TFBS Detection Algorithm Using Distance Metrics Based on Center of Mass and Polyphase Mapping

Mohammad Al Bataineh, Lun Huang, and Guillermo Atkin

Telecommunications Engineering Department, Yarmouk University
Yarmouk University, 21163, Irbid, Jordan
phone: + (962) 777949599, fax: + (962) 2 721 1129, email: mffbataineh@yahoo.com
web: http://hibit2012.ii.metu.edu.tr

ABSTRACT

Regulatory sequence detection is a fundamental challenge in computational biology. The transcription process in protein synthesis starts with the binding of the transcription factor to its binding site. Different sites can bind to the same factor. This variability in binding sequences increases the difficulty of their detection using computational algorithms. This paper proposes a novel algorithm for transcription factor binding site (TFBS) detection. The algorithm applies a polyphase mapping scheme to represent the four nucleobases in both the DNA sequence and the set of binding sites associated with a given transcription factor (TF). The center of mass (CoM) of each set of binding sites, which can be thought of as a consensus sequence, is then calculated. The algorithm then calculates distances between the CoM and each binding site belonging to a given TF. Same scenario is then applied to the genome sequence under study. The obtained distances are then utilized to detect new potential TFBSs based on their similitude of the set of binding sites that we already know. Analysis is applied to E. coli bacterial genomes. Simulation results verify the correctness and the biological relevance of the proposed algorithm.

1. INTRODUCTION

In bioinformatics, one can distinguish between two separate problems regarding DNA binding sites: searching for additional members of a known DNA binding motif (the site search problem) and discovering novel DNA binding motifs in collections of functionally related sequences (the sequence motif discovery problem) [1]. Many different methods have been proposed to search for binding sites. Most of them rely on the principles of information theory and have available web servers (Yellaboina)(Munch), while other authors have resorted to machine learning methods, such as artificial neural networks [2-4]. A plethora of algorithms is also available for sequence motif discovery. These methods rely on the hypothesis that a set of sequences share a binding motif for functional reasons. Binding motif discovery methods can be divided roughly into enumerative, deterministic and stochastic [5]. MEME [6] and Consensus [7] are classical examples of deterministic optimization, while the Gibbs sampler [8] is the conventional implementation of a purely stochastic method for DNA binding motif discovery. While enumerative methods often resort to regular expression representation of binding sites, PSFM and their formal treatment under information theory methods are the representation of choice for both deterministic and stochastic methods. Recent advances in sequencing have led to the introduction of comparative genomics approaches to DNA binding motif discovery, as exemplified by PhyloGibbs [9, 10].

Regulatory sequence detection is a fundamental challenge in computational biology. The transcription process in protein synthesis starts with the binding of the transcription factor to its binding site. Different sites can bind to the same factor. This variability in binding sequences increases the difficulty of their detection using computational algorithms. This paper proposes a novel algorithm for detecting transcription factor binding sites in the entire genomic structure by using a distance metric based on a center of mass concept. The proposed algorithm applies a polyphase mapping scheme to represent the four nucleobases. Not only can the proposed algorithm be used to investigate the detection problem, it also can help examine the distribution of regulatory sequences in coding and non-coding regions. This later knowledge can then be utilized in gene identification. Additionally, using the said algorithm allows discovering new potential motif sequences that need to be subjected to further biological analysis.

2. PROPOSED ALGORITHM

Figures 1 and 2 show a block diagram of the proposed algorithm.

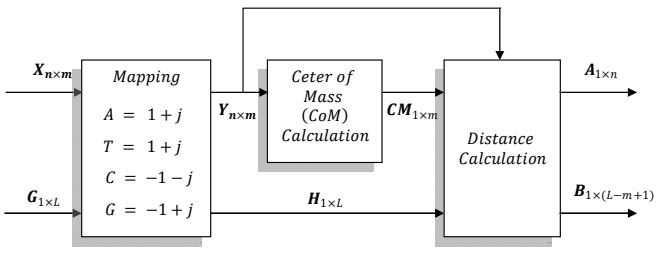

Figure 1. A block diagram of TFBS detection algorithm (Part 1)

2.1 Mathematical Description

The genome under study is given by

$$\boldsymbol{G}_{1\times L} = [g_1, g_2, g_3, \cdots, g_L], \quad (1)$$

where L is the length of the genome in nucleobases.

