ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Detection of Contiguous Pattern with Pattern Shift String Matcher.

L Mary Gladence*.

School of Computing Sathyabama University Chennai, Tamil Nadu, India.

ABSTRACT

Patterns appear repeatedly either inside a same string or over a set of strings. These repeated patterns are called motifs and their identification is called motif inference or motif extraction. This is an important problem in Computational biology. This problem involves in finding short patterns of interest from voluminous data. Three variation of this motif search problem have been identified in the literature. Mining structured motifs, wiil allow variable length gaps between simple motif components. we presented a new model called (L,M,S,K) based on Flexible and Accurate Motif Detector (FLAME). FLAME is a flexible suffix-tree-based algorithm which can be used to find frequent patterns with a variety of definitions. It is accurate, because it always finds the accurate pattern . Using sample DNA data set demonstrated the (L,M,S,K) Model and found how much percentage matched. In addition, based on (L,M,S,K) model ,addressed a pattern shift string matcher problem which is used to find out the exact match.

Data mining, Sequential pattern mining, Pattern Recognition.

Keywords: Sequential pattern mining, motif (Patterns), suffix tree, flexible and accurate motif detector,(L,M,S,K)model ,Pattern Shift string Matcher.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Discovering sequential patterns from a large database is a important problem in the field of knowledge discovery and data mining[1]to[17]. Aim of sequential pattern mining(SPAM) is to find complete set of frequent sequential pattern with the minimum support in the sequence database. SPAM attempts to find intersession patterns such as the presence of set of items followed by another item in a time ordered set of sessions. Mining frequent sequential pattern has many application such as market-basket analysis, telecommunication, web application, DNA analysis, stock prediction etc. SPAM algorithm mines the sequence database especially frequent sequences that can be used by end users or management to find associations between different items or events in their data and it is used for marketing campaigns, business reorganization, prediction and planning. Frequent pattern mining is the one of the main concept in sequential pattern mining which is used to find the frequent pattern presented in the complete set of sequence. The main task in bioinformatics is analyzing and interpreting the sequence data. One of the serious features of interpretation is to find the important patterns from the sequence datasets. There are two challenges occurs while extracting the pattern, they are

1. Extracting the frequent pattern which is used to design a flexible algorithm
2. Statistically legalize the pattern that are extracted and report the important pattern.

Motifs(Pattern) are basically classified in to two categories. They are simple motifs and structured motifs. If there is no variable gaps are allowed in the pattern, then it is referred as a single motif and whereas if any variable gaps are allowed then it is referred as the structured motifs. Planned set of simple motifs with gap limitation among each pair of adjacent simple motifs is also called as structured motif.

In most application of the sequential data mining, the purpose is to detect the continuously occurring patterns. For detecting such type of process initially a set of noise patterns are allowed. It may vary from one application to the other. In computational biology, sub-sequence mining issue is to detect the short sequence pattern of length between 6 – 15 which occur regularly in a given set of protein sequence or DNA.. We cannot assure that short sequence dataset will always be identical and a few of them differs from other. A complex similarity metrics should be used to find the distances.

From the above discussion we found that the problem of pattern mining is related to the problem of frequent sub sequences and the frequent item sets. Let us assume Q is a sub sequence of P, if Q can be build by using few of the elements from sequence P. Elements of sequence P is “a, b, a, c, b, a, c” and its sub sequence is constructed by choosing the selective elements from the sequence P and the sub sequence Q is formed as “a, b, b, c”. Here, only the continuous sub sequences mining issues are highlighted. It is motivated by the issue of detecting the frequent motifs in DNA sequences which has philosophical significance in the computational biology community and life sciences. Based on these many algorithms are created.They are MITRA[3], YMF[1], Random projections[4] and Weeder[2].

Best one discussed from here is the detection of association rules in the sequence data which is used to discover the best seeds for clustering the sequence data sets.From patients records of medical signals like respiratory data or ECG are mined to detect the signals which are used to find the possible dangerous conditions.The important part of gene rule is arbitrate through exact proteins i.e transcription factors which is used to manipulate the transcription of a specific gene by DNA sequences which are the transcription factor binding sites. Comparing with all algorithm FLAME is more powerful and flexible one. Based on FLAME we used(L,M,S,K)model to find the match using Suffixtree calculation. In addition to it Pattern Shift String Matcher is used to find the exact match with valid shifts. Here we can see Pattern Shift String Matcher is highly efficient.

