PROBABILISTIC FRAMEWORK FOR TRANSCRIPTION FACTOR BINDING PREDICTION

Harri Lähdesmäki and Ilya Shmulevich

Institute for Systems Biology, Seattle WA 98103, USA

ABSTRACT

We formulate a probabilistic framework for transcription factor (TF) binding prediction that is built on the standard position specific frequency matrix (PSFM) and higher order Markovian background models. Contrary to the traditional hypothesis testing based methods which report a significance (p) value of TF binding at every possible base pair position in a promoter sequence, we develop a probabilistic methodology to assess TF binding to whole promoter sequences. Performance of the proposed method is demonstrated via simulations.

1. INTRODUCTION

Transcriptional regulation is a central control mechanism for many biological processes, such as cell cycle and the immune response. Transcriptional regulation generally involves DNA-binding proteins, transcription factors (TF), that control expression of genes by binding to short regulatory sequence motifs in gene promoters. Experimental studies and computational approaches are extending our knowledge about TF binding specificities to different sequence motifs. Relatively little, however, is known about genome-wide TF binding to gene promoters. Thus, transcription factor binding site (TFBS) prediction remains an important problem in computational biology.

Computational approaches that make use of known TF binding specificities (e.g. TRANSFAC database [1]) and promoter sequences to predict putative TF binding sites are now widely used in computational biology. Binding site prediction tools have previously been formulated as hypothesis testing methods where a significance value of TF binding at a specific sequence position is obtained by comparing a test statistic to a null distribution. Predicted binding sites are then the ones that exceed a selected significance level. Here we formulate a probabilistic framework for predicting transcription factor binding that differs from the standard hypothesis testing approaches in two important ways. First, the proposed framework is probabilistic in nature and thus outputs a probability of binding (as opposed to a p-value), which directly assesses our belief of having a binding site. Secondly, the proposed method answers the question of whether the whole promoter has a binding site (as opposed to reporting a p-value for every possible position in the sequence), although it is straightforward to extend it to single base pair resolution.

The most commonly used probabilistic formulation for binding motifs is the position specific frequency matrix (PSFM) model [2]. Different modifications of the PSFM model based hypothesis testing have been proposed to predict binding sites [3, 4, 5, 6]. Our probabilistic framework is built on the same PSFM model. Similar probabilistic frameworks have been previously proposed in the context of motif discovery, e.g., in [7, 8, 9, 10]. Note, however, that our goal is not *de novo* motif discovery but probabilistic prediction of TFBS, given *a priori* information of TF binding specificities in form of a PSFM model.

2. METHODS

Let $S = (s_1, \ldots, s_L)$ denote a promoter sequence, where $s_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ and L is the length of the sequence. Let Q denote the number of (hidden) motif instances in sequence S. This is one of the key quantities estimated from the data. Further, let A denote the (unknown) start positions of non-overlapping motif instances in sequence S. For example, if $Q = c$, then $A = \{a_1, \ldots, a_c\}$. So, a promoter consists of c motif instances and $c+1$ background sequence chunks, some of which can be empty.

Non-binding background sequence locations are modeled by the commonly used dth order Markovian background model ϕ. That is, let $\phi(s_i) = P_\phi(s_i | s_{i-d}, s_{i-d+1}, \ldots, s_{i-1})$ denote the probability of observing nucleotide s_i at the ith position of a promoter sequence S in the background model ϕ given d previous nucleotides. For simplicity, we assume that for positions $i \leq d$ we have access to $s_{-d+1}, \ldots, s_0$. The likelihood of the background model, $A = \emptyset$, is thus $P(S|A = \emptyset, \phi) = P(S|\phi) = \prod_{i=1}^{L} \phi(s_i)$. Motifs are modeled using the standard PSFM model θ which is a product of independent multinomial distributions. Similarly as above, let $\theta(s_i, j) = P_\theta(s_i, j)$ denote the probability of observing nucleotide s_i at the jth ($j = 1, \ldots, \ell$) position of a motif model θ and ℓ is the length of the motif. The likelihood of sequence S, given non-overlapping motif positions and motif and background models, is

$$P(S|A,\theta,\phi) = \underbrace{\prod_{i_1=1}^{a_1-1} \phi(s_{i_1})}_{\text{background 1}} \underbrace{\prod_{j_1=a_1}^{a_1+\ell-1} \theta(s_{j_1}, j_1 - a_1 + 1)}_{\text{motif instance 1}} \times \cdots \times \underbrace{\prod_{j_c=a_c}^{a_c+\ell-1} \theta(s_{j_c}, j_c - a_c + 1)}_{\text{motif instance } c} \underbrace{\prod_{i_{c+1}=a_c+\ell}^{L} \phi(s_{i_{c+1}})}_{\text{background } c+1}. \quad (1)$$

1-4244-0999-3/07/$25.00 ©2007 IEEE.