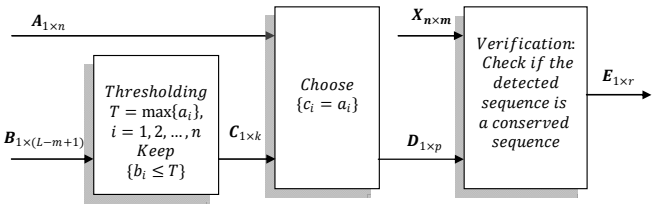

Figure 2. A block diagram of $TFBS$ detection algorithm (Part 2)

Each set of transcription factor binding sites ($TFBSs$) consists of n binding sites each of which is consisting of m nucleobases. Each set of binding sites can be represented by a matrix of the form

$$\boldsymbol{X}_{n\times m} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1m} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nm} \end{bmatrix}, \quad (2)$$

where $g_i \in \{A, G, C, T\}, \quad i = 1, 2, \dots, L.$
$x_{kj} \in \{A, G, C, T\};\ k = 1, 2, \dots, n;\ j = 1, 2, \dots, m.$
Using polyphase mapping, the genome $\boldsymbol{G}_{1\times L}$ becomes

$$\boldsymbol{H}_{1\times L} = [h_1, h_2, h_3, \dots, h_L], \quad (3)$$

and the nucleobases matrix $\boldsymbol{X}_{n\times m}$ becomes

$$\boldsymbol{Y}_{n\times m} = \begin{bmatrix} y_{11} & y_{12} & y_{13} & \dots & y_{1m} \\ y_{21} & y_{22} & y_{23} & \dots & y_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & y_{n3} & \dots & y_{nm} \end{bmatrix}, \quad (4)$$

where

$$h_i = \begin{cases} 1 + j & , \quad if\ g_i = A \\ 1 - j & , \quad if\ g_i = T \\ -1 + j & , \quad if\ g_i = C \\ -1 - j & , \quad if\ g_i = G \end{cases}. \quad (5)$$

Similarly,

$$y_{ij} = \begin{cases} 1 + j & , \quad if\ x_{ij} = A \\ 1 - j & , \quad if\ x_{ij} = T \\ -1 + j & , \quad if\ x_{ij} = C \\ -1 - j & , \quad if\ x_{ij} = G \end{cases}. \quad (6)$$

Given $\boldsymbol{Y}_{n\times m}$, the center of mass ($CoM$) can be obtained as the mean of all rows of $\boldsymbol{Y}_{n\times m}$ that correspond to all binding sites. Hence, the center of mass is given by

$$\boldsymbol{CM}_{1\times L} = [cm_1, cm_2, cm_3, \dots, cm_L], \quad (7)$$

where

$$cm_i = \frac{1}{n}\sum_{k=1}^{n} y_{ki}. \quad (8)$$

The next step is to calculate the Euclidean distance of all binding sites of a given $TFBS$ family ($\boldsymbol{Y}_{n\times m}$) from the center of mass ($\boldsymbol{CM}_{1\times L}$). The resulting distance vector is given by

$$\boldsymbol{A}_{1\times n} = [a_1, a_2, a_3, \dots, a_n], \quad (9)$$

where a_i is obtained by

$$a_i = \left|\sqrt{\sum_{i=1}^{m}(\boldsymbol{Y}_i - \boldsymbol{CM})^2}\right|. \quad (10)$$

$\boldsymbol{Y}_i$ is the i^{th} row of the matrix $\boldsymbol{Y}_{n\times m}$, i.e. the i^{th} binding site.
The same Euclidean distance measure is also used to calculate the distance between the center of mass ($\boldsymbol{CM}_{1\times L}$) and a window of the same length as the center of mass being slid all over the genome one base at a time. This will yield a distance vector $\boldsymbol{B}_{1\times L-m+1}$ of length equal to $(L - m + 1)$ given by

$$\boldsymbol{B}_{1\times L-m+1} = [b_1, b_2, b_3, \dots, b_{L-m+1}], \quad (11)$$

where b_i is obtained by

$$b_i = \left|\sqrt{\sum_{i=1}^{m}(\boldsymbol{H}_i - \boldsymbol{CM})^2}\right|. \quad (12)$$

