Data Mining Techniques for Informative Motif Discovery

Rashida Hasan
Dept. of Computer Science & Engineering
University of Dhaka
Bangladesh

Jainal Uddin
Dept. of Computer Science & Engineering
University of Dhaka
Bangladesh

ABSTRACT
The discovery of motifs in biological sequence is a much explored and still exploring area of research in functional genomics since they control the expression or regulation of a group of genes involved in a similar cellular function. This paper explores the use of data mining techniques by various researchers as a solution to discover motifs in biological sequence. Although data mining techniques has not been applied extensively by researchers as compared to other algorithms. But in recent years data mining techniques has caused a wide attention by the researchers to find motifs in biological sequence. This paper is an attempt towards exploring the effectiveness of data mining techniques for motif discovery.

Keywords
Motif, data mining, information content

1. INTRODUCTION
Motif discovery is one of the most well known problems, which is not yet totally solved. The near completion of the human genome in 2000 makes this interesting question more an urgent work in scientific community. A motif, in the context of biological sequence analysis, is a region or portion of a protein or RNA or DNA sequences that has specific structure and is functionally significant. The motif discovery problem can be simply formulated as the problem of finding short segments that are over represented among a set of long DNA/RNA or protein sequences [1]. These short segments are better conserved in evolution and therefore they occur more frequently than expected. DNA motifs are sometimes termed signals such as regulatory sequences, scaffold attachment sites and messenger RNA splices. Examples of protein motifs, which are also known as fingerprints, include enzyme active sites, structural domains and cellular localization tags.

Motif may occur in nucleotide and protein sequences that have been preserved through evolution because they are important to the structure or function of the molecule. Through the identification of protein sequence motifs, an unknown sequence can be quickly classified into its computationally predicted protein family/families for further biological analysis [2]. Conserved patterns in protein sequences can reveal the proteins functional role which can be a critical piece of information for drug designers [3].

With years of research and development, considerable approaches for discovering motifs have been developed in bioinformatics community. The algorithmic approaches to motif discovery exhibit surprising variety. Recently data mining techniques have been used for discovering motifs in DNA or protein sequence.

Most of the algorithms such as alignment algorithms, combinatorial methods, and probabilistic methods face an efficiency problem when they handle a quick growth of large scale sequence data. Probabilistic methods such as MEME [3], EM [4] is sensitive to noise and the fact they are not guaranteed to converge to a global maximum. Data mining techniques for motif discovery has certain advantages. It has been found that data mining techniques are able to find the exact motifs and require fewer parameters. The satisfactory experimental results show that these methods can handle a quick growth of large scale sequence data.

The remainder of this paper is divided into two sections. Section 2 presents the biological basis of motifs. Section 3 reviews data mining approaches for motif discovery.

2. BIOLOGICAL BASIS OF MOTIF
A motif is a nucleotide or amino-acid sequence pattern that is widespread and has, or is conjectured to have, a biological significance such as being DNA binding sites for a regulatory protein i.e., a transcription factor [5]. A nucleotide sequence is a string of letter (A, T, C, G) representing the sequence of nucleotide bases (Adenine, Cytosine, Guanine and Tyrosine) present with DNA and RNA molecules. A protein sequence is string of letters (D, E, K, R, H, N, Q, S, T, I, L, V, F, W, Y, C, M, A, G, P) representing the linear sequence of amino acids from which a protein is constructed [6].

3. DATA MINING TECHNIQUES FOR MOTIF DISCOVERY
With years of research and development, considerable and remarkable approaches for discovering sequential patterns have been developed in Bioinformatics community. Recently, the topic of sequential pattern mining has caused a wide attention in KDD community. The problem was first introduced by Agrawal and Srikant [7].

Dodd and Egan implemented a program called GYM which is based on the Aprirori method from data mining to detect helix-turn-helix motif that is common to many DNA binding proteins and plays a crucial role in their binding to DNA [8]. The algorithm find patterns generated from the sample training set and searches whether they are present in a new protein sequence. Their method is powerful in detecting approximately 50% more likely helix-turn-helix sequences without an increase in false predictions. But in this method, the choice of training set is a non-trivial problem and could determine the success or failure of a motif detection method. Another problem is the choice of minimum support threshold.

