Identification of Transcription Factor Binding Sites Based on the Chi-Square (χ^2) distance of a Probabilistic Vector Model

Lun Huang[1], Mohammad Al Bataineh[2], G. E. Atkin[3], Ismaeel Mohammed[4], Wei Zhang[5]
[1,2,3] are with ECE; [4,5]are with BCPS Departments
Illinois Institute of Technology
Chicago, U.S.A.
e-mail: lhuang13@iit.edu; albamoh@iit.edu; atkin@iit.edu; imuhamed@iit.edu; zhangw@iit.edu

Maria Parra, Maria del Mar Perez
ECE Department
Polytechnic University of Madrid
Madrid, Spain
e-mail: mparrah@iit.edu; mperezp@iit.edu

***Abstract*—This paper describes a new approach for locating signals, such as promoter sequences, in nucleic acid sequences. Transcription Factor (TF) binding to its DNA target site is a fundamental regulatory interaction. 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) distance model [2], which is based on the distance between the profiles of component vectors. It is a novel probabilistic method for modeling TF-DNA interactions. Our approach uses χ^2 distances to represent TF binding specificities. Simulation results show that the proposed approach identifies TF binding sites significantly better than the PWM model method.**

***Index Terms*—promoter, Transcription Factor, Chi-square distance**

I. 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 computational approaches is then to identify the PWM associated with each TF and use it to identify TFBS. A weight matrix is a two dimensional array of values that represent the score for finding each of the possible bases at each position in the TF for which we are looking. 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). 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 TFs 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. In theory, it is easy to see where this assumption fails. For example, the TFBS data contains only "CG" or "GC" in the center positions. Although the PWM learned from this data assigns high probability 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 models (FMM) [3] approach. Even though the FMM approach is better than the PWM, it involves the evaluation of complicated log likelihood function and objective function. Then, it significantly increases the computation complexity.

In this paper, a novel identification approach is proposed. It uses a statistical model based on the χ^2 distance. This approach does not require large computation complexity, and simulation results show that it can effectively identify the location of 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; in section III, the simulation results are analyzed; section IV presents the conclusion of the proposed approach.

II. SYSTEM MODEL

The system model is described in Figure 1. 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 is composed of 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

978-1-4244-4692-6/09/$25.00 ©2009 IEEE

contingency table. An example of the procedure to evaluate the χ^2 distance between the input sequence and the Center Of Gravity (COG) for the family composed of the conserved sequences in Table 1, are 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.

The row labeled ΣNC^T gives the total of each column. 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 in the conserved sequences. The weight of each column W^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. 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. In Table 2(b), the total number of matches is $N' = 75$.

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 COG of the rows, denoted C, 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 $W = (\mathrm{diag}\{C\})^{-1}$.

Table 1. The original conserved and input sequences

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

Table 2. Contingency table for the computation of 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		
C^T	0.28	0.107	0.28	0.333		
W^T	7.238	19	7.238	6.08		

(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	

With this weight matrix, variables which are used often contribute less to the distance between rows than variables which are used rarely. For example, from Table 2 (a), we find that the weight matrix is equal to

$$W = D_w = \mathrm{diag}\{w\} = \begin{bmatrix} 0.28^{-1} & 0 & 0 & 0 \\ 0 & 0.107^{-1} & 0 & 0 \\ 0 & 0 & 0.28^{-1} & 0 \\ 0 & 0 & 0 & 0.333^{-1} \end{bmatrix} = \begin{bmatrix} 7.238 & 0 & 0 & 0 \\ 0 & 19 & 0 & 0 \\ 0 & 0 & 7.238 & 0 \\ 0 & 0 & 0 & 6.08 \end{bmatrix} \quad (1)$$

Assume 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 reserved sequence i=1 and the input sequence is equal to

$$d^2(1) = 7.238 \cdot (0.237 - 0.375)^2 + 19 \cdot (0.132 - 0)^2 + 7.238 \cdot (0.263 - 0.375)^2 + 6.08 \cdot (0.368 - 0.25)^2 = 0.317 \quad (2)$$

In the same way, we have $d^2(2) = 0.219$, $d^2(3) = 0.549$, $d^2(4) = 0.088$.

If M is the number of conserved sequences; L is the length of conserved nucleotide sequence; the grand total of the contingency table is N, which is equal to $(M+1) \cdot L$ the total number of nucleotides in all of the conserved and input sequences, the distance from row i to the 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:

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

Where $i, i' \in \{\text{index of conserved and input sequences}\}$. r_i and $r_{i'}$ are the mass of each row, and it is the component of mass vector 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 COG for the conserved sequence family can be derived from (3) as

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

Where $i \in \{\text{index of conserved sequences}\}$. Since the index for input sequence can be fixed to M+1, this distance is independent of the index scheme, and can be denoted as d_g^2 instead of $d_g^2(M+1)$. 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 COG of the conserved sequences can be defined by

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

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

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

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

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

The lower threshold is

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

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

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

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 lower threshold function can be defined by

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

The metric function is defined as

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

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

III. SIMULATION

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 TTGACA 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 regions) are in [1]. The input nucleotide sequence (genome sequence) used in simulation can be found in [6].

First, we use -10 region of the conserved sequence to identify the -10 promoter, the result is shown in Figure 2. The red region marks the real location of the -10 promoter, the center of red region is at 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 -35 promoter identification, we use -35 region of the conserved sequences to form the conserved sequence table, the result is described in the Figure 3.

