Identification of Transcription Factor Binding Sites Using Statistical Distances Based on Correspondence And Discriminant Analyses

Lun Huang, Mohammad Al Bataineh, Alicia Fuente Acedo, Xiangyu Deng, Wei Zhang, G. E. Atkin, *senior Member, IEEE*

Abstract—This paper describes a new approach for locating signals related to Transcription Factor, such as promoter sequences, in nucleic acid sequences. Transcription Factor (TF) binding to its DNA target sites is a fundamental regulatory interaction in genome expression?. The most common model used to represent TF binding specificities is a position weight matrix (PWM) [1], which assumes independence between binding positions. However, in many cases, this simplifying assumption does not hold. In this paper, we present a Chi-Square (χ^2) and Mahalanobis distance models that are used in correspondence and discriminant analyses respectively [2]. These distances are based on the distances between the profiles of component vectors. It is a novel probabilistic method for modeling TF-DNA interactions. Our approach uses χ^2 and Mahalanobis distances to represent TF binding specificities. Simulation results show that the proposed approaches identifies TF binding sites significantly better than the PWM model method.

Index Terms—Transcription Factor, promoter, Chi-square, Mahalanobis distances.

1 INTRODUCTION

THE most common representation for sequence motifs is the position weight matrix (PWM), which specifies a separate probability distribution over nucleotides at each position of the Transcription Factor Binding Sites (TFBS). The goal of the computational approaches is then to identify the PWM associated with each TF and use it to identify the TFBS. A weight matrix is a two dimensional array of values that represent the scores for finding each of the possible bases at each position in the TF for which we are looking for. For DNA sequences the weight matrix will have a length equal to the length of the TF and depth of four (one row for each of A, C, G and T bases). Generally, we generate the frequency table for the TF and calculate the natural logarithms of the frequencies to get the position weight matrix.

Despite its successes, the PWM representation makes the strong assumption that the binding specificities of the TF's are position-independent. That is, the PWM assumes that for any given TF and TFBS, the contribution of a nucleotide at one position of the site to the overall binding affinity of the TF to the site does not depend on the nucleotides that appear in other positions of the site. It is easy to see where this assumption fails. For example, if the TFBS data contains only ''CG'' or ''GC'' in the center positions. Although the PWM learned from this data assigns high probabilities to these nucleotide pairs, it also undesirably (and unavoidably) assigns high probability to ''CC'' and ''GG'' in the center positions. However, if instead of the PWM representation, we allow ourselves to assign probabilities to multiple nucleotides at multiple positions; we could use the same number of parameters to specify the desired TF binding specificities. This observation leads to the feature motif model (FMM) [3] approach. Even though the FMM approach is better than the PWM, it involves the evaluation of complicated log likelihood and objective functions. Then, it significantly increases the computation complex-

ity.

In this paper, a novel identification approach is proposed. It uses a statistical model based on the Chi-Square χ^2 and Mahalanobis distances. This approach does not require large computation complexity, and simulation results show that it can effectively identify the location of the TFBS and other related signal sequences, such as promoters. This paper is organized as follows. In section II, the system model and underlying theories are described for the Chi-Squared and Mahalanobis distances; in section III, the simulation results are analyzed; section IV presents the conclusion of the proposed approach.

2 SYSTEM MODEL

2.1 χ^2 distance.

The system model is described in Fig. 1.

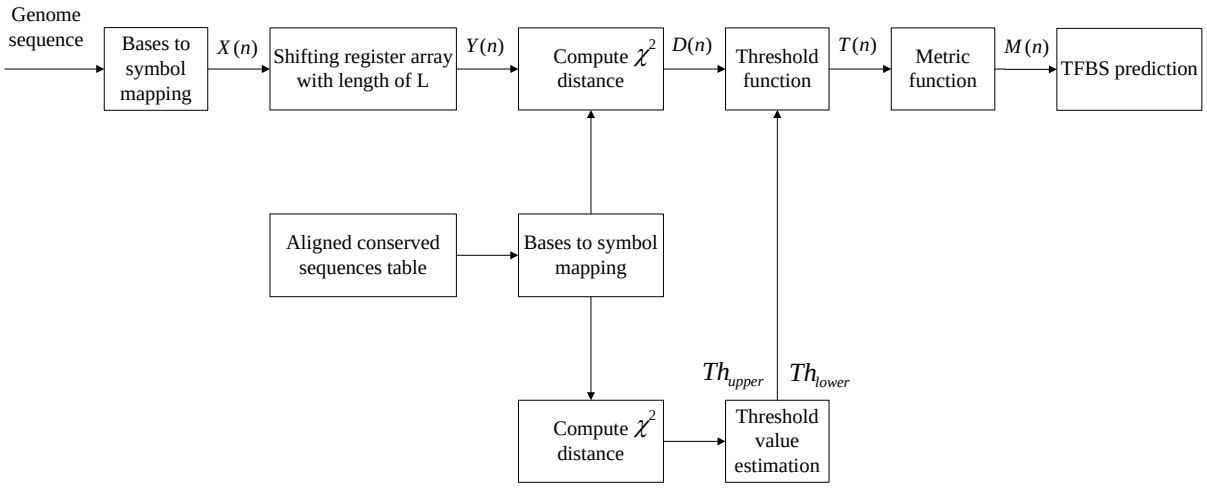

Fig. 1. System model for TFBS and signal detection based on χ^2 distance.

The output of the shifting register array is a vector Y(n) with L elements. The n is the location index of this vector on the nucleotide input sequence X(n). The χ^2 distance is originated from correspondence analysis. It is a distance between the statistical profiles of two different sequences or sets. A vector is called a profile when it comprises numbers greater or equal to zero whose sum is equal to one (such a vector is sometimes called a probabilistic vector). The χ^2 distance is defined for the rows (or the columns after transposition of the data table) of a contingency table. An example of the procedure to evaluate the χ^2 distance between the input sequence and the biological Center Of Gravity (COG) for the family composed of the conserved sequences in Table 1, is shown in Table 2.

It should be noted that the statistics of the row profiles for those conserved sequences are only taking into account the matched bases between the conserved sequences and input sequence [3] [4]. For example, for conserved sequence 3, it has 6 nucleotides coinciding with the input sequence, thus, only 6 of its bases would be used to evaluate its statistics and the total number of matches is 6.

TABLE 1
THE ORIGINAL CONSERVED AND INPUT SEQUENCES

Sequence name	Nucleotide sequences
Conserved 1	TCAATAGCAGTGTGAAATAACATAATTGAGCAACTGAA
Conserved 2	AGCGCACACTTGTGAATTATCTCAATAGCAGTGTGAAA
Conserved 3	TCAAGAAATAAACCACAAAATCGTAATCGAAAGATAAAA
Conserved 4	GTAATCGAAAGATAAAAATCTGTAATTGTTTTCCCCTG
Input Sequence	GTTTCCTGATGAACATTTTTCCAGCAATTACACCTCTG

TABLE 2
THE CONTINGENCY TABLE FOR COMPUTING THE χ^2 DISTANCE

(A) THE RAW DATA AND COLUMN STATISTICS

Raw data						
	A	G	C	T	$N' \cdot v$	$(N - N') \cdot r$
Conserved 1	3	0	3	2	8	30
Conserved 2	3	0	2	4	9	29
Conserved 3	3	0	2	1	6	32
Conserved 4	3	3	4	4	14	24
Input Sequence	9	5	10	14	38	0
ΣNC^T	21	8	21	25	75	
C^T	0.28	0.107	0.28	0.333		
D^T	3.571	9.375	3.571	3		

(B) THE STATISTICS OF ROW PROFILES

Row profiles					
	A	G	C	T	r_i
Conserved 1	0.375	0	0.375	0.25	0.261
Conserved 2	0.333	0	0.222	0.444	0.252
Conserved 3	0.5	0	0.333	0.167	0.278
Conserved 4	0.214	0.214	0.286	0.286	0.209
Input Sequence	0.237	0.132	0.263	0.368	

The row labeled ΣNC^T gives the total number of matches for the respective bases (sum of each column A, G, C and T). This is the total number of times this specific nucleotide was found in the matched bases. The centroid row C^T gives the proportion of each kind of nucleotide (A, G, C, T) in the conserved sequences with respect to all matches ($\sum_{A,G,C,T} \Sigma NC^T$). The weight of each column D^T is the inverse of the centroid. The column labeled $N' \cdot v$ gives the total number of matches used to evaluate the statistics by each sequence, where $N' = 75$ is the total number of matches, v is the vector composed of the ratio of matches, v_i, for each of the conserved sequences. The mass of each row, denoted by r_i, is the ratio of mismatches of this sequence to the total number of mismatches for all of the conserved sequences, r is the vector composed of the r_i's. In Table 2(b), N=190 is the total number of nucleotides in all of the conserved and input sequences.

