Mining Interesting and Frequent Contiguous Sequential Patterns in Large DNA Sequence Datasets

S. Jawahar
Research Scholar
PG & Research Department of Computer Science
Government Arts College, Coimbatore-18.
Tamilnadu. India
shivamjawahar@gmail.com

Dr. P. Sumathi
Assistant Professor
PG & Research Department of Computer Science
Government Arts College, Coimbatore-18.
Tamilnadu. India
sumathirajes@hotmail.com

***Abstract*— The most important task in data mining is sequential pattern mining from sequential databases. The genomic sequence analysis has led to pattern discovery algorithms for searching large DNA sequences. The proposed Frequent Contiguous Sequential Patterns (FCSP) mines set of contiguous patterns that contains same pattern than traditional pattern mining algorithms. The existing mining algorithms focus on reducing mining time and complexity but finding patterns in longer sequence is challenging. In this paper, we focus on mining contiguous patterns from large DNA sequence. The proposed algorithm is compared with traditional PrefixSpan, BIDE and CloSpan algorithms. The experimental results shows that proposed method is more efficient and mines frequent contiguous patterns less time period.**

Keywords*— *Sequential pattern mining, DNA, PrefixSpan, BIDE, CloSpan

I. INTRODUCTION

Pattern mining [2] is the important data mining topic for finding ordered patterns. These ordered patterns are called sequential patterns. Sequential Pattern Mining (SPM) proposed by Agrawa1 [1] is the process of discovering knowledge to find sequential patterns or frequent patterns in Sequence Database (SD). SPM is an important research in data mining. Mining ordered patterns in sequence database is a challenging task. A subsequent is defined as sequential pattern or frequent sequence if it appears frequently in the sequence database. The frequency is not less than the user specified minimum support threshold *minsup*.

SPM has many applications in fields such as DNA sequence analysis, analysis of customer pattern, diagnosing various diseases, purchase behavior, natural disaster. The SPM algorithms can be divided into three categories: 1. Apriori based algorithms, 2. Vertical format based algorithms and 3. Projection database based algorithms. The Apriori based algorithms was first introduced in 1995 by Agrawal and Srikanth [1]. The Apriori algorithm has three subclasses namely, 1. AprioriSome, 2.AprioriAll and 3. DynamicSome [1].

The Apriori method is bottom up approach which includes various other algorithms such as 1. GSP (Generalized Sequential Pattern) [4], 2. SPAM (Sequential Pattern Mining) [5] and 3. SPADE (Sequential Pattern Discovery using Equivalence)[5]. The GSP algorithm was introduced by srikanth is based on Apriori frequent pattern. It is a vertical format algorithm. The SPADE algorithm was proposed by Zaki in terms of equivalence classes based on projection database.

The projection based algorithms are Freespan [6] and PrefixSpan[7] proposed by Jiawei Han in 2000 and 2001 uses divide and conquer method for raw database into smaller databases. The CloSpan(Closed Sequential Patterns) [8] and BIDE[9] are two popular algorithms for mining closed sequential patterns. Clospan algorithm mines closed sequential pattern candidate set and performs post pruning for mined pattern. This algorithm needs more memory to store closed candidate sequence when mining long patterns. BIDE algorithm uses PrefixSpan framework for mining patterns and uses BackScan pruning technique for eliminating unnecessary patterns.

II. PROBLEM STATEMENT

In this section, we define the problem of Frequent Contiguous Sequential Pattern mining

Let Σ = {A, C, G, T} be DNA alphabets where A, C, G, and T are called DNA characters or four bases where A represents Adenine, C represents Cytosine, G represents Guanine, and T represents Thiamine. A DNA sequence S is an ordered list of DNA where S= {S1, S2…Sn>} and Si ∈ Σ and S denotes the length of sequence S.

Given a DNA Sequence Database (SDB) and minimum support threshold δ, the problem of frequent contiguous pattern mining is to find the complete set of frequent contiguous sequential patterns from the sequence database.

The biological DNA sequence contains repeating sequences of only four nucleotide namely, A, T, G and C respectively. An example relating to FCSP is
Itemset I = {A, C, G, T}
Sequence Database, SDB = {10, ***ATGGC*** TT),
20, A ***ATGGC*** A),
30, TT ***ATGGC***)}
Frequent Contiguous Sequential Pattern, FCSP is ***ATGGC***

In the above example three sequences are in SDB 10,20 and 30. Each SID contains sequences and FCSP is to find commonly repeated frequent contiguous sequence in the SDB where $1 \leq i \leq n$.

III. PROPOSED ALGORITHM FOR EFFICIENTLY MINING FREQUENT CONTIGUOUS SEQUENTIAL PATTERN (FCSP)

The traditional sequencing algorithms may generate many redundant patterns when mining biological sequences.
The definitions are:

Definition 1 [6]: (Sequential Pattern) Sequential pattern is a subsequence of pattern whose frequency of occurrence is not less than user defined minimum support (*min_sup*).

