Transcription Factor Binding Site Detection Algorithm Using Distance Metrics Based on a Position Frequency Matrix Concept

Mohammad F. Al Bataineh[1], Lun Huang[2], Guillermo E. Atkin[2]

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 (TF) to its binding site. These binding sites are short DNA segments that are called motifs. Different sites can bind to the same factor. This variability in binding sequences besides their low information content and low specificity increases the difficulty of their detection using computational algorithms. This paper proposes a novel algorithm for transcription factor binding sites (TFBSs) detection in the entire genomic structure and allow discovery of new motif sequences. This is achieved by using distance metrics based on a position frequency matrix (PFM) concept that quantify the similitude between the set of conserved sequences belonging to a particular TF and the entire DNA sequence under study. Hence, the PFM in this context can be thought of as a consensus sequence as it provides a representative measure of the said set of binding sites belonging to a particular TF. The algorithm then quantifies the correlation between the PFM and each binding site belonging to a given TF. Same scenario is then applied to the genome sequence under study. The obtained distance metrics are then utilized to discover new potential TFBSs based on their similitude of the set of binding sites investigated. Analysis is applied to Escherichia coli (E. coli) bacterial genomes. Simulation results verify the correctness and the biological relevance of the proposed algorithm.

Keywords: Transcription Factor Binding Site, Position Frequency Matrix, Consensus Sequence, Motif Discovery

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, 2]. 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 [3, 4], while other authors have resorted to machine learning methods, such as artificial neural

[1] Telecommunications Engineering Department, Hijjawi Faculty for Engineering Technology

[2] Electrical and Computer Engineering Department, Illinois Institute of Technology

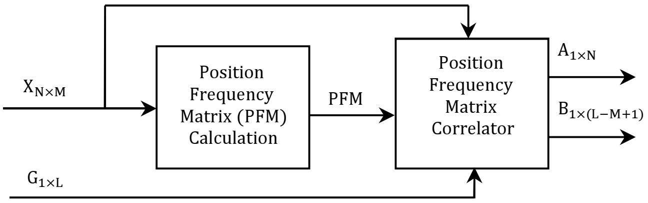
networks [5-7]. A plethora of algorithms is also available for sequence motif discovery [8, 9]. These methods rely on the hypothesis that a set of sequences share a binding motif for functional reasons. Binding motif discovery methods can be divided roughly into enumerative, deterministic and stochastic [10]. MEME [11] and CONSENSUS [12] are classical examples of deterministic optimization, while the Gibbs sampler [13] 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. Recent advances in sequencing have led to the introduction of comparative genomics approaches to DNA binding motif discovery, as exemplified by PhyloGibbs [14]. More complex methods for binding site search and motif discovery rely on the base stacking and other interactions between DNA bases. An example of such tool is the ULPB [15].

Unraveling the mechanisms that regulate gene expression is a major challenge in biology. An important task in this challenge is to identify regulatory elements, especially the binding sites in deoxyribonucleic acid (DNA) for transcription factors. These binding sites are short DNA segments that are called motifs. As different sites can bind to the same factor, this increases the difficulty of their detection using computational algorithms. Although traditional footprinting assays can accurately identify the precise binding sites of any factor, this low-throughput method is highly technical and can analyze only a single small region (< 1 kb) at a time. With the development of high-throughput sequencing technologies, a number of experimental techniques such as ChIP-chip and ChIP-seq have been used to identify the location of transcription factor binding sites. However, these methods are unable to resolve DNA-protein interactions at base pair resolution [16]. *In silico* identification of over-represented DNA motifs from the promoters of co-regulated or homologous genes as well as ChIP-enriched regions plays a significant role in locating binding sites in a high-throughput and high-resolution manner. This paper proposes a novel algorithm for detecting transcription factor binding sites in the entire genomic structure by using distance metrics based on a position frequency matrix (PFM) concept. The algorithm does not use any mapping for the four known nucleobases (A, T, C and G) unlike a previous work [1] where polyphase mapping was used to represent the four nucleobases. Not only can the proposed algorithms be used to investigate the detection problem of transcription factor binding sites, but also can help examine the distribution of regulatory sequences in coding and non-coding regions. This later knowledge can then be utilized in the gene identification problem. Moreover, the proposed algorithm allows discovering new potential motif sequences that need to be subjected to further biological analyses to verify their significance.