Hajime and Tomoki presents modified prefix span method for motif discovery in sequence databases [9]. This method extracts frequent patterns from an annotated sequence database that has such attributes as a sequence identifier, a sequence and a set of items. This method includes a function to extract frequent patterns together with gaps or wild character symbols. This method avoids generating large number of meaningless patterns extracted from the projected databases by restricting the range of each postfix-sub-

sequence. The user provides the number of gaps needed to extract meaningful frequent patterns and the underlined subsequence is designed as the projected database. The patterns satisfy minimum support threshold and maximum number of gaps are extracted from the projected database. This method reduces the required computational time to extract frequent sequence patterns. They provide a user interface that permits a database to be queried in different ways. The modified prefix span has been applied to the evaluation of three set of sequences that include the Zinc Finger, Cytochrome C and Kringle motifs. However, the main problem of this approach is the choice of minimum support threshold

Yun and Yangyong represent a novel protein sequential pattern mining algorithm based on prefix projected method called BioPM [7]. The major idea of BioPM is any frequent subsequences can always be found by growing frequent prefix, using the projection based on frequent prefixes to mine motifs, instead of considering all the possible subsequence. It greatly reduces the effort of candidate subsequence generation. As protein motifs are sometimes long, BioPM sets a window to control pattern growth width for each time. A scoring matrix BLOSUM is introduced to find a match degree among motifs such that the results of mining are suitable for biological meaning. After finding the complete set of sequential patterns, next step is to prune this set using BioPM-tree database in order to delete redundancy patterns. The satisfactory experimental results suggest that BioPM improves performance and overcome some shortcomings which traditional algorithm possesses.

L. Bilal and B. Brahim use pushdown automata to form a tree of sequences and then use a data mining technique to mine the tree structure [10]. Pushdown automata create a suitable grammar for biological data. A set of motif sequences from a biology bank transferred to a set of pushdown automata based on the grammar. They use apriroi algorithm to extract motifs from the pushdown automata.

H. ozer and W. C. Ray proposed an algorithm inspired by the classic apriori algorithm to find frequent residue motifs that are high in information content and outside of the family consensus, called informative motif [11]. This algorithm overcomes the limitation of other data mining techniques such as choice of minimum support threshold and avoids huge candidate generation but it is an iterative algorithm. In this method, a transaction refers to a sequence and an item set refers to residue motif. Each residue is subscripted with its position and Position Weight Matrix (PWM) is computed by calculating position specific probabilities of each residue. In apriori algorithm, minimum support is a critical user defined value. High value of minimum support may not find the rare but important ones. Low value of min support may discover many meaningless patterns. So motifs are mined based on information content.

F. Haque and N. Noman modified an algorithm to mine informative motifs proposed by Ozer and Ray [11]. Their algorithm based on Fp-growth method which requires only two scans to mine frequent pattern [12]. A study on the performance of the Fp-growth algorithm method shows that it is efficient for mining both long and short frequent patterns. In this method, they construct a compact data structure called FP-tree from database by one scan. Before applying algorithm, a position weight matrix is constructed. After generating all the candidates of different sizes, the candidates with information content less than 0.5 are eliminated to obtain the motifs. The experimental result shows that this method produces biologically significant motifs.

4. PERFORMANCE EVALUATIONS OF DATA MINING TECHNIQUES FOR MOTIF DISCOVERY

A large number of motif finding algorithms are available, but it is a challenging task to choose the best tool for motif finding. It is difficult to identify optimal test sets for benchmarking. It is also difficult to know whether a test result actually reflects the assumed methodological difference between alternative approaches. Many methods will require different degrees of parameter tuning such as motif length, expected number of motif occurrences, inter motif distances etc. Different implementations are optimized and fine tunes to different degree which makes it difficult to distinguish between the performances of underlying algorithmic approaches. In this section, we present a performance evaluation of data mining techniques comparing with other known motif discovery tools.

Table 1 shows the comparison of data mining techniques in respect of operating principle with some well known motif finding tools.

Table 1. Comparison of well know motif tools with data mining techniques in respect of operating principle

Algorithm	**Operating principle**
EM	Expectation maximization
MEME	Expectation maximization
AlignACE	Gibbs sampling
YMF	Enumeration
MITRA	Prefix tree/Graph
MotifSeeker	Data fusion and ranking
SMILE	Suffix tree
Consensus	Weight matrix
Bioprospector	Gibbs sampling
EC	Genetic algorithm
GYM	Data mining (apriori)
Modified Prefix Span	Data mining (prefix span)
BioPM	Data mining (prefix span with Blosum)
Pushdown automata with data mining	Data mining
Informative motifs using apriori	Data mining(modified apriori)
Informative motif using fp-growth	Data mining (Fp-growth)

The performance of modified prefix span was compared with variable and fixed gap-length methods using the prototype systems. All of the experiments were performed on a 500-MHz Pentium III PC computer. Modified prefix span method uses three sets includes Zinc Finger, Cytochrome C and Kringle for comparison. Fig 1 shows the comparison of variable gap-length method and fixed gap-length method when a mini-support threshold is 80%.