Definition 2 [6]: (Minimum Support) The minimum support of a subsequence *A* in a dataset *S* is the number of tuples in the dataset containing *A*.

support(*A*)=|{<*sequence_id, s*>|(<*sequence_id, s*>□*S*)_ (*A_S*)}|.

Definition 3: (Contiguous Sequence) A sequence a = { a1, a2, …. , an} is contiguous of another sequence b = {b1, b2,.....,bn} if there exists an integer $1 \leq i \leq k-j+1$. In this case *a* is also contained in *b* which is represented by $a \subseteq b$.

Proposed Frequent Contiguous Sequential Pattern (FCSP) Algorithm

Input: A Sequence Database *(SDB)* and minimum support *threshold min_sup*

Output: The set of frequent contiguous patterns.

Algorithm Frequent Contiguous Sequential Pattern ()

Step 1: The following variables are initialized

SID (Sequence ID): A1, A2, A3,.An

Sequence: S1, S2, S3…Sn where n is the length of sequence

Step 2: SID = 0; // To initialize sequence ID

Occurrence_Count = 0; // To initialize occurrence count of each contiguous pattern

Pointer_ptr = {NULL} // To initialize current position with matching position of each contiguous pattern

Item_I[i]={$i_1,i_2,i_3,\ldots\ldots i_k$} where $1 \leq I \leq k$ // Subsequence itemset values

FCSP: $p_1,p_2,p_3\ldots$.

Step 3: Identifying subsequences for SDB's

for i = 1 to k // i represents k items

for j = 1 to n // j represents n sequences

if (S_j[p] = = I[i]) //Check for sequence

item_I[i] = { SID, Occurrence_Count, Pointer_ptr} //store subsequence

Step 4: Finding Frequent Contiguous Sequential Patterns, FCSP

for j = 1 to k // j represents the k-sub SDBs

for i = 1 to n // i represents patterns in each sub SDBs

if (FCSP = = P1P2...Pl)

Print ("Frequent Contiguous Sequential Patterns, SID");

The Frequent Contiguous Sequential Patterns mines sequential patterns with user-defined minimum threshold value. The proposed algorithm is FCSP which is used for mining contiguous sequential patterns. The algorithm includes

following steps,
Step 1: Database is scanned only once to find out frequent subsequences (FS1) by dividing SDBs into sub-SDBs.

Step 2: The complete sequence is divided for each frequent subsequence (FS2).

Step 3: Find out the frequent contiguous patterns in the subsets.

The proposed FCSP algorithm reduces scanning time and searching time. The sequence search is not made in SDBs instead it is searched on sub-SDBs based on first sequence search.

IV. PERFORMANCE EVALUATION

The experiment environment is CPU i5 2.8GHZ processor, 4GB RAM, windows 7 operating system and algorithm code is written in JAVA language and compiled in eclipse tool. The comprehensive performance study of FCSP shows how proposed algorithm outperforms PrefixSpan, BIDE and CloSpan. The analysis is based on total contiguous sequence count, total memory (mb) and total time taken (ms). The table 1 represents the four input sequences for FCSP algorithm.

TABLE I DNA Sequence

SID	SEQUENCE
1	CAAGC
2	AGCG
3	CAGC
4	ATTCA

The table 2 represents the sequence count for the proposed algorithm FCSP and existing algorithms. The proposed algorithm FCSP has 4 sequence count, BIDE and CloSpan has 6 sequence count each and PrefixSpan has highest sequence count 15. The memory requirement is low for proposed FCSP algorithm than existing algorithms.

TABLE II Sequence count for various algorithms

Algorithm	Sequence Count
PrefixSpan	15
BIDE	6
CloSpan	6
FCSP	4

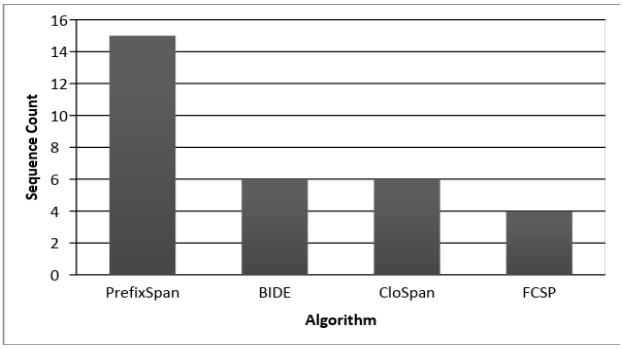

Figure 1 Various algorithm vs sequence count

The table 3 represents the total memory for the proposed algorithm FCSP and existing algorithms. The proposed algorithm FCSP has 0.95 mb of memory, BIDE algorithm consumes 1.68 mb of memory, CloSpan algorithm consumes 1.45 mb of memory and PrefixSpan algorithm consumes 0.96 mb of memory. The total memory requirement is low for proposed FCSP algorithm than existing algorithms.

