Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 11, November 2015, pg.272 – 276

SURVEY ARTICLE

SURVEY ON SEQUENCE DISCOVERY USING DNA SEQUENCE MINING DATA

S.Kalai Selvi, M.Sc.,M.Phil.,B.Ed.,MBA.,[1], **A.Meena**, M.Sc(CS) [2]

[1]Assistant Professor in B.Com(Computer Application), Dr.SNS.Rajalakshmi College of Arts and Science
Coimbatore, Tamil Nadu, India

[2]M.Sc.,M.Phil.,(Computer Science), Dr.SNS.Rajalakshmi College of Arts and Science,
Coimbatore, Tamil Nadu, India

[1] kulandaikalai@gmail.com; [2] mithra1710@gmail.com

Abstract— Sequence Mining is one of the most commonly used technique in data mining. Sequence mining is the process of mining frequent patterns from a large datasets. The exiting algorithms have some limitations in predicting frequent patterns, in terms of time, space complexity and accuracy. To overcome these drawbacks, in this paper made a study on existing sequence mining algorithms and generate a new algorithm for generating frequent patterns from the biological sequences (DNA). This paper attempt to locate all the tandem repeats in a DNA sequence. The future scope of this paper is not only predicting the frequent patterns; but will also satisfy some factors such as: space complexity, time and predict accurate solution to the required problem using some algorithms. With the help of these three things into consideration an effective algorithm can be used for predicting the tandem repeat in a given DNA sequence.

Keywords— "biological sequences, DNA - Deoxyribonucleic acid, frequent patterns, DNA Sequence, sequence mining algorithms"

I. INTRODUCTION

Data mining means "mining" knowledge from large data set. It is defined as "the process of discovering meaningful new interrelationship, patterns and mode by finding into large amounts of data stored in a data set". Data Mining is said to be KDD in Databases as data sets have grown in size and complexity, the new technologies of networks, computer and sensors have made data collection and its organization much simple and easy to handle. However, the data which has been stored needs to be converted into information and knowledge to make it more useful. Data Mining is the entire process of applying computer based techinques, including new methods for knowledge based discovery from data. Data Mining approaches seem ideally suited for Biological Data Mining, since it is data-rich, but lacks at the molecular level in finding comprehensive theory of life's organization. The comprehensive databases of biological information create both challenges and opportunities for development of novel Knowledge based discovery in databases methods. Mining biological data helps to extract useful knowledge from massive datasets stored in biology, and in other relative life science areas such as medicine and neuroscience. [1]The paper focuses on finding tandem repeats of all length in a given DNA sequence.

Sequences are an important kind of data which occur frequently in many fields such as medical, business, financial, customer behavior, educations, security, and other applications. In these applications, the analysis of the data needs to be carried out in different ways to satisfy different application requirements, and it needs to be carried out in an efficient manner [2].

SEQUENTIAL PATTERN MINING ALGORITHMS

[3]Sequential pattern mining is a very important mining technique with broad applications. It found very useful in various domain like natural disaster, sales record analysis, marketing strategy, shopping sequences, medical treatment and DNA sequences etc. It discovers the subsequence's and frequent relevant pattern from the given sequences. That we have provided the sequence database having sequences, in which each sequence is a list of the transactions ordered by the transaction time. Each transaction consists of the number of the items. The problem is to discover the all sequential pattern who satisfy the user specified constraint, from the given sequence database. There are various sequential pattern mining algorithms proposed earlier, some of them are GSP, SPADE, SPAM and PREFIX-SPAN. They are proposed to find the relevant frequent pattern from the sequences. In above algorithms the old dataset is deleted while some other dataset are updated. In these algorithm the timestamp is an important attribute of each dataset, and also it is important in the process of data mining for giving the more accurate and useful information.

A.GSP

The GSP algorithm described by Agrawal and Shrikant [1] makes multiple passes over the data. This algorithm is not a main memory algorithm. If the candidates do not fit in memory, the algorithm generates only as many candidates as will fit in memory and the data is scanned to count the support of these candidates. Frequent sequence produced from these candidates are captured, while those candidates without minimum support are removed. This procedure is repeated until all the candidates have been counted. In the first step GSP algorithm [2] finds all the length-1 candidates (using one database scan) and orders them with respect to their support the minimum support threshold, its entire super sequences will never pass the test.

B. SPADE AND SPAME

SPAM - SPAM combines the ideas of GSP, SPADE, and Free-Span. This algorithm uses the vertical bitmap data structure representation of database which is similar to the given id-list of SPADE. The whole algorithm with its data structure fits in the main memory. For the performance increase the SPAM use the depth-first traversal fashion. SPAM is similar to SPADE, but it uses the bitwise operations instead of the regular and temporal join when the comparison of SPAM and SPADE is consider the SPAM outperform more than SPADE, while the SPADE algorithm is more SPACE-efficient than SPAM.

C.PREFIX-SPAN

This algorithms presented by Jian Pei, Jiavei Han and Helen Pinto is the only projection based algorithms from all the sequencing pattern mining algorithms. It performs better than the algorithm like apriori, free-span, SPADE (vertical data format).This algorithm finds the frequent items by scanning the sequence database once. The database is projected into several smaller databases according to the frequent items. By recursively growing subsequence fragment in every projected database, we got the complete set of sequential pattern.

The main concept behind the prefix-span algorithm to successfully discovered patterns is employing the divide and conquer strategy. The prefix-span algorithm requires high memory space as compare to the other algorithms in the sense that it requires creation and processing of huge number of projected sub-databases.