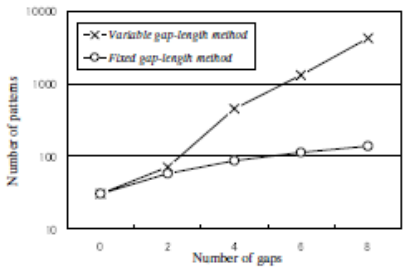

Fig 1: Number of patterns generated by variable gap-length method and fixed gap-length method [9]

The figure 1 suggests that fixed gap length method computes about 5 times faster than the variable gap length method. The fixed gap-length method produces 30 times fewer patterns than the variable gap-length method.

BioPM selects sequences of 10 protein families from Pfam [13] to test the efficiency. The average length of sequence is 800 and all the experiments are performed on 2 GHz Pentium PC computers with 256 megabytes main memory. BioPM compares its experimental results with L-tuple algorithm and apriori algorithm. It demonstrates BioPM based on projected strategy possesses better performance than others because it avoids candidate patterns generation. Fig 2 shows the comparison with L-tuple and apriori with different supports.

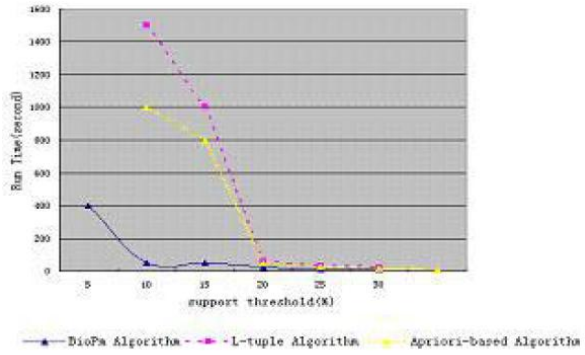

Fig 2: Comparison of BioPM, L-tuple and Apriori on Pfam dataset [7]

To test the superiority and effectiveness of informative motif using fp-growth and informative motif using apriori algorithm, three different data sets are used includes yst04r, yst08r and PDB structures 2ak3. Table 2 shows the experimental results on the performance of informative motif using fp-growth method in comparison with informative motif using apriori and MEME method. Table 3, 4 and 5 shows the runtime comparison. It demonstrates that method based on apriori and method based on fp-growth possesses better performance and much more efficient than existing MEME method.

Table 2: Motifs predicted for yst04r, yst08r, 2ak3

Data set	Method	Predicted Motif
yst04r	MEME	ACCGTGAAGGTGCCGTAGAG ACCAAGAAGATGCCGCCCTG ACGGTCAGGGTAGCGCCCTG AACATGTAGGTGGCGGAGGG
	Method based on apriori	ACCGTGAAGGTGCCGTAGAG ACCAAGAAGATGCCGCCTGG AAGGTCAGGGTAGCGCCCTG AACATGTAGGTGGCGGAGGG
	Method based on fp-growth	ACCGTGAAGGTGCCGTAGAG ACCAAGAAGATGCCGCCTGG AAGGTCAGGGTAGCGCCCTG AACATGTAGGTGGCGGAGGG
yst08r	MEME	GCGCCGCGCCGCCGCTCTC GCGCCGTCCGCCCTCTCTC GCACGTGCGATCATCGTGG CCACGCGCGATCGCCATGG
	Method based on apriori	GCGCCGCGCCGCCGCTCTC GCGCCGTCGCCCTCTCTC GCACGTCCGATCATCGTGG CCACGCGCGATCGCCATGG
	Method based on fp-growth	GCGCCGCGCCGCCGCTCTC GCGCCGTCGCCCTCTCTC GCACGTCCGATCATCGTGG CCACGCGCGATCGCCATGG
2ak3	MEME	D.24,T.27,E.30,E.36 P.4,S.6,R.8,I.23 D.24,T.27,E.30,P.31,V.33 H.3,P.4,S.6,R.8,E.14 H.3,S.6,R.8,V.10,N.12,E.30 I2,H.3,P.4,S.6,R.8,D.24
	Method based on apriori	D.24,T.27,E.30,E.36 P.4,S.6,R.8,I.23 D.24,T.27,E.30,P.31,V.33 H.3,P.4,S.6,R.8,E.14 H.3,S.6,R.8,V.10,N.12,E.30 I2,H.3,P.4,S.6,R.8,D.24
	Method based on fp-growth	D.24,T.27,E.30,E.36 P.4,S.6,R.8,I.23 D.24,T.27,E.30,P.31,V.33 H.3,P.4,S.6,R.8,E.14 H.3,S.6,R.8,V.10,N.12,E.30 I2,H.3,P.4,S.6,R.8,D.24

The results shows that existing method and the method based on apriori and method based on fp-growth predict the same motifs, which prove the biological significance of the motifs predicted by these two data mining methods.

