An Algorithm for Variable Length DNA Motif Discovery using *k*-means Clustering

Abdullah Al Yasin, S.M.Zobaed
Department of Computer Science and Engineering (CSE)
Islamic Univaersity of Technology (IUT)
BoardBazar, Gazipur-1704, Bangladesh
aay170@gmail.com
zobaed.skib@gmail.com

***Abstract*— Unravelling transcription factor binding sites, popularly known as motifs, is a prime concern in Bioinformatics. Several evolutionary algorithms like Linear PSO are used for motif discovery. However, Linear PSO searches linearly and appears slow. Here, we have used clustering algorithm based on *k* nearest neighbor for finding eligible sequence that can be a potential motif. The proposed algorithm needs no preprocessing like finding reference motif like Linear PSO. Hamming distance is taken as distance criteria for generate clusters that makes the algorithm faster. Primarily, the nearest neighbors gives potential motif containing clusters that reduces time to find motifs with higher validity and better efficiency.**

Keywords—clustering;k-nearestneighbour;motifs; evolutionary algorithms;

I. INTRODUCTION

Discovering the mechanisms that regulate gene expression is a major challenge in biology. A challenging task is to identify regulatory elements especially the binding sites in DNA sequence for transcription factors. DNA motif can be defined as 5 -20 base pair of nucleotides sequence. Motifs bear some biological significance such as being DNA binding sites for a regulatory protein. DNA motifs are often associated with structural motifs found in proteins. Motifs can be found on both strands of DNA. Motifs can bind protein directly on the double-stranded DNA. DNA Sequences have zero, one, or multiple copies of a motif.

The challenges present in motif identification include: 1. As the actual length of Motif sequence is not known. So, it's difficult to find out motif of unknown length.2. The motifs are never exactly the same as the actual conserved sequence. There is always a lot of sequence variability present with respect to a single motif. [1] 3.Motifs are very short signals as compared to the size of the DNA sequence under consideration. 4. The regulatory sequences containing the motifs may sometimes be located very far away from coding regions that they regulate. This makes it difficult to determine the portion of the DNA sequence that should be analyzed. 5. The regulatory sequences may, at times, be present on the opposite strand from the coding sequence they regulate.6. Mutation is also a serious challenge in motif finding.

As motif finding is a search problem so, particle swarm optimization and local search algorithm are used.

In this paper, we have proposed a clustering algorithm based on *k* -means clustering for improved search purpose. First, a set of cluster is found by applying the algorithm. Then, inter cluster analysis is performed. That's result in some potential motifs. In this way, no reference database is need not to be selected for finding reference motifs for finding target motifs. Algorithm can be applied on each individual database.

In this paper, Section 2 highlights our study on Linear-PSO using Index table approach. Section 3 briefly describes the proposed method for motif discovery. Section 4 describes the experimental results using Sus Scrofa DNA sequences and concludes the paper at section 5.

II. LINEAR PSO WITH INDEX TABLE APPROACH

Lailee et al [1] used a modified motif search algorithm based on a population based stochastic optimization technique called Linear Particle Swarm Optimization (PSO).

The PSO algorithms start by initializing a random population of individual particles in the search space. A fitness calculation is done for each particle. New position for each particle is calculated based on the current position and velocity values where velocity value for each particle uses random number generation. In motif discovery PSO was first introduced by B chang [4]. Later, the extended versions of the algorithm by integrating hybrid algorithm, de bruijn graph and stochastic local search concept has been used [5 - 8].

In our studied work on modified linear PSO, particle is selected from each iteration from the reference dataset. Selected particle length 6-20 bp is used as seed to search in other remaining datasets. The total number of target motifs can be calculated by

$$N = (L - ML) + 1 \qquad (1)$$

N is the total number of motifs, L is the length of the target set

Target Motif: **ATGATC**
CTT**A**TGTGGC**A**GT**A**G**A**GGCTG**AT**…..GT**A**TT**A**T
Target Motif: **ACTAGC**
CTTATGTGG**C**AGTAGAGG**C**TGAT…..GTATTAT

Creating Index Table:

Base	Position in 2nd sequence				
A	4	11	14	16	22
C	1	10	19	44	71

and ML is the length of the motif (6,7,8 is considered).In this approach, first an index table is created all the indices of the first base of motif are stores .those indices for motif Fig1:Linear PSO using Index table

sequence searching into other DNA sequences. The process of creating index table and motif sequence comparison with other DNA sequences are shown in fig1.

