MODELING REGULATORY SITES WITH HIGHER ORDER POSITION-DEPENDENT WEIGHT MATRICES

Hossein Zare, Mostafa Kaveh

Dept. of ECE, University of Minnesota, Minneapolis, MN
Email: hossein,mos@ece.umn.edu

Arkady B. Khodursky

Dept. of BBMB, University of Minnesota, St. Paul, MN
Email: khodu001@umn.edu

ABSTRACT

Identification of regulatory signals in DNA depends on the nature and quality of the patterns of representative sequences. These patterns are constructed from training sets of sequences by means of probabilistic models that either assume independence between positions or that suffer from considerable computational complexity.

We have developed and tested higher order models that account for significant dependent position pairs or triads, thereby capturing position-dependent information hidden in DNA binding sites. We have evaluated our algorithm on several data sets, including eukaryotic and bacterial transcription factor binding sites and shown that the scores from the higher order representation of binding sites have significant positive correlation to the binding affinity scores.

Index Terms— DNA Binding sites, Regulatory signal, Position weight matrix, Transcription factor

1. INTRODUCTION

Control of transcription and replication depends on the recognition of specialized DNA sequences by regulatory proteins. These specialized sequences, generally referred to as "binding sites", are relatively short segments of DNA embedded within larger regulatory regions. Over the years, considerable research effort has been applied to systematically identifying new binding sites for a given transcriptional regulator or transcription factor (TF) across a genome [1, 2, 3]. When the genome sequence of an organism is available and there are some known binding sites for a given TF, one can computationally predict additional sites by scanning the DNA sequences for short segments sharing common features with the known sites. The simplest and most widely used method to do so relies on a position-specific scoring matrix (PSSM), or a position weight matrix (PWM), surveyed in [1]. In PSSM, DNA binding sites are modeled in such a way that nucleotides at each position of the site contribute independently to the binding. The PSSM matrices are constructed from the alignment of known binding sites which have been identified experimentally. The PSSMs are $4 \times L$ matrices (L is the length of the sites) with rows indexed by nucleotide $i \in \{A, T, C, G\}$ and columns representing positions $j \in \{1, ..., L\}$. The entries of the matrix are the frequencies of the occurrences of each nucleotide at each position. The elements of position weight matrix, $\mathcal{M}_1$, are log-odd values which are calculated as $w_{i,j} = \log(\frac{f_{i,j}}{p_i})$, where $f_{i,j}$'s are entries of PSSM matrix and p_i is the probability of observing the symbol i in a genome or a background model. Then, the weight matrix can be used to calculate the score of DNA sequence, $Y = y_1, ..., y_L$, by $S(Y|\mathcal{M}_1) = \sum_{j=1}^{L} w_{y_j,j}$.

There are two main concerns with the above method. First, experimental evidence [4] suggests that the assumption of independent contribution of each position to the overall binding affinity is often not valid. Second, due to the choice of the score threshold, this method suffers from a high false positive rate and also misses some true sites. This indicates the necessity of constructing a comprehensive model which includes as much information as possible from both consensus and non consensus positions to represent the binding sites. The dependency assumption between nucleotides has been investigated through modification of the PWM in [2], and by means of probabilistic models in [3, 6]. It has been shown that accounting for the dependency structure of binding sites increases the specificity and prediction power of the pattern-matching algorithms,

THIS WORK WAS SUPPORTED IN PART BY NIH GRANT GM66098 (TO ABK)

and results in a more accurate prediction of protein-DNA binding affinity. In this paper, we present a new algorithm for constructing higher order position weight matrices which accounts for the position-specific dependencies between nucleotides in the sites. Higher order matrices can be constructed for dinucleotides or trinucleotides over the significant dependent adjacent and *non-adjacent* position pairs or triads. We used our model to analyze eukaryotic and bacterial transcription factor binding sites as well as confirming the improved performance of the model using independently obtained experimental data.

