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. 2, Issue. 11, November 2013, pg.129 – 134

SURVEY ARTICLE

Brief Survey on DNA Sequence Mining

NN Das[1], Poonam[2]

[1] *Computer Science and Engineering, Itm University, Gurgaon, India*
[2] *Computer Science and Engineering, Itm University, Gurgaon, India*

[1] nndas@itmindia.edu; [2] poonamr906@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, 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. A repeated substring is called a tandem repeat if each occurrence of the substring is directly adjacent to each other. 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. With the help of these three things into consideration an effective algorithm can be defined for predicting the tandem repeat in a given DNA sequence.

Keywords –Frequent patterns; DNA; Tandem Repeat; Motifs; KDD

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. The paper focuses on finding tandem repeats of all length in a given DNA sequence.

a. Sequence Mining

Sequence Mining means finding sequential patterns among the large dataset. It finds out frequent substring as patterns from a dataset. With massive amounts of data continuously being gathered , many industries are becoming interested in digging sequential patterns from their databases. Sequential pattern mining is one of the most well-known methods and has broad applications including web-based analysis, customer procure behavior analysis and medical record analysis. In the retailing business, sequential patterns can be mined from the transaction records of customers. For example, having bought a bread packet, a customer comes back to buy a butter and a milk packet next time. The seller can use all this information for analyzing the behavior of the customers procure , to understand their interests, to satisfy their demands, and to predict their requirements. In the medical field, sequential patterns of symptoms of any diseases exhibited by patients to identify strong symptom/disease correlations that can be a valuable source of information for medical diagnosis and preventive medicine. In Web log analysis, the exploring behavior of a user can be extracted from member records or log files. For example, having viewed a web page on "Data Extraction", user will return to evaluate "Business Perception" for new information next time. These sequential patterns gives huge profits, when acted upon, increases customer royalty.
The goal of sequential data mining is to discover frequently occurring patterns but not identical. The challenge in discovering such patterns is to allow for some *noise* in the matching process. To find such a method first is to find the definition of a pattern, and then definition of similarity between two patterns. This similarity definition of the two patterns can vary from one application to another.

b. DNA Sequence Mining
DNA sequence is an important mean to study the structure and function of the DNA sequence. In this paper, based on the characteristics of the DNA sequence an algorithm will proposed which uses the maximal frequent pattern segments based on adjacent maximal frequent pattern mining, to improve the efficiency and availability of the DNA sequence data mining. DNA sequences use an alphabet {A, C, G, T} representing the four nitrogenous bases Adenine, Cytosine, Guanine and Thymine . The Homo Sapiens (human) DNA sequence AX829174 [4] starts with TTCCTCCGCGA and contains 10,011 characters. The subsequence mining problem is of particular importance in reckoning biology, where the challenge is to find short repeated sequences, commonly of length 6- 15, that occur frequently in a given dataset of DNA sequences. These short sequences can provide clues regarding the locations of so called "restrictive regions," which are important repeated sequences in the biological dataset. The repeated occurrences of these tandems are not always identical, and few copies of these tandems may differ from others in some positions. The similarity cadent that is used here could be somehow complicated — for example, when comparing protein, a similarity matrix like PAM[7] or BLOSUM , may be used for comparing the "distance" between each symbol (protein) pair. These frequently occurring short sequences are called motif in reckoning biology. We use this term to describe frequently occurring approximate sequences. Different applications require different similar models to fit the kind of noise which they deal with. It is desirable for a tandem mining algorithm to be able to deal with a variety of concepts of similarity. In this paper, a powerful new model is presented for tandem mining that fits several applications with varying concepts of similarity, including the examples described above. We also present FLexible and Accurate Tandem dEtector —a novel tandem mining algorithm which can efficiently find motif that satisfy our model. In mining for DNA, we are interested in contiguous subsequences.

II LITERATURE WORK

In year 1995 Ramakrishanan Srikant and Rakesh Agrawal [1]proposed an algorithm called GSP[2]. In which he describes that input data is a set of sequences called data sequence .Each data sequence contains a list of transaction ,and each transaction have a list of literals known as items. There is also a transaction time binds with each transaction. A sequential pattern also consist of list of set of items. The problem is to discover all with a user specified minimum support where the support of sequential pattern is the percentage of data sequences that contains the pattern. It present GSP as a new algorithm finds generalized sequential pattern .Empirical evaluation using synthetic and real life indicates that GSP is much faster than the ApprioriAll algorithm. GSP scales linearly with the number of data sequences, and have very good scale up properties with respect to the average data sequence size.