The first step of the computation of the distance is to transform the raw data for row statistics into row profiles which is obtained by dividing each row by its total. There are I rows and J columns in a contingency table. The statistical COG of the rows, denoted by C^T, is computed by transforming the total of the columns into a row profile. For the χ^2 distance, the W matrix is diagonal which is equivalent to assigning a weight to each column. This weight is equal to the inverse of the relative frequency of the column; it can be expressed formally by $\mathrm{diag}\{W^{-1}\} = C^T$.

With this weight matrix, nucleotides which have more matches contribute less to the distance between rows than nucleotides which have fewer matches. For example, from Table 2 (a), we find that the weight matrix is equal to

$$D^T = \mathrm{diag}\{W\} = \begin{bmatrix} 0.28^{-1} & 0 & 0 & 0 \\ 0 & 0.107^{-1} & 0 & 0 \\ 0 & 0 & 0.28^{-1} & 0 \\ 0 & 0 & 0 & 0.333^{-1} \end{bmatrix} = \begin{bmatrix} 7.238 & 0 & 0 & 0 \\ 0 & 19 & 0 & 0 \\ 0 & 0 & 7.238 & 0 \\ 0 & 0 & 0 & 6.08 \end{bmatrix} \tag{1}$$

Assuming that $d^2(i)$ denotes the χ^2 distance between the conserved sequence i and the input sequence, then the χ^2 distance between the conserved sequence i=1 and the input sequence is equal to

$$\begin{aligned} d^2(1) &= (S_i - S_1)W(S_i - S_1)^T \\ &= 7.238 \cdot (0.237 - 0.375)^2 + 19 \cdot (0.132 - 0)^2 + \\ &\quad 7.238 \cdot (0.263 - 0.375)^2 + 6.08 \cdot (0.368 - 0.25)^2 \\ &= 0.317 \end{aligned} \tag{2}$$

Where S_i and S_1 are respectively the vectors corresponding to the input and the conserved sequence 1 for each base in Table 2(b).

In the same way, we obtain $d^2(2) = 0.219$, $d^2(3) = 0.549$, $d^2(4) = 0.088$. If M is the number of conserved sequences; L is the length of the conserved nucleotide sequences; N is equal to $(M+1) \cdot L$ which is the total number of nucleotides in all of the conserved and input sequences, the distance from conserved sequence i (row i) to the statistical COG of the family is denoted by $d_g^2(i)$, and the distance from row i to row i′ is denoted by $d^2(i,i')$, we obtain the following equality [2]:

$$\sum_{i=1}^{M} r_i d_g^2(i) = \sum_{i>i'} r_i r_{i'} d^2(i,i') \tag{3}$$

Where i, $i' \in$ (index of the conserved). r_i and $r_{i'}$ are the mass of each row, and are the components of the mass vectors r, which can be obtained by dividing the vector $(N - N') \times r$ by the scalar$(N - N')$.

An approximate estimation of the χ^2 distance between the input sequence and the biological COG for the conserved sequence family can be derived from (3) as

$$d_g^2 \cong \sum_{i=1}^{M} r_i d^2(i) \cong \frac{1}{M} \sum_{i=1}^{M} d^2(i) \tag{4}$$

Where $i \in$ {index of conserved sequences}, $d^2(i)$ is the χ^2 distance between the input sequence and the conserved sequence i. So, this distance d_g^2 depends on all of the conserved sequences. To further reflect the match degree between input sequence and the conserved sequences family, which is related to N', d_g^2 should be normalized by a factor: $A = (M+1) \cdot L / N' = N / N'$.

Then, the normalized χ^2 distance from the input sequence to the biological COG of the conserved sequences can be de-

fined by

$$D(n) = A \cdot d_g^2(n) \cong A\sum_{i=1}^{M} r_i d^2(i) \\ = \frac{(M+1)\cdot L}{N'}\sum_{i=1}^{M} r_i d^2(i) \cong \frac{N}{M \cdot N'}\sum_{i=1}^{M} d^2(i) \tag{5}$$

Where n denotes the index of the location on the input nucleotide sequence. Therefore, for the initial location of the input sequence (n = 0) in Table 2, we have

$$D(0) = \frac{5\cdot 38}{75}\sum_{i=1}^{4} r_i d^2(i) = \frac{190}{75}(0.261\cdot 0.317 \\ +0.252\cdot 0.219 + 0.278\cdot 0.549 + 0.209\cdot 0.088) = 0.783 \tag{6}$$

To define the dynamic range of the χ^2 distance D(n) for the input nucleotide sequence, we must evaluate lower and upper thresholds based on the conserved sequences. The upper threshold can be obtained with

$Th_{upper} = \max\{D_j(0)\}$, $j \in$ {index of conserved sequences}

The lower threshold is

$Th_{lower} = \min\{D_j(0)\}$, $j \in$ {index of conserved sequences}

Where $D_j(0)$ denotes the normalized χ^2 distance from the conserved sequence j to the biological COG of the conserved sequences.

For the conserved sequences given in Table 1, we find that the χ^2 distances ($D_i(0);\ i = 1,2,3,4$) for each of the four conserved sequences to the biological COG are

$D_1(0) = 0.191$, $D_2(0) = 0.62$, $D_3(0) = 0.174$, $D_4(0) = 0.66$

So, the upper threshold value is $Th_{upper} = \max\{D_j(0), j = 1,2,3,4\} = 0.66$; The lower threshold value is $Th_{lower} = \min\{D_j(0), j = 1,2,3,4\} = 0.174$. The output of the threshold function can be defined by

$$T(n) = \begin{cases} D(n), & Th_{lower} \le D(n) \le Th_{upper} \\ Th_{upper}, & \text{otherwise} \end{cases} \tag{7}$$

With T(n), the metric function that represents the probability of Transcription Factor Binding Site is defined as

$$M(n) = \frac{Th_{upper}}{T(n)} - 1 \tag{8}$$

The identification of the TFBS is based on the peak detection on the value of M(n).

2.2 The χ^2 distance based on the position weight matrix

To reflect the base bias on each position, we can use the position weight matrix (PWM) in the procedure of constructing the contingency table, which is then used to compute the χ^2 distance.

We still use the family and the input sequence in Table 1 to illustrate the approach of computing the χ^2 distance based on the PWM. The first step is to construct the PWM from the conserved sequences. Table 3 gives the PWM table for the family in Table 1. In fact, it describes the frequency of each base appearing on each position.

Then the contingency table for the computation of the χ^2 distance can be obtained by mapping the conserved sequences and input sequence with PWM. The PWM-mapping contingency table for the family in Table 1 is shown in Table 4. With Table 3, it is easy to obtain the corresponding vectors C^T, D^T, and $N' \cdot v$ in the same way as Table 2(a). The statistics of the row profiles for the conserved sequences in Table 4 can be obtained using the same approach as in Table 2(b). It should be noted is that the vector v instead of r is used in evaluating the PWM χ^2 distance. As an example, the distance $d^2(1)$ in (2) can then be re-evaluated as follows.

