FART Neural Network based Probabilistic Motif Discovery in Unaligned Biological Sequences

M. Hemalatha, P. Ranjit Jeba Thangaiah and K. Vivekanandan, Member IEEE

***Abstract* – Finding Motif in bio-sequences is the most primitive operation in computational Biology. There are many computational requirements for a motif discovery algorithm such as computer memory space requirement and computational complexity. To overcome the complexity of motif discovery, an alternative solution is proposed by integrating genetic algorithm and Fuzzy Art machine learning approaches for eliminating multiple sequence alignment process. *Problem statement:* More than hundred methods have been proposed for motif discovery in recent years, representing very large variation with respect to both algorithmic approaches as well as the underlying models of regulatory regions. The aim of this study is to develop an alternative solution for motif discovery, which benefits from both data mining and genetic algorithm, and which at the same time eliminates the cost caused by use of multiple sequence alignment. *Approach:* Genetic algorithm based probabilistic Motif discovery model is designed to solve the problem. The proposed algorithm is implemented using Matlab and also tested with large DNA sequence data sets and synthetic data sets. *Results:* Results obtained by the proposed model to find the motif in terms of speed and length are compared with the existing method. This proposed method finds length of 11 in 18 sec and length of 15 in 24 sec but the existing methods finds length of 11 in 34 sec. When compared with other techniques the proposed method outperforms the popular existing method. *Conclusion:* In this study, a model is proposed to discover motif in large set of unaligned sequences in considerably minimum time. Length of motif in this study is also long when compared with existing methods. The proposed algorithm is implemented using Matlab and is tested with large DNA sequence data sets and synthetic data sets.**

***Index Terms* - Bioinformatics, Genetic algorithm, Motif, DNA sequence, Multiple unaligned**

I. Introduction

The rapid advances in computing, data management systems and intelligent processing have contributed to the growth in the domain of bio-informatics. The representation and management of biological data has been made easier now. Many dimensions of processing from symbolic to machine learning are possible now. Motif discovery is one of those benefits of biological data, and naturally it is amongst fashionable bioinformatics topics. Motif discovery basically can be described as follows: for a given sample of sequences one can find the unknown pattern that is implanted in different positions of the given sequences [1].

Manuscript received December 13, 2008.

M. Hemalatha is with Karpagam University, Coinbatore, Tamil Nadu – 641 021, India (hema.bioinf@gmail.com).

P. Ranjit Jeba Thangaiah is with Karunya University, currently pursuing his Research at Bharathiar University, Coimbatore, Tamil Nadu – 641 046, India (thangaiah@gmail.com).

K. Vivekanandan is with Bharathiar University, Coimbatore, Tamil Nadu – 641 046, India (vivekbsmed@gmail.com).

Importance of these patterns for biology comes from the role of motifs at protein DNA binding sites. Furthermore, finding similar sequences can be used to reveal unknown evolutionary relationships between different species.

II. Materials and Methods

Lot of research is carried out to discover solutions for motif discovery. Many algorithms have developed to improve the existing popular motif discovery tools by means of performance, length of motifs and/or some other considerations. Stine *et al*. [5] employed genetic algorithm in their structured Genetic Algorithm (St-GA) to search and to discover highly conserved motifs amongst upstream sequences of co-regulated genes. Liu *et al*. [6] also employed genetic algorithm for finding potential motifs in the regions of Transcription Start Site (TSS). Pan *et al*. [7] developed MacosFSpan and MacosVSpan algorithms to mine maximal frequent sequences in biological data. While MacosFSpan and MacosVSpan underline inefficiency of apriori-like algorithms, and seeks a mining solution that works better in biological datasets [6]-[7], and combine genetic algorithm approach with multiple sequence alignment tools to discover motifs. St-GA [5] also works in similar fashion and needs to make multiple sequence alignment. Among those existing works; most recognized ones are: The Multiple Em for Motif Elicitation (MEME) system [2]-[4],[9], proposed a topdown mining method called as ToMMS, which is a promising approach for mining long sequential patterns. Classical mining methods use bottom-up strategy, and step by step go to the largest frequent itemset after finding shorter frequent ones first. On the contrary, top-down strategy starts with a predetermined length and from this large starting point it goes down to search smaller ones until finding a frequent one, then clearly that found one becomes the largest frequent itemset. The only weak point of topdown strategy is specifying its starting point which requires user knowledge. Baloglu and Kaya [12] proposed a GA-based topdown data mining approach for finding motifs in biosequences. It has combined a genetic algorithm and top-down data mining method. However, one of the motivations of motif discovery is to find bigger motifs since finding small ones has no use.

