A Structured Evolutionary Algorithm for Identification of Transcription Factor Binding Sites in Unaligned DNA Sequences

Shripal Vijayvargiya,
Department of Computer Science,
Birla Institute of Technology
Jaipur Campus, Jaipur
Email : shripalvijay@rediffmail.com

Pratyoosh Shukla
Department of Biotechnology,
Birla Institute of Technology,
Mesra (Ranchi)
Email: pratyoosh.shukla@gmail.com

Abstract

Identification of Transcription Factor Binding Sites (TFBS) also called as motifs, from the upstream region of genes remains a highly important and unsolved problem particularly in higher eukaryotic genomes. In this paper, we propose an evolutionary approach to identify transcription factor binding sites. This approach is based on the structured genetic algorithm. In this approach an individual is represented as a structured tree that help us to find variable length motifs. A simple GA can find only fix length motif where as proposed method can find variable length motifs. We applied this approach on various data sets and the results show that it can find correct result both effective and efficient for binding sites.

Keywords: Motif, Regulatory Binding Sites, and Genetic Algorithm

1. Introduction

Understanding the regulatory networks of higher organisms is one of the main challenges of functional genomics. Gene expression is regulated by transcription factors (TF) binding to a specific transcription factor binding sites (TFBS) also known as motif, in regulatory regions associated with genes. Identification of regulatory regions and binding sites is a prerequisite for understanding gene regulation [1] [2]. Experimental identification and verification of such elements is challenging and costly so much effort has been put into the development of computational approaches. Good computational methods can potentially provide high-quality prediction of binding sites and reduce the time needed for experimental verification.

Computational discovery of regulatory elements is mainly possible because they occur several times in the same genome, and because they may be evolutionary conserved. This means that searching for overrepresented motifs across regulatory regions may discover novel regulatory elements. However this simple looking problem is a particularly hard problem, made difficult by a low signal-to-noise ratio. This is because of the poor conservation and short length of transcription factor binding sites when compared to the length of promoter sequences. Recent reviews have noted some important limitations of existing tools for regulatory motif discovery: notably, the limited applicability of current nucleotide background models [3], rapid failure with increasing sequence length [4], and a tendency to report false positives rather than true transcription factor binding sites [3] [4].

We used a structured genetic algorithm for regulatory motif discovery. The algorithm uses multi – layer representation of chromosome thus enabling algorithm to find out variable length motifs. In section 2 we described the biological background of the problem. Section 3 contains a brief survey of various techniques and algorithms used to solve the motif-finding problem. Section 4 explains the method and it's components like representation, fitness score function, selection, crossover and mutation operators. Next section contains the simulation results followed by conclusion.

2. Biological Background

A motif, in the context of biological sequence analysis, is a pattern of nucleotide bases or amino acids, which captures a biologically meaningful feature common to a group of nucleic acid or protein sequences. Regulatory motifs capture the patterns of DNA bases responsible for controlling when and where a gene is expressed. Typically, regulatory motifs describe transcription factor binding sites (TFBSs) embedded in the DNA sequences upstream of a gene's transcription start site (TSS). More rarely, regulatory signals may occur downstream of the TSS and even within coding sequences. Many well-characterized motifs, such as the TATA box occur proximal to the TSS. DNA binding allows transcription factors to bind at TFBSs located kilo bases from the TSS to interact with the transcription complex.

Hence, regulatory motifs may be found large distances upstream or downstream of the TSS. This also means that, for most TFBSs, there are few constraints upon their spatial location within a DNA sequence. Most TFBSs have a span of 5-8 bp, although the footprint of a transcription factor typically spans 10-20 bp, placing constraints upon the bases surrounding the binding site [5].

