International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 01 | Jan-2018 www.irjet.net p-ISSN: 2395-0072

A Novel Approaches For Motif Discovery Using Data Mining Algorithm

M Mohamed Divan Masood[1], D Manjula[2]

[1] Research Scholar, Computer Science and Engineering, Anna University, Chennai.
[2]Head of the Department, Computer Science and Engineering, Anna University, Chennai.

Abstract - *Next generation sequences (NGS) analysis is a most widespread application in computation biology. NGS dataset contain massive amount of DNA sequences (ChIP-seq) so it will need efficient and intelligent method for analyzing the sequences. Motif discover one of the crucial task in sequences analysis. In our proposed method have been identified motif using MDWB (Motif Discover Word-based Algorithm). MapReduce function has been used in emerging mining step for handle the huge amount of data. In this motif discovery has been used to identify the transcription factor binding site (TFBS).*

Key Words: **ChIP-Seq, MapReduce, Word-Based Algorithm, Motif Discover, Transcription Factor Binding Site**

1. INTRODUCTION

Next generation sequences analysis is the one of the most important challenging task in computational biology. Diligent analytic methodology have been necessity for analyzing the DNA sequences, which contain features, function and structure [1]. Gene assembly established by adenine, guanine, cytosine and thymine in element of gene sequences. The DNA sequences methodology mainly used for drug discovery and medical research. Currently most research based on gene sequences alignment. There are lot of tools availability for gene alignment process. In computation biology consist of wide-ranging appropriate topics:

- Identifying the similarity between two kind of sequences (homologous)
- Analyzing the proper gene feature selection method based on data mining methodology.
- Identifying the sequence dissimilarity and modification such as mutation and particular nucleotide polymorphism (SNP) in sequencing market.
- Identification of molecular arrangement and assortment of gene activity from ChIP- Seq

In Molecular biology, Motif Discovery is most significant and challenging application. It has mainly used for locating TFBS in gene sequences. NP- complete has been proved for motif discovery [3]. There are three kind of most popular tools designed for identifying the motif: namely OOPS, ZOOP and TCM [3] with high accuracy. The motif discovery has used for identifying the TF binding site. TFBS means, where motif bind to gene sequences. Gene binding side are separate to other binding site instance of part of genome and bound by protein. It has frequently related using particular proteins known as transcription factors. The DNA binding sites are likewise includes other proteins, corresponding restriction enzymes, site-specific recombines and methyl transferees [4].

ChIP sequence is an approach used in analyses DNA- protein interaction. Chip stands for chromatin immune precipitation is a method used in extracting sequence from cell nucleus from any organism. ChIP sequence can be used in realize biological processes and disease state gene expression. Gene expression is regulate by DNA-protein interaction. It provides genome-wide maps of factor binding. Binding sites of DNA- protein are identified by ChIP sequence combine with massively parallel DNA sequencing. The application of ChIP-seq helps determining cancer progression as it reveal insight into gene regulation and also in biological research, drug development and molecular diagnostics. ChIP sequence comes under Next Generation Sequences (NGS). It has many advantages like it has complete genomes, the transcriptomes are whole and complete, from single organisms to complex pools like metagenomes. It doesn't contain sequence of sample organisms like other sequence do it has complete sequence of the original species [7]. The whole genome sequencing has sequencing divided according to with and without reference genome available. Motif are identified from the ChIP sequence. Motif means repetition of short sequence it can be one per sequence or zero per sequence. De novo sequencing had no reference genome sequencing and this sequencing is actually used in our case.

Next Generation Sequencing are also known as high-throughput sequencing there are some modern sequencing technologies which are Illumina sequencing, Roche 454 sequencing, Ion torrent: Proton/PGM sequencing, SOLiD sequencing, ChIP sequencing they sequencing are used in sequencing DNA and RNA more quicker and cheaper than the older sequencing. The researcher has been taken many transcription factor binding regions but it is very peak region. The motif discovery methodology has been taken accurate TF binding site [8]. The main application of ChIP-Seq dataset are spotless of the promoter region and also every sequences have high a great resolution. It appears easier for motif discovery approaches to gain a great identification result in ChIP-seq with good accuracy. Appropriately, exactly all systems intended for motifs

discover in promoter DNA gene sequences, whichever the pattern-driven method or statistical method grow into short time consuming sequence alignment. Main objective of my proposed system has identified the motif pattern using data mining algorithm with grate classification accuracy. MDP algorithm has based on optimization method [5, 6].

Our proposed methodology of this paper is as follows. Section 2 describes literature survey. Section 3 describes system architecture diagram of my proposed methodology. Section 4 describes Experimental Evaluation and finally section 5 describe conclusion.