The rest of the paper is organized as follows: Section 2 presents a mathematical description of the proposed algorithm. It also summarizes the list of steps that describe how the algorithm works. Section 3 presents the analysis and simulation results of applying the proposed algorithm to two different Escherichia coli bacterial genomes. Subsection 3.1 describes how the proposed algorithm can be utilized in the discovery of new motif sequences based on their similitude to the known TFBSs. Finally, conclusions are drawn in Section 4.

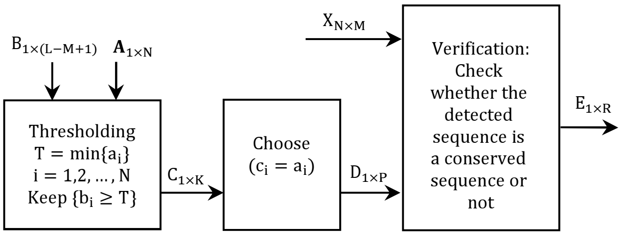
2 Proposed Algorithm

Figures 1 and 2 show a schematic system-like representation of the proposed algorithm. The input parameters to the algorithm are the genome under study $G_{1\times L}$ (L is the length of the genome in nucleobases) and the set of binding site sequences belonging to a particular transcription factor represented by the matrix $X_{N\times M}$. The distance metric vector $A_{1\times N}$ is assigned to the set of conserved binding sites belonging to the same transcription factor with a length equal to the number of binding sites (N), while $B_{1\times(L-M+1)}$ is another distance metric vector assigned to set of sequences in the genome with a length similar to the conserved sequence length (M). The output of the algorithm is a distance vector $E_{1\times R}$ that corresponds to the locations of actual binding sites investigated.



X N×M
Position Frequency Matrix (PFM) Calculation
PFM
Position Frequency Matrix Correlator
A 1×N
B 1×(L−M+1)
G 1×L

Fig. 1. TFBS Detection Algorithm (part 1)



B 1×(L−M+1)
A 1×N
Thresholding T = min{a_i} i = 1,2,…,N Keep {b_i ≥ T}
C 1×K
Choose (c_i = a_i)
D 1×P
X N×M
Verification: Check whether the detected sequence is a conserved sequence or not
E 1×R

Fig. 2. TFBS Detection Algorithm (part 2)

2.1 Mathematical Description

The genome under study is given by

$$G_{1\times L} = [g_1, g_2, g_3, \dots, g_L], \tag{1}$$

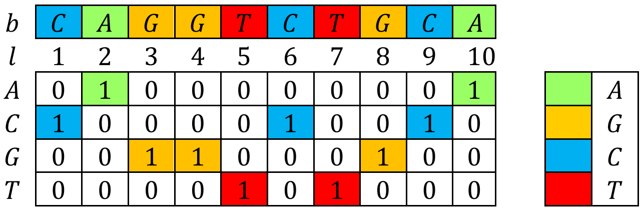
where $g_i \in \{A, G, C, T\}$ is the i^{th} nucleobase in the genome and $i = 1, 2, 3, \dots, L$ is the genome length in nucleobases.

Each set of binding sites belonging to a given transcription factor can be represented as a matrix consisting of N rows and M columns. The number of rows N corresponds to the number of binding sites conserved sequences, and the number of columns M corresponds to the number of nucleobases in each binding site. Hence, each one of the 124 different E. Coli transcription factors (see Table 1) investigated in this paper can be represented as a matrix of size (N×M) denoted as $X_{N\times M}$ and is given by

$$X_{N\times M} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1M} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & x_{N3} & \cdots & x_{NM} \end{bmatrix}, \tag{2}$$

where $x_{kj} \in \{A, G, C, T\}$; $k = 1, 2, \ldots, N$; $j = 1, 2, \ldots, M$.

Each row of the matrix $X_{N\times M}$ can be represented by a $(4 \times M)$ binary matrix (containing only 1's and 0's) where each row corresponds to one of the four nucleobases $\{A, G, C, T\}$. For example, if the 5th row of the matrix $X_{N\times M}$ is given by the sequence $5' - \mathrm{CAGGTCTGCA} - 3'$, then the corresponding binary matrix, denoted as S, will have the form shown in Figure 3.



b	*C*	*A*	*G*	*G*	*T*	*C*	*T*	*G*	*C*	*A*
l	1	2	3	4	5	6	7	8	9	10
A	0	1	0	0	0	0	0	0	0	1
C	1	0	0	0	0	1	0	0	1	0
G	0	0	1	1	0	0	0	1	0	0
T	0	0	0	0	1	0	1	0	0	0

🟩	*A*
🟨	*G*
🟦	*C*
🟥	*T*

Fig. 3. The **S** Matrix of the Fifth Row Sequence in $\mathbf{X}_{\mathbf{N\times M}}$.

