International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

An Advanced Algorithm for Finding Tandem Repeats in DNA Sequencing based on Text Mining

Anjali Saini, Poonam

Abstract*: *In this paper a new searching technique based on pattern matching is proposed. In bioinformatics, finding Tandem Repeats (TR) in DNA sequences is an critical issue. There exist many pattern matching algorithms and KMP (Knuth Morris Pratt) is one of the pattern matching algorithm that undergo deficiencies of runtime complexity and cost when size of the data set increases. The main aim of the paper is to generate an effective algorithm for detecting and identifying Tandem Repeats over a DNA sequence more efficiently. By introducing the concept of 2Dimensional matrix to minimize the purview scope and optimizing the problem, Tandem Repeat finding algorithm makes the detecting or identifying process more efficient and effective that improves the quality of results. The theoretical analysis and experimental results concludes that tandem repeat finding algorithm get equivalent results in less runtime. This algorithm is better to KMP for determining results, and it also reduces or weaken the runtime cost which is beneficial when DNA data becomes greater.

Keywords*: *Tandem Repeat, DNA sequence, Pattern matching, KMP.

I. INTRODUCTION

This paper concentrates on locating sequential patterns that occur in DNA known as tandem repeat. Tandem repeat appear in a string when a substring is repeated/redone for two or more times and where each repetition is directly adjoining to each other. For instance, a substring GGA occurs in the string Y=CGGAGGAGGAT for three times, and each occurrence of GGA is continuous , one after other. Then GGA is a tandem repeat of length 3 of Y. Tandem repeats are not extensible information ,but of either practical or progressive significance[1][2]. For instance, tandem repeats repeatedly occur within or in proximity of genes. Latest affirmation supports that tandem repeats(TR) in these regions play a very important role in managing gene expression and adjusting gene's function. This paper emphasis on finding tandem repeats in a given DNA of all length. This paper is structured in the following way: section II presents Tandem Repeats and its application. Next section III presents background, then section IV presents proposed work and experimental results and analysis are presented in section V. Section VI then presents our conclusion, followed by future prospects given in section VII.

Revised Manuscript Received on August 30, 2019.

* Correspondence Author

Anjali Saini*, Ph.D. in computer science &Engg., M.Tech., Assistant Professor of Computer Science, IP College for women, jhajjar, Haryana

Poonam, M.tech. ,Department of Computer Science and Engg.,NCU Gurugram

A) Sequence Mining and Text Mining

Text mining is all about obtaining patterns and associations from large text databases that are previously unknown. The sequence mining technique is used to find a set of items across time or position in a given database. A sequence $\alpha=\{i_1,i_2,i_3,\ldots,i_n\}$ is an ordered set of events. An event is an unordered set of items. For example in a web sequence the database record is the browsing activity of the web users in which a collection of data viewed by a web users on a single click represent the sequence transaction and index position, home page, contact information represent the sequence of events. The originated patterns are the sequenes of most frequently accessed pages of a particular web site.This type of information is used to again systematize the web site based on user accessed patterns.The work of finding all frequent sequences in large databases is challenging as the search space is exceedingly large. To find the sequential pattern in a database string matching algorithms are required.

B) DNA

DNA sequences consists of 4 characters: A, T, C, and G .It is assumed that all the secrets of our lives lies in these DNA sequences .Now how to fetch interesting patterns from those sequences and how it can help us in the discovery of remedies and therapies is extremely interesting task. The conspicuous characteristic of DNA sequence is the limit to which the subsequences of the nucleotides repeats in any genome[4]. Tandem repeats having strong repetitive nature are related to genetic disease, while structures having weaker repetitive nature are used for representing historical events related to sequential pattern repetition. Thus, it is essential to design responsive algorithm for detection & identification of repeated sequence. A new algorithm for determining Tandem repeats in DNA sequence is designed. The algorithm uses the inverted list transformation for computing the closest repeated sequence of fixed length at each position in a given DNA sequence to be evaluated.

II. TANDEM REPEAT

Regions in DNA sequences in which short sequences of DNA (nucleotides: adenine - A , thiamine – T , guanine – G , cytosine - C) which are repeated between tandem arrays. Within the genomic DNA which are at the same location, the number of times the sequence is repeated frequently changes among individual, among the population it basically occurs in DNA sequence when a pattern of two or more nucleotides is repeated and that repetition is directly adjoining with each other [6].

Most of the researchers used the basic string matching or pattern matching algorithm. Tandem Repeat finder algorithm needs two things Ist the pattern generation or pattern classification & other is pattern matching. Example: G-C-C--T-A-G-C-C-T-A-G-C-C-T-A. In this DNA sequence the pattern G-C-C-T-A is repeated three times known as tandem repeat [10]. Tandem repeat is challenging problem because of efficiency .Generally the DNA sequence are of thousands of symbols & the purpose of this algorithm is to identify the pattern of maximum length.

