Identification of Transcriptional Promoter Sequence Based on Statistical Filter Bank Model

Lun Huang, Mohammad Al Bataineh, Alicia Fuente Acedo, G. E. Atkin, Xiangyu Deng and Wei Zhang

***Abstract*—This paper describes a new approach for locating transcription related signals, such as promoter sequence in nucleic acid sequences. Transcription Factor (TF) and corresponding polymerase binding to their DNA target site is a fundamental regulatory interaction. The most common model used to represent TF and polymerase 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 statistical filter model based on Chi-Square (χ^2) distance [2], which is a statistical distance metric between the profiles of component vectors. It is a novel statistical method for modeling TF-DNA and polymerase-DNA interactions. Our approach also uses a generalized correlation algorithm to evaluate the combination coefficients for the filter bank. Simulation results show that the proposed approach identifies promoter sequences better than the PWM model method and Chi-Square (χ^2) distance model.**

***Index Terms*—Chi-Square Distance, Transcriptional Promoter Sequence, Transcription Factor Binding Sites**

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 Transcriptional Promoter Sequence (TPS). The goal of computational approaches is then to identify the PWM associated with each Transcriptional Promoter (TP) and use it to identify TPS. 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 TP for which we are looking. For DNA sequences the weight matrix will have a length equal to the length of the TP and depth of four (one row for each of A, C, G and T). Generally, we generate the frequency table for the TP 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 TPS are position-independent. That is, the PWM assumes that for any given TFBS and TPS, the contribution of a nucleotide at one position of the site to the overall binding affinity of the TF or polymerase to the site does not depend on the nucleotides that appear in other positions of the site. Theoretically, this assumption fails in certain case. For example, the TPS 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 TPS binding specificities. This observation leads to the feature motif models (FMM) [3] approach. Even though the FMM approach performs better than the PWM, it involves iterative evaluation of complicated log likelihood function and objective function. Then, it significantly increases the computation complexity. As an alternative method, considering the statistical dependency of the same type of nucleotide at multiple positions could also reveal the motif hidden in TPS and other related signal sequence.

In this paper, a novel identification approach is proposed. It uses a statistical filter bank model based on the χ^2 distance and generalized correlation. This approach does not require large computation complexity, and simulation results show that it can effectively identify the location of TPS and other related signal sequences. This paper is organized as follows. In section II, the system model and basic 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 Fig 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 in the nucleotide input sequence G(n).

The reconstruction filter bank is widely used in

L. Huang, PhD Candidate; Mohammad Al Bataineh, Ph.D. Candidate; and G. E. Atkin, Ph. D., Senior Member IEEE, Associate Professor; are with the Department of Electrical & Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616; 1-312-567-3417; 1-312-567-8976 Fax (e-mail: lhuang13@iit.edu; albamoh@iit.edu; atkin@iit.edu).

Alicia Fuente Acedo, M.S., is with the Department of Electrical and Computer Engineering, Polytechnic University of Madrid, Madrid, Spain; (e-mail: aliciafuenteacedo@gmail.com)

Xiangyu Deng, Ph.D. candidate; Wei Zhang, Ph. D., Assistant Professor; is with the Department of Biological, Chemical, Physical Sciences, Illinois Institute of Technology, Chicago, IL 60616; 312-567-3123; 312-567-3494 fax (e-mail: xdeng7@iit.edu, zhangw@iit.edu).

978-1-4244-6875-1/10/$26.00 ©2010 IEEE

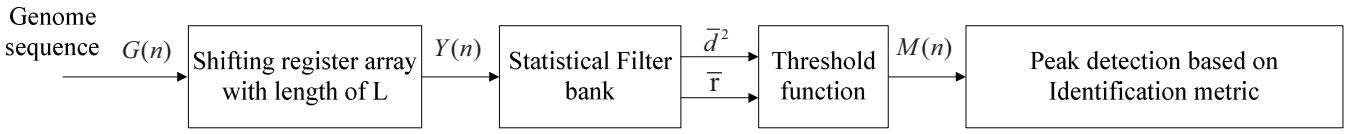

Fig. 1. System model for TPS identification based on statistical filter bank

communication and signal processing systems to process signals with multiple components. It turned out to be a very effective way to extract information from noise. The block diagram for the proposed statistical filter bank is shown in Fig 2.

