGGAGG	69.33334	71
		CGGGAG	69.33334	71
		CGGGGA	69.33334	71
		GCAGGG	69.33334	71
		GCGGAG	69.33334	71
		GCGGGA	69.33334	71
		GGACGG	69.33334	71
		GGAGCG	69.33334	71
		GGCGAG	69.33334	71
		GGGAGC	69.33334	71
		GGGCGA	69.33334	71
		GGAGGC	69.33334	71
		GGCAGG	69.33334	71
(5,1)	5	GTGGGG	73.16666	71
		GGGGTG	73.16666	71
		GGGGGT	73.16666	71
		GGGTGG	73.16666	71
		TGGGGG	73.16666	71

C. Identification of Cancer Motifs

To find the presence of cancer associated motifs, Levenshtein distance is applied on all the obtained k-mer clusters and the reference pattern of cancer associated human DNA motifs as discussed in section III(*C*). TABLE VII shows the identified cancer motifs in breast cancer data and stomach cancer data. Six cancer associated human DNA motifs are matched with breast cancer data and two cancer associated motifs are matched with stomach cancer data.

D.Time Taken for Identifying Cancer Associated Motifs

Time taken for identifying cancer associated motifs is analyzed in two different ways such as (i) after clustering k-mers (for each k-mer clusters) and (ii) before clustering k-mers (for whole separated k-mers). These two approaches show very less time difference. Time taken before clustering k-mers shows less amount of time than the time utilization after clustering k-mers, for both breast and stomach cancer data. But locating the k-mer which is matched with cancer associated motif is easy in k-mer clusters. Because number of k-mers within one cluster is lesser than the number of k-mers for whole separated k-mers. This will reduce the search in k-mer cluster. Tables VIII and IX show the time taken for identifying cancer associated motifs in breast cancer data and stomach cancer data respectively.

V.CONCLUSION

This proposed work is developed for analyzing and identifying the cancer-causing DNA motif in a human DNA sequence. K-mer separation reduces the search space of the human DNA sequence. Because repetition of k-mer is considered one time for processing. K-mer clustering is done using SOM, it calculates two features namely mean and median. It is effective one, because huge number of separated k-mers are grouped into small number clusters and it gives the transparent grouping results. Dynamic programming is used for identifying the cancer associated motifs and Levenshtein distance is used for comparison. It is suitable for finding the distance (i.e. number of transformations required to change a k-mer string to cancer associated motifs) between k-mer string and reference pattern of cancer associated human DNA motifs.

For identification of presence of cancer associated motifs, processing all k-mers without forming cluster takes less amount of time. But k-mer cluster seems to be effective for locating the cancer associated motifs, because k-mers count within each cluster is lesser than the k-mers count for whole separated k-mers. Experimental results of this proposed work indicate that both breast and stomach cancer data contain the cancer associated motifs. Thus, the proposed work will be an essential aiding tool for finding the cancer disease causing DNA motif from human DNA sequence.

TABLE VI
SAMPLE OF K-MER CLUSTER DETAILS OF STOMACH CANCER DATA

Node (Cluster) position in SOM	Number of instances (k-mers) in cluster	k-mer	Mean	Median
(0,1)	31	CCATCC	69.5	67
		CATCCC	69.5	67
		CTCACC	69.5	67
		CATGAA	69.5	66
		ACAAGT	69.5	66
		CCACCT	69.5	67
		CCTCAC	69.5	67
		TCCACC	69.5	67
		AAAGCT	69.5	66
		AACGAT	69.5	66
		ACATGA	69.5	66
		ACCCCT	69.5	67
		AGAACT	69.5	66
		AGACAT	69.5	66
		ATCGAA	69.5	66
		ATGAAC	69.5	66
		CACTCC	69.5	67
		CCACTC	69.5	67
		CCCATC	69.5	67
		CCCCTA	69.5	67
		CCCTAC	69.5	67
		CCCTCA	69.5	67
		CTACCC	69.5	67
		GATCAA	69.5	66
		TACAAG	69.5	66
		TACAGA	69.5	66
		TAGACA	69.5	66
		TCAAGA	69.5	66
		TCGAAA	69.5	66
		TGAACA	69.5	66
		TGCAAA	69.5	66
(7,3)	5	GGGTGG	73.1666	71.0
		GGTGGG	73.1666	71.0
		GTGGGG	73.1666	71.0
		TGGGGG	73.1666	71.0
		GGGGTG	73.1666	71.0