Based on this definition of the S binary matrix, the so-called Position Frequency Matrix, denoted as PFM, can be obtained by averaging the N corresponding S binary matrices that represent the N rows in the $X_{N\times M}$ matrix. This makes the Position Frequency Matrix be of a size $(4 \times M)$. The (b, l) element in the PFM matrix will then be equal to the frequency of the nucleobase b occurring at position l and hence can be defined as

$$\mathrm{PFM}(b, l) = \frac{1}{N} \sum_{j=1}^{N} S(b, l, j), \tag{3}$$

The $(4 \times M)$ Position Frequency Matrix can then be rewritten as

$$\mathrm{PFM} = \begin{bmatrix} r_{A1} & r_{A2} & r_{A3} & \cdots & r_{AM} \\ r_{G1} & r_{G2} & r_{G3} & \cdots & r_{GM} \\ r_{C1} & r_{C2} & r_{C3} & \cdots & r_{CM} \\ r_{T1} & r_{T2} & r_{T3} & \cdots & r_{TM} \end{bmatrix}. \tag{4}$$

where $r_{bl} = \mathrm{PFM}(b, l)$, $b \in \{A, G, C, T\}$, and $1 \le l \le M$. The (b, l) elements of the PFM matrix can be alternatively calculated as

$$r_{bl} = \frac{N_{bl}}{N}, \tag{5}$$

where $l = 1, 2, 3, \ldots, M$; N_{bi} stands for the number of times the base $b \in \{A, G, C, T\}$ occurs in column l of the matrix $X_{N\times M}$. Therefore, r_{bl} is the probability of the base b occurring at position l. This makes the elements of each row in the PFM matrix add up to one.

According to the schematic representation in Figures 1 and 2, the algorithm is able to detect TFBSs in the genome by the following steps:

Step 1: Position frequency matrix calculation

The Position frequency matrix (PFM) can be calculated using equation 3 or alternatively using equations 4 and 5.

Step 2: Distance metric vectors calculation

After calculating the position frequency matrix (PFM), the next step is to assign each conserved sequence (i.e. each row in $\mathrm{X_{N\times M}}$) a distance metric obtained by first aligning that particular row with the position frequency matrix and then adding up the values in the matrix that correspond to each nucleobase in that row. If any of these values is zero (this means that this particular base does not happen at that position), then we add up the values occurring before this later zero and stop and then move to the next row. Based on this, we get a distance vector $(\mathrm{A_{1\times N}})$ which can be written as

$$\mathrm{A_{1\times N}} = [\mathrm{a_1, a_2, a_3, \dots, a_N}], \tag{6}$$

where $\mathrm{a_i}$ is a distance metric associated with the $\mathrm{i^{th}}$ row in $\mathrm{X_{N\times M}}$ and is obtained by

$$\mathrm{a_i} = \sum_{\mathrm{j=1}}^{\mathrm{M' \leq M}} \mathrm{r_{x_{ij}j}}, \tag{7}$$

where $\mathrm{i = 1, 2, 3, \dots, N, x_{ij} \in \{A, G, C, T\}}$ is the $\mathrm{(i, j)}$ element in the $\mathrm{X_{N\times M}}$ matrix, and M' is the $\mathrm{j^{th}}$ index of the first base in the $\mathrm{i^{th}}$ row to map to a zero in the Position Frequency Matrix (PFM).

The next step is to compare the whole genome sequence to the set of conserved sequences represented by $\mathrm{X_{N\times M}}$. To achieve this, a sliding window of length equal to M (conserved sequence length) is translated all over the genome $(\mathrm{G_{1\times L}})$ one base at a time. This will divide the genome sequence into $(\mathrm{L - M + 1})$ subsequences each of length M. These subsequences are then assigned distance metrics following the same procedure used to get the distance vector $\mathrm{A_{1\times N}}$. Hence, this will yield another distance vector, $\mathrm{B_{1\times(L-M+1)}}$, which can be written as

$$\mathrm{B_{1\times(L-M+1)}} = [\mathrm{b_1, b_2, b_3, \dots, b_{L-M+1}}], \tag{8}$$

where $\mathrm{b_i}$ is a distance metric associated with a sequence of length M obtained at the $\mathrm{i^{th}}$ nucleobase position of the genome and is defined by