Let us rewrite Equation (1) more compactly as

$$P(S|A,\theta,\phi) = \prod_{i=1}^{L}\phi(s_i)\prod_{j=1}^{|A|}\prod_{k=0}^{\ell-1}\frac{\theta(s_{a_j+k},k+1)}{\phi(s_{a_j+k})} = P(S|\phi)\prod_{j=1}^{|A|}W_{a_j}, \quad (2)$$

where $|A| = Q = c$ and $W_{a_j} = \prod_{k=0}^{\ell-1}\frac{\theta(s_{a_j+k},k+1)}{\phi(s_{a_j+k})}$.

2.1. One motif model θ

Using Bayes' rule, the probability of c motif instances, given the sequence S, is [7]

$$P(Q=c|S,\theta,\phi) = \frac{P(S|Q=c,\theta,\phi)P(Q=c|\theta,\phi)}{P(S|\theta,\phi)}, \quad (3)$$

where the normalization factor has the following form

$$P(S|\theta,\phi) = \sum_{c=0}^{\lfloor\frac{L}{\ell}\rfloor}P(S|Q=c,\theta,\phi)P(Q=c|\theta,\phi) \quad (4)$$

and $\lfloor\frac{L}{\ell}\rfloor$ is the maximum number of non-overlapping ℓ-length motifs in an L-length sequence. Note that since the sum in Equation (4) has only $\lfloor\frac{L}{\ell}\rfloor + 1$ terms (instead of ∞) the normalization factor can be computed exactly. The likelihood of sequence S, given that it contains c motif instances, can be obtained by summing over all possible positions A of c motif instances [7]

$$\begin{aligned}&P(S|Q=c,\theta,\phi)\\ &= \sum_{A:|A|=c}P(S|A,Q=c,\theta,\phi)P(A|Q=c,\theta,\phi)\\ &= \sum_{a_1=1}^{L-c\ell+1}\cdots\sum_{a_c=a_{c-1}+\ell}^{L-\ell+1}P(S|\phi)\prod_{j=1}^{c}W_{a_j}\\ &\quad\times P(A|Q=c,\theta,\phi), \end{aligned} \quad (5)$$

where in the last equality we have used $P(S|A,Q=c,\theta,\phi) = P(S|A,\theta,\phi)$ and Equation (2). The above probabilistic formulation (Equations (3)–(5)) is practically identical to the one proposed by Thijs *et al.* in [7].

As in [7], let us assume that, for a fixed value of Q, the prior over motif positions A is uniform and is inversely proportional to the number of different motif positions, i.e., $P(A|Q=c,\theta,\phi) = \prod_{i=1}^{c}\frac{i}{L-c\ell+i}$. Let $R(S|Q=c,\theta,\phi)$ denote the sum in Equation (5) without the (constant) prior term $P(A|Q=c,\theta,\phi)$. The likelihood in Equation (5) can be computed efficiently using, e.g., the following recursion

$$\begin{aligned}R(S|Q=c,\theta,\phi) &= \sum_{a_1=1}^{L-c\ell+1}\cdots\sum_{a_c=a_{c-1}+\ell}^{L-\ell+1}P(S|\phi)\prod_{j=1}^{c}W_{a_j}\\ &= \sum_{a_1=1}^{L-c\ell+1}W_{a_1}\sum_{a_2=a_1+\ell}^{L-(c-1)\ell+1}\cdots\sum_{a_c=a_{c-1}+\ell}^{L-\ell+1}P(S|\phi)\prod_{j=2}^{c}W_{a_j}\\ &= \sum_{a_1=1}^{L-c\ell+1}W_{a_1}R(S_{a_1+\ell}|Q=c-1,\theta,\phi)\end{aligned}$$

where $S_{a_1+\ell} = (s_{a_1+\ell},\ldots,s_L)$ denotes a subsequence of S (note that $S_1 = S$). For the prior over the number of motif instances, we use a probability distribution motivated by previous studies [7]

$$P(Q=c|\theta,\phi) \sim \left[\frac{3}{4},\frac{\gamma}{C},\frac{\kappa\gamma}{C},\frac{\kappa^2\gamma}{C},\ldots,\frac{\kappa^{\lfloor\frac{L}{\ell}\rfloor-1}\gamma}{C}\right], \quad (6)$$

where $C = 4\sum_{i=0}^{\lfloor\frac{L}{\ell}\rfloor-1}\kappa^i\gamma$.

