International Journal of Engineering, Management, Humanities and Social Sciences Paradigms (IJEMHS)
(Volume 28, Issue 03) Publishing Month: August 2017
An Indexed and Referred Journal with Impact Factor: 2.75
ISSN: 2347-601X
www.ijemhs.com

Frequent Periodic Cryptic Sequence Mining in Biological Data: Experimental Analysis

[1]Sulochana Nanda
Gandhi Institute of Excellent Technocrats, Bhubaneswar, India
[2]Sanket Swarup Mohanty
Sanjay Memorial Institute of Technology, Berhampur, Odisha, India

Abstract

Amino acid sequences are known to constantly mutate and diverge unless there is a limiting condition that makes such a change deleterious. 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 Constraint Based Frequent Motif Mining (CBFMM). CBFMM is a flexible Frequent Pattern-tree-based algorithm that can be used to find frequent patterns with a variety of definitions of motif (pattern) models. They can play an active role in protein and nucleotide pattern mining, which ensure in identification of potentiating malfunction and disease. Therefore, insights into any aspect of the repeats – be it structure, function or evolution – would prove to be of some importance. This study aims to address the relationship between protein sequence and its three- dimensional structure, by examining if large cryptic sequence repeats have the same structure. We have tested the proposed algorithm on biological domains. The conducted comparative study demonstrates the applicability and effectiveness of the proposed algorithm.

***Keywords*:** *Motif, FP mining, FP tree, sequence mining, Repetition detection, data mining.*

I. INTRODUCTION

The approximate subsequence mining problem is of particular importance in computational biology, where the challenge is to detect short sequences, usually of length 6- 15, that occur frequently in a given set of DNA or protein sequences. These short sequences can provide clues regarding the locations of so called "regulatory regions," which are important repeated patterns along the biological sequence. The repeated occurrences of these short sequences are not always identical, and some copies of these sequences may differ from others in a few positions. A repeat is defined as two or more contiguous segments of amino acid (three or more) residues with identical and similar sequence. When such repeats are in high-complexity regions, they are called 'cryptic' [9]. Although low-complexity repeats are essential for evolutionary analysis and comprise a large section of the eukaryotic genome, high-complexity repeats are usually associated with a particular structure or function. This study considers large cryptic repeats comprising eight or more residues, as [26] fixed the length of a moderate-sized repeat as being between five and eight amino acids. The study of repeats is crucial because all but 5–6% of the high eukaryotic genome is repetitive [25]. Internal protein repeats are observed to be associated with structural motifs or domains. It is evolutionarily more 'economical' to evolve complex structures such as multiple domains by using 'modular plug-ins' [22] to fulfill a specific function. Furthermore, longer repeats normally act to enhance the stability of the native fold of the protein and, while small repeats interact with each other, larger repeats may either interact or remain isolated like beads on a string [22]. Three prominent reviews on repeats are those of [22], [11] and [33], and they concentrate on the relationship between structural repeats and their primary structure along with the characteristics of protein families. In [33] discuss the evolution of repeats as modules in the proteins. It is mentioned that

the number of repeats in a protein can vary between proteins, implying that the loss or gain of repeats is very rapid in evolution.

The remainder of the paper is organized as follows: Section 2 presents related works and Section 3 describes our model. In section 4, we present optimization strategy for our model and in Section 5 contains our experimental results. Section 6 contains our conclusions.

II. PREVIOUS WORK

There is a vast amount of literature on mining databases for frequent pattern [30], [17], [47]. The problem of mining for subsequence was introduced in [29]. Subsequence mining has several applications, and many algorithms like [23], [48], and [36] have been proposed to find patterns in the presence of noise. However, they primarily focus on subsequence mining, while we focus on contiguous patterns. A host of techniques have been developed have been developed to find sequence in a time series database that are similar to a given query sequence [29], [3], [31], [49]. The existing algorithm [5], [14], [20], [42], [24] requires the user to specify the repetition and patterns occurring with that repetition, otherwise which look for all possible repetitions in the time series. Some algorithms are classified based on the detection type of repetition for symbol, sequence or segment. Another algorithm that finds frequent trends in time series data was proposed in [1]. However, this algorithm is also limited to a simple mismatch based noise model. In addition, this is a probabilistic algorithm, and is not always guaranteed to find all existing patterns. The algorithms specified in [34], [35], [39], [13], looks for all possible repetitions by considering the range. COVN [34] fails to perform well when the time series contains insertion and deletion noise. *WARP* [35] can detect segment repetition; it cannot find symbol or sequence repetition. Sheng et al. [6], [8] developed algorithm based on *ParPer* [21] to detect repeated patterns in a section of the time series; their algorithm requires the user to provide the expected repetition value. COVN, *WARP* and *ParPer* are augmented to look for all possible repetitions, and which last till the very end of the time series. Cheung [7] used FP tree similar to *STNR* [13] which is not beneficial in terms of growth of tree. Huang and Chang [28] and *STNR* [13] presented their algorithm for finding repeated patterns, with allowable range along the time axis. Both finds all type of repetition by utilizing the time tolerance window and could function when noise is present. *STNR* [13] can detect patterns which are repeated only in a subsection of the time series. Repeated check in *STNR* last for all the positions of a particular pattern, which in our algorithm is been reduced.