D^T =[55.67, 55.67, 33.4, 33.4, 55.67 33.4, 55.67, 33.4, 55.67, 55.67, 55.67, 41.75, 33.4, 55.67, 20.875, 20.875, 33.4, 41.75, 33.4, 55.67, 33.4, 55.67, 33.4, 20.875, 20.875, 20.875, 55.67, 20.875, 55.67, 55.67, 83.5, 55.67, 55.67, 41.75, 83.5, 55.67, 33.4, 33.4];

S_1 =[0.022, 0.022, 0.034, 0.034, 0.022, 0.034, 0.022, 0.011, 0.022, 0.011, 0.022, 0.022, 0.034, 0.022, 0.045, 0.045, 0.034, 0.022, 0.034, 0.011, 0.034, 0.011, 0.034, 0.045, 0.045, 0.045, 0.022, 0.045, 0.022, 0.011, 0.011, 0.011, 0.022, 0.022, 0.011, 0.011, 0.034, 0.034];

S_i =[0.027, 0.027, 0, 0, 0.027, 0.027, 0, 0, 0.054, 0.027, 0.027, 0.054, 0, 0.027, 0.108, 0, 0.027, 0.054, 0.027, 0.054, 0.081, 0, 0, 0, 0, 0, 0.027, 0, 0.027, 0.054, 0.027, 0.027, 0.027, 0.054, 0.027, 0.027, 0.027, 0.027];

v = [0.266, 0.237, 0.263, 0.234];

Therefore,

$$d^2(1) = (S_i - S_1)W(S_i - S_1)^T = 0.177 \tag{9}$$

With $D(n) = d_g^2(n) \cong \sum_{i=1}^{M} r_i d^2(i)$

Then D(0) in (6) can be evaluated as

$$\begin{aligned} D(0) &= (0.266 \cdot 1.045 + 0.237 \cdot 0.958 \\ &+ 0.263 \cdot 1.024 + 0.234 \cdot 1.297) = 1.078 \end{aligned} \tag{10}$$

The threshold and metric function values can be obtained by using the same procedure given by (7) and (8).

2.3 Mahalanobis distance

The χ^2 method, used in our novel approach, is one of the most well-known generalized Euclidean distances. This algorithm shares important characteristic with another generalized Euclidean distance method known as the Mahalanobis distance.

This distance is widely used in many different fields when data mining and pattern recognition tasks involve calculating abstract distances between items or collections of items. It is based on correlations between variables by which these different patterns can be identified and analyzed.

The difference between the χ^2 distance and the Mahalanobis distance is the way of computing the weight matrix W. In the previous two χ^2 distance methods, the weight matrix W is a diagonal matrix with diagonal vector D^T, the other entries are 0. In the Mahalanobis distance, $W = S^{-1} = E^{-1}[F^T F]$, where the matrix F is comprised of the conserved sequences as shown in Table 2(a) and E[*] is the expectation operator, S is the covariance matrix between the columns of the data table. The weight matrix is obtained as the inverse of the columns of the covariance matrix. So, as an example, the matrix W for Table 2(a) would be a 4×4 matrix. Then, the Mahalanobis distance can be obtained in the same way as (9) and (10). The threshold and metric function value can be evaluated by using the same procedure used in (7) and (8).

Therefore the system model is the same as the one described previously in Fig. 1 except that the χ^2 distance is substituted with the Mahalanobis distance

TABLE 3
THE POSITION WEIGHT MATRIX (PWM)

p*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
N	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
A	1	0	3	3	0	3	1	3	2	2	1	2	0	1	4	4	3	2	3	1
C	0	2	1	0	1	1	1	1	1	0	0	0	1	1	0	0	0	0	0	1
G	1	1	0	1	1	0	2	0	0	1	1	2	0	2	0	0	0	0	0	0
T	2	1	0	0	2	0	0	0	1	1	2	0	3	0	0	0	1	2	1	2

p*	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
N	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
A	0	1	0	4	4	0	1	0	2	2	1	1	2	0	1	2	3	3
C	3	0	1	0	0	0	1	0	1	0	1	0	1	2	1	1	0	0
G	0	2	0	0	0	0	0	4	0	1	1	1	1	0	1	1	0	1
T	1	1	3	0	0	4	2	0	1	1	1	2	0	2	1	0	1	0

TABLE 4
THE PWM-MAPPING CONTINGENCY TABLE

Raw data													
p	1	2	3	4	5	6	7	8	9	10	11	12	13
Conserved 1	2	2	3	3	2	3	2	1	2	1	2	2	3
Conserved 2	1	1	1	1	1	3	1	3	1	1	2	2	3
Conserved 3	2	2	3	3	1	3	1	3	1	2	1	2	1
Conserved 4	1	1	3	3	2	1	2	3	2	2	1	2	3
ΣNC^T	6	6	10	10	6	10	6	10	6	6	6	8	10

Raw data													
p	14	15	16	17	18	19	20	21	22	23	24	25	26
Conserved 1	2	4	4	3	2	3	1	3	1	3	4	4	4
Conserved 2	2	4	4	1	2	3	2	3	1	1	4	4	4
Conserved 3	1	4	4	3	2	3	2	3	2	3	4	4	4
Conserved 4	1	4	4	3	2	1	1	1	2	3	4	4	4
ΣNC^T	6	16	16	10	8	10	6	10	6	10	16	16	16

Raw data												
p	27	28	29	30	31	32	33	34	35	36	37	38
Conserved 1	2	4	2	1	1	1	2	2	1	1	3	3
Conserved 2	1	4	1	2	1	2	1	2	1	2	3	3
Conserved 3	1	4	2	2	1	1	2	2	1	2	3	3
Conserved 4	2	4	1	1	1	2	1	2	1	1	1	1
ΣNC^T	6	16	6	6	4	6	6	8	4	6	10	10

* p is the index of position

metric.

The analysis proves to be interesting. Although the Mahalanobis distance produces similar results as those of the χ^2 distance, it takes into account the inter-correlations between each column vector in contingency table, which is not reflected in the χ^2 distance. This information was expected to provide us with more details about the profiles and support the χ^2 distance model.

The Mahalanobis distance of a multivariate $x = (x_1, x_2, x_3, \ldots, x_N)^T$ vector from a group of vectors with mean $\mu = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)^T$ and covariance matrix S is defined as

$$D_M(x) = \sqrt{(x-\mu)^T W (x-\mu)} = \sqrt{(x-\mu)^T S^{-1} (x-\mu)} \tag{11}$$

Continuing with the previous example (Table 2) considered for the χ^2 distance between the input sequence and the biological Center Of Gravity (COG) for the family comprised of the conserved sequences in Table 1, the two matrixes involved in the new procedure for evaluating the distances are:

$$x = \begin{bmatrix} 9 & 5 & 10 & 14 \end{bmatrix}, \quad \mu = \begin{bmatrix} 3 & 0 & 3 & 2 \\ 3 & 0 & 2 & 4 \\ 3 & 0 & 2 & 1 \\ 3 & 3 & 4 & 4 \end{bmatrix}$$

The mean values are obtained from the matched bases included in the conserved sequences to which the input sequence x, the multivariate input vector, is going to be compared with.

This method places higher weights to the components of input sequence which are more distant to the mean. Therefore, the final identification results locate genome sequences which have higher similarities to the fixed statistical COG vector of the conserved sequences family and the variance of the peak amplitude increased.

This method can be used with or without the PWM matrix and the results are consistent to the ones obtained with the χ^2 distance. The procedure to compute the position weight matrix (PWM) for the Mahalanobis distance is the same except that the matrix F is obtained only with the conserved sequence vectors in Table 4. The PWM Mahalanobis distance model provides a distance metric between genome sequences in terms of the bases in the same position.

3 SIMULATION

3.1 Simulation based on the χ^2 distance

The conserved sequences table used to locate promoters in E. coli sequences is taken from the compilation of such sequences produced by Hawley and McClure [5]. E. coli promoters have been shown to contain 2 regions of conserved sequence located about 10 and 35 bases upstream of the transcription start-site. Their consensuses are TATAAT and TTGA-

CA with an allowed spacing of 15 to 21 bases between. The spacing with maximum probability is 17 bases and all but 12 of the 112 sequences in the Hawley and McClure collection could be aligned with a separation of 17 + or -l bases. The spacing between the -10 region and the start-site is usually 6 or 7 bases but varies between 4 and 8 bases. Hawley and McClure also show a conserved section to exist around the +1 region. The range definitions for the three regions (the -35, -10 and +1 identify the -10 promoter, the result is shown in Fig. 2. The red region marks the real location of the -10 promoter, the center of red region is at position 101. It can be observed that the highest peak locates at 125, which is 24 bases from the center of red region; therefore, the predicted location of the -10 promoter should be 115 and the identification error is 14 bases. In the -35 promoter identification, we use the -35 region of the conserved sequences to form the conserved sequence table, the result is shown in Fig. 3.