TABLE VII
IDENTIFIED CANCER ASSOCIATED MOTIFS IN BREAST CANCER DATA AND STOMACH CANCER DATA

Identified cancer associated motifs in breast cancer data					
S. No.	Cancer motif no. as mentioned in Table I	Cancer motif	Node (cluster) position in SOM	K-mer location in corresponding node (cluster)	Matched k-mer
1	3	ACAAGA	(1,7)	32	ACAAGA
2	4	ACAAGT	(4,7)	53	ACAAGT
3	5	AGAACA	(1,7)	80	AGAACA
4	6	AGAACT	(4,7)	7	AGAACT
5	7	AGAAGA	(2,7)	15	AGAAGA
6	8	AGAAGT	(7,7)	40	AGAAGT
Identified cancer associated motifs in stomach cancer data					
1	4	ACAAGT	(0,1)	5	ACAAGT
2	6	AGAACT	(0,1)	13	AGAACT

TABLE VIII
TIME TAKEN FOR IDENTIFYING CANCER ASSOCIATED MOTIFS IN BREAST CANCER DATA: TIME UTILIZATION BEFORE CLUSTERING K-MERS (FOR WHOLE SEPARATED K-MERS) AND TIME UTILIZATION AFTER CLUSTERING K-MERS (FOR EACH K-MER CLUSTERS)

Total number of k-mers	Elapsed time (in seconds)
2267	5.216074

S.No.	Node (cluster) position in SOM	Number of instances (k-mers) in cluster	Elapsed time (in seconds)
1	(0,0)	65	0.286370
2	(0,1)	52	0.165626
3	(0,2)	36	0.109144
4	(0,3)	65	0.190538
5	(0,4)	27	0.084368
6	(0,5)	62	0.189373
7	(0,6)	47	0.144888
8	(0,7)	0	0
	Elapsed time for 0th layer		1.170307
9	(1,0)	14	0.042168
10	(1,1)	0	0
11	(1,2)	38	0.107069
12	(1,3)	0	0
13	(1,4)	0	0
14	(1,5)	76	0.218969
15	(1,6)	47	0.137617
16	(1,7)	113	0.327274
	Elapsed time for 1st layer		0.833097
17	(2,0)	0	0
18	(2,1)	0	0
19	(2,2)	33	0.093700
20	(2,3)	28	0.082005
21	(2,4)	77	0.210735
22	(2,5)	0	0
23	(2,6)	59	0.167684
24	(2,7)	48	0.157067
	Elapsed time for 2nd layer		0.711191
25	(3,0)	40	0.197100
26	(3,1)	66	0.225600
27	(3,2)	43	0.128946
28	(3,3)	0	0
29	(3,4)	1	0.005046
30	(3,5)	105	0.293890
31	(3,6)	78	0.219529
32	(3,7)	0	0
	Elapsed time for 3rd layer		1.070111
33	(4,0)	45	0.119311
34	(4,1)	45	0.125585
35	(4,2)	90	0.257835
36	(4,3)	0	0
37	(4,4)	64	0.164639
38	(4,5)	78	0.209304
39	(4,6)	80	0.222236
40	(4,7)	88	0.250105
	Elapsed time for 4th layer		1.349015
41	(5,0)	88	0.227403
42	(5,1)	5	0.013575
43	(5,2)	0	0
44	(5,3)	65	0.169711
45	(5,4)	11	0.031081
46	(5,5)	0	0
47	(5,6)	0	0
48	(5,7)	12	0.031677
	Elapsed time for 5th layer		0.473447
49	(6,0)	86	0.230100
50	(6,1)	0	0
51	(6,2)	22	0.069530
52	(6,3)	0	0
53	(6,4)	22	0.062232
54	(6,5)	2	0.006778
55	(6,6)	0	0
56	(6,7)	0	0
	Elapsed time for 6th layer		0.368640
57	(7,0)	0	0
58	(7,1)	83	0.210134
59	(7,2)	0	0
60	(7,3)	68	0.177690
61	(7,4)	0	0
62	(7,5)	0	0
63	(7,6)	7	0.020648
64	(7,7)	86	0.244197
	Elapsed time for 7th layer		0.652669
	Total elapsed time for clusters (from 0th layer to 7th layer)		6.628477