Several approaches described in the literature handle structured motif extraction problem [3], [2] and repetition among subsection of the time series. However, our approach described in this paper is capable of handling both motif extraction and reporting all type of repetition. In this paper, we present a flexible algorithm that handles general extended structured motif extraction problem and uses *CBFMM* to build Consensus tree. *CBFMM* is capable of reporting all types of repetitions with or without the presence of noise in the data up to a certain level. We believe that this is an interesting problem since it allows mining for useful motif patterns with all type of repetition, without requiring specific knowledge about the characteristics of the resulting motif. In this paper, we present a new model that is very general and applicable in many emerging applications. We demonstrate the power and flexibility of this model by applying it to data sets from several real applications. We describe a novel motif mining algorithm called *CBFMM* that uses a concurrent traversal of FP trees to efficiently explore the space of all motifs. We present a comparison of *CBFMM* with several existing algorithms (COVN [34], *WARP* [35], *STNR* [13], *ParPer*[21]). *CBFMM* never misses any matches (as opposed to some of these methods that apply heuristics). In fact, we show that *CBFMM* is able to identify many true biological motifs that existing algorithms miss. We show that our algorithm is scalable, accurate, and often faster than existing methods by more than an order of magnitude. We present an algorithm that uses *CBFMM* as a building block and can mine combinations of simple approximate motifs under relaxed constraints.

III. CONCLUSION

In this paper, we have presented a novel algorithm that uses FP tree as underlying structure. The algorithm can detect symbol, sequence and segment repetition as well as present the patterns that are repeated. It can also find repetition within a subsection of the biological data. It can detect the redundant repetitions which are pruned; before calculating confidence which in turn saves a significant amount of time. We took an initial step towards and understanding the constraints in the conservation of amino acid sequences by analyzing large cryptic identical and similar repeats. CBFMM is also superior to motif finding algorithms used in computational biology. We also presented experiments which show that CBFMM can scale to handle motif mining tasks that are much larger than attempted before. Our algorithm runs in *O (k. N)* in the worst case. In future, we are trying to extend our algorithm's working on online repetition detection. The algorithm to be experimented with streaming data using disk based tree [25].