III. Proposed Method of Motif Discovery

Our proposed idea is to use clustering algorithm in motif finding and evaluate this approach, To do so,

At first, we select a dataset for applying the clustering algorithm. Here, we have used Sus scrofa dataset where we have 4 Fasta formatted file. We take each file and based on the motif length whether it is 6,7 or 8, *k* –mers are found where k is the length of the motif to be searched.

A. Reduction of k -mers

After getting *k* –mers from each file, we have to discard those *k* –mers that present very often in the dataset. For this purpose, a threshold is set to check whether any *k* –mers is present quite often or not. It is done because, DNA sequence are not enriched with motifs. Motifs occur in a DNA sequence in a small number.

Then, we get some *k* –mers that are eligible for the next step of the algorithm.

B. Applying KNN

K- nearest neighbor is a popular algorithm for clustering data. After running KNN on the dataset of *k* –mers, we found some clusters. Cluster point will be updated on each iteration of KNN. When, cluster point updating is done, then KNN terminates and it results in clusters of *k* -mers.

C. Distance criteria

For running KNN algorithm, different distance measures are used like Euclidean distance, Manhattan distance, Hamming distance etc. As a distance measure,we have used hamming distance .Hamming distance provides the number of mismatch. So when we get exact match between two *k* –mers then hamming distance is 0. With the increase of hamming distance, number of mismatches increases.
Hamming distance makes the search process faster as there is no complex equation. It just checks whether in certain position there is a match or not. So, it is faster than Euclidean and Manhattan distance.

D. Inter cluster Analysis

The last phase is Inter cluster analysis. It is a comparison process that done among the clusters. Here, clusters are compared with one another to find those cluster that have minimum hamming distance. After finding clusters of minimum hamming distance, the *k* –mers in the cluster are the potential motifs after doing some refinement (eq 2).

According to the abovementioned method, a flowchart can visualize how our proposed method works and it is shown in figure:2

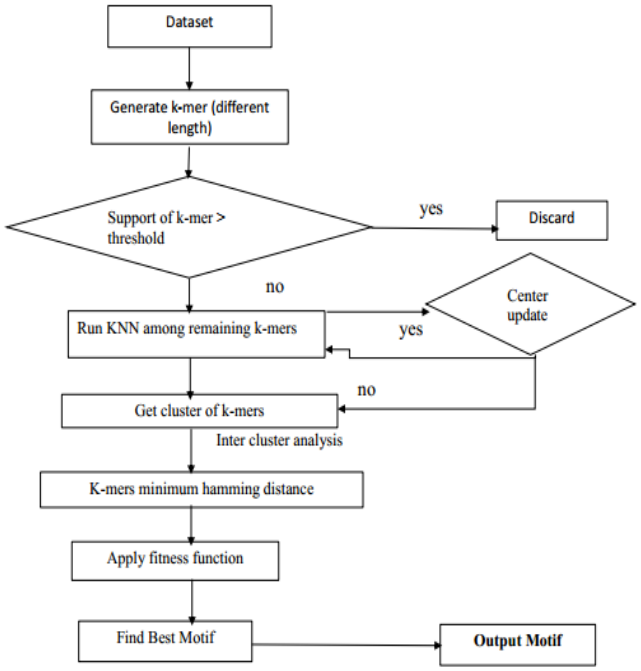

Figure 2: Flowchart of proposed algorithm

E. Design of Proposed Algorithm

Database is selected for motif discovery constraining that database files must contain DNA sequence of the same species.

Motif length is varied between 6 to 20 base pairs and *k* -mer length is same as the motif length. To discard redundant *k* -mers a threshold value is set for selecting some *k* -mers. A threshold is selected for discarding those *k* -mers that are present more than the threshold value. Because, motifs are not abundant in DNA sequence.

For motif discovery *k* -nearest neighbor algorithm is run among the remaining *k*-mers and Hamming distance is taken as

distance measure .Hamming distance is chosen for exact matching.

ALGORITHM: MOTIF TRACKER

INPUT: A set of DNA sequence

OUTPUT: possible set of motifs

1: Select the database for finding motifs, constraint is the database files must be DNA sequences of the same species.