Well-conserved motifs, such as CCATT and TATA, are defined by their consensus sequences or, where variation exists, by simple regular expressions. For many regulatory motifs, however, there exists considerable sequence variation both within and between species. Consequently, it is normal for regulatory motifs to be represented as position frequency matrices (PFMs, also known as profiles) or position weight matrices (PWMs), showing the likelihood of each base occurring at each position within the motif. Known regulatory motif profiles are cataloged in databases such as TRANSFAC [6] and JASPAR [7].

3. Existing Methods

Many studies were done to find solutions for motif discovery. According to survey two major strategies exist to discover repeating sequence patterns occurring in both DNA and protein sequences: enumeration and probabilistic sequence modeling [8]. Enumeration strategies rely on word counting to find words that are over-represented. Probabilistic model-based methods represent the pattern as a matrix, called a motif, consisting of nucleotide base (or amino-acid residue) multinomial probabilities for each position in the pattern and different probabilities for background positions outside the pattern. Among those previous works, most popular one is the Multiple Em for Motif Elicitation (MEME) system [9], Gibbs sampler [10] and CONSENSUS

[11]. Even with weak signals, methods such as MEME and Gibbs Motif Sampler effectively find motifs of variable width and occurrences in DNA and protein sequences.

Many other algorithms have been developed to improve these popular motif discovery tools by means of performance, length of motifs and/or some other considerations. Stine et. al. employed Structured Genetic Algorithm [12] to search and to discover highly conserved motifs amongst upstream sequences of co-regulated genes. Liu et. al. also employed genetic algorithm for finding potential motifs in the regions of transcription start site (TSS) [13].

Recently Algorithms based on promoter sequences of coregulated genes and phylogenetic footprinting have been suggested. These algorithms integrate two important aspects of a motif's significance, i.e., overrepresentation and cross-species conservation, into one probabilistic score. Based on the Consensus algorithm [14] Wang and Stormo developed the motif finding algorithm PhyloCon (Phylogenetic Consensus) [15] that takes into account both conservation among orthologous genes and coregulation of genes within a species. Sinha et al. developed the algorithm PhyME [16] based on a probabilistic approach that handles data from promoters of coregulated genes and orthologous sequences.

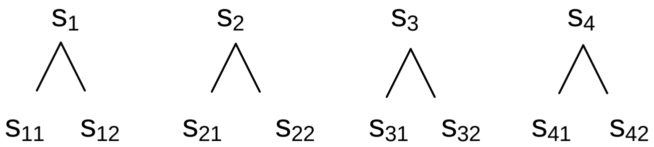
4. Proposed Method

A GA is a population-based method where each individual of the population represents a candidate solution for the target problem. This population of solutions is evolved throughout several generations, starting from a randomly generated one, in general. During each generation of the evolutionary process, each individual of the population is evaluated by a fitness function, which measures how good the solution represented by the individual is for the target problem. From a given generation to another, some parent individuals, usually those having the highest fitness produce “offsprings”, i.e., new individuals that inherit some features from their parents, whereas others (with low fitness) are discarded, following Darwin’s principle of natural selection. The selection of the parents is based on a probabilistic process, biased by their fitness value. Following this procedure, it is expected that, on average, the fitness of the population will not decrease every consecutive generation. The generation of new offsprings, from the selected parents of the current generation, is accomplished by means of genetic operators. This process is iteratively repeated until a satisfactory solution is found or some stop criterion is reached, such as the maximum number of generations.

A structured genetic algorithm is a multi- layered structure for the chromosomal representation that allows multiple bit changes to occur simultaneously at a different level. This leads to a large variation in the chromosome with a greater probability of maintaining high viability. In most cases of motif finding problem, motif length is fixed input. The structured genetic algorithm can find motif of varying length. The encoding of chromosomes and other genetic operators such as selection, crossover, mutation and the algorithm, that we used are described below.