$\boldsymbol{H}_i$ corresponds to the i^{th} window of the genome $\boldsymbol{H}_{1\times L}$ starting at position i and ending at position $i + m$.
At this point, the distance vector $\boldsymbol{B}_{1\times L-m+1}$ contains all the possible distances that the genome has from the center of mass ($\boldsymbol{CM}_{1\times L}$). To cut that down, a threshold T is calculated as the maximum distance in the $\boldsymbol{A}_{1\times n}$ vector obtained in the previous steps. All the distances greater than T are discarded. The resulting distance vector $\boldsymbol{C}_{1\times k}$ is defined as

$$\boldsymbol{C}_{1\times k} = \{c_i = b_j; b_j \le T\}, \quad (13)$$

where $i = 1,2,3, \dots, k; j = 1,2,3, \dots, L - m + 1.$
Now, the vector $\boldsymbol{C}_{1\times k}$ contains all the possible distances less than or equal to the threshold T. The next step is to discard all the values in $\boldsymbol{C}_{1\times k}$ that are different from the set of distances in the $\boldsymbol{A}_{1\times n}$ vector. The resulting distance vector $\boldsymbol{D}_{1\times p}$ is then defined as

$$\boldsymbol{D}_{1\times p} = \{d_i = c_j; c_j \in \boldsymbol{A}_{1\times n}\}, \quad (14)$$

where $i = 1,2,3, \dots, p; j = 1,2,3, \dots, L - m + 1$ and $p \le k$.
Finally, the vector $\boldsymbol{D}_{1\times p}$ contains all the possible distances that are exactly the same as the distances of the original binding sites from their center of mass. Some of these distances are false positive, i.e. they do not correspond to an actual binding site. To filter out these false positives, each one of the detected sequences is compared to the original binding sites and only the verified ones are kept. The resulting distance vector $\boldsymbol{E}_{1\times r}$ is defined as

$$\boldsymbol{E}_{1\times r} = \{e_i = d_j; \boldsymbol{S}(d_j) = \boldsymbol{X}(d_j)\}, \quad (15)$$

- where $i = 1, 2, 3, \dots, r; j = 1, 2, 3, \dots, p.$ $\boldsymbol{S}(d_j)$ is the sequence of length m in the genome corresponding to the j^{th} distance in the $\boldsymbol{D}_{1\times p}$ vector. $\boldsymbol{X}(d_j)$ is the binding site in $\boldsymbol{X}_{n\times m}$ having the same distance d_j.

3. ANALYSIS AND SIMULATION RESULTS

The proposed algorithm was applied to detect $TFBSs$ in two different *Escherichia coli* genomes (both positive strand and negative strand were used). Table 1 summarizes the obtained simulation results in terms of the total number of *TFBSs* detected in both *E.coli* genomes (forward and reverse strands).

Table 1. Simulation results using two E. coli genomes

E. coli Strain	*MG1655*		*O157:H7*	
Strand Orientation	*Positive Strand*	*Negative Strand*	*Positive Strand*	*Negative Strand*
TFBSs in the genome	956	1000	642	616
TFBSs in non − coding regions	817	59	540	527
TFBSs in coding regions	66	6.20	45	33
TFBSs overlapped in betw	73	79	57	56

In Figure 3-6, the green areas represent the coding regions. Figures 3 and 4 represent the detected $TFBSs$ in $MG1655\ E.coli$ genome (956 $TFBSs$ are detected in the positive strand as shown in Figure 3, and 1000 $TFBSs$ are detected in the negative strand as shown in Figure 4). The x-axis represents the base positions of the genome being considered, the y-axis represents the percentages that the de-

tected $TFBSs$ occur in the coding regions. In other words, a value of "100" at the y-axis means that the detected $TFBS$ is totally located in the coding regions. Likewise, a value of "0" means that the detected $TFBS$ totally occurs in the non-coding regions. For all other values between 0 and 100, the detected $TFBS$ occurs in between.

