International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-1 Issue-9 March-2013

A Review of Protein-DNA Binding Motif using Association Rule Mining

Virendra Kumar Tripathi[1], Hitesh Gupta[2]

Abstract

The survival of gene regulation and life mechanisms is pre-request of finding unknown pattern of transcription factor binding sites. The discovery motif of gene regulation in bioinformatics is challenging jobs for getting relation between transcription factors and transcription factor binding sites. The increasing size and length of string pattern of motif is issued a problem related to modeling and optimization of gene selection process. In this paper we give a survey of protein-DNA binding using association rule mining. Association rule mining well known data mining technique for pattern analysis. The capability of negative and positive pattern generation help full for discovering of new pattern in DNA binding bioinformatics data. The other data mining approach such as clustering and classification also applied the process of gene selection grouping for known and unknown pattern. But faced a problem of valid string of DNA data, the rule mining principle find a better relation between transcription factors and transcription factor binding sites.

Keywords

Protein-DNA, Motif, data mining, association rule mining

1. Introduction

The discovery of unknown motif is fundamental of gene regulation and life of survival. In process of unknown motif is found size and length of string is problem of binding in DNA-protein data [5, 6]. Proteins are vital functional units made of one or more polymer chains of amino acids. A protein can be considered as a sequence comprising 20 letters (residues) of amino acids, for example, "...AAAFVNQHLCGSHLVEALYLVCGERGFFYT..." DNA (Deoxyribonucleic acid) stores genetic instructions of living organisms. It consists of two long polymers called nucleotides.

Virendra Kumar Tripathi, Research Scholar, Department of Computer Science & Engineering, Patel College of Science and Technology, Bhopal, India.

Hitesh Gupta, Assistant Professor, Department of Computer Science & Engineering, Patel College of Science and Technology, Bhopal, India.

There are four bases which are denoted by the letters "A", "C", "G" and "T". A fragment of a DNA can be represented as a sequence comprising these four letters (residues). Protein binding to DNA plays a fundamental role in regulating cellular and viral functions. The mechanisms by which proteins and DNA interact to control transcription and replication are slowly being elucidated. DNA-protein interactions are studied using a variety of methods such as gel-shift assays, foot-printing, and transcriptional activation. While each of these methods may contribute distinct information about the location or effect of binding, they do not provide a simple way of quantitatively measuring specific binding. The studies of protein-DNA bindings between transcription factors (TFs) and transcription factor binding sites (TFBSs) are important bioinformatics topics. Discovering protein-DNA bindings between transcription factors (TFs) and transcription factor binding sites (TFBSs) is a fundamental biological problem[15,16]. A transcription factor (TF) is a special type of proteins that can bind to a region of DNA called transcription factor binding site (TFBS) to regulate (activate and control) the expression of a gene. Only short regions (length <10 residues) of a TF and a TFBS are actually involved in the binding. TF-TFBS binding cores are traditionally identified through expensive, labor-intensive and time-consuming 3D structure experiments. However, due to the complexity, the numbers of 3D structures extracted are far less than the number of records in sequence representation [12, 13, 14]. The problem of mining protein-DNA approximate association rules over biological databases, to solve this problem a novel algorithm for efficient mining using two special data structures called Frequent Sequence Tree(FS-Tree) and Frequent Sequence Class Tree(FSC-Tree) has been proposed. The association rule mining is used to predict approximate protein-DNA binding cores. The problem of mining association rules was first proposed in aiming to efficiently mine association rules between large itemsets over a transaction database. Given a transaction database, the problem is to find all co-existing itemsets which both frequently appear independently (support) and dependently (confidence) in the database. Association rule mining finds all rules in the database that satisfies some minimum support and minimum

confidence constraints. Unlike a transactional database normally used in association rule mining that does not have many associations, classification data tends to contain a huge number of associations. Adaptation of the existing association rule mining algorithm to mine only the CARs is needed so as to reduce the number of rules generated, thus avoiding combinatorial explosion. The rest of paper is organized as follows. In Section 2 discuss related technique of motif discovery. The Section 3 discusses problem, Followed by a conclusion in Section 4.