$$\mathrm{b_i} = \sum_{\mathrm{j=1}}^{\mathrm{M' \leq M}} \mathrm{r_{g_ij}}, \tag{9}$$

where $\mathrm{i = 1, 2, 3, \dots, L - M + 1; g_i \in \{A, G, C, T\}}$ is a base in the genome being considered. $\mathrm{r_{g_ij}}$ is the $\mathrm{j^{th}}$ element in the PFM matrix that corresponds to the $\mathrm{i^{th}}$ base in the genome $(\mathrm{g_i})$, and M' is the $\mathrm{j^{th}}$ index of the first base in the $\mathrm{i^{th}}$ subsequence in the genome to map to a zero in the Position Frequency Matrix.

Step 3: Thresholding of the distance vector $\mathrm{B_{1\times(L-M+1)}}$.

At this point, the weighting vector $\mathrm{B_{1\times(L-M+1)}}$ contains all the possible weights corresponding to all possible conserved sequences in the genome sequence. The higher the weight value of a given sequence, the higher is the probability of that particular sequence to be a real conserved sequence. To eliminate those subsequences in the genome with low weights, a threshold T is calculated as the minimum weight in the distance vector $A_{1\times N}$ vector obtained in step 2. All the weights less than T are discarded. The resulting weighting vector after Thresholding ($C_{1\times k}$) is defined as

$$C_{1\times K} = \{c_i = b_j; b_j \geq T\}, \tag{6}$$

where $i = 1, 2, 3, \dots, K;\ j = 1, 2, 3, \dots, L - M + 1$. The threshold T is defined as

$$T = \min\{a_i\}, \tag{7}$$

where $i = 1, 2, 3, \dots, N$. Now, the vector $C_{1\times K}$ contains all the possible weights greater than or equal to the threshold T. In other words, the vector $C_{1\times K}$ gives the locations of the subsequences in the genome that are of either exact or of high similarity to the given set of binding sites. As we want to only keep the locations of the subsequences in the genome that are of total similarity to the set of conserved sequences, all the values in $C_{1\times K}$ that are different from the set of weights in the $A_{1\times N}$ vector are discarded. The resulting weighting vector can be written as

$$D_{1\times P} = \{d_i = c_j; c_j = a_j\}, \tag{8}$$

where $i = 1, 2, 3, \dots, P; j = 1, 2, 3, \dots, K$ and $P \leq K$.

Step 4: TFBS Identification.

Although at this point of the algorithm we have drastically decreased the number of possible conserved sequences, we still have too many possible sequences belonging to the genome that are not real conserved sequences (false positives). The vector $D_{1\times P}$ contains all the possible weights that are exactly the same as the weights of the original conserved sequences. Some of these weights are false positive, i.e. they do not correspond to an actual conserved sequence. To filter these false positives out, each one of the detected sequences is compared to the original conserved sequences and only the verified ones are kept. The resulting distance vector $E_{1\times R}$ is evaluated as

$$E_{1\times R} = \{e_i = d_j; S(d_j) = X(d_j), \}, \tag{9}$$

where $i = 1, 2, 3, \dots, R;\ j = 1, 2, 3, \dots, P;\ S(d_j)$ is the sequence of length M in the genome corresponding to the j^{th} weight in the $D_{1\times p}$ vector. $X(d_j)$ is the conserved sequence in $X_{N\times M}$ having the same weight d_j since if these two sequences are the same, they have to have the same weight value.

3 Analysis and Simulation Results

In order to be able to handle the conserved sequences of all transcription factors investigated in this paper, a database (a cell array in MATLAB) is built where each TF is assigned a number. Since Escherichia coli (E. coli) is a well-studied organism with several highly accurately annotated genome sequences, it is used here as a test case. As such, a database of E. coli transcription factor binding sequences is produced. Table 1 shows 124 different E. coli transcription factors obtained from Regulon database [17], and Ecogene database [18]. Sequences were collated, redundancies were eliminated, and validated by identifying them in the E. coli genome in Ecogene.

Table 1. E. Coli Transcription Factors.