TABLE IX
TIME TAKEN FOR IDENTIFYING CANCER ASSOCIATED MOTIFS IN STOMACH CANCER DATA: TIME TAKEN BEFORE CLUSTERING K-MERS (FOR WHOLE SEPARATED K-MERS) AND TIME UTILIZATION AFTER CLUSTERING K-MERS (FOR EACH K-MER CLUSTERS)

Total number of k-mers	Elapsed time (in seconds)
1256	3.086345

S.No.	Node (cluster) position in SOM	Number of instances (k-mers) in cluster	Elapsed time (in seconds)
1	(0,0)	0	0
2	(0,1)	31	0.155338
3	(0,2)	27	0.101214
4	(0,3)	44	0.147568
5	(0,4)	41	0.128624
6	(0,5)	52	0.148558
7	(0,6)	34	0.098779
8	(0,7)	92	0.244968
	Elapsed time for 0th layer		1.025049
9	(1,0)	0	0
10	(1,1)	18	0.060271
11	(1,2)	7	0.019075
12	(1,3)	23	0.070542
13	(1,4)	0	0
14	(1,5)	0	0
15	(1,6)	0	0
16	(1,7)	0	0
	Elapsed time for 1st layer		0.149888
17	(2,0)	66	0.185137
18	(2,1)	3	0.009387
19	(2,2)	0	0
20	(2,3)	46	0.127351
21	(2,4)	15	0.039764
22	(2,5)	0	0
23	(2,6)	57	0.157449
24	(2,7)	0	0
	Elapsed time for 2nd layer		0.519088
25	(3,0)	6	0.018816
26	(3,1)	0	0
27	(3,2)	41	0.133838
28	(3,3)	1	0.004347
29	(3,4)	0	0
30	(3,5)	60	0.163426
31	(3,6)	2	0.008072
32	(3,7)	0	0
	Elapsed time for 3rd layer		0.328499
33	(4,0)	47	0.130447
34	(4,1)	52	0.153686
35	(4,2)	35	0.105823
36	(4,3)	9	0.027266
37	(4,4)	0	0
38	(4,5)	0	0
39	(4,6)	33	0.096343
40	(4,7)	23	0.066653
	Elapsed time for 4th layer		0.580218
41	(5,0)	74	0.200858
42	(5,1)	31	0.091123
43	(5,2)	0	0
44	(5,3)	0	0
45	(5,4)	27	0.106580
46	(5,5)	0	0
47	(5,6)	27	0.080713
48	(5,7)	0	0
	Elapsed time for 5th layer		0.479274
49	(6,0)	0	0
50	(6,1)	0	0
51	(6,2)	0	0
52	(6,3)	20	0.059014
53	(6,4)	39	0.113959
54	(6,5)	9	0.032070
55	(6,6)	42	0.116160
56	(6,7)	0	0
	Elapsed time for 6th layer		0.321203
57	(7,0)	0	0
58	(7,1)	14	0.045669
59	(7,2)	17	0.057360
60	(7,3)	5	0.013987
61	(7,4)	0	0
62	(7,5)	22	0.068615
63	(7,6)	29	0.084266
64	(7,7)	35	0.099474
	Elapsed time for 7th layer		0.369371
	Total elapsed time for clusters (from 0th layer to 7th layer)		4.267590

Future enhancement of this work can be done for considering the mutation in disease causing motifs. Different type of clustering and distance computing method can be used in future work. It can be extended for finding other diseases such as sickle cell anemia, Huntington's disease.

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 financial support - grant number EEQ/2016/000255.

REFERENCES

[1] Modan K Das, Ho-Kwok Dai, "A survey of DNA motif finding algorithms", BMC Bioinformatics, 8 (Suppl 7): S 21, 2007.

[2] Tong Ihn Lee, Richard G.Jenner, LauireA.Boyer, Matthew G.Guenther, Stuart S.Levine, RoshanM.Kumar, et al., "Control of Developmental Regulators by Polycomb inHuman Embryonic Stem Cells", Cell, vol. 125, no. 2, pp. 301– 313, 2006.

[3] Bling Ren, Francois Robert, John J.Wyrick, Oscar Aparicio, Ezra G.Jennings, Itamar Simon, et al., "Genome-wide Location and Function of DNA Binding Proteins", Science, vol. 290, no. 5500, pp. 2306–2309, 2000.