IV. REFERENCES

[1] A. Udechukwu, K. Barker and R. Alhajj, Maintaining Knowledge-Bases of Navigational Patterns from Streams of Navigational Sequences, RIDE, 2005,37-44.
[2] A.A. Ptitsyn, Computational analysis of gene expression space associated with metastatic cancer, BMC Bioinformatics (BMCBI), 10(S-11), 2009, 6.
[3] A. W.C. Fu, J. Li and P. Fahey, Efficient discovery of risk patterns in medical data, Artificial Intelligence in Medicine, 45(1), 2009, 77-89.
[4] B. Balamurugan et al., PSAP: protein structure analysis package, J. Appl. Crystalogr., 2007, 40773–777.
[5] C. Berberidis and G. Tzanis, Mining for Mutually Exclusive Items in Transaction Databases, Database Technologies: Concepts, Methodologies, Tools, and Applications, 2009, 2192-2203
[6] C. Sheng, W. Hsu, M. L. Lee, J. C. Tong and S. K. Ng, Mining mutation chains in biological sequences, ICDE, 2010. 473-484
[7] C.F. Cheung, J.X. Yu and H. Lu, Constructing FP Tree for Gigabyte Sequences with Megabyte Memory, IEEE Trans. Knowledge and Data Engg., 17(1), 2005, 90-105.
[8] C.Sheng, W. Hsu and M.L. Lee, Mining Dense Periodic Patterns in Time Series Data, In Proceedings of 22nd IEEE Int'l Conf. Data Eng. (ICDE), Atlanta, Georgia, USA, 2006, 115.
[9] D. Tantz, D. Trick and G.A. Dover, Cryptic simplicity in DNA is a major source of genetic variation, Nature (London), 1986, 322652–656.
[10] E. Eskin, N. A. Furlotte, H. M. Kang and C. Ye, Mixed-model coexpression: calculating gene coexpression while accounting for expression heterogeneity, Bioinformatics, 27(13),2011, 288-294.
[11] E. M. Marcotte and Y. Park, Revisiting the negative example sampling problem for predicting protein-protein interactions, Bioinformatics, 27(21). 2011, 3024-3028.
[12] F. Rasheed and R. Alhajj, Using FP Trees for Periodicity Detection in Time Series Databases, In Proceedings of IEEE Int'l Conf. Intelligent Systems, September 2008.
[13] F.Rasheed, M. Al-Shalalfa and R. Alhajj, Efficient Periodicity Mining in Time Series Databases Using FP Trees, IEEE Trans. Knowl. Data Engg. (TKDE), 23(1), 2011, 79-94.
[14] Fei Chen, Jie Yuan and Fusheng Yu, Finding periodicity in pseudo periodic time series and forecasting, GrC 2006, 2006, 534-537.
[15] G. Das , C. A. Ratanamahatana, J. Lin, D. Gunopulos, E. J. Keogh and M. Vlachos, Mining Time Series Data, Data Mining and Knowledge Discovery Handbook, 2010,1049-1077
[16] G. Pavesi, P. Mereghetti, G. Mauri and G. Pesole, Weeder Web: Discovery of Transcription Factor Binding Sites in a Set of Sequences From Co-Regulated Gene, Nucleic Acids Research, 32(2004), W199-W203.
[17] G.K. Sandve and F. Drablos, A Survey of Motif Discovery Methods in an Integrated Framework, Biology Direct, 1, 2006, 11-26.
[18] H. Wu, B. Salzberg, G.C. Sharp, S.B. Jiang, H. Shirato, and D. Kaeli, Subsequence Matching on Structured Time Series Data, Proc. ACM SIGMOD, 2005, 682-693.
[19] J. Buhler and M. Tompa, Finding Motifs Using Random Projections, J. Computational Biology, 9(2), 2002, 225-242.
[20] J. Han, W. Gong and Y. Yin, Mining Segment-Wise Periodic Patterns in Time Related Databases, In Proc. of ACM Int'l Conf. Knowledge Discovery and Data Mining, New York City, New York, USA, 1998, 214-218.
[21] J. Han, Y. Yin and G. Dong, Efficient Mining of Partial Periodic Patterns in Time Series Database, In Proc. of 15th IEEE International Conference in Data Engineering, Sydney, Australia, 1999, 106.