FN	TF	FN	TF	FN	TF	FN	TF
1	AcrR	32	ExuR	63	MalI	94	PspF
2	Ada	33	FNR	64	MalT	95	PurR
3	AgaR	34	FabR	65	MarA	96	QseB
4	AlaS	35	FadR	66	MarR	97	RbsR
5	AllR	36	FhlA	67	MelR	98	RcsAB
6	AraC	37	Fis	68	MetJ	99	RhaR
7	ArcA	38	FlhDC	69	MetR	100	RhaS
8	ArcR	39	FruR	70	MhpR	101	Rob
9	ArgR	40	Fur	71	MngR	102	RstA
10	AtoC	41	GadE	72	ModE	103	RutR
11	BaeR	42	GalR	73	MprA	104	SdiA
12	BetI	43	GcvA	74	MtlR	105	SgrR
13	BirA	44	GlcC	75	Nac	106	SlyA
14	CRP	45	GlpR	76	NagC	107	SoxR
15	CaiF	46	GntR	77	NanR	108	SoxS
16	Cbl	47	H-NS	78	NarL	109	TdcA
17	ChbR	48	HU	79	NarP	110	TdcR
18	CpxR	49	HcaR	80	NhaR	111	TorR
19	CsgD	50	HipB	81	NikR	112	TreR
20	CspA	51	HyfR	82	NorR	113	TrpR
21	CueR	52	IHF	83	NrdR	114	TyrR
22	CusR	53	IclR	84	NsrR	115	UhpA
23	CynR	54	IdnR	85	NtrC	116	UidR
24	CysB	55	IscR	86	OmpR	117	UlaR
25	CytR	56	KdgR	87	OxyR	118	UxuR
26	DcuR	57	KdpE	88	PaaX	119	XapR
27	DeoR	58	LacI	89	PdhR	120	XylR
28	DgsA	59	LexA	90	PepA	121	YiaJ
29	DnaA	60	LldR	91	PhoB	122	ZntR
30	EnvY	61	LrhA	92	PhoP	123	ZraR
31	EvgA	62	Lrp	93	PrpR	124	Zur

The proposed algorithm for TFBS detection is applied to two Escherichia coli bacterial genomes namely: MG1655 and O157:H7 E. coli strains (both forward and reverse strands were investigated). Figures 4-7 show the simulation results obtained when the algorithm was applied to MG1655 positive strand, MG1655 negative strand, O157:H7 positive strand and O157:H7 negative strand, respectively. The red color in these figures corresponds to the set of TFBSs detected in the non-coding regions, the blue color to the ones detected in the coding regions, and the green color to the ones overlapping between non-coding and coding regions.

The y-axis in Figures 4-7 represents the transcription factor number with some offset values introduced to distinguish the three set of detected TFBSs as located in the non-coding regions (marked in red or the ones below the 124 horizontal line) or in the coding regions (marked in green or above the 248 horizontal line) or overlapping in between (marked in blue or between 124 and 248 horizontal lines). The offset values are 124 and 248. In other words, any horizontal line below 124 will pass through all the TFBSs related to the transcription factor indexed by the y-axis value. If this horizontal line is between 124 and 248 (i.e. the middle region) a value of 124 should be subtracted from the y-axis value to know what transcription factor is being referred to. Finally, if the horizontal line is above 248 then a value of 248 should be subtracted. In this way, Figures 4-7 not only classify the detected TFBSs into three different sets but also tell which transcription factor is being referred to at each level.

According to the simulation results shown in Figures 4-7, Table 2 gives some statistical information of the detected TFBSs in terms of percentages. As can be observed, most of the detected TFBSs are located in the non-coding regions which is biologically relevant and agrees with theory [9] as the transcription factors which bind to the detected sites regulate the transcription of the adjacent genes located ahead.