2. Related Work

In [1] relationship between DNA binding and protein sequence composition is analyzed. As a result it is found that sequence composition provides sufficient information to predict the probability of its binding to DNA with nearly 69% sensitivity at 64% accuracy for the considered proteins. In this analytical survey it is found that residue composition of DNA-binding proteins has two levels of specificity. One of them is at the sequence level, which can be used to classify sequences as binding or non-binding. The other is at the binding site level, which, when coupled with residue neighborhood information and local structural information (particularly solvent accessibility), can be helpful for locating binding sites in a totally new sequence, even if there is no homology with known binding proteins.

A method for DNA binding classification according to the structural motif was proposed [2]. The present results show that some proteins with the same motif are classified into different clusters whereas different proteins with distinct motifs are classified into the same cluster, suggesting that the motif-based classification of DNA-binding proteins may not necessarily correspond to structural and functional properties characterizing protein-DNA recognition.

A method to approximate associated TF-TFBS pattern discovery employing a probabilistic model for more biologically appealing patterns than simple counts in association rule mining was proposed[3]. In this method similar TFBS motif has been grouped into a consensus group Ci according to a similarity threshold TY, where Ci indicates the consensus group identifier, the TRANSFAC TFBS motif identifier of the first member used to represent the group.

A methodology to identify de novo individual and interacting pairs of binding site motifs from ChIP-chip data, using an algorithm that integrates localization data directly into the motif discovery process was proposed [4]. To identify binding site motifs this method uses a strategy of generating candidates using sequence and localization data, determining how well the candidates can predict the localization data (alone or in pairs), and focusing the search once more on sequence regions near high scoring candidates to identify additional, possibly more subtle motifs that co-localize with a high scoring candidate. A comprehensive method for identifying bind-ing site motifs and motif pairs from ChIP-chip data that incorporates several features that are new to ChIP-chip analysis is proposed.

A method to use TF classifier as a good support in TF annotations as the amount of protein sequences increases rapidly in the post-genomic era based on functional domain composition was proposed [5]. Transcription factor (TF) is often termed as the major regulator of transcription. Most of them (or dimers) bind to specific DNA fragments using their DNA-binding domain and modulate nearby genes' transcription through their trans-activating/repressing domains. Generally, transcription factors can be classified into four major classes: (1) Basic domains. (2) Zinc-coordinating DNA-binding domains. (3) Helix-turn-helix. (4) b-Scaffold factors with Minor Groove Contacts.

An algorithms for generation of frequent item sets by successive construction of the nodes of a lexicographic tree of item sets was proposed [6]. In this paper different strategies in generation and traversal of the lexicographic tree such as breadth first search, depth-first search or a combination of the two is proposed by the authors. These techniques provide different trade-offs in terms of the I/O, memory and computational time requirements. This paper demonstrated the power of using transaction projection in conjunction with lexicographic tree structures in order to generate frequent itemsets required for association rules. The advantage of visualizing itemset generation in terms of a lexicographic tree is that it provides us with the flexibility of picking the correct strategy during the tree generation phase as well as transaction projection phase. The depth-first projection technique provides locality of data access, which can exploit multiple levels of cache. Authors have also demonstrated the parallelizability of the Tree Projection technique, and

the advantages of its parallel implementation over the parallel implementation of the Apriori algorithm.

A novel, OpportuneProject, algorithm or mining complete set of frequent item sets by projecting databases to grow a frequent item set tree was proposed [7]. This algorithm is fundamentally different from those proposed in the past in that it opportunistically chooses between two different structures, array-based or tree-based, to represent projected transaction subsets, and heuristically decides to build unfiltered pseudo projection or to make a filtered copy according to features of the subsets. Authors have presented novel pseudo projection methods for tree-based representations in the depth first search, which greatly improves the efficiency of counting and projecting operations in dense transaction subsets. Authors have proposed an array-based data structure that is the most space efficient and the simplest for sparse datasets. Authors define heuristics that adapts the algorithm to the features of the projected transaction subsets by integrating array-based and tree-based representations, and employing different projecting and counting methods opportunistically.