The aim of this study is to develop an alternative solution for motif discovery, which benefits from both data mining and genetic algorithm, and at the same time eliminates the cost caused by use of multiple sequence alignment. The computational cost of multiple sequence alignment is also emphasized in [5] which suffers from use of time consuming BLAST [8]. A combination of machine learning approach and genetic algorithm is not time consuming is emphasized using this study. This study not only considers the computational cost of alignment and how to eliminate, but also tried to find the most efficient way to handle mining part. However, one of the motivations of motif discovery is

to find bigger motifs since finding small ones has no use. This condition gives a meaning to design a Hybrid model for motif discovery. In this study, a hybrid model is used of GA with Fuzzy Art for motif discovery.

The solution of this research work is based on a combination of genetic algorithm and Fuzzy Art. It is used to discover motifs in biological sequence datasets. There are two main motivations for this approach. First, genetic algorithm is used to find the possible motifs. This is done by selecting two or more potentially matching motif regions M1, M2...Mn of length 'W' in one or more gene sequence using Genetic algorithm. Second, training FART Neural Network to recognize the 'n' previously found potential motifs M1, M2...Mn as 'n' different classes. Classify all the possible segments of window length 'W' of the sequences using trained FART Neural Network. Group the detected motifs into 'n' groups based on the class label. Finally it will have 'n' sets of potential Motif in the sequence. Change the Expected length of Motif continue the search if necessary.

The proposed GA based motif discovery algorithm:
Step 1: The initial population is two sets of string represented by binary numbers. The selection is made randomly, which contains a bit string which represents the size and the location of two or more sub sequences for example $P = p_1, p_1, \cdots, p_s$. $Q_1 = q_1, q_1, \cdots, q_s$ where, s = the size of the population. The two sets of dissimilar locations in the sequence G = pointed by the two sets of strings.

Step 2: Evaluation: After the generation of population is formed, the initial step is to compute the fitness value of each member in the population $P = p_1, p_1, \cdots, p_s$. $Q_1 = q_1, q_1, \cdots, q_s$. That is, the fitness of each corresponding subsequence depends on the similarity of the corresponding subsequence pairs. The fitness evaluation process for a chromosome involves the following steps:

- Converting the chromosome's genotype to its phenotype. This refers to the conversion of the binary string into corresponding real values
- The objective function is evaluated
- The value of objective function is converted into fitness. In this case, to generate positive fitness values, the fitness of each chromosome is calculated as the difference in values of the maximized objective function and the objective function evaluated for every chromosome in the population as given in (1).

$$F\_total = \sum_{i=1}^{pop\_size} Eval(V_i) \quad (1)$$

Step 3: Objective function values and fitness: The objective function values F and the fitness values Eval of above chromosomes (the first population) are computed. Here, the fitness function is nothing but a suitable gene subsequence matching policy such as hamming distance or more sophisticated score matrix based distance measurement algorithm.