2. Literature Survey

The Hirschberg [5] develop a new method for solves the frequency associated to data mining requests of strings in ideal period. In this structure has been used frequent string and emerging string. The result of this process represented to pattern matching on DNA sequences. The main approaches are based on suffix and lcp arrays. The application of this research are array based data assemblies and also great locality conduct and extensibility of memory. MapReduce by Jeffrey et al [7] developed a new methodology for analyzing the DNA sequence in optimal time period. The map function processes of key/value brace to make a set of intermediary key and merge function connected with intermediate key values. The MapReduce function mainly focuses on reduce the computational time period.

A MapReduce process is collected of map() technique that achieves cleaning and arrangement of the DNA sequences. A Reduce () technique that achieves a precipitate. Daniel et al. Developed [17] a. EXTREME tools for identifying the motif pattern. It has been related to expectation-maximization processes to identifying the motifs. Several SW/NW problems only score. By instance trackback process could not performed at the same time space difficulty is linear. DREME by Bailey [11] is considerably quicker than many ordinarily utilized algorithms for it can dissect extensive dataset. They can just discover short theme. DREME is substantially quicker than many usually utilized algorithms, scales straightly in dataset estimate, finds different, non-excess themes and reports a solid measure of factual centrality for every theme found. DREME is accessible as a major aspect of the MEME Suite of theme based succession tools used in sequence analysis.

An effective multicore algorithm, PMS6MC developed by Tompa [10], for the motif(l,d)exploration to discover all strings of length l that is present in each string of a given arrangement of strings with at most d dissimilarity. The speedup, in respect to PMS6, achieved by our multicore algorithm varies from a high of 6.62 for the (17, 6) testing occasions to a low of 2.75 for the (13, 4) testing examples on an Intel 6-center framework. Evaluation of PMS6MC for larger instances of motif is 2 to 4 times quicker than other parallel calculations used.

In the trending years, genetic algorithms are implied in various association rule mining techniques [15]. Utilizes weighted things to characterize novel pairs. Estimation of various principles is calculated using the fitness function present in the weighted things. This algorithm can discover satisfactory limits for association rule mining. Eberhart [16] proposed a unique technique for enhancing separated principles, utilizing genetics algorithm. The significance of this work is to figure out the negative values used in the process.

Hu et al. [13] introduced a PSO based strategy for self-detection of value falling below the threshold value. Their work demonstrates that fundamental PSO can discover values rapidly and superior to the genetic algorithms used. The additionally offered a strategy at for self-detection of values falling below the particular range using weighted PSO. His outcomes indicates high effectiveness of PSO for associative rule mining. This approach likewise can increase better estimation of limits with respect to the previous results.

3. System Architecture

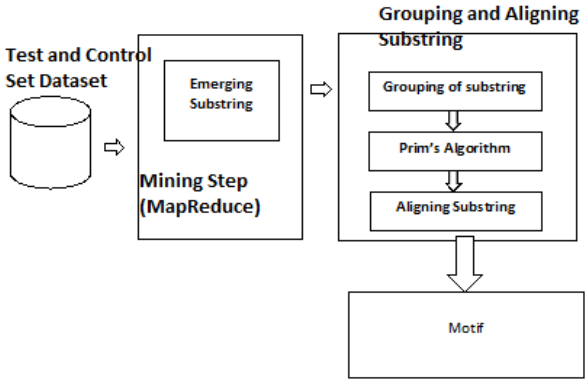

Fig 1 System Architecture of MDWD

In Fig 1 Preprocessing step combine the suffix array (SA) and the longest common prefix array (LCP). Mainly two kind of dataset have been used such as test set and control set. Fischer et al [6] developed a new method for identifying motif, which is based on SA and LCP.SA. LCP- array C'as calculated by labeling of control and test datasets. It decide threshold values as same time least evaluation rate. Every substring have occurs frequently in emerging substring stage. All the substring extracted by using of SA and LCP. 'L' for a Substring traversal and bottom up substring traversal to each string for all branches substring.

Subsequently each substring sequences call by suffix tree traversal using SA and LCP-array. Fast calculation process has been done in labeling phase. As a final point combine step reoccurrence the all strings transformed for frequency based condition. In MapReduce process occurred in mining stage. The MapReduce process done to reduce memory and

time difficulties. Reduce process undertaking to merge the all control and test emerging substring datasets. Frequently the map function combine to all substring based on reduce function. The map and reduce function algorithm given as,

Algorithm 1 MapReduce Function

Input: test (Dt) and control (Dc) set dataset.