A novel sequential pattern mining method, called PrefixSpan was proposed [8]. This method explores prefix-projection in sequential pattern mining. PrefixSpan mines the complete set of patterns but greatly reduces the efforts of candidate subsequence generation. Its general idea is to examine only the prefix subsequences and project only their corresponding postfix subsequences into projected databases. To further improve mining efficiency, two kinds of database projections are explored: level-by-level projection and bi-level projection, and an optimization technique which explores psuedo-projection is developed. PrefixSpan represents a new and promising methodology at efficient mining of sequential patterns in large databases. It is interesting to extend it towards mining sequential patterns with time constraints, time windows and/or taxonomy, and other kinds of time-related knowledge. Also, it is important to explore how to further develop such a pattern growth-based sequential pattern mining methodology for effectively mining DNA databases.

A novel SPADE algorithm was proposed [9]. SPADE algorithm is presented for fast discovery of Sequential Patterns. SPADE utilizes combinatorial properties to decompose the original problem into smaller sub-problems that can be independently solved in main-memory using efficient lattice search techniques, and using simple join operations. All sequences are discovered in only three database scans. Experiments show that SPADE outperforms the best previous algorithm by a factor of two, and by an order of magnitude with some pre-processed data. It also has linear scalability with respect to the number of input-sequences, and a number of other database parameters. Unlike previous approaches which make multiple database scans and use complex hash-tree structures that tend to have sub-optimal locality, SPADE decomposes the original problem into smaller sub-problems using equivalence classes on frequent sequences. Not only can each equivalence class be solved independently, but it is also very likely that it can be processed in main-memory. Thus SPADE usually makes only three database scans—one for frequent 1-sequences, another for frequent 2-sequences, and one more for generating all other frequent sequences. If the supports of 2-sequences are available then only one scan is required. SPADE uses only simple temporal join operations, and is thus ideally suited for direct integration with a DBMS.

An algorithm SPAM (Sequential PAttern Mining) for mining sequential patterns was proposed [10]. This algorithm is especially efficient when the sequential patterns in the database are very long. This algorithm introduces a novel depth-first search strategy that integrates a depth-first traversal of the search space with effective pruning mechanisms. A salient feature of our algorithm is that it incrementally outputs new frequent itemsets in an online fashion. SPAM assumes that the entire database (and all data structures used for the algorithm) completely fit into main memory. With the size of cur-rent main memories reaching gigabytes and growing, many moderate-sized to large databases will soon become completely memory-resident. SPAM is the first depth-first search strategy for mining sequential patterns. An additional salient feature of SPAM is its property of online outputting sequential patterns of different length compare this to a breadth-first search strategy that first outputs all patterns of length one, then all patterns of length two, and so on. This algorithm utilizes a depth-first traversal of the search space combined with a vertical bitmap representation to store each sequence. Experimental results demonstrated that our algorithm out-performs SPADE and PrefixSpan on large datasets by over an order of magnitude.

3. Problem Formulation of DNA Data Binding

In the process of review we found that some discovery problem related to the DNA-protein data mining. These problem are affected the performance and accuracy of mining of unknown pattern of TF and TFBS in gene regulation. The increasing size and length, decrease the accuracy and performance of rule mining. The problems mentioned in [4, 6, 9, 10] are Consequence of motif discovery, Instant based motif discovery, Size and length of unknown motif, Noise in biological data and Error Correcting Code.