RELATED WORK

There is a vast amount of literature on mining databases for frequent patterns [5], [6], [7]. Early work focused on mining association rules [8]. The problem of mining for subsequences was introduced in [4]. Subsequence mining has several applications, and many algorithms like SPADE , BIDE , and CloSpan and several others have been proposed as improvements over [8]. In the beginning all the algothms focussed on subsequence mining, while we focus on contiguous patterns. Some subsequence mining algorithms allow

ISSN: 0975-8585

certain constraints. Constraints which limit the maximum gap between two items in the subsequence make it possible to use these algorithms to mine for contiguous patterns. Algorithms such as cSPADE [9], CloSpan [10], and Pei et al. [11], [12] can be adapted to mine for exact contiguous motifs. An obvious reason why these are unsuitable for approximate frequent pattern mining is that these algorithms do not include a notion of noise or an approximate match. Furthermore, they tend to be inefficient even when used for exact substring mining. FLAME, on the other hand is extremely efficient even for approximate substrings. The vast body of work in bioinformatics for finding patterns in long noisy DNA sequences can be divided into two classes—pattern-based and statistical. The pattern based algorithms typically search through the space of potential patterns and find a motif that satisfies the minimum support. Marsan and Sagot [13] proposed a suffix-tree-based algorithm to find structured motifs tolerating a few mismatches as noise. This method is primarily focused at finding pairs (or sets) of motifs that co-occur in the data set within a short distance of each other. This method only considers a simple mismatch-based definition of noise, and does not consider other more complex motif models such as a substitution matrix or a compatibility matrix. Similarly, Rajasekaran et al. [14] propose an algorithm for solving an instance of the motif mining problem where wildcard characters are allowed but it also uses the Hamming distance model. Several other algorithms such as the Yeast Motif Finder [9] (YMF), Weeder [10], and MITRA [11] have been used for finding motifs. YMF is a simple algorithm that computes the statistical significance of each motif. YMF scales very poorly with increasing complexity of motifs, and thus cannot be easily adapted to other applications. Weeder is a suffix-tree-based algorithm that makes certain assumptions about the way the mismatches in an instance of the motif are distributed. This makes Weeder extremely fast, but it is not guaranteed to always find the motif. Weeder too, cannot be adapted for other motif models. MITRA is a mismatchtree- based algorithm which uses clever heuristics to prune the large space of possible motifs. MITRA is very resourceintensive and requires large amounts of memory.

Statistical approaches use techniques like Expectation Maximization , Sampling , Random Projections etc., to search frequent patterns in the data. All of these heuristic approaches run the risk of finishing at a local optimum, and may not be able to find the right motif. All these methods require a training set of known motifs as an input. For these reason we need to spend much time and get the required output,eventhough we are not sure about result.FLAME [15]doesn't require any prior knowledge about the motifs that may appear in the data set.FLAME used as the backbone for my work since knowledge has been retrieved from refering that paper.Suffix tree calculation always look for the forward string.First we need to find the distinct character for the given string,then we need to find the subsequences.

PROPOSED WORK

System Architecture

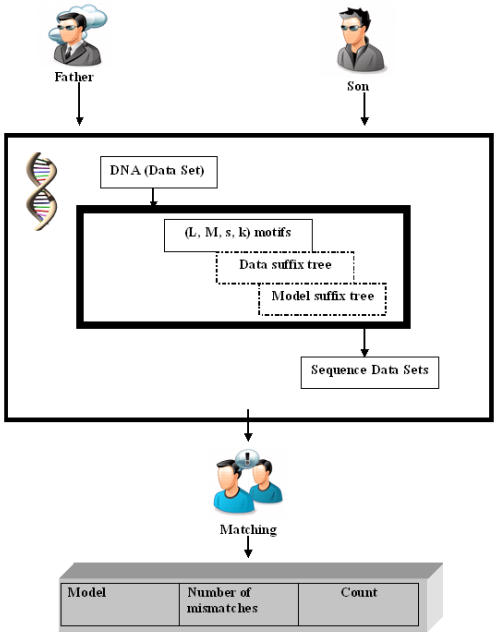

Fig 1: System Architecture

"Figure 1"shows the entire concept of the proposed work.Father and Mothers DNA data set has been taken to to find out the match using (L,M,S,K) model.Steps are given below.