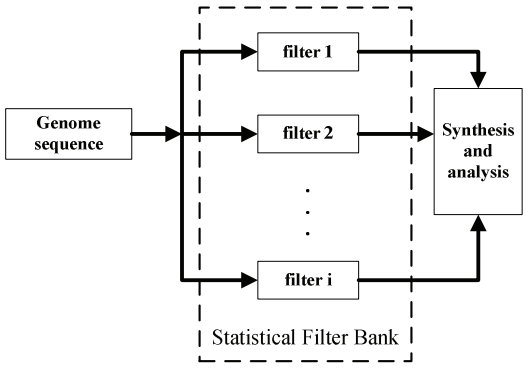

Fig. 2. Block diagram for the statistical filter bank

What should be noted is that the "filter" used in our scheme are not limited to classical filter, we can generalize them with a wide range of algorithms. The general model for the filter is shown in Fig 3. The filter compare the input sequence to the conserved sequence, and output a distance metric and a matching metric. Without loss of generality, we can define a matching metric vector $\mathrm{v} = \{\mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_\mathrm{i}\}$, where v_i is the value of the matching metric function $\mathrm{S}(\mathrm{x}_\mathrm{i}, \mathrm{y})$, where x_i is the conserved sequence i and y is the input genome sequence. With this matching metric vector, the weight vector generation module can generate the weight vector $\mathrm{r} = \{\mathrm{r}_1, \mathrm{r}_2, \ldots, \mathrm{r}_\mathrm{i}\}$ by using the algorithm discussed in section B, so that we can evaluate a generalized identification metric. Based on this identification metric, we can determine the accurate location of the TPS and related signal sequence.

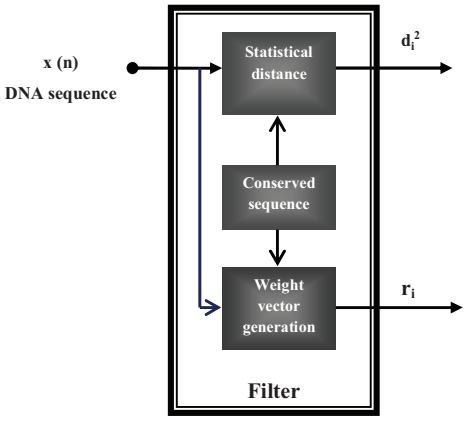

Fig. 3. Filter model

A. Evaluation of Statistical distance

The statistical distance we used here is χ^2 distance. 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 of a contingency table. 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, is described in this section. An example of conserved sequence family is given in TABLE I.

TABLE I
THE ORIGINAL CONSERVED SEQUENCES

Sequence name	Nucleotide sequences
Conserved 1	AATTGAGCAACTATAGCAGTGTGAAATAACAT
Conserved 2	AATAGCAGTGTGGCACACTTGTGAATTATCTC
Conserved 3	GATCGAAAGATAAGAAATAAACCAAAAATCGT
Conserved 4	AATTGTTTTCCCATCGAAAGATAAAAATCTGT

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, if the input sequence to statistical filter bank is "ATCGAAAGATAAAAATCTGTAATTGTTTTCCC", for conserved sequence 2, it has 10 nucleotides coinciding with the input sequence, thus, only 10 of its bases would be used to evaluate the statistics and the number of matches for it is 10. A match number function $\mathrm{f}^\mathrm{m}(\mathrm{x}_\mathrm{j}, \mathrm{y})$ can be defined for each of the four types of nucleotides A, T, C, G, where y is the input sequence and x_j denotes the conserved sequence i. Then, $f_A^m(x_j, y) + f_T^m(x_j, y) + f_C^m(x_j, y) + f_G^m(x_j, y) = 10$. The Total number of matches for a family with i conserved sequences is

$$\mathrm{N_m} = \sum_{j=1}^{i} \left[f_A^m(x_j, y) + f_T^m(x_j, y) + f_C^m(x_j, y) + f_G^m(x_j, y) \right] \quad (1)$$

Let p^m be the proportion of each kind of nucleotide in the total number of matches for all of the conserved sequences. Then we have

$$\mathrm{p_k^m} = \frac{\sum_{j=1}^{i} f_k^m(x_j, y) + f_k^m(y, y)}{\mathrm{N_m} + \mathrm{L}}, \mathrm{k} = \mathrm{A, T, C, G.} \quad (2)$$

The conserved and input sequences are also needed to be transformed into profile vector. For conserved sequence i, its profile vector can be defined as

$$S_i = \{S_i^A, S_i^T, S_i^C, S_i^G\} \quad (3)$$

where

$$S_i^k = \frac{f_k^m(x_i, y)}{f_A^m(x_i, y) + f_T^m(x_i, y) + f_C^m(x_i, y) + f_G^m(x_i, y)}$$
$$, k = A, T, C, G.$$

The coefficient matrix used to evaluate the χ^2 distance can be obtained by

$$C = \begin{bmatrix} (p_A^m)^{-1} & 0 & 0 & 0 \\ 0 & (p_T^m)^{-1} & 0 & 0 \\ 0 & 0 & (p_C^m)^{-1} & 0 \\ 0 & 0 & 0 & (p_G^m)^{-1} \end{bmatrix} \quad (4)$$

Thus, the χ^2 distance between the input sequence and the conserved sequence i can be evaluated with

$$d^2(i) = (S_i - S_0)C(S_i - S_0)^T \quad (5)$$

where S_i and S_0 are respectively the profile vectors to the conserved sequence i and the input sequence.