In Year 2006, Panagiotis Papapetrou performed a work[5].The aim of this work is to discover the regions of highly occurrence items in a given sequence and thus proposed two efficient algorithms. First one is known as entropy based algorithm which perform a recursive segmentation to divide the input sequence into number of segments.

In Year 2008, Jing Hu performed ," Sequence Mining [6] for the prediction of DNA binding site. In this author uses a greedy search method to identify the DNA binding sites.There was 534 features out of which 5 were selected.After that Naïve Bayes method obtained 0.31 DNA binding sites in a protein sequence given as input.

In Year 2011, Kwong-Sak Leung[7] performed a work," DNA Sequence Mining on Hepatitis B Virus.In this framework of Data Mining an analysis on molecular evolution, feature selection, clustering ,classifier learning and classification is discussed. That research group has gathered HBV DNA patterns either genotype B or C obtained from over 200 patients for this project.Algorithm proposed is Rule Learning which is based on evolutionary techniques. Which in turn gives important information about the mutated sites and their interaction increasing interest towards classification.

In Year 2009, Bolin Ding performed a work," Mining on Closed Repetitive Gapped Subsequences in a Database". This paper describes the problem of finding gapped repetitive subsequences and proposes a solution for this[8].A database sequence is given where each sequence contains an ordered list of event. Author gives a description of the concept of repetitive support threshold to measure how frequently a subsequence repeats in the database. And study the finding of closed subsequence. Efficient algorithms are proposed to find the complete set of patterns.

In Year 2010, Chien-Chih Wang performed a work," whose aim is to discover bindings of DNA and their orientation. They used Knowledge-based Learning [12]. This paper proposes a learning procedure to detect the location of the bound groove by considering geometric propensity between protein side chains and DNA bases.

In Year 2011, Avrilia Floratou performed a work," Efficient and Accurate Discovery of Sequence Patterns in a given Data Base" and a new algorithm is presented called Flexible and Accurate Motif Detector (FLAME)[10].

FLAME is a flexible suffix-tree-based algorithm that can be used to find frequent patterns with a variety of definitions of motif models. It is also accurate, as it always finds the pattern if present. Using this dataset an algorithm is described called FLAME which is efficient, effective, scalable, and can be suitable for a variety of performance metrics.

III. CLASSIFICATION OF SEQUENTIAL PATTERN MINING ALGORITHM

Basically, there are two main research issues in sequential pattern mining:

1. The first is to improve the efficiency in sequential pattern mining process while the other one is to
2. Extend the mining of sequential pattern to other time-related patterns.

Based on these two conditions sequential pattern mining can be divided as:

Apriori Based
Pattern Growth Based

Apriori-Based Algorithms:

The Apriori and AprioriAll set the basis for a breed of algorithms that depend largely on the apriori property and use the Apriori-generate join procedure to generate candidate sequences. The apriori property states that — All nonempty subsets of a frequent itemset must also be frequent. It is also described as antimonotonic , in that if a sequence cannot pass 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 ignoring ones for which support < min_sup. Then for each level , this algorithm scans datasets to collect support count for each candidate sequence and generates candidate length (k+1) sequences from length-k frequent sequences using Apriori algorithm. This is repeated until frequent sequence or no candidate can be found.

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.

SPADE:

Besides the horizontal formatting method (GSP)[2], the sequence database can be transformed into a vertical format consisting of items' id-lists., is a list of (sequence-id, timestamp) pairs indicating the occurring timestamps of the item in that sequence. Searching in the lattice of the datasets formed by id-list intersections, the SPADE[3] (Sequential Pattern Discovery using Equivalence classes) algorithm presented by M.J.Jaki completes the mining in three passes of scanning the databases. However, additional computation time is required to transform this database of horizontal layout to vertical layout, which also requires additional storage space which is larger than that of the original sequence database.

Pattern Growth Based Algorithms

Just after the design of apriori-based algorithm that was designed in the mid-1990s, the pattern growth-method developed in the early 2000s, and gives solution to generate-and-test problem.The idea behind this algorithm is

to eliminate the numerous candidate generation steps as well as focus the search on a portion of the initial database which is restricted one. In pattern growth algorithm partitioning of search space feature palys a major role. In each pattern growth algorithm working starts by representing the database sequence and then provides the way to partition the database search space and thus producing as less candidate patterns as possible by growing on the already mined frequent sequences, and applying the apriori algorithm as the search space is being traversed recursively looking for frequent sequences.

Features Of Pattern Growth Algorithm:

*(a) Search space partitioning***:** It allows the generated search space having candidate sequence to be partitioned to make the efficient use of memory. Several ways are available for the partitioning of the search space. Among which first is to partition the search space in smaller blocks in parallel. Modern techniques include projected and conditional search.