A) Applications of Tandem Repeats:

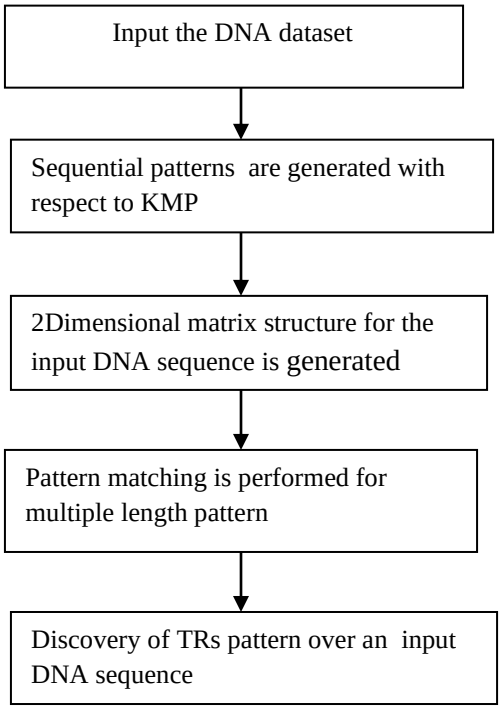
i. Tandem Repeats are used for human identification, it is very important to have DNA markers that display the highest possible variation in order to discriminate between different samples.
ii. Tandem Repeats specifies the pattern which are useful in determining individual inherited traits as every individual carry one copy of tandem repeats from every parent, which might or might not carry similar repeat sizes. The number of repeated patterns in tandem repeats markers can be highly inconsistent among different individuals which makes these tandem repeat effective for the determination of inherited traits.
iii. Tandem Repeats are useful in determining parentage.
iv. Tandem Repeats are used in forensics, matching finger prints & genetic biology.

III. BACKGROUND

Boyer- Moore algorithm (BM) was developed by R.S.Boyer and J.C.Moore in 1977.The BM algorithm[3] scaned the characters of the pattern form right to left beginning from the rightmost one and performs the comparisons from right to left. In case of a mismatch it uses two pre-computed functions in order to move the window to right hand side. Time complexity is O(m +n). 1974 the Knuth-Morris-Pratt Algorithm (KMP) was developed by D.Knuth, J.Morris and V.Pratt [5] . It compares the string pattern with the given text from left hand side to the right hand side . In case of occurrence of a mismatch or complete match it uses the notion border of the string. It reduces the time of searching a pattern compared to the Brute Force algorithm. KMP algorithm uses automata to find all the occurrences of a pattern in a text. Time Complexity given by KMP is O(m+ n). G. Lakshmi Priya (2011) has presented a comparative study on different approaches to perform the pattern prediction over the DNA sequence. Author defined a process to perform the extraction of pattern over biological dataset. Author defined the association analysis based on gene expression. Author performed the frequent pattern analysis and prediction so that the crucial problem was identified. Author defined multifactor analysis based algorithm to perform the prediction of pattern over the biological sequence[8]. Shuang Bai (2009) has presented Joint Maximum Pattern identification approach over the DNA sequence. Author defined work on basic DNA sequence dataset. Author performed the analysis over the sequence pattern and performed the redundant pattern analysis over the DNA sequence. Author defined the functional and structural analysis over the sequence so that characteristic analysis over the pattern will be performed [11]. Quanwei Zhang (2008) has defined Genetic based K mode approach for DNA sequence and feature analysis. Author defined an adjacent sequence analysis approach using conservation feature and sequence analysis. Author defined a biological knowledge based estimation over DNA sequence. Author also used Hidden Markov Model approach to improve the discovery algorithm[12].

IV. PROPOSED WORK

A) Research methodology



B) Two Dimensional Matrix Algorithm:

Matrix based work is the one which is used for pattern matching in order to find tandem repeat over a DNA sequence. The main aim is to search the tandem repeat patterns/sequences over the DNA sequence so that we can search the DNA pattern. In this work a two dimensional model or the structure is been represented in which the input DNA sequence will be filled in such a way so that the search of any DNA pattern can be done accurately and efficiently. The presented work is defined as the algorithm that can be implemented and integrated with any text based application. Where in matrix algorithm it will define a frequency matrix so that the individual symbol and patterns occurrences will be identified. It gives all possible tandem repeats i.e patterns which are directly adjacent to each other in a DNA dataset along with their count and positions.