Data Representation model

Input as a DNA Data Sets. We find the (L, M, s, k) motifs for the input data sets .

L → the length of the motif,
M → distance matrix which is used to compute the similarity between two strings
S → the maximum distance threshold within which two strings are considered similar
K→ the minimum support required for a pattern to qualify as a motif

Suffix Tree Calculation Model

After data representation from the paricular motif, we derive two suffix trees

1.Model suffix tree
2.Data suffix tree

Model suffix tree first perform Pruning. Then set data on the set of all possible model strings. Data suffix tree set the data on the actual data set, which contains counts in each node.

Motif Extraction And Matching Model

From the suffix tree, we perform extended structured motif extraction its P-structured occurrence. We call the resulting array F-Existential array. Then FLAME returns a set of motifs that match the given model. It produces the Results. The result contains

- Model
- Number of mismatches
- Count

Pattern Shift String Matcher Model

From the suffix tree, we perform extended structured motif extraction its P-structured occurrence. We call the resulting array F-Existential array. Then apply Pattern shift String Matcher. Its string-matching problem is the problem of finding all valid shifts with which a given pattern P occurs in a given text T. It produces the Results. The result contains

- Model
- Number of mismatches
- Count

Comparison Module

Using (L,M,S,K) Model we will come to know how much percentage father and mother are related based on subsequences used in the model. Once found the percentage using (L,M,S,K) model compare it with the Pattern Shift String Matcher. We can clearly see the difference by checking the percentage(raise in percentage match)because all the valid shifts are used in Pattern Shift String Matcher. In Results and Discussion section it is clearly displayed using graph.