Similarly, the red region indicates the real location of -35 promoter, it can be observed that the highest peak appears around the 62 position, so the predicted location for the -35 promoter should be 27. Since the center of the red region is at 78, the identification error is 51 bases. However, if we consider the secondly highest peak, which locates at 121 and 43 bases away from the center of the red region, the predicted location is 86 and the identification error is 8 bases. In fact, there is a restriction enzyme Taq-I recognition site at location 63, that it might be responsible for the highest peak. We are currently researching on this interesting result that indicates the potential of the algorithm to identify other signal sequences.

using both -35 and -10 region of the conserved sequences to form the conserved sequence table, we obtain a more accurate identification result as shown in the following figure. The red region indicates the real location of the -35 promoter and the green region indicates the -10 promoter. In the same way, with the highest peak located at 116, it can be discovered that the identification errors for -35 and -10 promoter are 3 and 5 respectively. They are far smaller than those corresponding identification errors in the two previous cases. This is due to the use of longer conserved sequences. The use of longer conserved sequences provides more reliable statistical information, thus a better identification performance.

As a comparison, similar results are described in Fig. 5c in [1]. In case of identifying -10 promoter with the PWM model method, since the peak around the real location, which is 111, is not higher than the subsidiary peak 40 base-pairs upstream, the identification result of -10 promoter must combine with that of -35 promoter to achieve the real location. It involves a complex procedure of optimization and the choice of certain criteria. However, our proposed approach does not require any optimization operation, and it is faster and more accurate.

We also used the proposed approach to a Listeria genome. The conserved and input nucleotide sequences can be found in [7]. By the same way used in [5], [6], we can construct a conserved sequence table, which is given in the Appendix. The

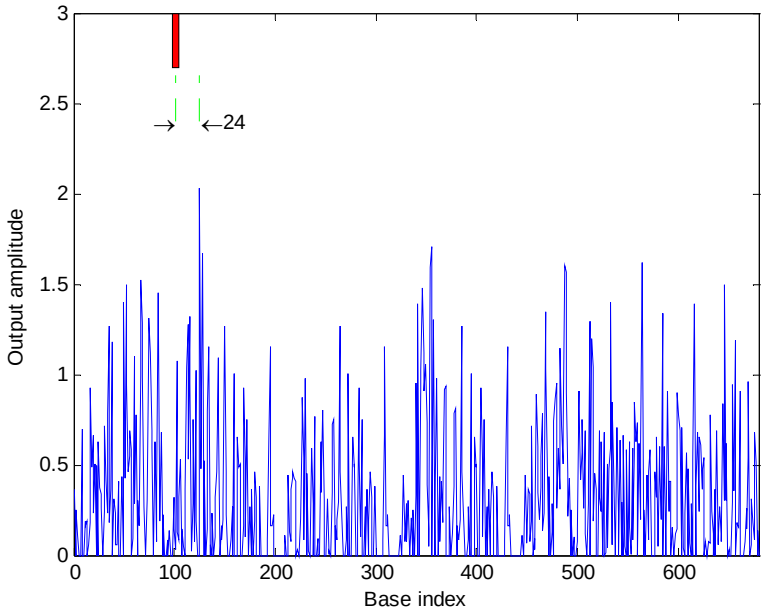

Fig. 2. -10 promotor identification.

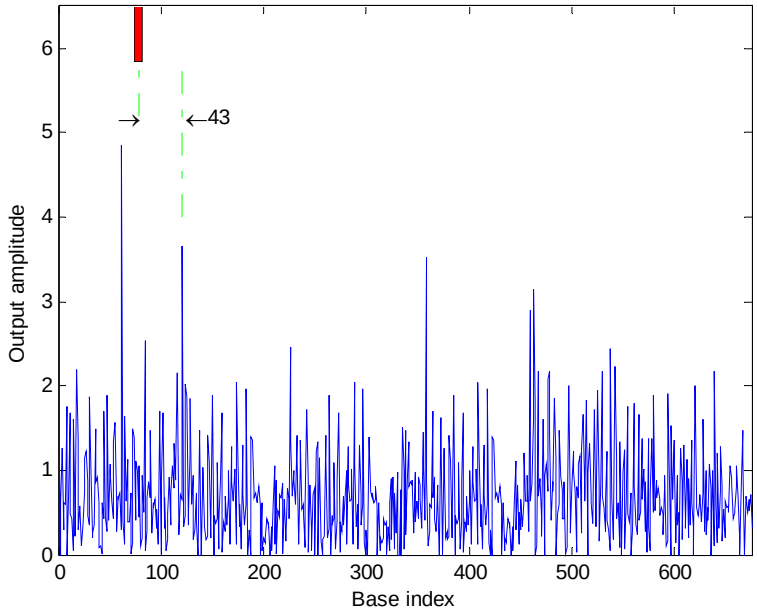

Fig. 3. -35 promotor identification.

conserved sequences are extracted from the non-coding area of the operons and aligned with [8]. The genome sequence used in our simulation includes a 'plcA' gene and a -10 promoter. The range of conserved sequences for detecting the -10 region is from 44 to 79. The result is show in Fig. 5.

The red region indicates the real location of -10 promoter in the genome, the highest peak is 4 bases from the center of red region, so the identification error is 6 bases.

3.2 Simulation based on the position weight matrix (PWM) χ^2 distance

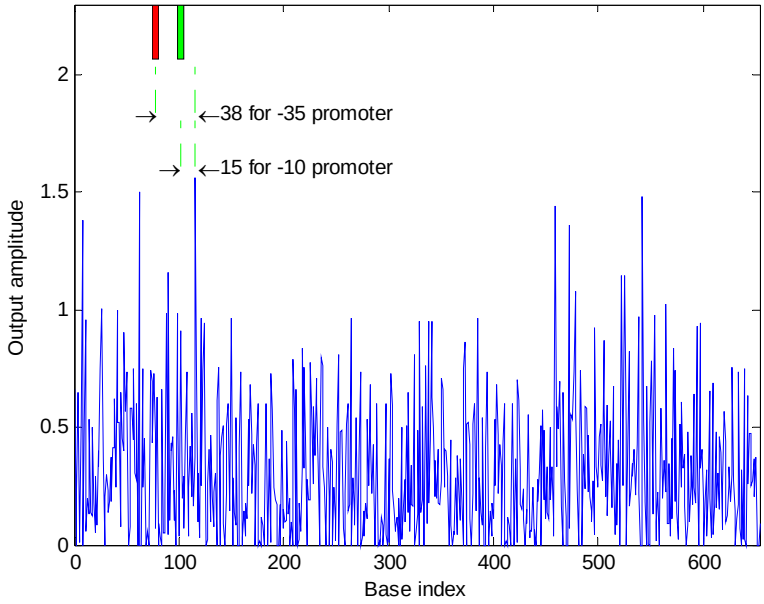

Fig. 4. -35 and -10 promotors identification.

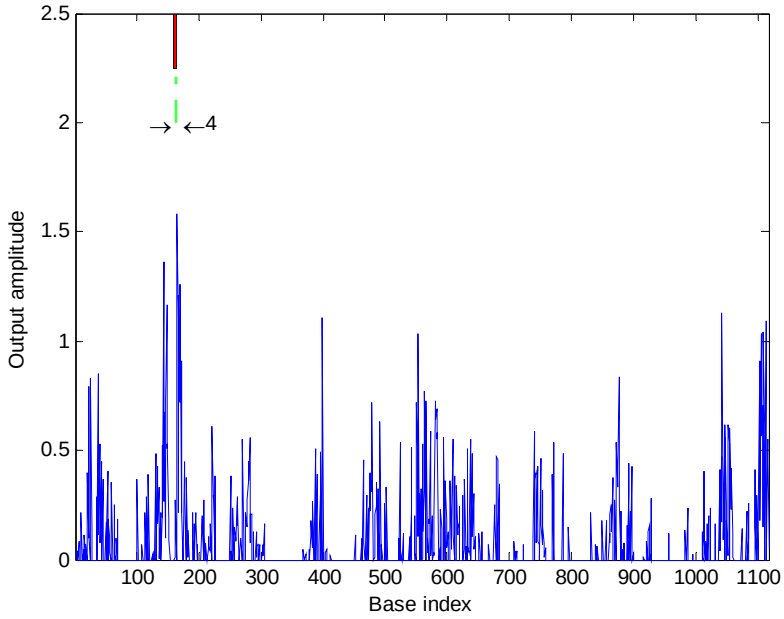

Fig. 5. -10 promotor identification for Listeria.