2. MODELING BINDING SITES USING HIGHER ORDER PWMS

We model sequences of binding sites using PWMs of the first-, second- and third- order. Hereafter, by the second and third-order matrices, or models, we mean that position weight matrices are defined for dinucleotides and trinucleotides. In constructing these matrices we consider not only the dependency between adjacent nucleotides but also dependencies between *non-adjacent* nucleotides. A second-order PWM is a $16 \times L_2$ matrix, $\mathcal{M}_2$, where L_2 is the number of pairs of dependent positions among the total number of $\binom{L}{2}$ pairs. Similarly, a third-order PWM is $64 \times L_3$ matrix, $\mathcal{M}_3$, where L_3 is the number of triads of nucleotides of dependent positions having significant dependency chosen from the total number of $\binom{L}{3}$ triads. Pearson's χ^2, Chi-square, test statistic is used to find which pairs or triads are significantly dependent. Therefore, the Null hypothesis for our test is that nucleotides at positions i and j or for triads nucleotides at positions i, j and k are independent. Let $f_i(x)$ be the observed count of nucleotide x at position i for a given training set of N sequences and $G_{i,j}(x_1, x_2)$ be the joint observed count of occurrence of nucleotide x_1 at position i and nucleotide x_2 at position j. Then the expected count of nucleotides x_1 and x_2 occurring jointly at positions i and j is $E_{i,j}(x_1, x_2) = f_i(x_1)f_j(x_2)/N$. Let $X = \{A, T, C, G\}$ then, the χ^2 value for positions i and j is defined as

$$\chi^2(i,j) = \sum_{x_1 \in X} \sum_{x_2 \in X} \frac{(G_{i,j}(x_1,x_2) - E_{i,j}(x_1,x_2))^2}{E_{i,j}(x_1,x_2)}.$$

The χ^2 value for triad (i, j, k) can be defined in a similar way using joint observed count and expected count of trinucleotides occurring at triad (i, j, k).

We compute the p-values for χ^2 values to choose significant pair and triad candidates to form PWMs. Low p-values corresponding to large χ^2 values indicate some sort of dependency between nucleotide positions forming respective pairs and triads. In all simulation cases we used the p-value of 0.05 to select significant dependent position pairs and triads.

Having chosen the candidate pairs or triads, the entries of matrices $\mathcal{M}_2$ and $\mathcal{M}3$ will be log-odd values of the observed frequency of dinucleotides or trinucleotides in dependent positions calculated from training set and that of background model. Then, depending on the information content of each matrix one can select the matrix with higher information content to compute the score for a given sequence Y. One can also build a combined model using weighted average of the normalized scores calculated from each matrices,

$$S(Y|\mathcal{M}) = \textstyle\sum_{i=1}^{3} \omega_i \hat{S}(Y|\mathcal{M}_i)$$

where $\mathcal{M}$ is a combined model, $\hat{S}(Y|\mathcal{M}_i)$ is the normalized score of sequence Y from position matrix of order i, and $\omega_i \geq 0$ with $\sum_{i=1}^{3} \omega_i = 1$, are the coefficients weights, which can be estimated from training data as follows. Let X be the set of m known binding sites for the transcription factor F and $\omega = [\omega_1, \omega_2, \omega_3]$ be the vector of coefficients weights (we only included matrices up to order 3) such that, $\sum_{i=1}^{3} \omega_i = 1$. Then one can choose ω^* to be

$$\omega^* = \arg\max_{\omega} \sum_{y \in X} \sum_{i=1}^{3} \omega_i \hat{S}(y|\mathcal{M}_i^y).$$

Here $\mathcal{M}_i^y$ is the position weight matrix of order i constructed from all known binding sites in X excluding y.

Our analysis has revealed that when one of the models performs substantially better on the training set, that model can be used for prediction of new sites in the genome and it would have prediction power comparable to that of the combined model. In the following due to space limitation we only compare the performance of individual matrices.

3. RESULTS

To assess the performance of our algorithm and to check the richness of the higher order models in capturing the dependency structure of binding sites we used the JASPAR [7] data set of eukaryotic transcription factor binding sites matrices and E. coli transcription factor LRP's

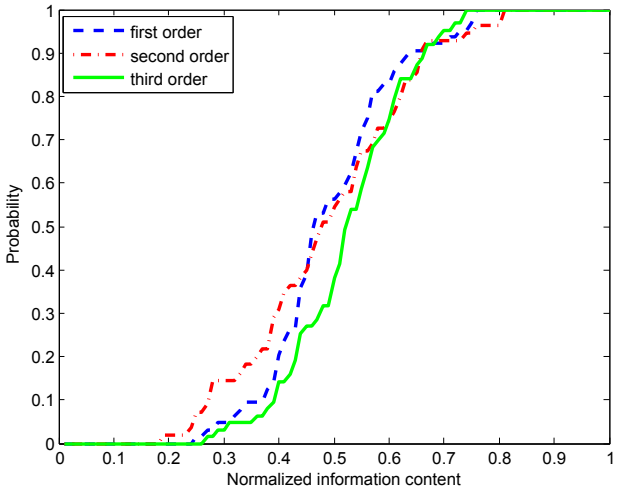

Fig. 1. Cumulative distribution function of information content of the PWM of the first, second and third-order for selected TFs in Jaspar data set.