As an example, the χ^2 distance between the reserved sequence i=1 and the input sequence mentioned above is equal to

$$\begin{aligned} d^2(1) &= (S_1 - S_0)C(S_1 - S_0)^T \\ &= 2.24 \cdot (0.5 - 0.375)^2 + 10.83 \cdot (0.1 - 0.125)^2 + \\ &\quad 6.5 \cdot (0.2 - 0.156)^2 + 3.25 \cdot (0.2 - 0.344)^2 \\ &= 0.122 \end{aligned} \quad (6)$$

In the same way, we have $d^2(2) = 0.042$, $d^2(3) = 0.421$, $d^2(4) = 0.672$.

B. Evaluation of weight vector for filter bank

The process of deciding weight vector $r = \{r_1, r_2, \dots, r_i\}$ is based on generalized correlation and free energy distance metric. In telecommunication, a correlation filter or matched filter is obtained by correlating a known signal, or template, with an unknown signal to detect the presence of the template in the unknown signal. This is equivalent to convolving the unknown signal with a time-flipped version of the template. The generalized correlation filter used in the proposed scheme is shown in Fig 4, it is the optimal filter for maximizing the signal to noise ratio (SNR) in the presence of additive stochastic noise. It can be implemented by using a matched filter with impulse response equal to x(-n) , which is the time-flipped version of conserved sequence x(n), y(n) is the input DNA sequence. Generalized correlation between k-th conserved sequence $x_k(n)$ and input sequence y(n) can be defined by

$$z(k) = x_k(n) \otimes y(n) = \sum_{j=1}^{L-1} E(j) \quad (7)$$

where ($\otimes$) denotes generalized correlation, L is the length of conserved sequence $x_k(n)$.

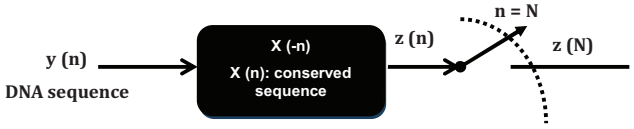

Fig. 4. Generalized correlation filter

TABLE II
ENERGY DOUBLETS [5]

Pairs of bases Energy	
AA -0.9	GA -2.3
AU -0.9	GU -2.1
UA -1.1	CA -1.8
UU -0.9	CU -1.7
AG -2.3	GG -2.9
AC -1.8	GC -3.4
UG -2.1	CG -3.4
UC -1.7	CC -2.9

We can use the free energy table (see TABLE II) to calculate a free energy distance metric in kcal/mol [5]. This metric z(n) is the sum of the energy component calculated at each doublet between the input DNA sequence and the conserved sequence under study. At the position j of the input sequence, the free energy component can be evaluated by:

$$E(j) = ED(x_j x_{j+1}) \cdot \delta(j) \quad (8)$$

where $x_j x_{j+1}$ denotes the doublet at position j and j+1 of the conserved sequence. $ED(x_j x_{j+1})$ is the energy dissipated on the nucleotide doublet $x_j x_{j+1}$, and it is from TABLE II. Let $y_j y_{j+1}$ be the nucleotide doublet in the input DNA sequence, which binds to $x_j x_{j+1}$. $\delta(j)$ is given by:

$$\delta(j) = \begin{cases} 1, & x_j x_{j+1} = y_j y_{j+1} (match) \\ 0, & x_j x_{j+1} \neq y_j y_{j+1} (mismatch) \end{cases} \quad (9)$$

If the total number of conserved sequence is i, the matching and weight vector $v = \{v_1, v_2, \dots, v_i\}$ and $r = \{r_1, r_2, \dots, r_i\}$ can be evaluated using

$$v_j = \frac{Z[j]}{\sum_{m=1}^{i} Z[m]} \quad (10)$$

$$and\ r_j = \frac{(1 - v_j)}{\sum_{m=1}^{i} (1 - v_m)} \quad (11)$$

where j=1,2, … , i; v represents the normalized matching metric vector; weight vector r is composed of the normalized mismatching metrics, it can be used to obtain the normalized generalized distance D(n) and the identification metric function.