Assuming a single binding position is sufficient for transcriptional regulation, the probability that a given transcription factor (defined by θ) regulates a gene having promoter sequence S, denoted by $\theta \to S$, can be computed as

$$\begin{aligned}P(\theta\to S|S,\theta,\phi) &= P(Q>0|S,\theta,\phi) &(7)\\ &= \sum_{c=1}^{\lfloor\frac{L}{\ell}\rfloor}P(Q=c|S,\theta,\phi) &(8)\\ &= 1-P(Q=0|S,\theta,\phi), &(9)\end{aligned}$$

where $P(Q=c|S,\theta,\phi)$ can be obtained using Equations (3)–(6). The above construction also allows to compute the expected number of motif instances $\mathbb{E}[Q|S,\theta,\phi] = \sum_{c=0}^{\lfloor\frac{L}{\ell}\rfloor}c\cdot P(Q=c|S,\theta,\phi)$ as proposed in [7].

2.2. Multiple motif models Θ

A transcription factor typically recognizes several binding sites and is therefore characterized by several motif models (i.e., matrices) $\Theta = (\theta^{(1)},\ldots,\theta^{(m)})$ each having length ℓ_i. Let $\pi \in \{1,\ldots,m\}^c$ denote a configuration of motif models from Θ in A. That is, π_i specifies the motif model at location a_i. For notational convenience, define

$$W_{a_j}^{(\pi_j)} = \begin{cases}\prod_{k=0}^{\ell_{\pi_j}-1}\frac{\theta^{(\pi_j)}(s_{a_j+k},k+1)}{\phi(s_{a_j+k})}, & \text{if } 1\le a_j\le L-\ell_j+1,\\ 0, & \text{else,}\end{cases}$$

and note that (see also Equation (2))

$$P(S|A,\pi,\Theta,\phi) = P(S|\phi)\prod_{j=1}^{c}W_{a_j}^{(\pi_j)}. \quad (10)$$

The probability of c motif instances can be obtained using Bayes' rule as in Equations (3)–(4) but θ replaced with Θ. Further, following Equation (5), the likelihood of sequence S given c motif instances can be obtained by summing over all possible positions and configurations

$$\begin{aligned}&P(S|Q=c,\Theta,\phi)\\ &= \sum_{\pi\in\{1,\ldots,m\}^c}\sum_{A:|A|=c}P(S|A,\pi,Q=c,\theta,\phi)\\ &\quad\times P(A,\pi|Q=c,\Theta,\phi)\\ &= \sum_{\pi\in\{1,\ldots,m\}^c}\sum_{a_1=1}^{L-c\ell_{\min}+1}\cdots\sum_{a_c=a_{c-1}+\ell_{\pi_{c-1}}}^{L-\ell_{\min}+1}P(S|\phi)\\ &\quad\times\prod_{j=1}^{c}W_{a_j}^{(\pi_j)}P(A,\pi|Q=c,\Theta,\phi),\end{aligned} \quad (11)$$

where $\ell_{\min} = \min\{\ell_1, \ldots, \ell_m\}$. Let us again assume a uniform prior over motif positions A and configurations π (for each fixed value of Q), and let $R(S|Q = c, \Theta, \phi)$ denote the sum in Equation (11) without the (constant) prior term $P(A, \pi|Q = c, \theta, \phi)$. A computationally efficient recursive formula can be written, e.g., as

$$
\begin{aligned}
& R(S|Q = c, \Theta, \phi) \\
&= \sum_{\pi \in \{1,\ldots,m\}^c} \sum_{a_1=1}^{L-c\ell_{\min}+1} \cdots \sum_{a_c=a_{c-1}+\ell_{\pi_{c-1}}}^{L-\ell_{\min}+1} P(S|\phi) \\
&\quad \times \prod_{j=1}^{c} W_{a_j}^{(\pi_j)} \\
&= \sum_{\pi_1 \in \{1,\ldots,m\}} \sum_{a_1=1}^{L-c\ell_{\min}+1} W_{a_1}^{(\pi_1)} \sum_{(\pi_2,\ldots,\pi_c) \in \{1,\ldots,m\}^{c-1}} \\
&\quad \times \sum_{a_2=a_1+\ell_{\pi_1}}^{L-(c-1)\ell_{\min}+1} \cdots \sum_{a_c=a_{c-1}+\ell_{\pi_{c-1}}}^{L-\ell_{\min}+1} P(S|\phi) \prod_{j=2}^{c} W_{a_j}^{(\pi_j)} \\
&= \sum_{\pi_1 \in \{1,\ldots,m\}} \sum_{a_1=1}^{L-c\ell_{\min}+1} W_{a_1}^{(\pi_1)} \\
&\quad \times R(S_{a_1+\ell_{\pi_1}}|Q = c-1, \Theta, \phi).
\end{aligned}
$$

A closed form formula for uniform $P(A, \pi|Q = c, \Theta, \phi)$ is more difficult to obtain in general, but it can be computed numerically using a similar recursion as the one above.