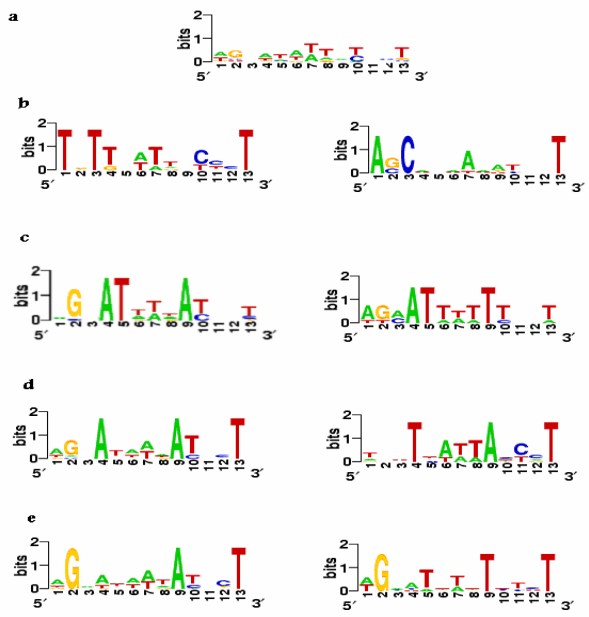

Fig. 2. Sequence Logo of Lrp binding sites, (a) Logo from whole sites, (b) Logo of 2 groups of binding sites which showed significant dependencies at positions 1,3,13 with two different trinucleotides, (c) significant dependencies at positions 4,5,9 (d) significant dependencies at positions 4,9,13 (e) significant dependencies at positions 2,9,13.

expression and binding data. Since the information content for higher order models of the sequence elements with very small number of known sites is low, we applied our model to 77 TF's with the number of known sites greater than 15 in JASPAR data set.

3.1. The learning procedures for JASPAR data set

We compared the performance of the models for each TF in data set separately by the following procedure. We ranked the scores of the known sites among the scores for random segments. This provided us with the measure of falsely discovered sites, which had higher scores than the known sites. For each transcription factor's binding site we calculated the rank of its score for each model. Let $r(Y|\mathcal{M}_i)$ be the rank of the site Y when the model i is used. For each TF, we computed the representative rank by averaging the ranks of all known sites, $R_{(\mathcal{M}_i)} = \frac{1}{N_{F_j}} \sum_{k=1}^{N_{F_j}} r(Y_k|\mathcal{M}_i)$ for transcription factor F_j having N_{F_j} known sites. A model is considered to be better for a TF if its corresponding average rank is smaller than that of other models. We assumed the rank difference is significant if for one model the average rank is more than 3 fold smaller than that of other models.

In 32 cases out of 77 TFs the second-order model outperformed the first-order one, the third-order model performed better than the first-order in 45 cases, and the third-order model was better than the second-order for 60 TFs. When there were no second- or third-order matrices, we assumed that they performed worse than first-order model in our comparison. This happened for some TFs, since no position pairs or triads passed the significance test by the χ^2 statistic. We also computed the information content of all three models (three matrices) for the selected TFs. Figure 1 depicts the cumulative distribution function of normalized information content for three models. It can be seen from the figure that the normalized information content of the third-order model is higher than that of the second- and first-order model for the majority of the TFs. This increase in average information content is due to dependency between adjacent and non adjacent positions which fully cannot be captured using first order PWM and simple Markov models.