C. Evaluation of thresholds and identification metric

If i is the number of conserved sequences; L is the length of conserved nucleotide sequence; the total number of nucleotides in all of the conserved and input sequences is equal to (i+1)L, an approximate estimation of the generalized statistical distance between the input sequence and the Center Of Gravity (COG) for the conserved sequence family can be defined as

$$D(n) = \frac{(i+1)L}{N_m + L} \sum_{j=1}^{i} r_j \, d_n^2(j) \quad (12)$$

where n denotes the index of the location on the input

nucleotide sequence, and

$$d_n^2(j) = \left[\mathrm{S_j} - \mathrm{y(n)}\right]\mathrm{C}\left[\mathrm{S_j} - \mathrm{y(n)}\right]^T \quad (13)$$

To define the dynamic range of the generalized distance D(n) for the input nucleotide sequence, we must evaluate lower and upper thresholds based on the conserved sequences.

The threshold set can be obtained with function

$$T(k) = \frac{(i+1)L}{N_m + L}\sum_{j=1}^{i} r_j\,(\mathrm{S_j} - \mathrm{S_k})\mathrm{C}(\mathrm{S_j} - \mathrm{S_k})^T, \quad \mathrm{k} = 1, 2, \ldots, \mathrm{i} \quad (14)$$

Then, the upper threshold is

$$Th_{upper} = \max\{T(k), \mathrm{k} = 1, 2, \ldots, \mathrm{i}\} \quad (15)$$

And the lower threshold is

$$Th_{lower} = \min\{T(k), \mathrm{k} = 1, 2, \ldots, \mathrm{i}\} \quad (16)$$

The output identification metric can be defined by

$$\mathrm{M(n)} = \begin{cases} \dfrac{Th_{upper}}{\mathrm{D(n)}} - 1, & Th_{lower} \le D(n) \le Th_{upper} \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

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

In peak detection, a window with length L is applied. When we slide this window on the identification metric output sequence M(n), at each time n = k, we can obtain a maximum value at each time n = k, and

$$Max(k) = \max\left\{M\left(k + i - \frac{L}{2}\right), \mathrm{i} = 1, 2, \ldots, \mathrm{L}\right\} \quad (18)$$

If $Max(k) = M(k)$, it can be judged that there is a peak at n = k. The steps to determine the optimum value of L based on minimum peak interval variance are as follows.

1. Initiate L_0 and L_1 with $L_x/3$, and set $p_0 = 2$, whereL_0 is initial window length, L_1 is current window length, p_0 is original peak number.
2. Use window with length of L_1 to detect the peaks of the identification metric output sequence M(n).
3. If the number of detected peaks is $p_1 < p_0 + 1$, $L_1 = L_1 - 1$ and go to step 2.; otherwise, set the minimum value of variance $V_{min} = \infty$.
4. For this p_1 peak distribution, find out the peak interval sequence $\{I_j\}$, $j = 1, 2, \ldots, p_1 - 1$.
5. Find out the variance of $\{I_j\}$ with

$$V_1 = \frac{1}{p_1 - 1}\sum_{j=1}^{p_1 - 1}\left(\mathrm{I_j} - \bar{\mathrm{I}}\right)^2, \quad (19)$$

$$\text{where } \bar{\mathrm{I}} = \frac{1}{p_1 - 1}\sum_{j=1}^{p_1 - 1}\mathrm{I_j}$$

6. If $V_1 < V_{min}$, $V_{min} = V_1$, $p_0 = p_1$, $L_0 = L_1$ and go back to step 4.; otherwise, $L_1 = L_0$. In the same ways as step 2. and 3., minimize L_1 for peak number $p_1 = p_0$, and $L = 2\cdot\left\lceil \frac{1}{3}Min(L_1)\right\rceil + 1$, where $Min(\bullet)$ is the minimization function, and $\lceil \bullet \rceil$ denotes the function that rounds its argument up to the next integer.

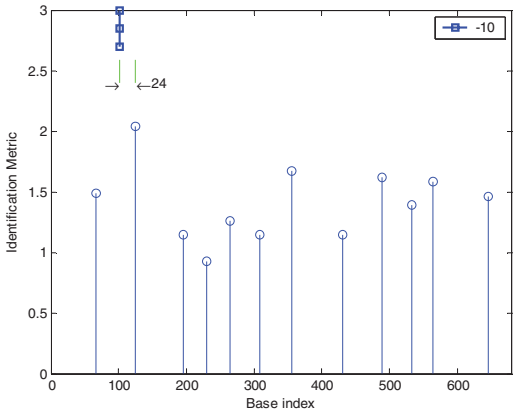

Fig. 5. -10 promotor sequence identification.