Table 3: comparison of runtimes for yst04r

Length	Time	
	MEME	**Method based on fp-growth**
10	1.00 sec	0.00 sec
11	6.00 secs	0.00 sec
13	13.00 secs	1.00 sec
14	1.13 mins	13.00 secs
15	4.87 mins	1.85 mins
18	90.00 mins	77.10 mins
20	190.00 mins	146.49 mins

Table 4: comparison of runtimes for yst08r

Length	Time	
	MEME	**Method based on fp-growth**
10	1.00	0.00 sec
11	8.00	0.00 sec
13	51.00	4.00 secs
14	3.017 min	36.00 secs
15	21.00 min	12.05 mins
18	84.28 min	48.2 mins
20	264.00 mins	210.6 mins

Table 5: comparison of runtimes for 2ak3

Length	Time	
	MEME	**Method based on fp-growth**
10	21.00 secs	5.00 secs
11	25.00 secs	10.00 secs
13	71.00 secs	15.00 secs
14	3.35 mins	50.00 secs
15	25.07 mins	4.98 mins
18	90.6 mins	50.2 mins
20	170.05 mins	140.0 mins

5. CONCLUSION

Motif discovery is one of the most attempted problems in the area of bioinformatics. Despite considerable efforts to date, motif finding remains a complex challenge for biologists and computer scientists. In this paper, we presented a survey of motif discovery algorithms using data mining techniques. The experimental results show that data mining techniques produces biologically significant motifs and has reasonable efficiency. There are number of directions where improvement can take place in future studies. Currently available some motif algorithms based on data mining techniques cannot deal with gapped motifs; some algorithm needs minimum support threshold. The researchers can improve methods with a combination of allowing gapped motifs as well as avoiding support threshold so that it can find motifs from all kinds of sequences.

6. ACKNOWLEDGEMENTS

The authors would like to gratefully acknowledge the University of Dhaka for providing necessary facilities.

7. REFERENCES

[1] Chen, X. and Jiang, T. 2006. An improved gibbs sampling method for motif discovery via sequence weighting. In Proceedings of Computational System Bioinformatics Conference.

[2] Jahanian, K. 2011. Using sequential pattern mining in protein sequences discovery with gap. Aust. J. Basic and App. Sci. 5(12). 1476-1480.

[3] Helden, J. and Rios, A. 2000. Discovering regulatory elements in non-coding sequences by analysis of apaced dyads. Nuc. Acids. Res. 28(8). 1808

[4] Lawrence, C. E. and Reilly, A. A. 1990. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins: Struct. Func. And Gen. 7(1). 41-51.

[5] Rigoutsos, I and Flortos, A. 1998. Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. Bioinformatics. 14. 55-67.

[6] Lones, M. A. and Tyrell. A. M. 2005. In Proceedings of GECCO Workshop on Genetic and Evolutionary Computation. 1-11.

[7] Yun, X and Z. Yangyoung. 2007. BioPM: An efficient algorithm for protein motif mining. In Proceedings of Bioinformatics and Biomedical Engineering.

[8] Dodd, I. B. and Egan, J.B. 1990. Improved detection of helix-turn-helix DNA binding motifs in protein sequences. Nuc. Acids, Res. 18(17).

[9] Kitakami, H., Kanbara. T., Mori, Y., Kuroki, S. and Yamazaki, Y. 2002. Modified prefix span method for motif discovery in sequence databases. Lecture Notes Comp. Sci. 482-491.

[10] Bilal, L., Brahim, B. and Abdelouahab, M. 2013. Biological motif discovery algorithm based on mining tree structure. Int. J. Comp. App. 69(4). 35-39.

[11] Ozer, H. G. and Ray, W. C. 2007. Informative motifs in protein fanily alignments. Algorithms in bioinformatics. Lecture notes Comp. Sci. 161-170.

[12] Haque, F. A., Mohebujjaman, M. and Noman, N. 2011. Informative motif detection using data mining. Res. J. Inf. Tech. 3(1). 23-32.

[13] Bateman, A. Birney, E. and C, Lorenzo. Et al. 2002. The Pfam protein families database. Nuc. Acids. Res. 30(1). 276-280.