ISSN: 0975-8585

RESULTS AND DISCUSSION

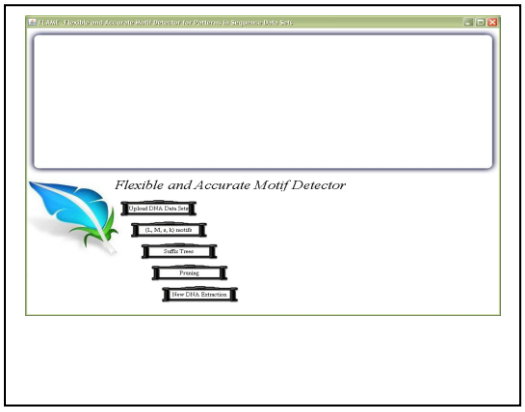

Fig 2:Flexible and Accurate Motif Detector Representation

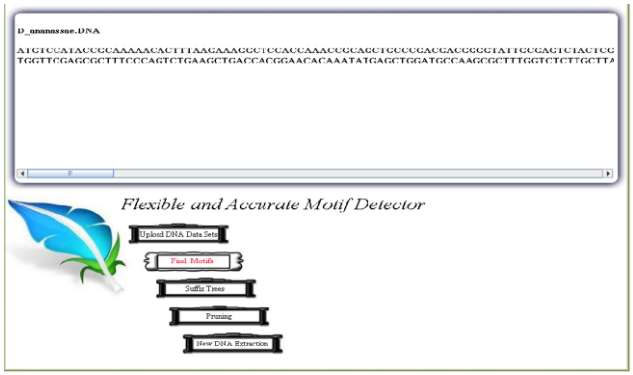

Fig 3: Detection of Motif for the given input

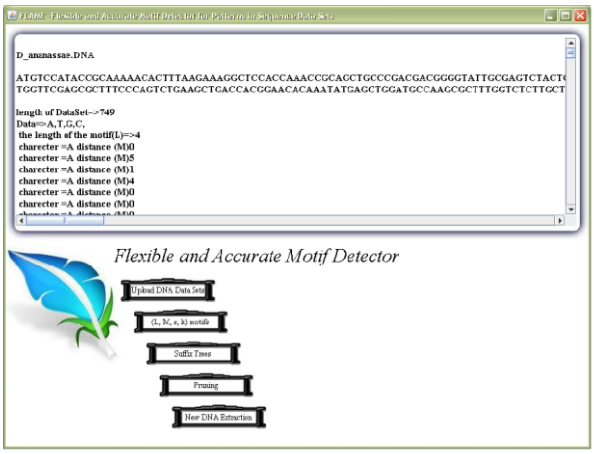

Fig3:Detection of (L,M,s,k) for the given input

ISSN: 0975-8585

Fig4:Results of (L,M,s,k) Model

This paragraph is a repeat of 3.1

Fig 5: Model Suffix Tree

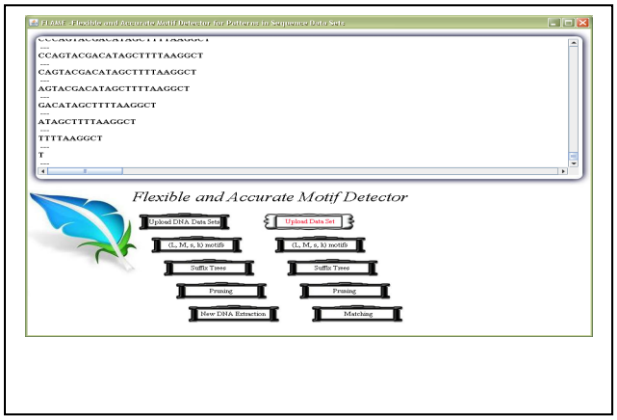

Fig 6: New Data Extraction

Above displayed images shows how Pattern Matching is done using (L,M,s,k) Model and these are compared with the Proposed Pattern Shift String Matcher.

ISSN: 0975-8585

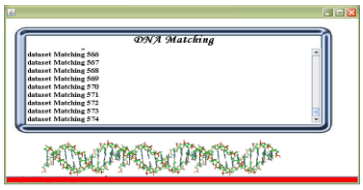

Fig 7:Results of DNA Matching

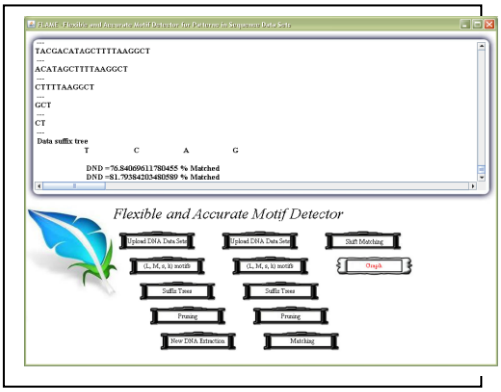

Fig 8: Final Output of Flame

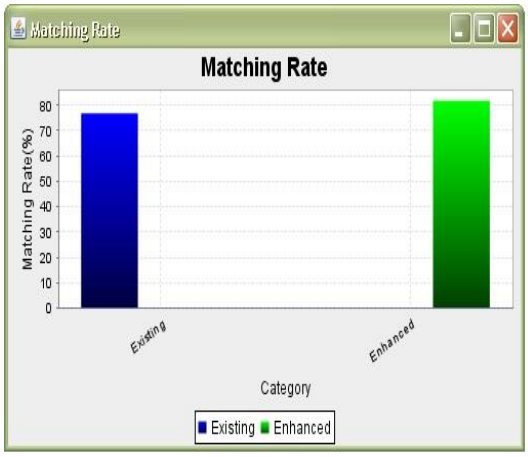

Fig 9: Comparison between Existing and Proposed work

From the "Fig.7","Fig.8","Fig.9" displayed images we can clearly undestand how the matching process is done using both (L,M,S,K) and Pattern Shift String Matcher. From this we can clearly say Pattern Shift String Matcher gives the optimal results with all valid shifts.

CONCLUSION

Here, a significant new model (L, M, S, K) for motif mining in sequence database is presented. This model consider various existing models and present extra flexibility which makes the model good in a large diversity of data mining applications. In addition Flexible and Accurate Motif Detector is used to obtain accuracy and flexibility for detecting the (L, M, S, K) motifs. By using sample DNA data set (L,M,S,K) model has been verified and found how much percentage both the samples are matched. In addition to this Pattern Shift String Matcher algorithma is used with all valid shifts.Finally we compares both model and found Pattern Shift

String Matcher has high efficiency compared with (L,M,S,K) model. It is also proved that Pattern Shift String Matcher can hold larger dataset compared with other algorithms.