Output: Emerging Substring EMr

EMr← Ω

for i←Dt& Dc

merge i and EMr

freq (Ω, Dt) ← α/| Dt|

freq (Ω, Dc) ← β/| Dc|

growth (Ω, Dc, Dt) ← α/| Dt| / β/| Dc|

if Freq (Ω, Dt)>pf and growth (Ω, Dc, Dt)>pg then

add Ω to EMr

return EMr

In our proposed methodology has been used Word based algorithm. It designed for identification of motif pattern from DNA sequences datasets. Most of data mining algorithm has been used for motif selection such machine learning algorithm, deep learning, genetic algorithm, word based algorithm[7] [8] [9].van Helden et al. [8] implemented a word-based approaches for the motif discover methodology. While theoretically easy methodology for understanding the string, in this method has showed effectively identifying the motifs from sequences dataset. These result has been found before lab test analysis. Also alleged novel binding sites were identified related concurrent gene. Word based algorithm given as,

Algorithm 2: Word-Based Algorithm

Input: Emerging Substring of DNA sequences

Output: the set motif M

Each string set parameter

if ||Dt|| + ||Dc||<= z' then

mining step combine into each string

else

if M = Ω then

pf ←pf/2

return M

4. EXPERIMENTAL EVALUVATION

The experimental analysis have been completed on Fujitsu system with Intel core i5, 2.33 GHz processor with 16GB RAM. Our proposed method has been implemented on R – language version 3.4.3. The motif result has been stored on image format. The dataset has been taken from National Center for Biotechnology information (NCBI). The ChIP-Seq dataset have huge volume of datasets. Motif discover methodology is a very huge challenging process in computational biology. The result has been much better then previse methodology. Our proposed system have implemented a complete set of performance action in motif level and analytically calculated single motif algorithm in sequences datasets. The performance of our algorithm, which only used ChIP-Seq dataset for identification of motif. The motif level accuracy found great level compare to previse methodology.

Sequence and motif level accuracy

To evaluate the performance to identified any one common motif pattern from sequence dataset and explain the correct motif discover pattern 'M'. The total number of sequences 'N' of string, as set as follows:

$$mSr= Ns/N$$

M means motif pattern, S means string, r means Accuracy rate, N means Total number of sequences.

The overall accuracy rate of in our algorithm average is mSr from input sequence.

Hence, the proposed method is called MDWB is compared with the MCES algorithm for time complexity measure. In which the set of all DNA sequences are tested on both the algorithm for finding the presences of motif in DNA sequences. The uses of finding motif in DNA sequences are helped to find the transcription factor binding site with gene mutation feature in faster manner to know the states of the diseases presents in the DNA sequences. Even, when the number of sequence are increased the time taken to find the motif in the DNA are less in the proposed method. The graphical representation of this result are shown fig 2.

Number of Sequences	Algorithms	
	MCES	Proposed MDWB
10000	18	16
20000	35	28
30000	44	35
40000	57	48
50000	85	60
60000	128	80

Table 1 Time taken for Motif Discover in MCES vs MDWB

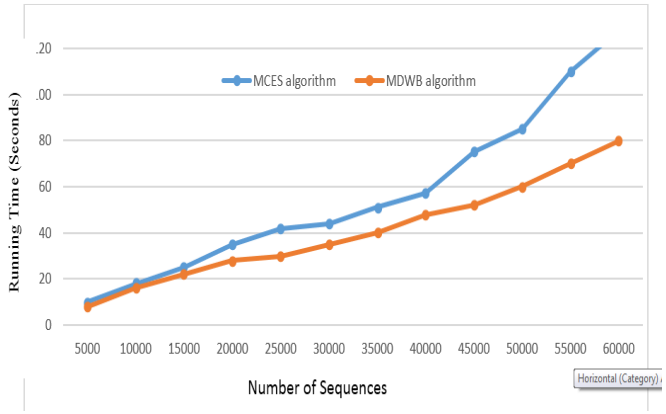

Fig 2. Difference between number of sequence vs Running time (Seconds)

In Table 2, describe the accuracy measure between the proposed MDWB and MCES method. The prominent value of proposed value of proposed method has higher accuracy for identification motif in the DNA sequences for finding diseases presence in the dataset. The graphical representation of this result are shown fig 2

Number of Sequences	Algorithms (Accuracy)	
	MCES	Proposed MDWB
10000	0.89	0.92
20000	0.91	0.93
30000	0.94	0.96
40000	0.88	0.92
50000	0.85	0.9
60000	0.95	0.97