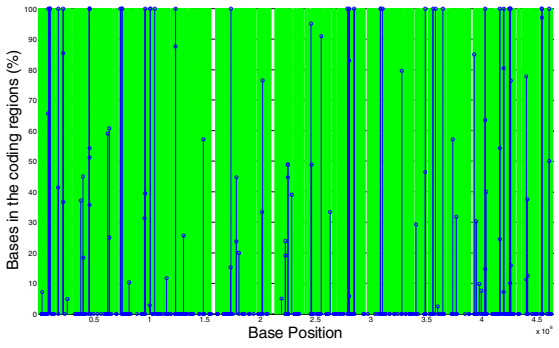

Figure 3. TFBS detection using MG1655 positive strand

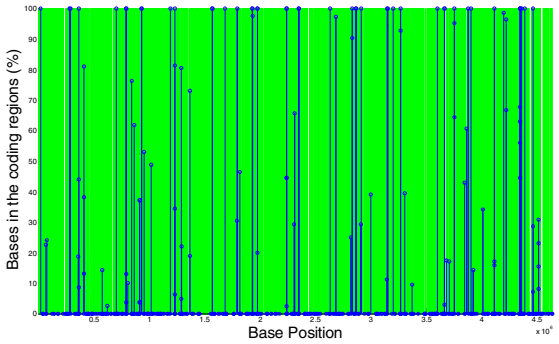

Figure 4. TFBS detection using MG1655 negative strand

The same analysis was also applied to O157:H7 *E. coil* genome (both positive and negative strands). The corresponding simulation results are shown in Figures 5-6 (642 $TFBSs$ are detected in the positive strand, and 616 $TFBSs$ are detected in the negative strand).

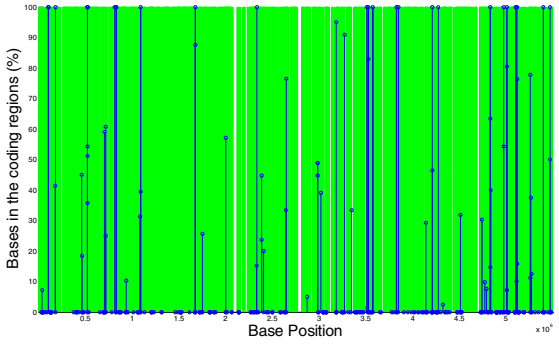

Figure 5. TFBS detection using O157:H7 Positive strand

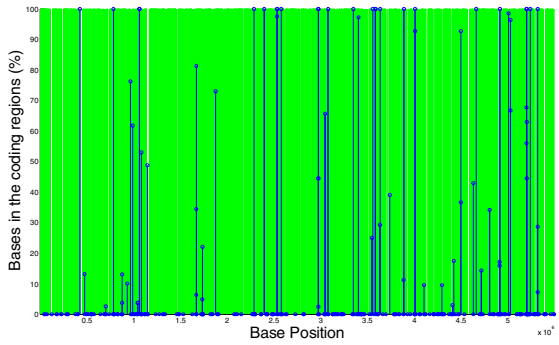

Figure 6. TFBS detection using O157:H7 negative strand

3.1 Discovery of New Motif Sequences

The Motif finding problem involves finding short DNA sequences or consensus patterns that occur surprisingly often, without any prior knowledge of what these patterns looks like. A visualization technique and algorithm for finding motifs in DNA sequences is proposed in [11]. Due to the fact that human vision is not able to discern patterns from millions of strings of the four nucleobases $\{A, G, C \text{ and } T\}$; we need to develop a method capable of doing that. This proposed algorithm described before can be used to identify any substring of bases that occurs surprisingly very often in the DNA sequence under study. If a sequence occurs more frequently than expected, then it can be assumed that this sequence is more probable to have a particular significance in the biological sense. Such sequences can be subjected to further biological research to verify their significance.

Out of the set of all possible $TFBSs$ resulted from applying the detection algorithm described in section III (12150 possible sequences), this extra step searches if there is any sequence in the whole set of possible $TFBSs$ appears more than once. Based on this, the algorithm classifies each possible sequence depending on the number of times it appears in the DNA sequence as well as the TF that this particular sequence belongs to.

Once the sequences that appear more than once are identified, they can be considered as possible motif sequences. These sequences can be tested to see if they appear in other E. coli genomes. If yes, this will increase the chance that these sequences are potential TFBSs. Hence, they will need to be

subjected to biological experiments to verify their significance. This type of analysis was applied to MG1655 and O157:H7 $E. coli$ strains. Sequences that appeared five, four, and three times in $MG1655\ E. coil$ genome (the positive strand is considered here) along with their corresponding locations were obtained.

Table 2. Possible potential motif sequences in *E.coli* genome

#	Sequence	TF	Frequency of occurrence	
			in MG1655	in O157:H7
1	UUUGUUCAUGCCGGAUGCGGCGUGAACGCC	*MalT*	5	2
2	AUGGGUCAGCGACUUAUAUUCUGUAGC	*NarL*	5	5
3	UGAAGUCGGAAUCGCUAGUAAUCGUGG	*NarL*	5	5
4	AAAAAAUUGAUGGAACAUAUUUCUAUU	*IclR*	4	3
5	CUGUUGACUAUUUGUGCCGUUAUUUCUG	*MetJ*	4	1
6	AAAUGGAACUUAAAAAAAUUGAUGGAAC	*NarL*	4	3
7	AAUUCGCAUUUUAUGUUUAAAAAUUGA	*NarL*	4	2
8	UUUUAUGUUUAAAAAUUGAGAUAUUCC	*NarL*	4	1
9	UAACAUGCUGGAUAUUAAAGGAAAAAU	*NarL*	4	1