In the same way, the simulation results of the promoter identification based on the PWM χ^2 distance are shown in Fig. 6 and Fig. 7. Fig. 6 (a) describes the identification result of the -10 promoter in the same E.coli genome as in Fig. 2, and Fig. 6 (b) for the -35 promoter.

It can be observed that the results are very similar to those in Fig. 2 and Fig. 3. The identification errors are 5 and 8 bases respectively. The identification of -10 promoter is more accurate than the previous approach, while the same interference

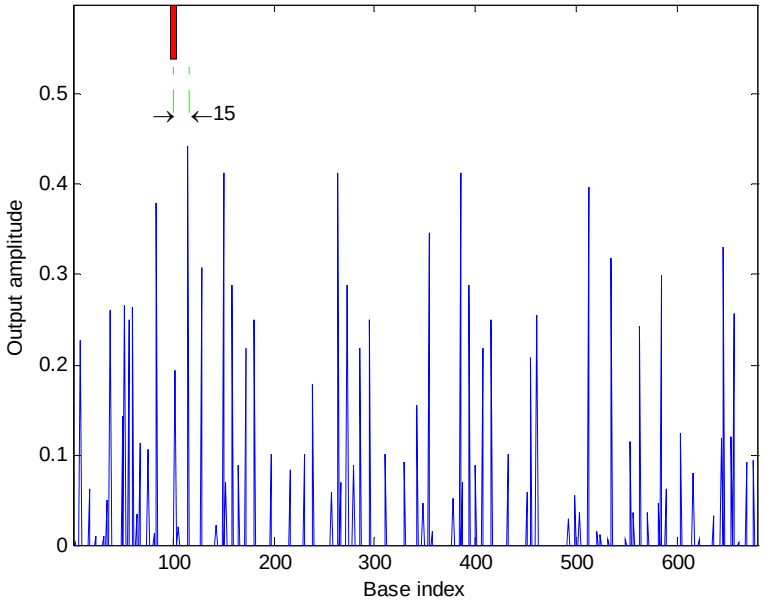

Fig. 6(a). -10 promotor identification.

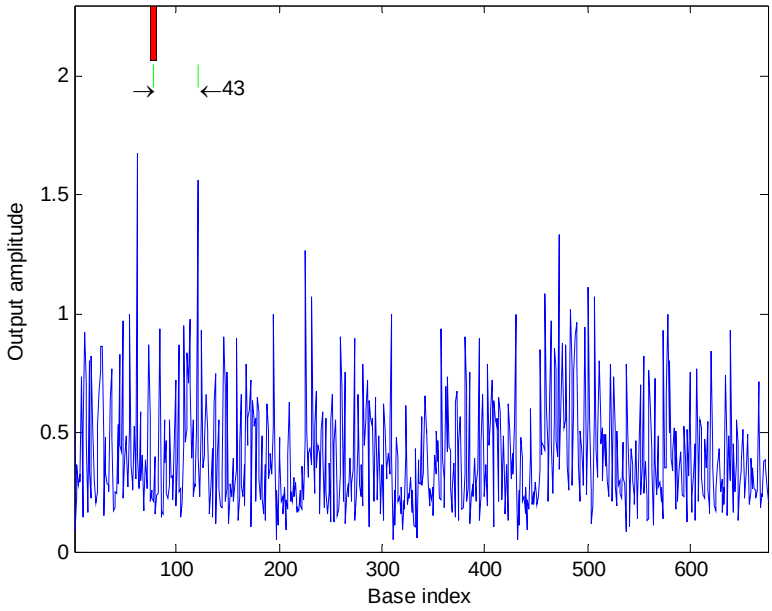

Fig. 6(b). -35 promotor identification.

peak still exists in -35 promoter identification result.

The promoter identification result for Listeria based on PWM χ^2 distance is as Fig. 7.

Since the highest peak locates 7 bases away from the center of -10 region, the error is 3 bases.

3.3 Simulation based on the Mahalanobis distance

The results of the simulation corroborate the previously obtained peaks with the χ^2 distance model, and the identifica-

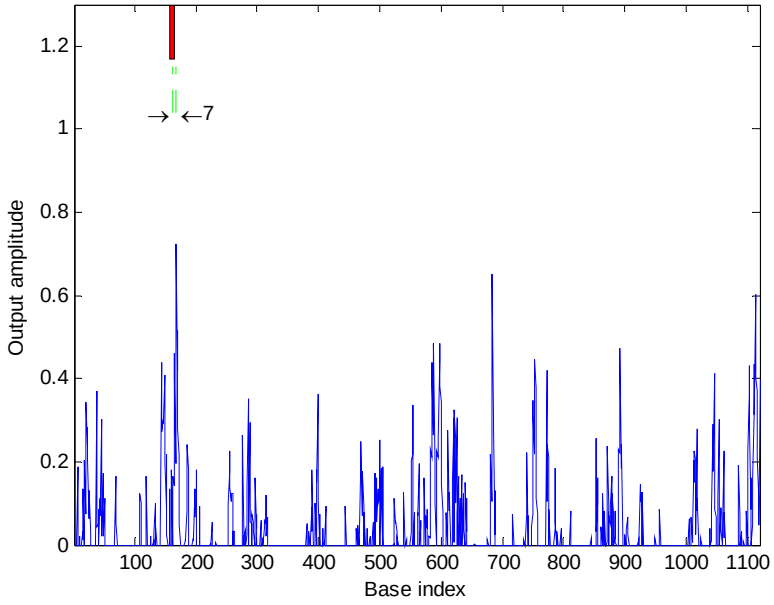

Fig. 7. -10 promotor identification for Listeria.

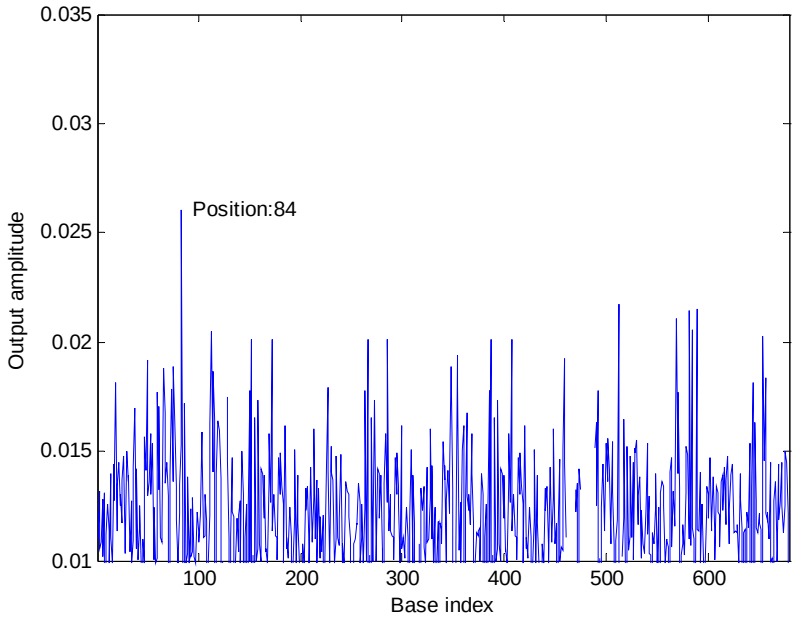

Fig. 8(a). -10 promotor identification.

tion errors are maintained. The main difference between these two methods is mainly the amplitude of the peaks obtained, and they strongly related to the correlation between bases in different positions for the Mahalanobis method and therefore affecting directly to the distance measure.