3.2. Modeling binding sites of E-coli transcription factor, leucine responsive protein(Lrp)

We also applied our model on known binding sites for E-coli transcription factor Lrp to construct the second and the third-order matrices by selecting significant pairs and triads. To show how the higher order model increases the specificity, we identified the top most significant triad positions and the corresponding trinu-

cleotides, which contributed to the dependency test. For each dependent triad, we chose two groups of binding sites each containing a selected trinucleotide. Figure 2 shows the sequence logo of Lrp binding sites and the sequence logos of subgroups for the top 4 triads. It is clear that several positions, which have low information content based on the first-order model, have very high information content in those subgroups that can be captured with the third-order matrices.

Next for each gene, we scanned 500bp upstream of the gene and selected a site in a corresponding strand with the maximal score and ranked all selected sites for the first and third order models. The median rank of the collection of the known sites was calculated for both models. For the known target genes which have more than one binding site, we ranked only the site which had the maximal score. From gene expression microarray data [8], the median rank of 17 Lrp targets is 126. Thus, while the scores determined by the first-order model (median rank=1064) were clearly inconsistent with the observed transcriptional activity of the set, the scores from the third-order model (median rnak=50) supported the transcription data. The probability that such consistency between the median transcription and site scores occurred by chance is less than 1 in 100,000. Moreover, we found that for sensitivity of above 80%, the lists of genes were very significantly enriched for transcriptionally affected genes, when we examined the lists of genes with corresponding sites scored by the third-order model at different sensitivity cut-offs (Table 1).

We also were interested in seeing if our higher order model was assigning scores which better capture differential affinity of a regulator to the sites. To that end we used the genome-wide binding data for the Lrp protein. The relative signal intensities for each microarray probe were obtained as a result of the comparative two-color hybridization between the DNA sample bound by Lrp and specifically precipitated by Lrp antibodies and the DNA sample recovered from the cells lacking Lrp protein (manuscript in preparation). The normalized log ratios of the signal intensities from two channels were calculated and used as binding affinity scores of the sequences located upstream of the known target genes. Assuming that Lrp binds in the vicinity of the known target genes in vivo, we wanted to determine whether there was any correlation between the binding affinity score and the corresponding site score for these genes. For 17 known target genes, interestingly and consistent with our hypothesis, the scores from the first-order model did not correlate with the affinity scores, whereas the scores from the third-order model showed significant correlation with the affinities (r=0.41, with a p-value of 2.2×10^{-3}). Moreover, when we removed the two least transcriptionally responsive genes from the list, dadA and ompC, the correlation between the site scores from the 3rd-order model and affinity scores increased to 0.6, while it did not improve correlation with the 1st order affinity scores.

Table 1. Significance analysis of predicted sites using gene expression data

Sensitivity	Fraction of expressed genes (%)	p-value
1	30	3.3E-18
0.96	25	2.1E-11
0.92	20	1.1E10
0.88	17	1E-7
0.84	15	3.6E-6
0.80	14	1.8E-5

4. REFERENCES

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

[2] Y. Barash, G. Elidan, N. Friedman, T. Kaplan, "Modeling dependencies in protein-DNA binding sites," *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology* pp.28-37, Berlin ,Germany, ACM press, NY 2003.

[3] G.Yeo, C. Burge, "Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals," *Journal of Computational Biology*, vol. 11, pp. 377-394, 2004.

[4] M.L.T. Lee, M.L. Bulyk, G.A. Whitmore, G.M. Church, "A statistical model for investigating binding probabilities of DNA nucleotide sequences using microarrays," *Biometrics*, 2003, vol. 58, pp. 981-988, 2003.

[5] T.K. Man, G.D. Stormo, "Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity(QuMFRA) assay," *Nucleic Acids Research*, vol. 29, pp. 2471-2478, 2003.

[6] K. Ellrott, C. Yang, F.M. Sladek, T. Jiang, "Identifying transcription factor binding sites through Markov chain optimization," *Bioinformatics*, vol. 18, pp. 100-109, 2002.

[7] S. Albin, A. Wynand, E. Par, W. Wyeth, L. Boris, "JASPAR: an open access database for eukaryotic transcription factor binding profiles," *Nucleic Acids Res.* Vol. 32, no.1 Database Issue, 2003.

[8] T.H. Tani, A.B. Khodursky, R.M. Blumenthal, P.O. Brown, R.G. Matthews, "Adaptation to famine: a family of stationary-phase genes revealed by microarray analysis," *Proc Natl Acad Sci U S A* vol. 99, pp. 13471-13476, 2002.