In order to check whether the obtained sequences are potential motif sequences or not, a different $E.coil$ genome (O157H7) was scanned to see if any of those sequences appear any time. Based on obtained results, it can be observed that 1) all of sequences that appeared five times in the $E.coli\ MG1655$ genome, appeared in the $E.coli\ O157H7$ genome as well. In addition, they also appear several times, what increases the chance that these sequences are more probable to be considered potential motif sequences. 2) Most of the sequences that appeared 4 times in MG1655 also appeared in O157:H7. 3) Some of the sequences that appeared 3 times in MG1655 also appeared in O157:H7.

In conclusion, we can state that while the sequences that appeared five or four times have a higher probability of being in other genomes and hence must be considered as potential motif sequences, the ones that appear three, two and one times can be discarded. Table 2 shows the set of sequences that can be considered as potential motif sequences based on their frequency of occurrence in both $E.coli$ genomes considered in this analysis. These sequences are recommended for further laboratory experimentation to see if they possess any biological significance.

4. CONCLUSIONS

In this paper, a novel algorithm for transcription factor binding site (TFBS) detection is proposed. The algorithm is based on using a distance metric based on a "center of mass (CoM)" concept. The four known nucleobases were represented using a polyphase mapping scheme. The center of mass, analogous to the consensus sequence here, for each set of binding sites associated with a particular transcription factor was calculated. Distances of these binding sites from their corresponding CoM is obtained. A thresholding scheme is then followed to identify sequences in the genome under study that are totally or highly similar to the set of available binding sites belonging to a particular transcription factor. The developed algorithm was applied to two different *E. coli* bacterial genomes (MG1655 and O157:H7). Simulation results show that the algorithm is not only able to efficiently identify and accurately locate the known TFBSs, but also can be utilized in the discovery of new motif sequences based on 1) their frequency of occurrence in the genome sequence under study, and 2) their similitude to the know TFBSs.

Having analyzed only two *E. coli* genomes, some potential motif sequences were detected (as shown in Table 2). Therefore, if this method of motif finding is applied to more other genomes, other potential motif sequences can be identified. Hence, the previous analysis of using the frequency weight to detect and identify new motif sequences can efficiently yield other potential motif candidates that are recommended to be subjected for further analysis.

Future work can utilize the results obtained in this paper to help distinguish coding and non-coding regions. In other words, the developed analysis here can help in gene identification.

REFERENCES

[1] I. Erill and M. C. O'Neill, "A reexamination of information theory-based methods for DNA-binding site identification," *BMC Bioinformatics,* vol. 10, p. 57, 2009.

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

[3] D. Bisant and J. Maizel, "Identification of ribosome binding sites in Escherichia coli using neural network models," *Nucleic Acids Res,* vol. 23, pp. 1632-9, May 11 1995.

[4] M. C. O'Neill, "Training back-propagation neural networks to define and detect DNA-binding sites," *Nucleic Acids Res,* vol. 19, pp. 313-8, Jan 25 1991.

[5] T. L. Bailey, "Discovering sequence motifs," *Methods Mol Biol,* vol. 452, pp. 231-51, 2008.

[6] T. L. Bailey, "Discovering novel sequence motifs with MEME," *Curr Protoc Bioinformatics,* vol. Chapter 2, p. Unit 2 4, Nov 2002.

[7] G. D. Stormo and G. W. Hartzell, 3rd, "Identifying protein-binding sites from unaligned DNA fragments," *Proc Natl Acad Sci U S A,* vol. 86, pp. 1183-7, Feb 1989.

[8] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment," *Science,* vol. 262, pp. 208-14, Oct 8 1993.

[9] M. K. Das and H. K. Dai, "A survey of DNA motif finding algorithms," *BMC Bioinformatics,* vol. 8 Suppl 7, p. S21, 2007.

[10] R. Siddharthan, E. D. Siggia, and E. van Nimwegen, "PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny," *PLoS Comput Biol,* vol. 1, p. e67, Dec 2005.

[11] G. Rambally, "A Visualization Approach to Motif Discovery in DNA Sequences," *Praire View A&M University: Pairie View, Texas,* 2007.