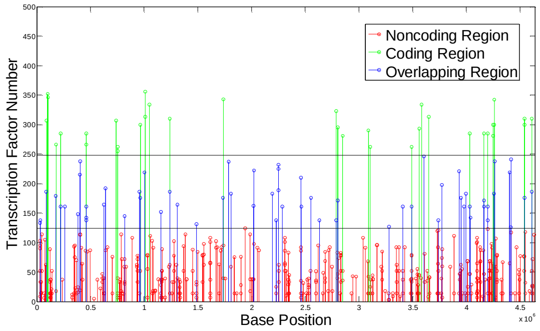


Fig. 4. TFBS detection using MG1655 positive strand

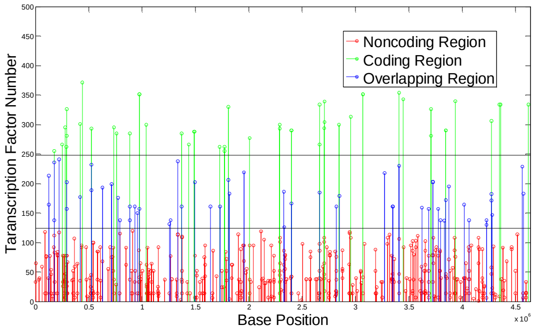


Fig. 5. TFBS detection using MG1655 negative strand

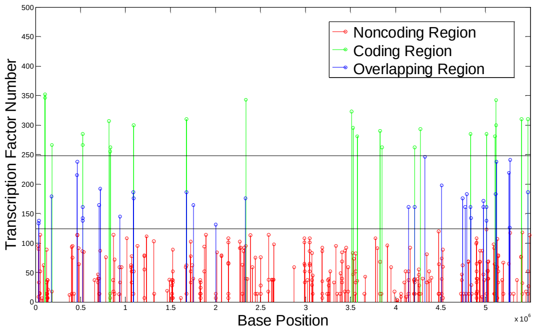


Fig. 6. TFBS detection using O157:H7 positive strand

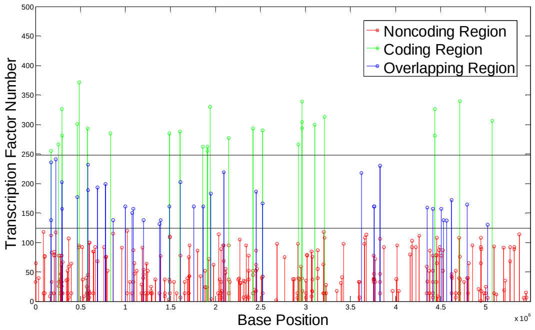


Fig. 7. TFBS detection using $O157$: $H7$ negative strand

Table 2. TFBSs detected in MG1655 and O175:H7 E. coli strains

E. coli Strain	MG1655		O157:H7	
Strand Orientation	Positive Strand	Negative Strand	Positive Strand	Negative Strand
Total number of TFBSs	956	1000	642	616
% of TFBSs in non-coding regions	85.46	85.90	86.15	85.23
% of TFBSs in the coding regions	6.49	6.20	6.85	5.84
% of TFBSs overlapping in between	8.05	7.90	7.00	8.93

3.1 Discovery of New Motif Sequences

In the gene expression process, it is very common that a single transcription factor binds not to an exact motif sequence but rather to a consensus motif. This fact is due to some bases belonging to the motif sequence are not as important as others when a transcription factor binds in the binding site. To consider this in the developed algorithm, the search for possible TFBSs was modified such that if the subsequence being tested is more that 80% (this number can be modified as required) similar to the original corresponding set of TFBSs, then it can be considered as a possible TFBS. For example, applying the algorithm to MG16655 E. coli genome (forward strand) to locate subsequences of similitude greater than or equal to 80% of their original corresponding sets of TFBSs, yields six other possible sequences. Table 3 gives a detailed description of these six sequences with their corresponding location in the genome, the transcription factor they belong to, the number of matching bases and the percentage of similitude of these sequences with their original corresponding transcription factor binding sites.

Table 4 show similar simulation results but with the percentage of similitude being set to 70%. By so doing, thirty two subsequences result. Therefore, this algorithm does not only provide locations of the sets of verified conserved sequences but also locations of all subsequences of the genome that are partially similar as well (with a predefined percentage of similitude as required). This freedom in introducing this percentage of similitude between the target sequences to be located and the original transcription factor binding sites allows for more flexibility in detection.

Table 3. Possible motif sequences based on greater than 80% similitude to their original sets of TF conserved sequences

#	Possible **TFBSs**	Base Position	TF #	TF	# of Matches	Similarity Percentage
1	'AGGCCUACGUUAAUUCUGCAAUAUAUUGAAUCU'	5593	52	FNR	29	87.879
2	'AGGCCUACGUUAUUUCAGCAAUAUAUUGAAUUU'	4.3021e+005	52	FNR	28	84.848
3	'AGGCCUACGUGAACUCUGCAAUAUAUUGAAUUU'	8.3153e+005	52	FNR	29	87.879
4	'AGGCCUACAUGAUUUCUGCAAUAUAUUGAAUUU'	3.7386e+006	52	FNR	28	84.848
5	'UGGUAUAACAGGUAUAAAGGUAUACA'	4.5012e+006	77	DcuR	24	92.308
6	'UUAUACCUGUUAUACCAGAUCAAUUA'	3.3713e+006	77	DcuR	24	92.308