4.1 Representation

We used the binary encoding to represent the DNA sequence motif. To represent the four nucleotides (A, T, C, G) we need two bits. The following binary code is used:
[A, T , G , C : 00 , 11 , 01 , 10]
In a double stranded DNA adenine (A) always pairs with thymine (T) and guanine (G) always pairs with cytosine(C). Based on this relationship the binary code of A & T and G & C are complimentary.
We used a two level binary chromosomal structure. First level gene acts as a switch that can activate or deactivate the genes at second level. If a first level bit is 1 then the corresponding second level gene is expressed. The chromosome is divided in two parts: left side bit represents the activation level and the right side bit represents the expressed motif. Following diagram explains the representation.



(a) A two level structure of St-GA

$(s_1\ \ s_2\ \ s_3\ \ s_4\quad s_{11}\,s_{12}\quad s_{21}\,s_{22}\quad s_{31}\,s_{32}\quad s_{41}\,s_{42}\ \ )$

(1 0 1 1 01 10 11 01)

Extracted pattern is : GTG

(b) Encoding process of St-GA representing a chromosome and corresponding binary coding

The chromosome we used is of the form as follows:

$C = (S1, S2) = ([s_i], [s_{ij}])$

C represents an ordered set containing string S1 at level 1 and S2 at level 2. s_i , s_{ij} represents the genes in chromosomes at first level and second level. M is the interpreted motif that can be constructed from C by joining each s_{ij} together when $s_i = 1$. Although C is a fixed length string that is used in all genetic operations in St-GA the predicted motif M may be of variable length depending upon the number of 1's in C's sub-string $s_{i.}$

4.2 Fitness score function

The fitness function is to evaluate how good the individuals are. To compute the fitness of each individual in population we used the fitness score function similar to as defined in FMGA [13]. First, we consider the computation of fitness score of a candidate motif for a single sequence then we compute the total fitness score for the candidate motif. Given a motif pattern, there may have several regions in the sequence that match the motif pattern and each has a fitness score according to how good is the match. First, one motif pattern will check the score of every possible position for one single sequence by comparing all letters. If the letter in motif pattern and letter in sequence match, the score will be incremented by one. In this way region with the highest matches is identified in a single sequence.

$$FS(S_m, P_n) = \max_{j} \left\{ \sum_{i=1}^{k} match(S_{mji}, P_{ni})/k\} \right\} \qquad \text{……………}(1)$$

where

$$match(S_{mji}, P_{ni}) = \begin{cases} 1 & if \quad S_{mji} = P_{ni} \\ 0 & if \quad S_{mji} \neq P_{ni} \end{cases} \qquad \text{……………}(2)$$

For example, suppose the motif pattern P_1 and promoter sequence S_1 are as follows:

P_1: ACGGCGTA
Promoter S1: ATACGGTAGGCCAGTGCGGACGGTGTAGATCCCG
Fitness_Score: 7

This way for a candidate motif highest score is calculated for all the sequences. Second, the total fitness score of a candidate motif is computed. The total fitness score function of a motif pattern is the summation of fitness score function for all sequences. It represents the score of a motif pattern in each generation of the genetic algorithm. Since structured GA can identify the variable length motifs, so to favor the motifs of maximum length we included one more term in total fitness score function. Here we have tried to maximize both component of a candidate motif, the length and the conservation. We used the total fitness score function as follows.

$$Total\_Fitness(P) = w_1 * L + w_2 * \Sigma FS(S_{m,} P_n) \qquad \text{…………}(3)$$

where Total_Fitness is total fitness score function, L is the normalized length of motif. The weight given to the length and similarity is w_1 and w_2 respectively.

4.3 Selection

Maintaining population diversity and selective pressure is the key issue while using a selection method. We used elitism to retain best members and remaining is selected using the stochastic tournament selection model. Every time, randomly two individuals are selected and the one with higher fitness score is preserved.

4.4 Crossover and Mutation

To generate new offspring from their parents we used uniform crossover method. In this method a crossover mask is generated of the chromosome length as a random bit string. Then taking the bit from first parent if the corresponding mask bit is 1 and from the second parent if the corresponding mask bit is 0 produces an offspring. For the second offspring the scheme is reversed.