REFERENCES

[1] S. Sinha and M. Tompa, "YMF: A Program for Discovery of NovelTranscription Factor Binding Sites by Statistical Overrepresentation,".

[2] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole, "Weeder Web:Discovery of Transcription Factor Binding Sites in a Set of Sequences From Co-Regulated Genes," Nucleic Acids Research,vol. 32, pp. W199-W203, 2004

[3] L.MaryGladence,M.Karthi,T.Ravi,A Novel Technique for Multi-Class Ordinal Regression-APDC"Indian Journal of Science&Technology,Vol.9,Issue10,DOI: 10.17485/ijst/2016/v9i10/88890 indexed in Scopus.

[4] E. Eskin and P.A. Pevzner, "Finding Composite Regulatory Patterns in DNA Sequences," Proc. 10th Int'l Conf. Intelligent Systems for Molecular Biology (ISMB), pp. S354-S363, 2002

[5] J. Buhler and M. Tompa, "Finding Motifs Using Random Projections," J. Computational Biology, vol. 9, no. 2, pp. 225-242,2002

[6] W. Wang and J. Yang, Mining Sequential Patterns from Large Data Sets, vol. 28, Springer-Verlag, 2005.

[7] M. Das and H.K. Dai, "A Survey of DNA Motif Finding Algorithms," BMC Bioinformatics, vol. 8, p. S21-S33, 2007.

[8] G.K. Sandve and F. Drabløs, "A Survey of Motif Discovery Methods in an Integrated Framework," Biology Direct, vol. 1,pp. 11-26, 2006.

[9] R. Agrawal and R. Srikant, "Fast Algorithms for

[10] Mining Association Rules," Proc. Int'l Conf. Very

[11] Large Data Bases (VLDB),pp. 487-499, 1994.

[12] [10] L.MaryGladence,M.Karthi,V.MariaAnu" A Statistical Comparison of Logistic Regression and different Bayes Classification Methods for Machine Learning "ARPN Journal of Engineering and Applied SciencesISSN 1819-6608 in Volume 10,Number 14 August2015P.No 5947-5953.

[13] M.J. Zaki, "Sequence Mining in Categorical Domains: Incorporating Constrains," Proc. Ninth Int'l Conf. Information and Knowledge Management (CIKM), pp. 442-429, 2000.

[14] X. Yan, J. Han, and R. Afshar, "CloSpan: Mining ClosedSequential Patterns in Large Datasets," Proc. SIAM Int'l Conf.Data Mining (SDM), 2003.

[15] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "PrefixSpan: Mining Sequential Patterns by Prefix-Projected Growth," Proc. IEEE Int'l Conf. Data Eng. (ICDE),pp. 215-224, 2001.

[16] J. Pei, J. Han, and W. Wang, "Mining Sequential Patterns with Constraints in Large Databases," Proc. 11th Int'l Conf. Information and Knowledge Management (CIKM), pp. 18-25, 2002.

[17] V.Maria Anu "Preserving Privacy in Distributed Medical-Healthcare Systems", Research Journal of Pharmaceutical, Biological and Chemical Sciences ISSN:0975-8585 in Vol 7 issue 2 March 2016, Pg:1291-1296 .

[18] L. Marsan and M.-F. Sagot, "Algorithms for Extracting Structured Motifs Using a Suffix Tree with Application

[19] to Promoter and Regulatory Site Consensus Identification," J. Computational Biology,vol. 7, nos. 3/4, pp. 345-360, 2000.

[20] S. Rajasekaran, S. Balla, C.-H. Huang, V. Thapar, M.R. Gryk, M.W.Maciejewski, and M.R. Schiller, "Exact Algorithms for Motif Search," Proc. Asia-Pacific Bioinformatics Conf. (APBC), pp. 239-248,2005

[21] Avrilia Floratou, Sandeep Tata, and Jignesh M. Patel, Member, IEEE,"Efficient and Accurate Discovery of Patterns in Sequence Dataset" IEEE Transactions On Knowledge And Data Engineering, Vol. 23, No. 8, August 2011