Fig. 8.a and 8.b shows the results obtained using the Mahalanobis distance focusing specifically on the matched bases between the conserved sequences and the input sequence (no PWM weighting). We observe that the highest peaks appear around the position of restriction enzyme Taq-I recognition site as happened in Fig. 3. It means that there is significant

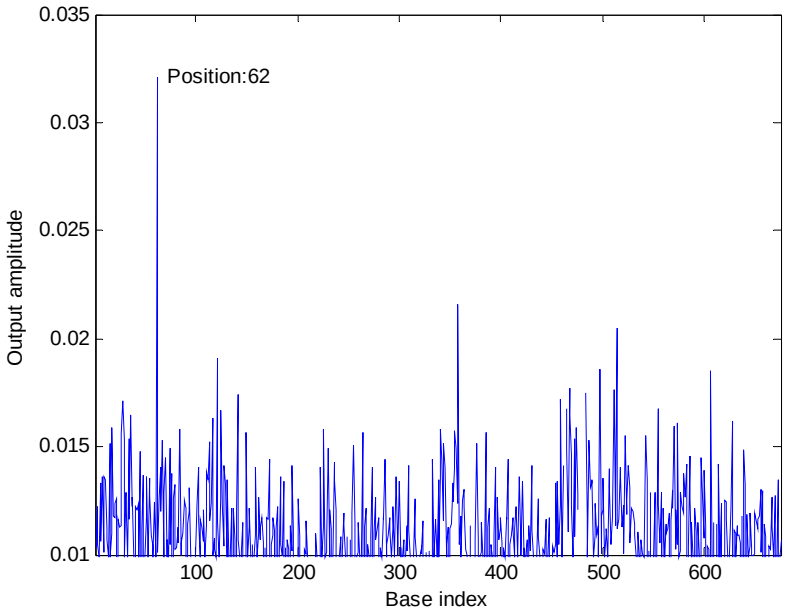

Fig. 8(b). -35 promotor identification.

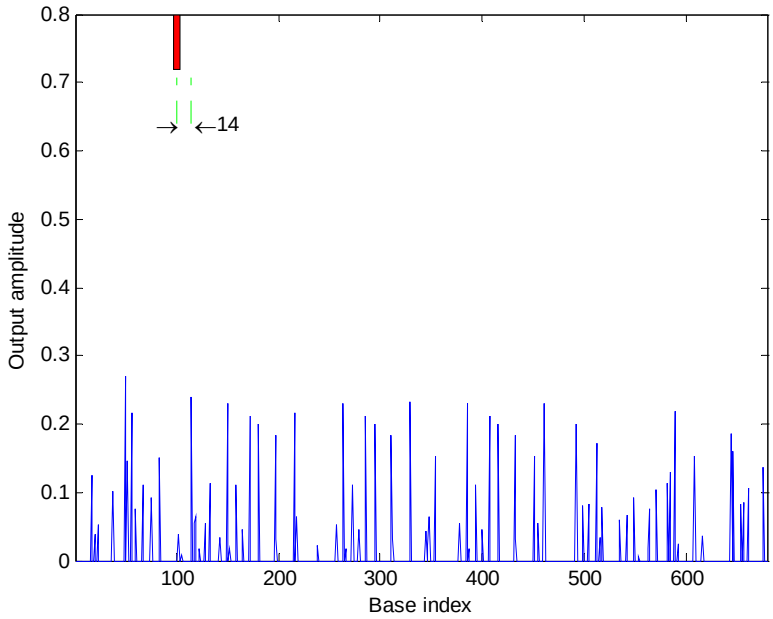

Fig. 9(a). -10 promotor identification.

similarity between this recognition site and the conserved region of the regulon. This phenomenon is also corroborated by the observation of the simulation results provided above.

3.4 Simulation based on the position weight matrix (PWM) Mahalanobis distance

To reflect the base bias on each position, the position weight matrix (PWM) can be used in the procedure of constructing the contingency table, which will then be used to compute the Mahalanobis distance, as it was previously done with the χ^2 distance. Therefore the contingency table for the computation of the new distance can be obtained by mapping the

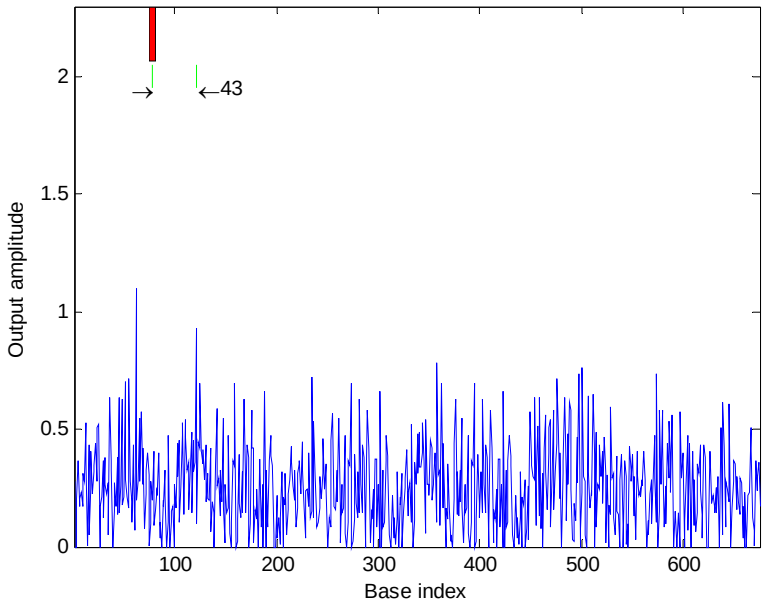

Fig. 9(b). -35 promotor identification.

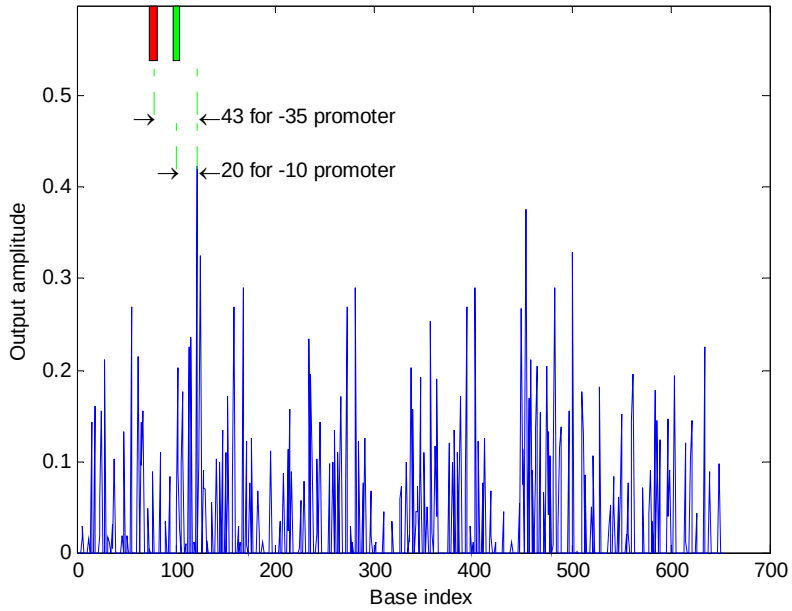

Fig. 10. 35 and -10 promotors identification.

conserved sequences and input sequence with PWM.

The simulation results of promoter identification based on the PWM Mahalanobis distance are shown in Fig. 9 and Fig. 10. In Fig. 9(a) and 9(b), there are interference peaks around position 62. They might be also due to the enzyme Taq-I recognition site as happened in the previous simulation cases. If we consider the second largest peak, the error for -10 and -35 identification would be 4 and 8 respectively. Fig. 10 shows the case of using both -35 and -10 region of the conserved sequences to form the conserved sequence table, the error for -10 and -35 identification would become 10 and 8, but the in-

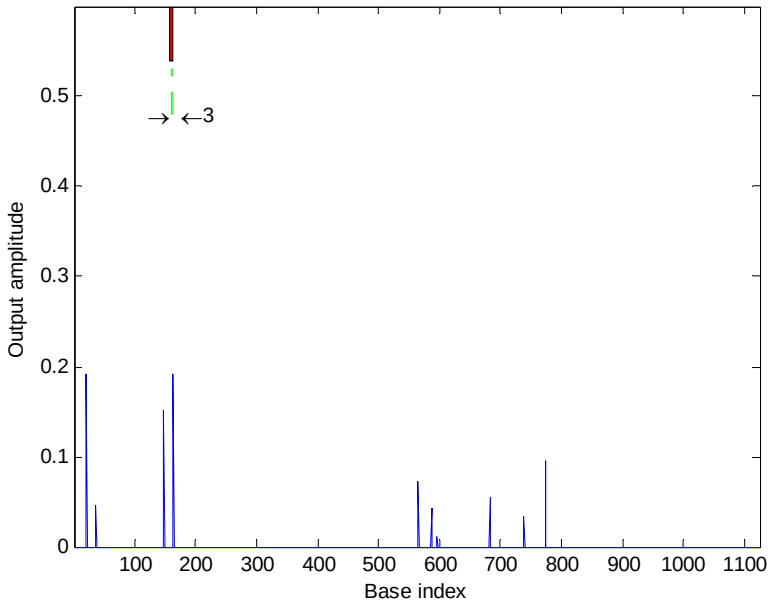

Fig. 11. 10 promotor identification for Listeria.

terference peak disappeared.