II. RELATED WORK

DISCOVERY OF PATTERNS IN SEQUENCE DATA

[4]Existing sequence mining algorithms mostly focus on mining for sub-sequences. However, a large class of applications, such as biological DNA and protein motif mining, require efficient mining of “approximate” patterns that are contiguous. The few existing algorithms that can be applied to find such contiguous approximate pattern mining have drawbacks like poor scalability, lack of guarantees in finding the pattern, and difficulty in adapting to other applications. In this paper, we present a new algorithm called Flexible and Accurate Motif Detector (FLAME).

A.FLAME

It is flexible suffix-tree-based algorithm that can be used to find frequent patterns with a variety of definitions of motif (pattern) models. It is also accurate, as it always finds the pattern if it exists. Using both real and synthetic data sets, we demonstrate that FLAME is fast, scalable, and outperforms existing algorithms on a variety of performance metrics. In addition, based on FLAME, we also address a more general problem, named extended structured motif extraction, which allows mining frequent combinations of motifs under relaxed constraints.

DISCOVERY ALGORITHM OF GENETIC MOTIF FOR SEQUENTIAL DATA

[5]Motif discovery in sequential data is a problem of great interest and with many applications. However, previous methods have been unable to combine exhaustive search

with complex motif representations and are each typically only applicable to a certain class of problems.

RESULTS: Here we present a generic motif discovery algorithm (Gemoda) for sequential data. Gemoda can be applied to any dataset with a sequential character, including both categorical and real-valued data. As we show, Gemoda deterministically discovers motifs that are maximal in composition and length. As well, the algorithm allows any choice of similarity metric for finding motifs. Finally, Gemoda's output motifs are representation-agnostic: they can be represented using regular expressions, position weight matrices or any number of other models for any type of sequential data. We demonstrate a number of applications of the algorithm, including the discovery of motifs in amino acids sequences, a new solution to the (l,d)-motif problem in DNA sequences and the discovery of conserved protein substructures.

DISCOVERY OF DNA AND PROTEIN SEQUENCE MOTIFS

[6]MEME (Multiple EM for Motif Elicitation) is one of the most widely used tools for searching for novel ‘signals’ in sets of biological sequences. Applications include the discovery of new transcription factor binding sites and protein domains. MEME works by searching for repeated, un-gapped sequence patterns that occur in the DNA or protein sequences provided by the user. Users can perform MEME searches via the web server hosted by the National Biomedical Computation Resource and several mirror sites. Through the same web server, users can also access the Motif Alignment and Search Tool to search sequence databases for matches to motifs encoded in several popular formats. By clicking on buttons in the MEME output, users can compare the motifs discovered in their input sequences with databases of known motifs, search sequence databases for matches to the motifs and display the motifs in various formats. This article describes the freely accessible web server and its architecture, and discusses ways to use MEME effectively to find new sequence patterns in biological sequences and analyze their significance.

MINING SEQUENTIAL PATTERNS

[7]A large database of customer transactions, where each transaction consists of customer-id, transaction time, and the items bought in the transaction. We introduce the problem of mining sequential patterns over such databases. We present three algorithms to solve this problem, and empirically evaluate their performance using synthetic data. Two of the proposed algorithms, Apriori some and Apriori All, have comparable performance, albeit Apriori some performs a little better when the minimum number of customers that must support a sequential pattern is low. Scale-up experiments show that both Apriori some and Apriori All scale linearly with the number of customer transactions. They also have excellent scale-up properties with respect to the number of transactions per customer and the number of items in a transaction.

III.FUTURE WORK

Many patterns are hard to find, because they are very sparse and/or variable. Also they might not be specific enough to distinguish them from random noise only based on local sequence information. Small conserved regions that were before interspersed through protein and thus hard to find can occur close to each other in secondary or tertiary structure conformation. Also DNA in eukaryotic nucleus is closely packed in a kind of tertiary structure and this structure can have influence on position of regulatory element binding sites. Therefore it seems that the future approaches to pattern discovery will also use other information available about the sequences.

IV. CONCLUSION

Pattern discovery is an important area of bioinformatics. The algorithms for pattern discovery use wide range of computer science techniques, ranging from exhaustive search elaborate pruning techniques, efficient data structures, to machine learning learning methods and iterative heuristics.

The tools developed by computer scientists are today commonly used in many biological laboratories. They are important to handle large scale data, for example in annotation of newly sequenced genomes, and organization of proteins into families of related sequences. They can be used to detect possible sites of interest and assign putative structure or function to proteins. Thus they can be used to guide biological experiments, decreasing the time and money spent in discovering new biological.

REFERENCES

[1] Agrawal, R. & Srikant, R. (1995). Mining sequential patterns. In 11th Intl. Conf. on Data Engineering.

[2] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast discovery of association rules.

[3] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufman publishers, 2001.

[4] M.O. Dayhoff, R.M. Schwartz, and B. Orcutt, "A Model for Evolutionary Changes in Proteins," *Atlas of Protein Sequence and Structure*, vol. 5, pp. 345-352, Nat'l Biomedical Research Foundation, 1978

[5] Department of Chemical Engineering, Massachusetts Institute of Technology Cambridge, MA 02139, USA October 24, 2005

[6] SDSC, UCSD, La Jolla CA, USA Institute of Molecular Bioscience, The University of Queensland St Lucia, QLD 4072, Australia March 21, 2006

[7] Agrawal, R , Srikant, R. Data Engineering, 1995. Proceedings of the Eleventh International Conference on 6-10 Mar 1995