2: Select randomly any one of the DNA sequences and find all *k*-mer where *k* =6 to 20.

3: Select a threshold, if any occurrence of *k* -mers in a DNA sequence > threshold discard those *k* -mers.

4: Take remaining *k* -mers and run KNN among them.

5: After running KNN some clusters are found of *k*- mers.6: Find the cluster containing those *k*-mers whose have less hamming distance among all the cluster.

7: Resultant motifs are refined using refined function.(eq. 2).

Here, it is seen that after running KNN for several times center point of cluster will no longer be updated. Inter cluster analysis is run among the clusters and clusters having less hamming distance are regarded as potential motifs.

F. Motif refinement

Here, in motif refinement some low complexity sequence like "TTTTTT" has lower complexity than "ATTTTA". For this purpose, one equation from [4] Dianhui Wang, Sarwar Tapan is followed.

$$\frac{4}{3}[1 - 1/k^2 \sum_{\forall bi \in x} \quad \sum_{i=1}^{k} k(bi,i))^2] \quad (2)$$

Here, k = motif length;

b = type of nucleotide e.g. 'A','G','C','T';

After applying this equation low complexity sequence are discarded from the obtained motifs.

IV. EXPERIMENTAL DATA AND RESULTS

In this work, we have tested Sus Scrofa , Mus musculus, Homo sapiens, Aedes aegypti , Felis catus dataset.but in this paper,we have shown the result of Sus scrofa dataset. The data was taken from GeneBank database.The lengths of selected sequences are from 829 based to 850 bases. Table shows the dataset we have used.

TABLE I. COLLECTED DNA SEQUENCES

Sus Scrofa	*Felis catus*	*Mus musculus*	*Homo sapiens*	*Aedes aegypti*
BW970508	KJ933256	JK707008.1	NM_001166002.2	KJ73686
BW971304	KJ933223	JK707007.1	NM_001166004.2	KJ73687
BW972295	KJ933191	JK707011.1	NM_001166003.2	KJ73685
BW973679	KJ933160	JK707010.1	NM_181773.4	KJ73684

Variable length motifs are discovered in these approach. *k* -mer size can be 6 to 20 bp. We have experimented 6,7,8 length *k* -mer and apply *k* -means clustering algorithm

Here, in table 3 we have considered the motif length as six without mutation (6, 0), and get some clusters under each cluster some motifs are found. Likewise 7 length (7,0) and 8 length (8,0) motifs are shown in the table

TABLE II. RESULTS FOR MOTIF DISCOVERY (WITHOUT MUTATION)

Motif length	center	Motifs	Hamming distance
6,0	CCGCTC	CCACTC	1
	CCTGCT	CCTGCT	0
	CTCTCG	CTCTCG	0
7,0	CCACTCG	CCACTCG	0
	CCTGCTG	CCTGCTG	0
8,0	AGCTCCGC	AGCTCCGC	0
	CACTCGCG	CACTCGCG	0
	CCCGCTCT	CGCGCTTT	2

We, have also considered mutations in DNA sequences. Here, we allowed one mutations i.e (6,1) and it is seen that number of motifs is increased significantly. It is also applied for length 7 (7,1) and length 8 (8,1) -mers allowing one mutation just changing one parameter in the code.

In Table III, the result for motifs allowing 1 mutation is shown. It is evident from the output of this table that with the increase in the number of mutation, the frequency of motifs existed in DNA sequence also increases

TABLE III. RESULTS FOR MOTIF DISCOVERY (WITH 1 MUTATION)

Motif length	center	Motifs	Hamming distance
6,1	CCGCTC	CCGCTC, CCACTC	0,1
	CCTGCT	CCTGCT, CCCGCT	0,1
	CTCTCG	CTCTCG,CACTCG	0,1
7,1	CTCCGCT	CTCCGCT,CTCCACT	0,1
	TCCTGCT	TCCTGCT,TCCCGCT	0,1
8,1	AGCTCCGC	AGCTCCGC	0
	CACTCGCG	CACTCGCG	0
	TAGCTCCG	TAGCTCCG	0

Here, it is experimented that the clustering algorithm runs faster than Linear PSO approach. As, no preliminary motif selection is needed by sliding window approach. So, it's take less processing time than linear PSO.