4. Conclusion and Future Work

In this paper we present survey of DNA-protein data binding with data mining approach. Here we discuss some problem related to DNA-protein data and also discuss their minimization technique. Some problem such as length and size of motif degraded the performance of pattern evaluation process through rule mining technique. We also discuss some model based approach for finding valid motif in biological data. The main contributions of this paper as bringing this formulated the problem related to motif discovery process, the database community by showing the feasibility of applying association rule mining to predict approximate protein-DNA binding cores. We hope that future works in this community can further improve the algorithm so that eventually scientists can benefit from this computational approach.

References

[1] Shandar Ahmad, M. Michael Gromiha and Akinori Sarai1" Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information" Vol. 20 no. 4 2004, pp477–486.

[2] Ponraj Prabakaran, Shandar Ahmad and M. Michael Gromiha" Classification of Protein-DNA Complexes Based on Structural Descriptors" in IEEE 2010.

[3] Tak-Ming Chan , Kwong-Sak Leung , Kin-Hong Lee , Man-Hon Wong , Terrence Chi-Kong Lau and Stephen Kwok-Wing Tsui "Supplementary Materials for Subtypes of Associated Protein-DNA Patterns" 2006.

[4] Andrew D. Smith, Pavel Sumazin, Debopriya Das, and Michael Q. Zhang "Mining ChIP-chip data for transcription factor and cofactor binding sites", Vol. 21 Suppl. 1 2005.

[5] Ziliang Qian, Yu-Dong Cai and Yixue Li" Automatic transcription factor classifier based on functional domain composition" in Biochemical and Biophysical Research Communications 347 (2006) 141–144.

[6] Ramesh C. Agarwal, Charu C. Aggarwal And V.V.V. Prasad "A Tree Projection Algorithm For Generation Of Frequent Item Sets," Journal of Parellel And Distributed Computing (2001), Vol. 61, No. 3, Pages 350-371.

[7] Junqiang Liu, Yunhe Pan Ke and Wang Jiawei Han "Mining Frequent Item Sets by Opportunistic Projection" in SIGKDD '02, July 23-26, 2002.

[8] Jian Pei Jiawei, Han Behzad Mortazavi,Asl Helen Pinto" PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth" in IEEE2010.

[9] Mohammed J. Zaki "Spade: An Efficient Algorithm For Mining Frequent Sequences" In Kluwer Academic Publishers. Manufactured In The Netherlands, Machine Learning, 42, 31–60, 2001.

[10] Jay Ayres, Johannes Gehrke, Tomi Yiu, and Jason Flannick" Sequential PAttern Mining using A Bitmap Representation" in IEEE, 2009.

[11] Ke Wang, Yabo Xu and Jeffrey Xu Yu "Scalable Sequential Pattern Mining for Biological Sequences" in ACM 2004.

[12] Y. S. Ong and A. Keane, "Meta-Lamarckian learning in memetic algorithms," IEEE Trans. Evol. Computat., vol. 8, no. 2, pp. 99–110, Apr. 2004.

[13] J. Smith, "Coevolving memetic algorithms: A review and progress report," IEEE Trans. Syst., Man, Cybern., Part B: Cybern., vol. 37, no. 1, pp. 6–17, Feb. 2007.

[14] O. J. Mengshoel and D. E. Goldberg, "The crowding approach to niching in genetic algorithms," Evol. Comput., vol. 16, pp. 315–354, Sep. 2008.

[15] T.-M. Chan, K.-C. Wong, K.-H. Lee, M.-H. Wong, C.K. Lau, S. K. Tsui, and K.-S. Leung, "Discovering approximate associated sequence patterns for protein-DNA interactions," vol. 27, no. 4, pp. 471–478, Feb. 2011.

[16] K.-S. Leung, K.-C. Wong, T.-M. Chan, M.-H. Wong, K.-H. Lee, C.-K. Lau, and S. K. W. Tsui, "Discovering protein-DNA binding sequence patterns using association rule mining," Nucleic Acids Res., vol. 38, no. 19, pp. 6324–6337, Oct. 2010.