TABLE III Total memory consumption

Algorithm	Total Memory (mb)
PrefixSpan	0.96
BIDE	1.68
CloSpan	1.45
FCSP	0.95

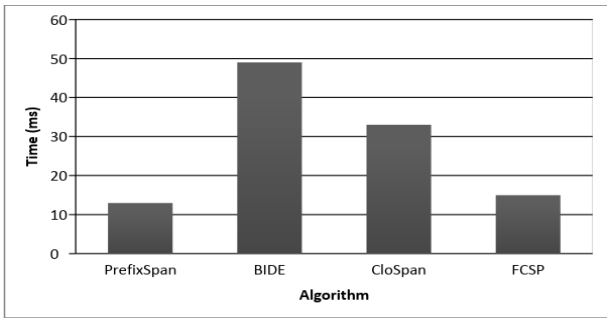

Figure 2 Various algorithm vs memory

The table 4 represents the total time for the proposed algorithm FCSP and existing algorithms. The proposed algorithm FCSP has 15 ms of time period, BIDE algorithm takes 49 ms of time period, CloSpan algorithm takes 33 ms of time period and PrefixSpan algorithm takes 13 ms of time period. The total time taken for proposed FCSP algorithm is less than existing algorithms, but PrefixSpan algorithm is faster than other algorithms.

TABLE IV Total Time taken

Algorithm	Total Time (ms)
PrefixSpan	13
BIDE	49
CloSpan	33
FCSP	15

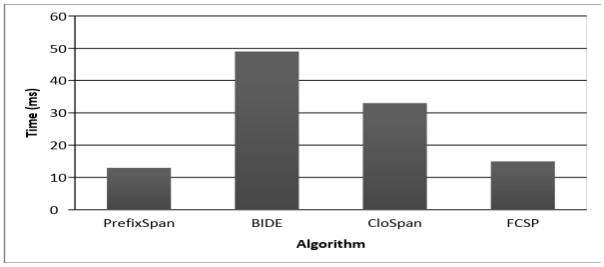

Figure 3 Various algorithm vs time taken

The frequent contiguous sequential patterns achieves reasonable performance especially in biological sequence analysis. The FCSP patterns have smaller volume and shorter length of contiguous sequential patterns. The FCSP algorithm shows better performance in terms of both efficiency and scalability compared to existing algorithms.

V. CONCLUSION

In this paper, we proposed frequent contiguous sequential pattern algorithm (FCSP) an efficient approach for mining complete set of frequent contiguous patterns. This method eliminates number of inefficient and redundant patterns efficiently. We evaluated the performance of FCSP algorithm with real life datasets and experiment results shows set of frequent contiguous patterns are more compact in memory and run time than traditional algorithms (PrefixSpan, BIDE & CloSpan). The proposed algorithm has greater retrieval performance and reduced access to sequence database. The advantage is entire sequence database is scanned only once for mining contiguous patterns. It generally reduces the space and time complexity of the mining process.

References

[1] R. Agrawal, R. Srikant. Mining Sequential Pattern Pro. of the 11st Int. Conf. on Data Engineering , Taipei,1995,3:3~14.

[2] Agrawal, R., Imieliskiand, T., Swami, A.: Mining association rules between sets of items in large databases. In: ACM SIGMOD, pp. 207–216 (1993)

[3] Pei, J. , J. W. , and Wang, J. Y. et al. Mining sequential patterns by pattern-growth: The prefixspan approach. IEEE Transactions On Knowledge And Data Engineering, Volume 16, 2004:1-17.

[4] Srikant R,Agrawal R. Mining sequential pattems: generali-zations and performance improvements [C]. EDBT 96:Proceeding of the 5th International Conference on Extending Database Technology: Advance in Database Technology .UK,London :Springer-Verlag,1996:3-17.

[5] Zaki M J. SPADE: An efficient algorithm for mining frequent sequences[J]. Machine learning, 2001, 42(1-2): 31-60.

[6] Han J, Pei J, Mortazavi-Asl B, et al. Freespan：freguent patternprojected seguential pattern mining ［A］.In：Proceedings of the International Conference on Knowledge Discovery and Data Mining ACM［C］.Montreal, Canada, 2000. 355 359.

[7] Pei J,Han J,Mortazavi-Asl B,et al. Mining sequential patterns by patterngrowth: the prefixspan approach [J]. IEEE Transaction On Knowledge and Data Engineering ,2004,16(11):1424-1440.

[8] X. Yan, J. Han, and R. Afshar, “CloSpan: Mining closed sequential patterns in large databases,” Proceedings of SIAM’s SDM ’03, 2003.

[9] J. Wang, J. Han, and Chun Li, “Frequent closed sequence mining without candidate maintenance,” IEEE TKDE, Aug. 2007.http://m.youtube.com/results?q=arduino%20connection%20with%20GPS%20and%20GSM&sm=12