Hamming distance is a fast distance measure criteria in k-means clustering. It is computationally fast as no complex mathematical equation or operation e.g squaring is needed when Euclidean distance is used, But in case of measuring hamming distance it's like Boolean operation match or non match.

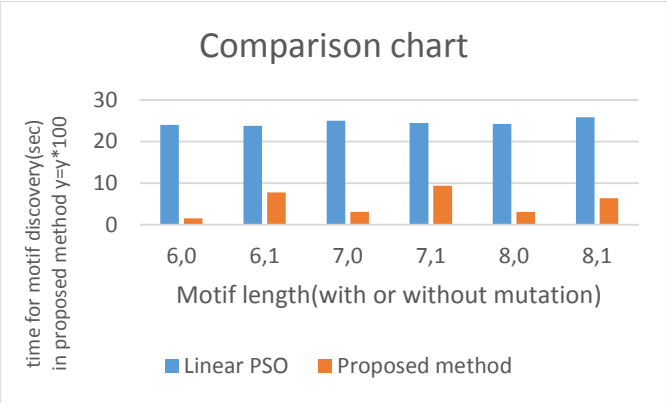

Fig 3. Time comparison (seconds) between Linear PSO and Proposed clustering method of motif discovery

Here, Fig 3 depicts the time comparison between the algorithm that used Linear-PSO [10] for motif discovery and the proposed method based on clustering for 6,7 and 8 length motif.

From the figure it is clearly understood the substantial improvement in processing time for DNA motif discovery for length 6, 7 and 8 in both considering mutation and without allowing mutation.

V. CONCLUSION

From experimental result, it is shown that effective progress in time and speed of processing DNA sequence for finding potential motifs is gained by proposed clustering method. This method not only alleviates the cumbersome process of reference motif selection and feed into other datasets, but also make the process faster and improvise the validity of the result.

REFERENCES

[1] Sharifa, L., S., A., Harun, H., and Taib, M., N.: A Modified Algorithm for Species Specific Motif Discovery. In International Conference on Science and Social Research (CSSR 2010), Kuala Lumpur, Malaysia, Dec 5-7, 2010.

[2] Chang, B., C., H., Ratnaweera, A., and Halagmuge, S., K.,: Particle Swarm Optimization for Protein Motif Discovery. In Genetic Programming and Evolvable Machine, vol. 5, pp. 203-214. (2004).

[3] Zhou, W., Zhu, H., Liu, G., Huang, Y., Wang, Y., Han, D., and Zhou C.: A Novel Computational Based Method for Discovery of Sequence Motifs from Coexpressed Genes. In International Journal of Information Technology, vol. 11 (2005).

[4] Islam, S.M.S.; Asger, M.R. ; Hasan, M.A. ; Mottalib M.A. : A modified algorithm for variable length DNA motif discovery, Smart Instrumentation, Measurement and Applications (ICSIMA), 2013 IEEE International Conference on 25-27 Nov. 2013, Pages 1-4.

[5] James M. Keller ; Michael R. Gray ; James A. Givens : A fuzzy K-nearest neighbor algorithm, in IEEE July,1985

[6] Dianhui Wang, Sarwar Tapan. : MISCORE: a new scoring function for characterizing DNA regulatory motifs in promoter sequences. 23rd International Conference on Genome Informatics (GIW 2012) Tainan, Taiwan. 12-14 December 2012.

[7] Miriam Manevitz , Moshe Samson :De-Novo Motif Finding using Genetic Algorithm . 2014 IEEE 28-th Convention of Electrical and Electronics Engineers in Israel.

[8] Modan K Das and Ho-Kwok Dai : A survey of DNA motif finding algorithms. Fourth Annual MCBIOS Conference. Computational Frontiers in Biomedicine .New Orleans, LA, USA. 1–3 February 2007

[9] I Zhou, W., Zhu, H., Liu, G., Huang, Y., Wang, Y., Han, D., and Zhou C.: A Novel Computational Based Method for Discovery of Sequence Motifs from Coexpressed Genes. In International Journal of Information Technology, vol. 11 (2005)

[10] Hardin, C., T., and Rouchka, E., C.: DNA Motif Detection Using Particle Swarm Optimization and Expectation-Maximization. In IEEE Symposium on Swarm Intelligence, 2005.