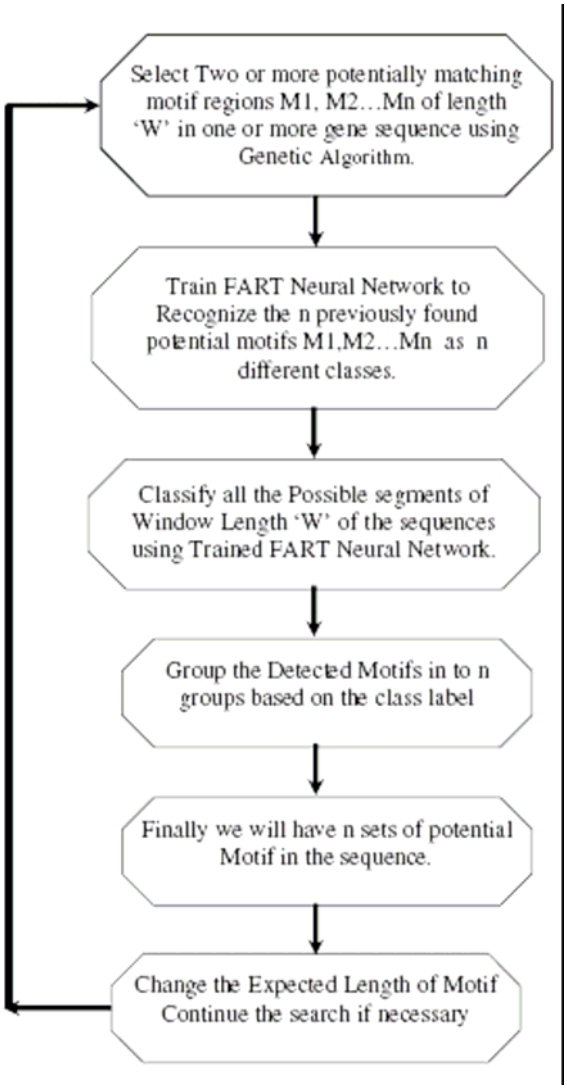

Fig. 1: Block diagram of GA based FART Model

Step 4: Create a new population: After the process of evaluation, a new population should be created from the current generation. In this case the three operators (reproduction, crossover, and mutation) are employed. The size of the population is fixed with regards to the convergence factors. This process also considers previously selected potential motifs.

Reproduction: The two chromosomes (strings) having the best fitness and the second best fitness are permitted to live and produce offspring in the next generation. The first two-best matching sub sequence pairs are selected as new parents.

The one-cut-point method of crossover is implemented in this case. In this method one cut-point is selected randomly and the right parts of two parents are inter-changed to produce the offspring. The selection of the crossover point can be performed in a selective manner considering the convergence factors.

Step 5: Mutation: After the crossover, mutation process is performed. The convergence factor is considered for the selection of the mutation level. In this process one or more genes are altered with a probability equal to the mutation rate.

- A sequence of random numbers r_k is generated. (In this case, this is the number of bits in the whole population).
- In case r_i is 1, the i^{th} bit in the whole population is altered from 1-0 or from 0-1.
- The chromosomes reproduced are not subjected to mutation, so after the mutation process, the chromosomes should be restored.

The output for a single iteration of the genetic algorithm is the creation of a new population. Proceed to Step 2 and continue the process.

The procedure (iterations) can be repeated for any number of times as desired. The best value of the objective function for the population of every generation is computed. The whole process is repeated for the desired number of times.

Ultimately, two final set of potential motifs $P = p_1, p_1, \cdots, p_s$. $Q_1 = q_1, q_1, \cdots, q_s$ are obtained. By the use of any one of these two sets other similar patterns in the sequence G can be identified using the sliding window operation. As this operation involves only one pass and matching of S subsequence at all possible window positions (presuming uniform length of motifs in P_1 or P_2), clearly, this technique will consume lesser time when compared with the other brute-force approaches of motif discovery. Fig. 1 explains the proposed model for motif discovery using genetic algorithm. By applying this method a set of potential motifs in a sequence is detected. Using the detected motifs as seed all the similar patterns in overall sequence can be found by sliding window operation.

III. Results

A dataset of 300 E. coli promoter sequences is used for the experiments. This dataset is previously used in Baloglu, U.B., Kaya, M [12].

The proposed GA based motif discovery model (described in the above diagram) has been implemented using Matlab on windows XP on a normal desktop PC (Intel Pentium 2G.Hz, 512 MB RAM). The built in toolbox in Matlab is not utilized for the customization purposes, instead a custom model for genetic algorithms is developed to solve the motif discovery problem. The developed system effectively detected potential motifs in a remarkably minimum period of time.

The optimum parameters to enhance the system performance are found out by altering the GA parameters on trial and error basis. Real time data sets are used to test the proposed model. The proposed research makes use of the gene sequence of E. coli (EcoliPromoters1_300.seq) as is used in the research works of [12]. The length of each sequence present in the dataset is 100 bases. The data set consists of 300 sequences in all. The sum of lengths of all the sequences in the dataset is 30300. The system is programmed so as to discover 5 motifs. The following GA parameters are assumed.

The GA Parameters:

The Total Population Size: 100
The Total Number of Generations: 20
The Mutation Level: 0.2
The Crossover Rate: 0.20

Table I Illustrates the performance of the GA based method to find the Motif in terms of time and length:

Table I: Performance of GA to find the Motif in terms of time and length

Sl. No	Motif Length	Time Taken for GA to find 5 Motifs (sec)
1	7	13.76
2	9	15.39
3	11	18.34
4	13	20.36
5	15	24.26

Fig. 2 shows the performance of GA based method in terms of Time and Motif length is presented below. It is evident that the Top-Down GA method outperforms the basic motif discovery methods such as the MEME and the Gibbs algorithm. The Fig. 3 depicts the same.