[22] J. Heringa and Bernd W. Brandt, Protein analysis tools and services at IBIVU, Journal Integrative Bioinformatics (JIB), 8(2), 2011.
[23] J. Wang and J. Han, BIDE: Efficient Mining of Frequent Closed Sequences, In Proceedings of 20th IEEE Int'l Conf. Data Eng. (ICDE), 2004,79-90.
[24] J. Yang, W. Wang and P. Yu, InfoMiner: Mining Partial Periodic Patterns with Gap Penalties, In Proceedings of Second IEEE Int'l Conf. Data Mining, Maebashi City, Japan, 2002.
[25] J.M. Hancock and M. Simon, Simple sequence repeats in proteins and their significance for network evolution, Gene, 2005, 345113–118.
[26] J.M. Hancock, E.A. Worthey and M.F. Santibanez-Koref, A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in human and mice, Mol. Biol. Evol., 2001, 181014–1023.
[27] K. Sumathi, P. Ananthalakshmi, M.N.A.M. Roshan and K. Sekar, 3dss: 3-dimensional structural superposition, Nucleic Acids Res., 2006, 34W128–W134.
[28] K.Y. Huang and C.H. Chang, SMCA: A General Model for Mining Asynchronous Periodic Patterns in Temporal Databases, IEEE Trans. Knowledge and Data Engg., June 17(6), 2005, 774-785.
[29] L. Chen, M. Tamer Ozsu and V. Oria, Robust and Fast Similarity Search for Moving Object Trajectories, In Proceedings of ACM SIGMOD, 2005,491-502.
[30] M. Das and H.K. Dai, A Survey of DNA Motif Finding Algorithms, BMC Bioinformatics, 8, 2007, S21-S33.
[31] M. Vlachos, G. Kollios and D. Gunopulos, Discovering Similar Multidimensional Trajectories, In Proceedings of 18th IEEE Int'l Conf. Data Eng. (ICDE), San Jose, California, USA, 2002, 673-684.
[32] M.A. Andrade and P. Bork, Heat repeats in the Huntington's disease protein, Nat. Genet, 1995, 11115–116.
[33] M.A. Andrade, C. Perez-Iratxeta and C.P. Ponting, Protein repeats: structure, functions and evolution, J. Struct. Biol., 2001, 134117–131.
[34] M.G. Elfeky, W.G. Aref and A.K. Elmagarmid, Periodicity Detection in Time Series Databases, IEEE Trans. Knowledge and Data Eng, 17(7), 2005,875-887.
[35] M.G. Elfeky, W.G. Aref and A.K. Elmagarmid, WARP: Time WARPing for Periodicity Detection, In Proceedings of Fifth IEEE Int'l Conf. Data Mining, Houston, Texas, USA, November 2005.
[36] M.J. Zaki, SPADE: An Efficient Algorithm for Mining Frequent Sequences, Machine Learning, 42(1), 2001, 31-60.
[37] M.V. Katti, R. Sami-subbu, P.K. Rajekar and V.S. Gupta, Amino Acid Repeat Patterns in Protein Sequences: Their Diversity and Structural-Function Implications, Protein Science, 9(6), 2000, 1203-1209.
[38] P. Djian, Evolution of simple repeats in DNA and their relation to human disease, Cell, 1998, 94155–160.
[39] P. Indyk, N. Koudas and S. Muthukrishnan, Identifying Representative Trends in Massive Time Series Data Sets Using Sketches, In Proceedings of Int'l Conf. Very Large Data Bases, Cairo, Egypt, 2000.
[40] P.Patel, E. Koegh, J. Lin and S. Lonardi, Mining Motifs in Massive Time Series Databases, Proc. IEEE Int'l Conf. Data Mining (ICDM), Maebashi City, Japan, 2002, 370-377.
[41] S. Hoppner, Discovery of Temporal Patterns - Learning Rules about the Qualitative Behaviour of Time Series, In Proc. Fifth European Conf. Principle and Practice of Knowledge Discovery in Databases, Freiburg, Germany, 2001, 192-203.
[42] S. Ma and J. Hellerstein, Mining Partially Periodic Event Patterns with Unknown Periods, In Proceedings of 17th IEEE Int'l Conf. Data Engg., Heidelberg, Germany, 2001.
[43] S. Papadimitriou, A. Brockwell and C. Faloutsos, Adaptive, Hands Off-Stream Mining, In Proceedings of Int'l Conf. Very Large Data Bases (VLDB), Berlin, Germany, 2003, 560-571.
[44] S.F. Altschul, T.L. Madden, A.A. Schaer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res., 1997, 253389–253402.
[45] S.Sinha and M. Tompa, YMF: A Program for Discovery of Novel Transcription Factor Binding Sites by Statistical Over representation, Nucleic Acids Research, 31(13), 2003, 3586- 3588.
[46] T.F. Smith, C.G. Gaitatzes, K. Saxena and E.J. Neer, The WD-repeat: a common architecture for diverse functions, Trends Biochem. Sci., 1999,24181–185.
[47] W. Wang and J. Yang, Mining Sequential Patterns from Large Datasets. Springer-Verlag, 28, 2005.
[48] X.Yan, J. Han and R. Afshar, CloSpan: Mining Closed Sequential Patterns in Large Datasets, In Proceedings of SIAM Int'l Conf. Data Mining (SDM), San Francisco, CA, USA, 2003.
[49] Y. Zhu and D. Shasha, WARPing Indexes with Envelope Transforms for Query by Humming, In Proceedings of ACM SIGMOD, 2003, 181-192.
[50] Y.C. Liou, A. Tocilj, P.L. Davies and Z. Jia, Mimicry of ice structure by surface hydroxyls and water of a beta-helix antifreeze protein, Nature (London), 2000, 406322–324.