There may be chances of being trapped in a local optimum and getting the false motif. To avoid this we used mutation. Mutation also help in maintaining population diversity and fast convergence of GA. Mutation is done by changing a randomly selected position’s binary value of the individual.

4.5 Algorithm

```
//Initialization
$n$ ← Number of individuals
import promoter sequences $S_1$ - $S_m$
// Evaluation of Fitness
for $i$ = 1 to $n$ do
    create candidate chromosomes randomely: $C_1$ - $C_n$
    extract candidate motifs from chromosomes : $P_1$ - $P_n$
    evaluate $TFS(S, P_i)$for each candidate motif
end for
// Generation cycle
while specified number of generations not complete
    //Selection : elitism
    get best of n individuals
    //Tournament Selection
    for $j$ = 1 to n do
        get two individual randomely $P_a$ and $P_b$
        if $TFS(S, P_a) > TFS(S, P_b)$ then
            retain $P_a$
        else
            retain $P_b$
        end if
    end for
    //Crossover
    for $k$ =1 to n/2 do
        make pairs of individuals
        generate mask for each pair
        produce offsprings from each pair
    end for
    //Mutation
    randomely find the victim individual
    randomely modify the victim bit
    // Insertion & Evaluation
    for $l$ = 1 to $n$ do
        replace current individuals by newly produced offsprings
        extract candidate motifs from new chromosomes : $P_1$ - $P_n$
        evaluate $TFS(S, P_l)$ for each candidate motif
    end for
end while
```

5. Simulation Results

In order to evaluate the performance of our algorithm for TFBS identification, synthetic datasets with sequence length 200 to 400 bp are generated with the various combinations of scenarios: (1) motif width: short (less than 8 bp) medium (between 8bp & 12bp) or long (greater than 12bp); (2) number of sequences: small (10) or large (30); (3) motif conservation: high or low. For each combination, datasets are generated and embedded with the instances of a random motif.

For each simulated dataset, to evaluate the performance of our algorithm we computed the hit ratio that is representing the percentage of successful identification of embedded motif in synthetic datasets. To compute the hit ratio we run the algorithm for a given combination number of times and then average is taken.

Results of various scenarios the number of sequences, length of sequences, average motif width, average conservation, generation cycles and hit ratio for each simulation condition are shown in Table 1 given below.

Table 1: Results of various scenarios

SNo.	*(N)*	*(L)*	*(W)*	*(C)*	*(GC)*	*HR*
1.	10	200	S	H	100	81%
2.	10	300	S	L	100	69%
3.	15	400	S	H	150	82%
4.	15	250	M	L	100	65%
5.	20	300	M	H	150	78%
6.	20	350	M	H	100	80%
7.	20	400	M	H	200	78%
8.	20	250	Ln	H	100	76%
9.	25	300	Ln	L	150	68%
10.	30	400	Ln	H	200	77%

N : *number of sequences* ***L:*** *length of sequence* ***W:*** *predicted motif width*
C : *motif conservation* ***GC :*** *generation cycles* ***HR :*** *Hit Ratio*
S: small M: medium Ln: long H : high L: low

6. Conclusion

Identification of transcription factor binding sites is an important and difficult problem. Most of the existing methods such as Gibbs sampling algorithm are local search methods, so they may suffer from the problem of local optima. Genetic algorithm provides a good approach to solve this problem. Genetic algorithm solves the optimal problem based on the biological characteristics. In this paper, we used a structured representation for individuals of the population that enable our algorithm to predict transcription factor binding sites of variable length. A lot of biological messages are hidden in promoter, and motif is one of the important messages. The motifs have the possibilities to be the binding sites of transcription factors. If the motifs can be predicted accurately, the biologists can then explore which transcription factors activate genes.