Table II and Fig. 4 shows the performance results by the various Genetic Algorithm based techniques:

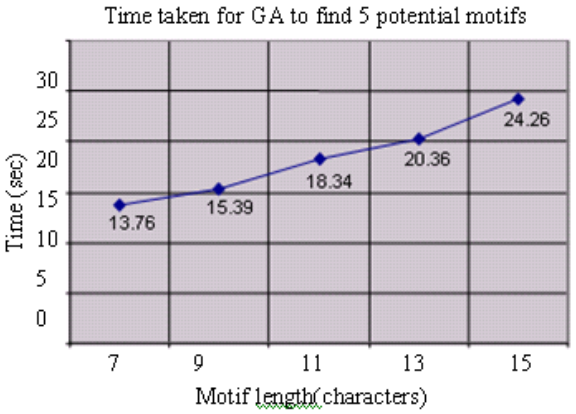

Fig. 2: Performance of GA in terms of Time and Motif length

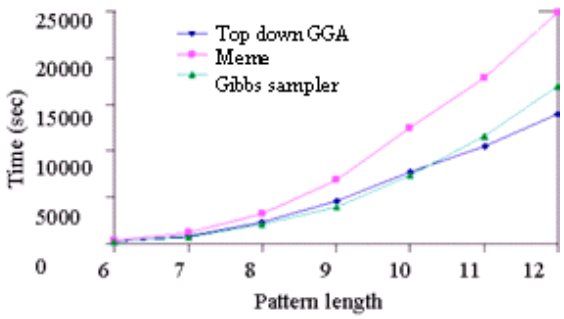

Fig. 3: Chart showing the comparative results of Top Down GGA, MEME and Gibbs Sampler

Table II: Performance of GA Based FART in terms of Length and Time compared with other Methods

Sl. No	Motif Length	Time Taken for Finding All Other Matches		
		GA Based Exhaustive Search	GA Based FART Method	Top-Down Based GA
1	6	333.37	93.00	199
2	7	349.43	93.88	793
3	8	373.59	96.62	1937
4	9	376.35	98.82	3723
5	10	391.12	101.82	7102
7	11	406.68	104.15	10471
8	12	430.50	106.71	18902

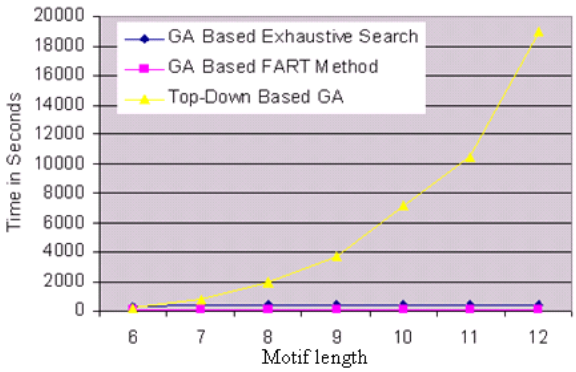

Fig. 4: Comparison between proposed and existing method

IV. Discussion

The proposed GA based Fuzzy ART method is compared with the implemented GA based exhaustive search method and the Top-Down GA method to prove that the proposed motif discovery algorithm outperforms the existing techniques.

From the above Table 2 and Fig. 4, it is obvious that the proposed GA based Fuzzy Art method outperforms the GA based exhaustive search and Top-Down based GA methods.

In order to overcome the complexity of motif discovery an integration of genetic algorithm and Fuzzy Art mining approach is proposed which eliminates multiple sequence alignment process. From the experimental results, it can be inferred that the proposed method of combination of genetic algorithm and fuzzy art mining outperforms other renown motif discovery algorithms, such as MEME and Gibbs Sampler and Genetic algorithm.

The results thus obtained are promising. The proposed model yields improved performance over the brute force approaches. 5 likely motifs are detected within a minute's time using the proposed model from a sequence of length 30300. The same model can be applied to detect motifs in any sequence apart from gene sequence such as a time series data. Thus this research does not focus on the biological significance of the detected motifs. Focus on the biological significance of the motifs can be developed in the future to address this issue.