2. Proposed Algorithm

```
set MaxCount=0
for i=1 to length(patternDNA)
set Count=0
for j=1 to length(DNASequence)
if(DNASequence(j)=patternDNA(i))
count=count+1
else if (Count>MaxCount)
set MaxCount=Count
end
[Identify the Number of Columns in
DNASequenceMatrix]
generate DNASeqMatrix(length
(patternDNA),MaxCount*2]
assign Index to DNAPattern
alphabets suchas A=>1, G=>2 ..
to represent the DNA pattern rows
for i=1 to length(DNASequence)
index=IndexofAlphabet
(DNASequence(i))
nextAlpha=DNA Sequence(i+1)
count=frequency(DNA
Sequence(i))
count1=frequency(DNA
Sequence(i+1))
DNASequenceMatrix(Index,Count*2)
=nextAlpha
DNASequenceMatrix(Index,
Count*2+1)= Count1
return DNA sequence matrix
```

```
L=length(DNAPattern)
    if (L>2)
  index=DNAPattern(1)
  index1=2;
for j=1 to Length(DNASeqMatrix
   (Index,:)
while (Index1<=L)
  if (DNASeqMatrix(index,j)=
DNA
     Pattern(index1)
index= index1
index1=index1+1
     else
     break
if(index1=L)
    count1=count1+1
    return count1
```

IV. MODELING RESULTS AND ANALYSIS

Here in this section we study the performance/work of Tandem Repeat(TR) finding algorithm by differing different length of tandem repeats and by comparing its pattern matching technique with KMP algorithm used for matching the patterns . KMP was implemented as described in Knuth Morris Pratt [5] . Tandem repeat finding algorithm gives all possible tandem repeats(TR) of each and every length in a DNA sequence that are given. Bar graphs given below shows the graphical comparsion between KMP and Proposed approach. The analysis is performed on DNA sequence of 19461 length characters. The results are driven in milliseconds.

) Tandem Repeat finding Algorithm

1. Preprocessing Phase:

The first step of the proposed algorithm is to generate the inverted list table for almost all characters in the dataset .The next step is to reads each and every character from the given pattern and then updates the inverted list. The searching phase utilises the navigator variable m as current comparison position; SHIFT as the shift window; pos as the required position for current matching; “life” as the control loop

```
Inverted-List Table (p=c1,c2,c3,...cm)
    Create table for all characters of the dataset
                    j=1
            while (j<=m)
    Create inverted list and add to table
        at alphabet char (Cj)
```

variable used in each of search window; and SETA, SETB, and SETG as the temporary variables used in the matching. The first character of each and every search window is compared with last character in the text which are then followed by communicating the inverted list to SETG for all the reference. If SETG is not empty and will matches with the end character, we have to scan that for comparing the text from the first to the end character, or if SETG does not contain the end character, we study the farthest character matching the SETG and scan for matching again. Each comparison takes the inverted list to the temporary variable SETA or SETB, among holding the inverted list to these variables. SETA and SETB must also be operate. The focus of the operation is to find the sequential pattern and check the matching.

A) Results

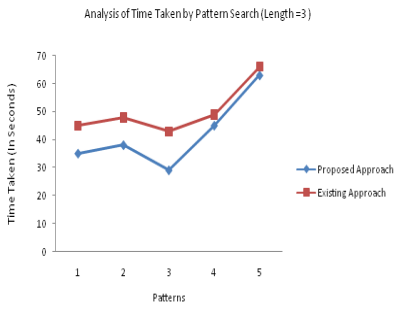


Fig5.1. Time analysis for pattern length=3

Here the comparison of time efficiency of presented and existing approach is shown. The results are driven in milliseconds

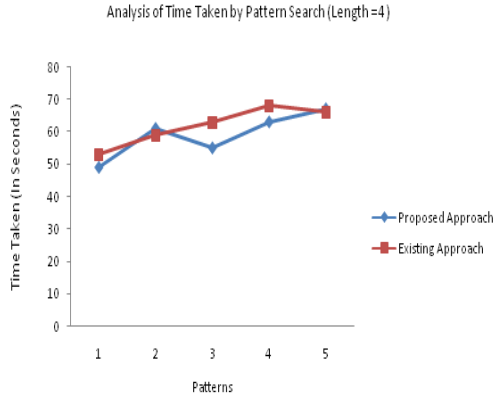


Fig5.2.Time analysis for pattern of length=4

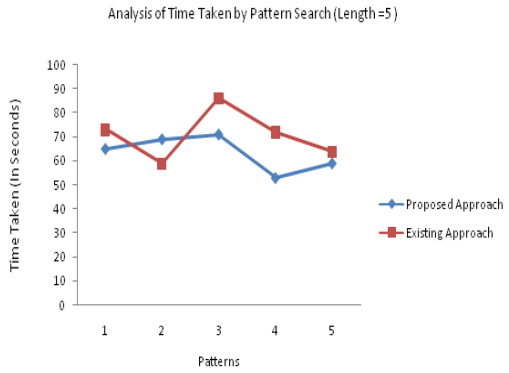


Fig5.3. Time analysis for pattern of length=5

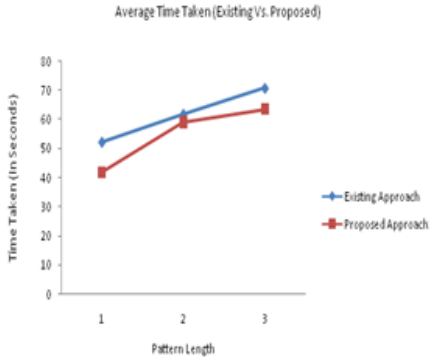


Fig5.4. Average time taken (Existing vs Proposed)