Simulation results of the algorithm on synthetic data including various scenario shows that the algorithm is able to predict the motifs with average hit ratio greater than 76% - 82% for the highly conserved regulatory transcription factor binding sites. However when conservation of regulatory transcription factor binding sites is poor the hit ratio is below 70%. The performance of this approach can probably be improved using more intelligent operators for selection, crossover and mutation. On the other hand, the fitness evaluation can be improved if we are able to additionally incorporate terms that reflect the biological messages behind the structural similarities among motifs.

References

[1] Lockhart, D., Winzeler, E., "Genomics, Gene Expression and DNA Arrays," Nature, 405 (2000) 827-836

[2] Stormo G. "DNA binding sites: representation and discovery," Bioinformatics 2000, 16:16-23.

[3] M. Tompa, N. Li, T.L. Bailey, G.M. Church, B.D. Moor, E. Eskin, A.V. Favorov, M.C. Frith, Y. Fu, W.J. Kent, V.J. Makeev, A.A. Mironov, W.S. Noble, G. Pavesi, G. Pesole, M. Rgnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu, "Assessing Computational Tools for the Discovery of Transcription Factor Binding Sites," Nature Biotechnology, vol. 23, no. 1, pp. 137-144, Jan. 2005

[4] J. Hu, B. Li, and D. Kihara, "Limitations and Potentials of Current Motif Discovery Algorithms," Nucleic Acids Research, vol. 33,no. 15, pp. 4899-4913, 2005.

[5] G.A. Wray, M.W. Hahn, E. Abouheif, J.P. Balhoff, M. Pizer, M.V. Rockman, and L.A. Romano, "The Evolution of Transcriptional Regulation in Eukaryotes," Molecular Biology and Evolution, vol. 20, no. 9, pp. 1377-1419, Sept. 2003.

[6] V. Matys, E. Fricke, R. Geffers, E. Gssling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A.E. Kel, O.V. Kel-Margoulis, D.-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Mu¨ nch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, "TRANSFAC: Transcriptional Regulation, from Patterns to Profiles," Nucleic Acids Research, vol. 31, no. 1, pp. 374-378, Jan. 2003.

[7] A. Sandelin, W. Alkema, P. Engstro¨m, W.W. Wasserman, and B. Lenhard, "JASPAR: An Open-Access Database for Eukaryotic Transcription Factor Binding Profiles," Nucleic Acids Research, vol. 32, pp. D91-D94, Jan. 2004.

[8] Modan K Das and Ho-Kwok Dai, "A survey of DNA motif finding algorithms," BMC Bioinformatics 2007, 8(Suppl 7):S21

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

[10] Thompson, W., Rouchka E.C. and Lawrence C.E. "Gibbs Recursive Sampler: Finding transcription factor binding sites", J. Nucleic Acids Research, Vol.31, pp. 3580-3585, 2003.

[11] Hertz, G.Z., Hartzell, G.W. and Stormo, G.D. "Identification of consensus patterns in unaligned DNA sequences known to be functionally related", Bioinformatics, Vol.6, pp. 81-92, 1990.

[12] Stine M., Dasgupta, D. and Mukatira, S. "Motif Discovery in Upstream Sequences of Coordinately Expressed Genes", The 2003 Congress on Evolutionary Computation, pp.1596-1603, 2003.

[13] Liu, F.F.M et al. "FMGA: Finding Motifs by Genetic Algorithm", Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioengineering, pp.459-466, 2004.

[14] Hertz GZ, Hartzell GW, Stormo GD. "Identification of consensus patterns in unaligned DNA sequences known to be functionally related," Comput Appl. Biosci 1990, 6:81-92.

[15] Wang T, Stormo GD. "Combining phylogenetic data with coregulated genes to identify regulatory motifs," Bioinformatics 2003, 19:2369-2380.

[16] Sinha S, Blanchette M, Tompa M. "PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences," BMC Bioinformatics 2004, 5:170.