Table 2 Accuracy between MCES vs MDWB

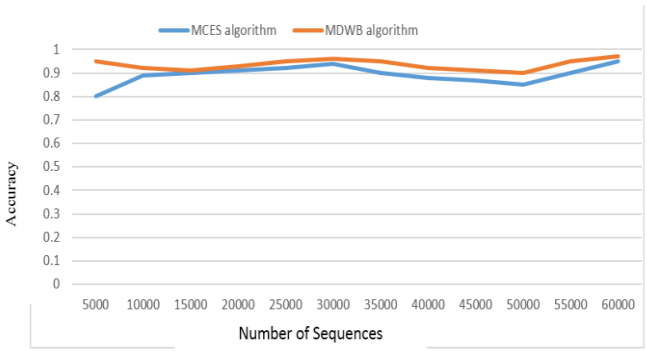

Fig 3 classification accuracy between MCES vs MDWB

5. CONCLUSION

The proposed MDWB algorithm is worked based on the word count for identification the presence of motif in DNA Sequences dataset. In this algorithm every sub string in the DNA sequence with the variation in length are mined and the information are extracted. With the effect the proposed method is capable to discover the motif sequences from ChIP sequences. Our method proposed higher classification accuracy with minimum time taken to discover the presence of motif in ChIP sequences.

REFERENCES

[1] Durbin, M. Richard, Eddy, R. Sean, Krogh, Anders, Mitchison and Graeme, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, doi: 10.2277/0521629713, ISBN 0-521-62971-3, OCLC 593254083, 1998.

[2] P. A. Evans, A. D. Smith, and H. T. Wareham, "On the complexity of finding common approximate substrings," Theoretical Computer. Sci., vol. 306, pp. 407–430, 2003.

[3] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," in Proc. ISMB, pp. 28–36, 1994.

[4] E. S. Halford,J. F. Marko. "How do site-specific DNA-binding proteins find their targets"? Nucleic Acids Research. 32 (10): 3040–3052. doi:10.1093/nar/gkh624. PMC 434431? Freely accessible. PMID 15178741, 2004

[5] D. S. Hirschberg. "A linear space algorithm for computing maximal common subsequences". Communications of the ACM 18, 6, 341–343. DOI:http://dx.doi.org/10.1145/360825.360861, 1975.

[6] J. Fischer, V. Heun, and S. Kramer, "Optimal string mining under frequency constraints". In Proc. PKDD, pp. 139–150, 2006.

[7] Jeffrey, Huo, Qiang, Xiaoyang, Haitao and Huan. "An efficient algorithm for discovering motifs in large DNA data sets." IEEE transactions on nanobioscience 14.5, 535-544, 2015.

[8] J. V. Helden, B. Andrea and C. J. Vides. "Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies". J Mol Biol. 281:827–842, 1998.

[9] J. V. Helden, A. F. Rios and J. V. Collado. "Discovering regulatory elements in non-coding sequences by analysis of spaced dyads". Nucleic Acids Res., 28:1808–1818, 2000.

[10] Tompa. M. "An exact method for finding short motifs in sequences with application to the ribosome binding site problem". Proceedings of the Seventh International Conference on Intelligent Systems on Molecular Biology, pp. 262–271, 1999.

[11] T. L. Bailey, P. Krajewski, I. Ladunga, C. Lefebvre, Q. Li, T. Liu, P. Madrigal, C. Taslim, and J. Zhang, "Practical guidelines for the comprehensive analysis of ChIP-seq data," PLoS Comput. Biol., vol. 9, no. 11, p. E1003326, 2013.

[12] P. Machanick and T. L. Bailey, "MEME-ChIP: motif analysis of large DNA datasets," Bioinformatics, vol. 27, no. 12, pp. 1696–1697, 2011.

[13] M. Hu, J. Yu, J. M. Taylor, A. M. Chinnaiyan, and Z. Qin, "On the detection and refinement of transcription factor binding sites using ChIP-seq data," Nucleic Acids Res., vol. 38, no. 7, pp. 2154–2167, 2010.

[14] D. Quang and X. Xie, "EXTREME: an online EM algorithm for motif discovery doi:10.1093/bioinformatics/btu093," Bioinformatics, 2014. P. Machanick and T. L. Bailey, "MEME-ChIP: motif analysis of large DNA datasets," Bioinformatics, vol. 27, no. 12, pp. 1696–1697, 2011.

[15] Jahanian. K. Using sequential pattern mining in protein sequences discovery with gap. Aust. J. Basic and App. Sci. 5(12). 1476-1480, 2000.

[16] R. C. Eberhart and J. Kennedy. A new optimizer using particle swarm theory. 6th Int. Symp. Micro machine Human Sci., Nagoya, Japan, 1995, pp. 39–43.

[17] D. Quang and X. Xie, "EXTREME: an online EM algorithm for motif discovery doi:10.1093/bioinformatics/btu093 ," Bioinformatics, 2014.