A comparative study on the time and length for finding the motifs and performance is made on popular methods such as [3], [7], [12], along with GA based exhaustive search and the proposed GA based FART is done. The results from the GA based FART method outperformed the others by a considerable margin. The overall result including the factor of speed and length of finding motif by the use of the proposed method is found to be satisfactory. Even though the proposed model discovers a given number of N motifs of length L, the issue of discovering the total number of motifs of all possible lengths remains unaddressed and can be considered as a scope of enhancement in future.

V. Conclusion

To find the motif of DNA sequence, the proposed GA based model has been successfully designed and implemented on MATLAB under Windows operating system using normal desktop computers. The performance of the proposed model is tested with the very large synthetic numeric data sets and DNA sequence data sets. Several tests are made on the model and overall significant results are achieved. While considering existing approaches, the performance of the proposed model is very much appreciating. In this 30300-character long gene sequence, it has detected 5 probabilistic motifs in less than a min. The proposed model has discovered only a given number of N motifs of length L each. But, still there are lots of issues such as finding total number of motifs of all possible lengths can be addressed in future. The same model can be applied to detect motifs in any sequence apart from gene sequence such as a time series data. Thus in this research the biological significance of the detected motifs is not mentioned. Future research works in biological domain which will be very much particular about the biological significance of motifs can address these types of issues.

Acknowledgment

The author acknowledge the management of Karpagam University for the support and motivation to fulfill the project.

References

[1] U. Keich and P. A. Pevzner, "*Finding motifs in the twilight zone. Bioinformatics*", DOI:10.1093/bioinformatics/18.10.1374, 18(10), 2002, pp. 1374-1381.

[2] T. L. Bailey and C. Elkan, "*Fitting a mixture model by expectation maximization to discover motifs in biopolymers*", Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, California, 1994, pp. 28-36.

[3] W. Thompson, E. C. Rouchka and C. E. Lawrence, "*Gibbs recursive sampler: Finding transcription factor binding sites*", J. Nucl. Acids Res., Vol. 31, 2003, pp. 3580-3585. Available: http://nar.oxfordjournals.org/cgi/content/abstract/31/13/3580.

[4] G. Z. Hertz, G. W. Hartzell, and G. D. Stormo, "*Identification of consensus patterns in unaligned DNA sequences known to be functionally related*", Bioinformatics, 6(2), 1990, pp. 81-92. Available: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/6/2/81.

[5] M. Stine, D. Dasgupta and S. Mukatira, "*Motif discovery in upstream sequences of coordinately expressed genes*", The 2003 Congress on Evolutionary Computation, DOI:10.1109/CEC.2003.1299863, 2003, pp.1596-1603.

[6] Liu, F.F.M. Tsai, J.J.P. Chen, R.M. Chen, S.N. Shih, S.H.., "*FMGA: Finding Motifs by Genetic Algorithm*", Proceedings of the 4th IEEE Symposium on Bioinformatics and Bioengineering, 2004, pp. 459-466. Available: http://www2.computer.org/portal/web/csdl/doi/10.1109/BIBE.2004.1317378,

[7] Jin Pan; Peng Wang; Wei Wang; Baile Shi; Genxing Yang.,"*Efficient algorithms for mining maximal frequent concatenate sequences in biological datasets*", Proceedings of the 5th International Conference on Computer and Information Technology, DOI: 10.1109/CIT.2005.106, 2005, pp. 98-104.

[8] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "*Basic local alignment search tool*", J. Mol. Biol., Vol. 215, 1990, pp. 403-410.

[9] M. Ester and X. Zhang, "*A Top-Down Method for Mining Most Specific Frequent Patterns in Biological Sequence Data*", Proc. Siam int. Conf. on Data Mining, SDM, ISBN: 0898715687, 2004, pp. 537.

[10] M. Scherf, A. Klingenhoff and T. Werner, "*Highly specific localization of promoter regions in large genomic sequences by promoterInspector: A novel context analysis approach*", J. Mol. Biol., Vol. 297, DOI: 10.1006/jmbi.2000.3589, 2000, pp. 599-606.

[11] P. Horton, "*Tsukuba BB: A branch and bound algorithm for local multiple alignment of DNA and protein sequences*", J. Comput. Bio., Vol. 8, DOI:10.1089/10665270152530854, 2001, pp. 283-303.

[12] U. B. Baloglu and M. Kaya, "*Top-down motif discovery in biological sequence datasets by genetic algorithm*", International Conference on Hybrid Information Technology, 2(9):103 –107, DOI: 10.1109/ICHIT.2006.253597, 2006, pp. 103-107.