Here figure 5.4 is showing the results of input pattern of length 3,4 and 5 over the DNA sequence where, the time taken by existing approach more than proposed approach so that the work provided the effective results and hence reducing the runtime cost.

B) Complexity Analysis

Runtime Analysis of KMP

Computation of the prefix function values is O(m) . O(n) is to compare the pattern to the text.

Total function values = O(n + m) run time

Runtime Analysis of Proposed Algorithm

A DNA sequence is taken of N length as an input. Here, a search of pattern of length M is search over the sequence. The complexity of the working pattern search algorithm is *O(N/4+M).*

V. CONCLUSION

DNA sequence mining is one of the most increasing bioinformatics applications, that can be applied to perform the knowledge discovery over the DNA sequence. These patterns represent the characterstics of a living thing so that the characterstic discovery and matching are the common operations of DNA sequence mining. The focus of the work was to perform the pattern match over the sequence. This stage was decomposed into two parts. In first part, a 2D pattern was generated based on the DNA sequence.The representation of this structure is done in the form of an algorithm. The given algorithm gives the effective complexity i.e. O(N/4+M).The obtained results highlighted that the proposed work has provided the results in effective time.

VI. FUTURE PROSPECTS

The presented work is about to perform the DNA pattern search over the DNA sequence. This presented search mechanism can be improved in future with different aspects:

- In this work, the DNA sequence mining is been presented. In future, same algorithms can be implemented in some other application areas such as web content mining or window based text mining etc.
- In this work, repeated sub pattern based multiple pattern generation is defined. In future the robustness of system can be improved by generating some other patterns.
- In this work, the sequential search is performed. In future, the parallel pipelined search can be defined to improve the search efficiency.

REFERENCES

1. Patrick C. H. Ma," An Iterative Data Mining Approach for Mining Overlapping Coexpression Patterns in Noisy Gene Expression Data", IEEE Transactions On Nanobioscience 1536-1241 , 2009.
2. Jing Hu," Mining sequence features for DNA-binding site prediction", 978-1-4244-1779-7/08 , 2008 IEEE
3. Boyer R. S., and J. S. Moore, ‘‘A fast string searching algorithm‘Communications of the ACM 20, 762- 772 , 1977.
4. Bolin Ding," Efficient Mining of Closed Repetitive Gapped Subsequences from a Sequence Database", IEEE International Conference on Data Engineering 1084-4627/09 , 2009 IEEE
5. Knuth D.,Morris.J Pratt, V Fast pattern matching in strings, SIAM Journal on computing, vol 6(1),323-350, 2005.
6. Avrilia Floratou“Efficient and Accurate Discovery of Patterns in Sequence Datasets”,ICDE conference 2010 978-1-4244-5446-4/10 , 2010 IEEE
7. Xinyan Zha and Sartaj Sahni “Multipattern String Matching On A GPU”,IEEE,2011,pp. 277-282.
8. G. Lakshmi Priya," A Comparative study on existing methodologies to Predict Dominating Patterns amongst Biological Sequences", IEEE-ICoAC 2011 978-1-4673-0671-3/11©2011 IEEE
9. Po-Yuen Wong,"Predicting Approximate Protein-DNA Binding Cores Using Association Rule Mining",2012 IEEE 28th International Conference on Data Engineering 1084-4627/12 , 2012 IEEE

10. Sheng Li," An Optimized Algorithm for Finding Approximate Tandem Repeats in DNA Sequences", 2010 Second International Workshop on Education Technology and Computer Science 978-0-7695-3987-4/10 , 2010 IEEE
11. Shuang Bai," The Maximal Frequent Pattern Mining of DNA Sequence".
12. Quanwei Zhang," Genetic K-modes based DNA Splice Site Adjacent sequence_ Feature Analysis", Proceedings of the 7th World Congress on Intelligent Control and Automation 978-1-4244-2114-5/08, 2008 IEEE

AUTHORS PROFILE

Anjali saini , Ph.D. in computer science &Engg., M.Tech., Assistant Professor of Computer Science, IP College for women, jhajjar, Haryana

Poonam, M.tech. ,Department of Computer Science and Engg.,NCU Gurugram