[4] FedricoZambelli, GrazianoPesole, GiulioPavesi, "Motif discovery and transcription factor binding sites before and after the next-generation sequencing era", Briefings Bioinformatics., vol.14, no. 2, pp. 225–237, 2013.

[5] NejatMahdieh, BaharehRabbani, "An Overview of Mutation Detection Methods in Genetic Disorders", Iran J Pediatr, vol 23. No.4, pp: 375-388, 2013.

[6] Jeremy Buhler, Martin Tompa, "Finding Motifs Using Random Projections", Journal of computational biology, vol. 9, no.2, pp. 225-242, 2002.

[7] ShripalVijayvargiya, Pratyoosh Shukla, "A Genetic Algorithm with Clustering for Finding Regulatory Motifs in DNA Sequences", IJCA Special Issue on "Artificial Intelligence Techniques- Novel approaches & Practical Applications, pp. 6-10, AIT 2011.

[8] Rui Chen, Yun Peng, Byron Choi, JilanliangXu, Haibo Hu, "A private DNA motif finding algorithm", Journal of Biomedical Informatics, vol. 50, pp. 122-132, 2014.

[9] Dianhui Wang, SarwarTapan, "Robust Elicitation Algorithm for Discovering DNA Motifs Using Fuzzy Self-Organizing Maps", IEEE Transactions on neural networks and learning systems, vol. 24, no. 10, pp.1677-1688, 2013.

[10] Yetian Fan, Wei Wu, Rongrong Liu, Wenyu Yang, "An iterative algorithm for motif discovery", 17th Asia Pacific Symposium on Intelligent and Evolutionary Systems, vol. 24, pp. 25-29, 2013.

[11] David L.Gonzalez-Alvarez, Miguel A. Veg-Rodriguez, Juan A. Gomez-Pulido, Juan M. Sanchez-Perez, "Comparing multiobjective swarm intelligence metaheuristics for DNA motif discovery", Engineering Applications of Artificial Intelligence, vol. 26, pp. 314-326, 2013.

[12] Robert J.Pantazes, Jack Reifert, Joel Bozekowski, Kelly N.Ibsen, Joseph A. Murry, Patrick S.Daugherty, "Identifaction of disease specific motifs in the antibody specificity repertoire via next-generation sequencing", Sci. Rep.6,pp. 1-11, 2016.

[13] Oleg V.Vishnevsky, AndreyV.Bocharnikov, Nikolay A. Kolchanov, "Argo_CUDA: Exhaustive GPU based approach for motif discovery in large DNA datasets", Journal of Bioinformatics and Computational Biology, vol.16, no.1, pp. 1740012: 1-23, 2017.

[14] NungKion Lee, Allen ChiengHoonChoong, "Filtering of background DNA sequences improves DNA motif prediction using clustering techniques", Procedia- Social and behavioural Sciences, vol.97, pp. 602-611, 2013.

[15] Jian-Jun SHU, "Identification of DNA Motif with Mutation", Procedia Computer Science, vol. 51, pp. 602-609, 2015.

[16] Shaun Mahony, Panayiotis V.Benos, Terry J. Smith, Aaron Golden, "Self-organizing neural networks to support the discovery of DNA-binding motifs", Neural networks, vol. 19, pp. 950-962, 2006.

[17] 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.

[18] TeuvoKohonen, PanuSomervuo,"Self-organizing maps of symbol strings", Elsevier, Neurocomputing, vol. 21, pp.19-30, 1998.

[19] Marghny Mohamed, AbeerA.Al-Mehdhar, Mohamed Bamatraf, Moheb R.Girgis,"Enhanced Self-Organizing Map Neural Network for DNA Sequence Classification", Intelligent Information Management, vol.5, pp.25-33, 2013.

[20] Igor Fischer, Andreas Zell, "String averages and self-organizing maps for strings", Proceedings of the ICSC Symposia on Neural Computing, pp. 208-215, 2000.

[21] Nassiri, Azadian, Nejad, "A Sequence Associated with Intrinsic Mutation Hot-Spots in Human DNA", Journal of Poteomics and Bioinformatics, vol. 6, 2013.

[22] PatrikDhaseleer, "What are DNA sequence motifs?", Nature Biotechnology, vol.24,no.4,pp.423-425,2006.