Fig. 11 describes the -10 promoters identification on Listeria. The identification error is 7. Although there are less interference peaks, it can only be obtained when the conserved sequence is shortened. It means that Mahalanobis distance is more likely to be influenced by noise (randomly distributed bases).

4 Conclusion

In this paper, we have introduced a novel TFBS identification algorithm, which is based on the χ^2 and Mahalanobis distance model. The proposed approach is more efficient than traditional methods based on log-linear model, such as FMM [3], because it waives the need to evaluate highly complicated log likelihood function and objective function. Since this new approach takes into account the position-dependence of the TF motifs in computing the χ^2 and Mahalanobis distance, it also brings about significant performance improvement over the PWM [1] model method. By analyzing the simulation result, we have shown that, it is possible to obtain an accurate identification of the TFBS and related signals in the genome.

Appendix

Gene name*

```
dnaA      '-------AGCTTTTTCTATCTGTGGATAACTTTATAGCATCCATTTACATTACATAAAAAGGGGGGGTACTA-------'
lmo0266   '-------GGCTTTTTTCTTGCTTTTAATACAGTTTAGTACTAAACTA-ATAATATCAAGAGGAGGGA-ACGAAC-----'
lmo0096   '-------GGCACAGAACTTGCAATAAATATAGCG-GGTAGCAAAATAAATGAATTATATTAGGAGGGAAAAAG------'
lmo0234   '--------TATATGTATATTTGGTAATAGCCCAATTATACATATAAATTTAAAAATGGAAAGGAGGAAACTAA------'
lmo0276   '--------CATATATAACTATTCCGATTCGCACATGGTATACTTAAGTTATTAACTAAAAAGGATGATTTTAT------'
lmo0031   '-------GTTATTTACCC-ATAA-ATACGCTACACTTA---AAATAGTTAAAAATGCTTACGAGGTGAACATTTCGA--'
```

```
lmo0288    '----------GTCTACCCCATAACATATGGTATACTAATTTAAATATATTACCCCGAATAGGAGTAGAACGACTT----'
qoxA       '------CGGTAATGGTATTTTTGCGCA-AAAATACACATTAGAAAACATCAAAAGGAAAAGAGGGGATTCGG-------'
lmo0108    '-------CCATTC--TTTTTCAAAATTAGGAGTGTAAACTAAATGGATATATGAATATTAAAGGATGTGAAAAA-----'
lmo0291    '-------GCATGCAATTTTGCTGAATTAGCGTTACAA-TGAAAGATATATATAAGCCTTTAGGGGGATAAAAA------'
lmo0070    '-------ATACTTTTCACTAAAAAAGTTAGACTATAATAGCAAAAAATAAAAAAAT-TTAAGGAGAGTATAAT------'
lmo0190    '--------CTACATATTTTTTCTTTCATATGATAGAATAGAAAAAGA--GATTTTGAACGCGGAGGCCAAGAAAA----'
cysE       '--------AGGGGCTTTTTTTGTTGGTTAGGTTCCAATAAAAAATGTACGAACTTGAATAAGGAGGCAACAAA------'
lmo0004    '-------TGTATTATTGGCAAAACTTTAGTAAAATAGAAGGTAGTGATAAAAAAAGATTGGGCGTGAAATTT-------'
lmo0229    '---------TATAATT----AAAGTCAAAAATAGTCAAAGTCAATGATTTGCAAAAAACAGAAAGGAGATTCCTTATA-'
lmo0052    '-------TATGTTATGATATAGATACTCAAATT-CAGGCAAAAGGTATGTAAAAAAATGGGGAGCTGGATGAA------'
lmo0217    '------AAACGTTTTTTTCTCTACCGAAAATCA-TAGACAATAAGCAAATAAGAACATGGGAGGGACGATGT-------'
lmo0038    '------GGTGATTGTTGGTGAAATATGTATTTGGCTTGC--TGGCAAAAGAAAAGACAAAGGAGAGATAAAAA------'
lmo0244    '-----------ATATGGATGAAATGTGCCTATTTTTTGGCATGGCATATTTAATGATTGCGGAGTTGAGATAATTT---'
lmo0157    '----TATCGAATTGTTG--CGAACGTAGGTTCTGTGT-TATAATTCAACTGGATTGTAAAGGAGGACTATTG-------'
rplJ       '---------TGTTGCTGATCAAACGGAACTTGTGTCTGTTTAATTAGATTTAATTGTTGGACGGAGGTGGAAAA-----'
lmo0135    '-------GATTCTGACCAAAAAACAGAGGAAGCGTTA-TTTTTTTAGCGCTTAAAGAGGGGAGTTTTTGTTAG------'
lmo0271    '------GAATATACTTTAAAATATATTAGTAGAGTTT-TTTAATAAAAATCGCTTTAATGGAGGTTATATTA-------'
lmo0212    '-----TTCGCATTT-GTTTGCGATATCTTCCATTTTT-TTTGTTATAATAGTTTCGAGAAGGAGGCGTTTAT-------'
sul        '-------GGTTTTT-TTGTGCGCTAAAATCACTAAAT-TCTGCTATAATAAATCTCGGAAAGAGGTGTTTGGGC-----'
lmo0008    '---------CGCTT-TTTGGATGTATTTTATGTATAA-TATCCT-TAATACGTCTAAAAAAGGGAAGAGTTGAAAAA--'
lmo0048    '--CATTTCAT-TACATTTTTGGTTATTATGGGTAAAT-TCGTTGTAAAAT-ATTAGTGGAGGTGAATTAG---------'
gcaD       '---AAAAAATGTAAATTTTACGTTATCATTCATAG---TGGATGAATAGTGAATATTGGAGGTTCTATATA---------'
lmo0010    '---------AAAAATTAT---GGTATCATGA-AG-TATTAAAGTTTGTATTATTTTGATGGAGAAAGGATTGGCGTGCA'
lmo0302    '-----ATTTAAAGGTTACTAGGATATGATGATAG-TTAGAGAGAAAATATTACATAGGAGATGAATAGATG--------'
lmo0221    '---GGATTTTTAAGTAAA--ACATGTTATGATAC-GTTAGAAAAATCTTCTA-AAGGACGGTATTTATACTT-------'
lmo0304    'TCTACATTCTTAGGTATA--ATGTAT-ACGAGAA-GATGAAAATACATTATATAAAAAGGATGTTTGTG----------'
lmo0079    '---------GAAAATAAGTGGAATTTCAAAGTATCTAATAATTTACTACATGATATACAAAAGGAGTTGTTTCA-----'
lmo0084    '--------ATAAACTAAATACAACATCCTGATATTTCTAATTTAAAAATTTGGTTTAAAAGGAGCATGTTTTT------'
lmo0162    '-----ATTCCGAAGCTATTTAAGCTTGGTCATGG-TATAATATTACAATATGGATAGAAAGGAAGTTTTTT--------'
lmo0252    '-----GGCTTTTTTTTATTGGTATTACATGCTAG-TATAAAGAGATGGATTAGTTTTCTGGGGGTTTATAT--------'
lmo0176    '------TGTTATTTCAGCTATAATGAAGTAATTGAAAACTA-AATTAAAGGATATACAAGGAGGAAAACGAT-------'
lmo0285    '---AAGTTTTTTTATTGCTTTCATGAATAAATCTGGAT----AATCACACAACATACTAGGGAGGAAAAAAG-------'
lmo0169    '-------AATAGGAATGATTTCATGAGGAAAAGGGTATAGACCATAAGCAGAAGAATTAGGGGGAATAAAAA-------'
lmo0193    '-----TGAATTCTATCGTTGCTACATAAAATTCAAGGCTTCTAGACATTAGGAGCAGAAGGAGAACATTC---------'
lmo0067    '--------AGTTACGGAACTTAACGGCAAGATAGAAATTATAGTTGACTTATTATAACGAAGGAGTGAAGAGG------'
lmo0109    '------GTAAGCATGGACAATGTTATATAATTAAAAGATGAATGTAAGC-GTTATAATTAGGAGGTGAAGTG-------'
lmo0187    '-----TTTTTTCTTTGTCCAAAATGTGGTATGAT-GTATTTACTTA-TTTTATAAAAAAGGATGT--AGTAATT-----'
mpl        '--------TCTCTCTGTCAGATTAGTTGTAGGTG-GCTTAAACTTAGTTTTACGAATTAAAAAGG--AGCGGTGAA---'
rpoB       '--------TTTTGTGCCTAAAAGTGAAGTCGGTGTGCTTATAAATTTTTTAATTAATTTGAGGGGTGAATAGT------'
lmaD       '-------GTACCGATAGCACCAAATTGGTGACAAAACTACTATATTAATTTATGAGATGGAAGTGGGAATGG-------'
lmo0133    '--------TTTTCTTTTGGTTGATGAGTGGAATAGATGGAAGGTAGAGAAACTTTGGCGAAGGAGGTAATCAG------'
lmo0315    '------CTGTTTCTTTGCTTATGCTAAGGGACAGTAGGCGCTTTTTTTATTGTTCAAAAGGAGGAGTTTGA--------'
lmo0114    '------AAAAAGTTCCGTTTTGGTGCTAAAAGTGCTATACTGAAGCCATAAACATATGGAGGACAATTATT--------'
lmo0196    '---------GTGTACTTAAGTTGCATGTGAGCATTGTTTACTAGCAGCGCAACTAGATTAAGGTGGTGAAGATA-----'
rpsF       '----------CAGGGATATGGTATGGTTTCATGCTATATCAGAGCCGCTTAAGACCACAAGGAGGTGTAGAGTAG----'
rplK       '---------TTTAACGTGGGAGGGGAAATATCAGCCCCAGTCAACCACATCACGGACTTAAGGAGGTATGTCTC-----'
lmo0279    '------AAAAAAACTTCATCTTGTGTTATATTATTAAGGCAGACACAATATATAGTTAGTGGGGGATTATC--------'
lmo0020    '-------ATCATGGTACCATTAACCTAAAACTTAACGTGTTAACGCGCGGGATGCTTTAGGGGGAACTTGTT-------'
```