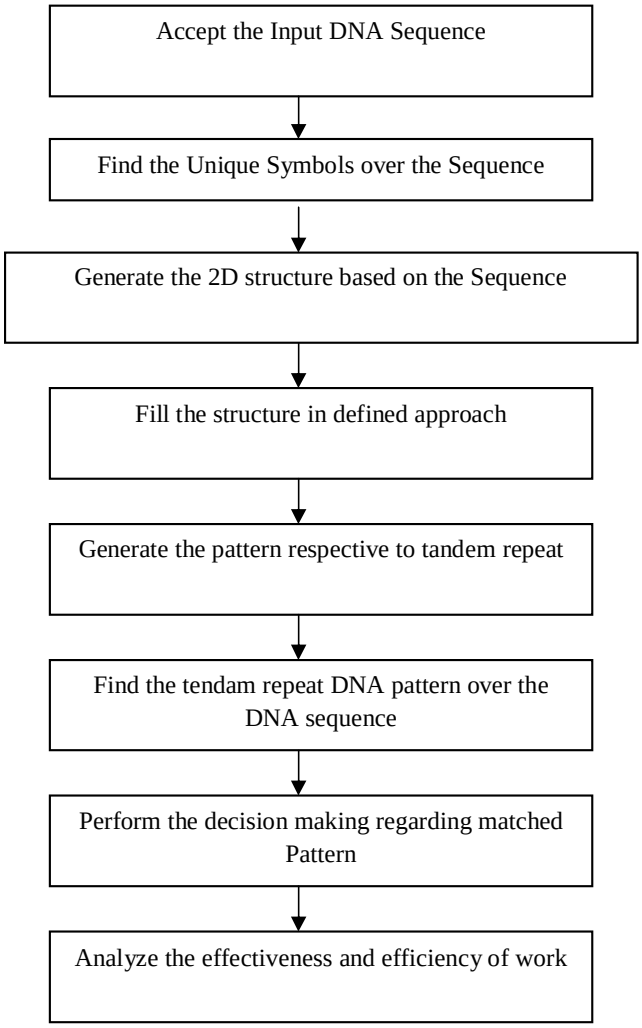
(b) Tree projection:. In this algorithms implement a physical tree data structure representation of the search space, which searches the frequent patterns either using breadth first or depth first search and an apriori algorithm is used for minimizing the sequence length.

*(c) Depth-first traversal***:** That depth-first search of the search space makes a big difference in performance, and also helps in the early pruning of candidate sequences as well as mining of closed sequences . The fact that depth-first traversal used is that utilizes less memory, more directed search space, and thus less candidate sequence generation than breadth-first algorithms.

IV PROPOSED WORK

At present, more popular method of the sequential pattern mining can be divided into two categories:

The first kind, the frequent patterns mining base on sequence alignment, Such as FASTA, BLAST[6] etc. The second is the pattern mining base on the frequent pattern mining algorithm in the field of data mining.
Now the work is done as:



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] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proc. of the ACM SIGMOD Conference on Management of Data, pages 207{216, Washington, D.C., May 1993..

[4] Patrick Hoffman," DNA Visual And Analytic Data Mining", Proceedings of the 8th IEEE Visualization '97 Conference 1070-2385/97 © 1997 IEEE

[5] Panagiotis Papapetrou," Discovering Frequent Poly-Regions in DNA Sequences", Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06) 0-7695-2702-7/06 © 2006 IEEE

[6] Jing Hu," Mining sequence features for DNA-binding site prediction", 978-1-4244-1779-7/08 ©2008 IEEE

[7] Kwong-Sak Leung," Data Mining on DNA Sequences of Hepatitis B Virus", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 8, NO. 2, MARCH/APRIL 2011

[8] Bolin Ding," Efficient Mining of Closed Repetitive Gapped Subsequences from a Sequence Database", IEEE International Conference on Data Engineering 1084-4627/09 © 2009 IEEE

[9] Shuang Bai," The Maximal Frequent Pattern Mining of DNA Sequence".

[10] Avrilia Floratou," Efficient and Accurate Discovery of Patterns in Sequence Datasets", ICDE Conference 2010 978-1-4244-5446-4/10@ 2010 IEEE

[11] 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

[12] Chien-Chih Wang," Predicting DNA-binding Locations and Orientation on Proteins Using Knowledge-based Learning of Geometric Properties", 2010 IEEE International Conference on Bioinformatics and Biomedicine 978-1-4244-8305-1/10 ©2010 IEEE

[13] Avrilia Floratou," Efficient and Accurate Discovery of Patterns in Sequence Data Sets", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 8, AUGUST 2011