Table 4. Possible motif sequences based on greater than 70% similitude to their original sets of binding sites

#	*Possible TFBSs*	*Base Position*	*TF #*	*TF*	*# of Matches*	*Similarity Percentage*
1	′AGGCCUACGUUAAUUCUGCAAUAUAUUGAAUCU′	5593	52	*IHF*	29	87.879
2	′AGGCCUACGUUAUUUCAGCAAUAUAUUGAAUUU′	4.3021*e* + 005	52	*IHF*	28	84.848
3	′AGGCCUACAUGAUCUCUGCAAUAUAUUGAGUUU′	7.0708*e* + 005	52	*IHF*	26	78.788
4	′AGGCCUACAUUUUCUCCGCAAUAUAUUGAAUUU′	8.1483*e* + 005	52	*IHF*	25	75.758
5	′AGGCCUACGUGAACUCUGCAAUAUAUUGAAUUU′	8.3153*e* + 005	52	*IHF*	29	87.879
6	′AGGCCUACAUGAUUUCUGCAAUAUAUUGAAUUU′	3.7386*e* + 006	52	*IHF*	28	84.848
7	′AGGCCAACGGUAGAAUUGUAAUCUAUUGAAUUU′	4.0476*e* + 006	52	*IHF*	24	72.727
8	′UUAAUUAAAAUGUUACGUGUUUAAUGU′	5.2862*e* + 005	53	*IclR*	19	70.37
9	′UAUUUAAAUUUUUUGUGCUUUUGUUUU′	1.2187*e* + 006	53	*IclR*	19	70.37
10	′ACAUUAAAAAUGAAACUUAUUAAAUUG′	1.5962*e* + 006	53	*IclR*	19	70.37
11	′UUGAUUAAAAAGGUAAAUAUUUAAAAU′	2.9025*e* + 006	53	*IclR*	19	70.37
12	′UAUUUUUUGUUUUUAUUUUUUAAAGGA′	3.4946*e* + 006	53	*IclR*	20	74.074
13	′UUAAUAAAAAUGAUGAAUGAUUUAGAC′	3.5807*e* + 006	53	*IclR*	19	70.37
14	′CUGUACAUAUCAUAGACUAAACGGAUAC′	1.7994*e* + 006	68	*MetJ*	20	71.429
15	′UGGUAUAACAGGUAUAAAGGUAUACA′	4.5012*e* + 006	77	*NanR*	24	92.308
16	′UUAUACCUGUUAUACCAGAUCAAUUA′	3.3713*e* + 006	77	*NanR*	24	92.308
17	′CAUAGAGGUUUAAUCCUUAUUCAGAGU′	2.4967*e* + 006	78	*NarL*	19	70.37
18	′GUCACUAUACAACGGACGGGGGAAGGA′	4.363*e* + 006	78	*NarL*	19	70.37
19	′AAUUGCAUUUAAAAAAUAUGUUCUGUG′	3.8409*e* + 005	78	*NarL*	19	70.37
20	′UGAUAUCACUAUAGAUAUUGAUCAUUA′	1.3899*e* + 006	78	*NarL*	19	70.37
21	′UAUUACAAUGUAAUCAUUAAUUGCUAA′	1.9563*e* + 006	78	*NarL*	19	70.37
22	′UGAACUCAGUAAGAGCAGUGAUAAUCA′	2.1632*e* + 006	78	*NarL*	19	70.37
23	′CUUAAUGAAGUACUCAAUAGAUUUGUU′	2.4831*e* + 006	78	*NarL*	19	70.37
24	′AUUGCUAAUGAAAAACAUCAAUCCAAC′	4.2313*e* + 006	78	*NarL*	19	70.37
25	′UUAGGUAUUGAUAACAAUCAAUAGUAC′	4.5015*e* + 006	78	*NarL*	19	70.37
26	′GUCAUGAUGGCGCAUUAUUUUGUGGUG′	5.3802*e* + 005	78	*NarL*	19	70.37
27	′UUACCGCUGGUGCCGCAGGUCAGUUUC′	1.3878*e* + 006	78	*NarL*	19	70.37
28	′UUACCCAUGAAGCGGUAGGUAAAUGUG′	1.8717*e* + 006	78	*NarL*	20	74.074

29	'GGUAUCAAACUUCUCUUUAAACAGAUA'	$4.2953e+006$	78	$NarL$	19	70.37
30	'CUAUUGCUUGUGCGGUAUUUGCCAAAA'	$1.8083e+006$	78	$NarL$	20	74.074
31	'UUUAAUAAAAAAAAGAUUAAGGGAUGA'	$5.834e+005$	78	$NarL$	19	70.37
32	'AUACUAUCACUACCCUUUUUUUACACA'	$3.6487e+006$	79	$NarP$	19	70.37

Based on the high similitude of the sequences detected in Tables 3 and 4 of their original corresponding sets of TFBSs, 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.