*The gene is the first gene in the operon

Genome sequence including ‘plcA’ gene*.

CCGTTTGCCACCCCTCTCTTTTGATAATTATAATATTGGCGAAATTCGCTTCTAAAGATGAAACGCAATATTATATGCTTGCTTTATAGCTTTAT
TCTAGTCCTGCTGTCCCTTTATCGTCGTTAACAAATGTTAATGCCTCAACATAAAAGTCACTTTAAGATAGGAATATACTAATCAAAGGAGGGGG
CCATTTTGTATAAGAATTATTTGCAACGCACATTAGTTTTATTACTCTGTTTTATTTTATACTTTTTTACTTTCCCATTAGGCGGAAAAGCATAT
TCGCTTAATAACTGGAATAAGCCAATAAAGAACTCTGTAACTACAAAACAATGGATGTCCGCTCTACCTGACACAACAAACCTAGCAGCGCTCTC
TATACCAGGTACACATGATACGATGAGCTATAACGGAGACATAACATGGACATTAACCAAACCACTAGCTCAAACACAAACGATGTCATTGTACC
AACAACTAGAAGCAGGAATACGGTACATCGATATTAGAGCAAAAGACAATCTCAACATTTACCATGGGCCAATTTTTTTAAATGCATCACTTTCA
GGTGTATTAGAAACGATTACTCAATTTTTAAAGAAAAATCCAAAAGAAACCATTATTATGCGTTTAAAAGACGAGCAAAACAGCAACGATAGTTT

TGATTATCGGATCCAACCACTAATCAACATTTATAAAGATTATTTTTACACTACTCCCAGAACTGACACGAGCAATAAAATCCCTACATTAAAAG
ATGTCCGCGGAAAAATATTATTACTTTCAGAGAACCACACAAAAAAGCCATTAGTCATTAACTCACGCAAATTCGGCATGCAGTTCGGCGCACCT
AACCAAGTAATTCAAGATGACTACAATGGTCCGAGTGTGAAAACAAAATTCAAAGAGATTGTCCAGACTGCTTATCAAGCTTCCAAAGCGGACAA
TAAACTTTTTCTTAACCATATTAGCGCCACTTCATTAACATTCACACCTCGTCAGTATGCTGCAGCATTAAACAACAAAGTAGAGCAATTCGTAC
TCAACTTAACATCGGAAAAAGTTCGAGGATTAGGCATACTAATCATGGACTTCCCCGAAAAACAAACAATTAAAAACATCATAAAAAACAATAAA
TTCAACTAACATA

*The ‘plcA’ gene is underscored with ‘ ’; the -10 box is underscored with ‘ ’.

REFERENCES

[1] Rodger Staden, “Computer methods to locate signals in nucleic acid sequence”, Nucleic Acids Research, Vol. 12, No. 1 Part2, 1984, pp. 505-519.
[2] Abdi, H., “Distance”, Encyclopedia of Measurement and Statistics, Thousand Oaks (CA): Sage, 2007, pp 280-284.
[3] Sharon E, Lubliner S, Segal E, “A Feature-Based Approach to Modeling Protein–DNA Interactions”, PLoS Comput Biol 4(8): e1000154. doi:10.1371/journal.pcbi.1000154, 2008.
[4] Eden E, Lipson D, Yogev S, Yakhini Z, Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol 3: e39. doi:10.1371/journal.pcbi. 0030039, 2007.
[5] Diane K.Hawley, William R.McClure, “Compilation and analysis of Escherichia coli promoter DNA sequences”, Nucleic Acids Research, Vol. 11, No. 8, 1983, pp 2237-2255.
[6] Gregg Duester, Renee K.Campen, W.Michael Holmes, “Nucleotide sequence of an Escherichia coli tRNA (Leu 1) operon and identification of the transcription promoter signal”, Nucleic Acids Research, Vol. 9, No. 9, 1981, pp 2121-2139.
[7] http://genolist.pasteur.fr/ListiList/index.html.
[8] http://www.ebi.ac.uk/Tools/clustalw2/index.html.
[9] Lun Huang, Mohammad Al Bataineh, G. E. Atkin, Maria Parra, Maria del Mar Perez, Ismaeel Mohammed, Wei Zhang, "Identification of Transcription Factor Binding Sites Based on the Chi-Square (X^2) Distance of a Probabilistic Vector Model". 2009 International Conference on Future BioMedical Information Engineering, Sanya, China, December 13 - 15, 2009.
[10] Lun Huang, Mohammad Al Bataineh, Guillermo Atkin, Siyun Wang and WeiZhang, "A Novel Gene Detection Method Based on Period-3 Property", 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'09), Minneapolis, Minnesota, USA, 2nd - 6th September, 2009.
[11] Mohammad Al Bataineh, Lun Huang, Ismaeel Muhamed, Nick Menhart, and Guillermo Atkin, "Gene Expression Analysis using Communications, Coding and Information Theory Based Models", BIOCOMP'09 - The 2009 International Conference on Bioinformatics & Computational Biology, Las Vegas, Nevada, USA, July 13-16, 2009.
[12] Mohammad Al Bataineh, Maria Alonso, Lun Huang, Nick Menhart, Guillermo Atkin, "Effect of Mutations on the Detection of Translational Signals Based on Communications Theory Concepts", 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'09), Minneapolis, Minnesota, USA, 2nd - 6th September, 2009.

Huang Lun received the M.S.E. degree in electrical engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2004. He is currently working toward the Ph.D. degree at the Electrical and Computer Engineering department, Illinois Institute of Technology, Chicago, IL. His research interests include bioinformatics, digital signal processing, stochastic process, and wireless communication systems.