Similarly, the red region indicates the real location of -35 promoter, it can be discovered that the highest peak appears around the location of 62, so the predicted location for -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 mere 8 bases. In fact, there is a restriction enzyme Taq-I recognition site at location 63, it might be responsible for the highest peak. We are currently researching on this interesting result that points out the ability of the algorithm to identify other signal sequences.

By 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 Figure 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.

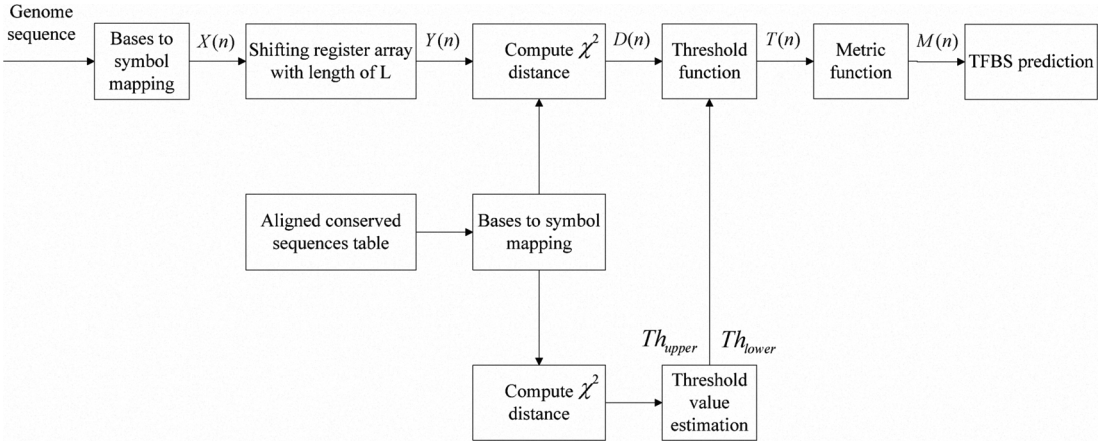

Figure 1. System model for TFBS and signal detection based on χ^2 distance

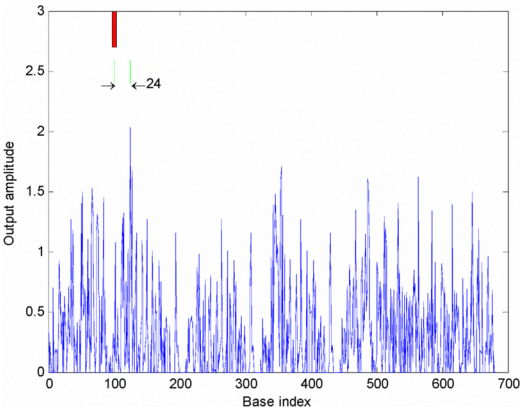

Figure 2. -10 promotor identification

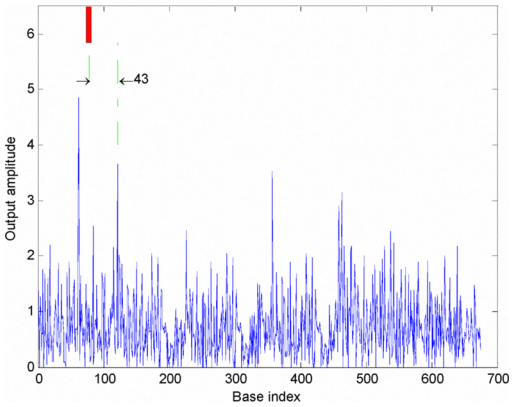

Figure 3. -35 promotor identification

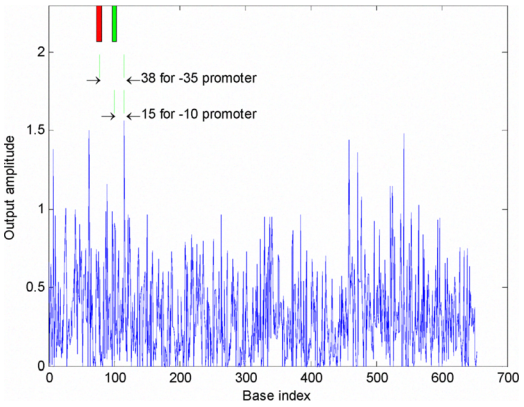

Figure 4. -35 and -10 promotors identification

IV. CONCLUSION

In this paper, we have introduced a novel TFBS identification algorithm, which is based on χ^2 distance model. The proposed approach is more efficient than traditional method 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 TF motifs in computing χ^2 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.

REFERENCES

[1] Rodger Staden, "Computer methods to locate signals in nucleic acid sequence", Nucleic Acids Research, Vol. 12, No. 1 Part2, pp 505-519, 1984.

[2] Abdi, H., "Distance", Encyclopedia of Measurement and Statistics, Thousand Oaks (CA): Sage. pp 280-284, 2007.

[3] Sharon E, Lubliner S, Segal E (2008) A Feature-Based Approach to Modeling Protein–DNA Interactions. PLoS Comput Biol 4(8): e1000154. doi:10.1371/journal.pcbi.1000154.

[4] Eden E, Lipson D, Yogev S, Yakhini Z (2007) Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol 3: e39. doi:10.1371/journal.pcbi. 0030039.

[5] Diane K.Hawley, William R.McClure, "Compilation and analysis of Escherichia coli promoter DNA sequences", Nucleic Acids Research, Vol. 11, No. 8, pp 2237-2255, 1983.

[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, pp 2121-2139, 1981.