4 Conclusion

In this paper, a novel algorithm for transcription factor binding site (TFBS) detection is proposed. The algorithm is based on a position frequency matrix concept to design distance metrics that quantify the similitude between the set of conserved sequences and the entire DNA sequence under study. This algorithm does not use any type of mapping for the nucleobases and hence independent of mapping. The developed algorithm is applied to two different E. coli bacterial genomes (namely: MG1655 and O157:H7). Simulation results show that around 85% of the detected TFBSs are located in the non-coding region, 6.5% are located in the coding regions, and only 8.5% are overlapped between coding and non-coding regions. This shows that the proposed algorithm is not only able to efficiently identify and accurately locate the known TFBSs in their exact positions in the entire genome sequence, but also can be utilized in the discovery of new motif sequences based on their similitude to the known TFBSs with a predefined percentage of similitude as required.

Having analyzed only two E. coli genomes, six possible motif sequences were detected (as shown in Table 3) based on an 80% similitude constraint, while thirty two possible motif sequences were detected (as shown in Table 3) based on a 70% similarity constraint. Therefore, if this method of motif finding is applied to more other genomes as done in a previous work [1], other potential motif sequences can be identified. Hence, the previous analysis of using the position frequency matrix based algorithm to detect and identify new motif sequences can efficiently yield other potential motif candidates that are recommended 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 analyses here can help in the gene identification problem.

REFERENCES

1. Mohammad Al Bataineh, Lun Huang, and G. Atkin, "TFBS detection algorithm using distance metrics based on center of mass and polyphase mapping," in *2012 7th International Symposium on Health Informatics and Bioinformatics (HIBIT)*, Cappadocia, Turkey, 2012, pp. 37-40.

2. 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.
3. Sailu Yellaboina, Jayashree Seshadri, M. Senthil Kumar, and A. Ranjan, "PredictRegulon: a web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes," *Nucleic Acids Research 32 (Web-Server-Issue),* pp. 318-320, 2004.
4. Richard Münch, Karsten Hiller, Andreas Grote, Maurice Scheer, Johannes Klein, M. Schobert, and D. Jahn, "Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes," *Bioinformatics,* vol. 21, pp. 4187-4189, 2005.
5. G. D. Stormo, "DNA binding sites: representation and discovery," *Bioinformatics,* vol. 16, pp. 16-23, Jan 2000.
6. 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.
7. 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.
8. M. K. Das and H.-K. Dai, "A survey of DNA motif finding algorithms," *BMC Bioinformatics,* 2007.
9. C. Jia, M. B. Carson, and J. Yu, "A Fast Weak Motif-Finding Algorithm Based on Community Detection in Graphs," *BMC Bioinformatics,* vol. 14, 2013.
10. T. L. Bailey, "Discovering sequence motifs," *Methods Mol Biol,* vol. 452, pp. 231-51, 2008.
11. T. L. Bailey and C. Elkan, "The value of prior knowledge in discovering motifs with MEME," in *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, CA, 1995, pp. 21–29.
12. G. Z. Hertz and G. D. Stormo, "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences," *Bioinformatics,* vol. 15, pp. 563–577, 1999.
13. 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.
14. 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.
15. S. RA and S. DJ, "Inclusion of Neighboring Base Interdependencies Substantially Improves Genome-Wide Prokaryotic Transcription Factor Binding Site Prediction," *Nucleic Acids Research,* vol. 38, 2010.
16. E. L, J. VX, F. PJ, and J. SJ, "Locating Mammalian Transcription Factor Binding Sites: A Survey of Computational and Experimental Techniques," *Genome Res,* vol. 16, pp. 1455-1464, 2006.
17. E. van Nimwegen, "Scaling laws in the functional content of genomes," *Trends Genet,* vol. 19, pp. 479-84, Sep 2003.
18. K. E. Rudd, "EcoGene: a genome sequence database for Escherichia coli K-12," *Nucleic Acids Res,* vol. 28, pp. 60-4, Jan 1 2000.