III. Simulation

E. coli promoter sequences 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 conserved sequences table used to locate promoter sequences in E. coli sequences is taken from the compilation of such sequences produced by Hawley and McClure [6]. 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 [7].

First, we use the -10 region of the conserved sequence to identify the -10 promoter sequence, the result is shown in Fig 5. The center of -10 promoter sequence is at 101. It can be observed that the highest peak locates at 125, which is 24 bases from the center of -10 promoter sequence; therefore, the predicted location of the -10 promoter sequence should be 115 and the identification error is 14 bases. The ratio of the highest peak to the second highest peak increases compared to the simulation results in [8], it means the location of TPS would be more reliable to identify.

In -35 promoter sequence identification, we use -35 region of the conserved sequences to form the conserved sequence table, the result is shown in the Fig 6. The highest peak appears around the location of 62, so the predicted location for -35 promoter sequence should be 27. Since the center of the -35 promoter sequence is at 78, the identification error is 51 bases. However, if we consider the third highest peak, which

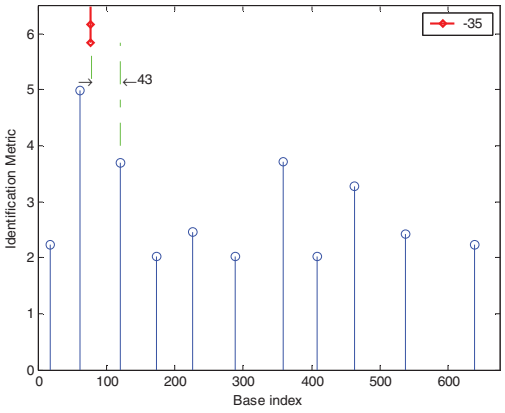

Fig. 6. -35 promotor sequence identification.

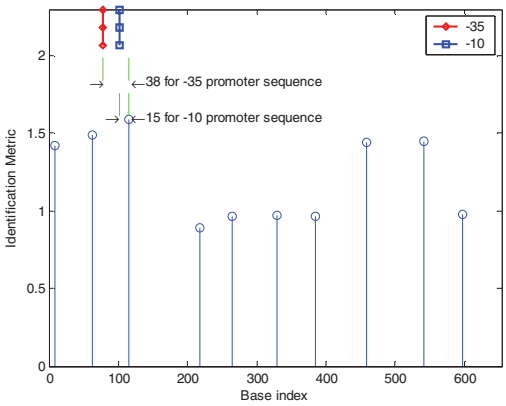

Fig. 7. -35 and -10 promotor sequences identification.

locates at 121 and 43 bases away from the center of the -35 promoter sequence, the predicted location is 86 and the identification error is mere 8 bases. Compared to the corresponding results in [8], the peak locates at position 121 is lowered. This interesting phenomenon may point out certain signal sequence or motif in the promoter sequence. In fact, there is a restriction enzyme Taq-I recognition site at location 63, that might be responsible for the highest peak.

In the same way, 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 Fig 7. With the highest peak located at 116, it can be discovered that the identification errors for -35 and -10 promoter sequence are 3 and 5 respectively. They are far smaller than those corresponding identification errors in the previous two cases. This is due to the use of longer conserved sequences, which provides more reliable statistical information, thus a better identification performance. Compared to [8], the ratio of highest to second highest peak in this case is also enhanced.

Similar simulations are described in Fig 5c in [1]. In case of identifying -10 promoter sequence 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 sequence must combine with that of -35 promoter sequence to achieve the real location. It involves a complex procedure of optimization and the choice of certain criteria. The proposed approach does not require any complicated optimization operation, and it is faster and more accurate.

IV. Conclusion

In this paper, we have introduced a novel TPS identification algorithm, which is based on statistical filter bank 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 iteratively evaluate highly complicated log likelihood function and objective function. Since this new approach takes into account the statistical position-dependency of TPS motifs in computing χ^2 distance and identification metric, 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 TPS 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.

[7] D. Rosnick, “Free Energy Periodicity and Memory Model for Genetic Coding”, PhD thesis, North Carolina State Univesity, Raleigh, 2001.

[8] Lun Huang, Mohammad Al Bataineh, Guillermo Atkin, Maria Parra, Maria del Mar Perez, Ismaeel Mohammed, and Wei Zhang "Identification of Transcription Factor Binding Sites Based on the Chi-Square (χ^2) distance of a Probabilistic Vector Model", International Conference on Future BioMedical Information Engineering (FBIE 2009), December 13-14, 2009.