The prior $P(Q = c|\Theta, \phi)$ depends now on Θ and thus can be adjusted for multiple motifs. However, it is unrealistic to assume that different motif models $(\theta^{(1)}, \ldots, \theta^{(m)})$ are independent. Indeed, it is likely that they are strongly dependent. Therefore, we use the same prior as in the case of a single motif model as a first approximation.

Let us assume that a TF characterized by Θ can transcriptionally regulate a gene having promoter S if at least one of the motifs in Θ has a binding site in S. Then the final probabilities $P(\Theta \to S|S, \Theta, \phi)$ can be computed as in Equations (7)–(9) but θ replaced with Θ.

The above probabilistic modeling framework that incorporates multiple motif models can be viewed as an extension of a framework proposed in [7]. Note that the proposed framework is also similar to a hidden Markov model (HMM) proposed together with a so called w-score in the context of motif discovery by Sinha [9]. A HMM is defined by motif and background models Θ and ϕ and so called transition probabilities (between states of HMM) whereas the proposed modeling framework is built on motif and background models alone, with additional information brought into the computation via the priors $P(Q|\Theta, \phi)$ and $P(A, \pi|Q, \Theta, \phi)$.

3. SIMULATIONS

In this section we demonstrate the performance of the proposed computational methods. Parameters of the background model are estimated from a separate background sequence data set using the maximum-likelihood principle. Higher-order background models are typically found to perform better [11], but the number of parameters needed to describe a model increases exponentially with the model order. We found that $d = 2$ provides a good trade-off and use it here. Parameters of motif models are taken directly from TRANSFAC (professional version 10.3) and each column is normalized separately to yield a PSFM model. With these parameter settings, we applied our computational methods to a validation sequence set that is explained below.

We focus on three mouse transcription factors `Ddit3`, `Hoxa9`, and `Tcf1` which are associated with 1, 2, and 4 motif models (i.e., matrices), respectively. We use similar artificially generated data as in [12] except here each promoter can contain more than one motif instance. Promoter sequences are first set to contain a genomic "junk" sequence from mouse, each of length $L = 1000$. These promoters are assumed to correspond to the background model and contain no real binding sites. Different versions of these background promoter sequences are then obtained by artificially inserting motif instances: 1 for `Ddit3`, 1 and 2 for `Hoxa9`, and 1, 2 and 4 for `Tcf1`. Positions of non-overlapping motif instances, as well as the order in which multiple motifs are inserted, are chosen uniformly randomly. Motif instances are randomly sampled from the known motif models (products of multinomial distributions). The same motif models are used both in data generation and inference.

We report histograms of the binding probabilities for different TFs, i.e., $1 - P(Q = 0|S, \Theta, \phi)$, each obtained from 100 promoter sequences. Figure 1 shows the histograms of binding probabilities for `Ddit3`. Blue (resp. green) graph shows the histogram for background sequences (resp. after randomly adding one motif instance). Figure 2 shows the histograms of binding probabilities for `Hoxa9` for background sequences (blue) and sequences containing one/two randomly added motif instances (green/red). Finally, Figure 3 shows the histograms of binding probabilities for `Tcf1` for background sequences (blue) and sequences containing one/ two/four randomly added motif instances (green/red/black).

Figures 1–3 show that the proposed method is well capable of distinguishing background sequences (that do not contain any real binding sites) from the ones that do contain at least one binding site. Naturally, perfect separation is impossible as the motifs are randomly sampled from the probabilistic motif models. The degree of separation also depends on motif model(s) associated with a TF. PSFMs that have higher information content are generally easier to detect and hence provide better separation. Further, better separation is achieved for sequences that contain multiple motif instances. Note that although we demonstrate the proposed method by its ability to distinguish background sequences from sequences that contain a binding site, the method is fundamentally generative and the full potential lies in its probabilistic nature. In other words, the methodology outputs probabilities (not p-values) that

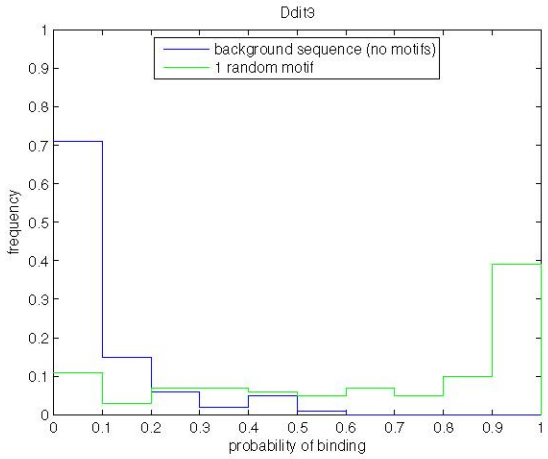

Figure 1. Histograms of binding probabilities for `Ddit3`.

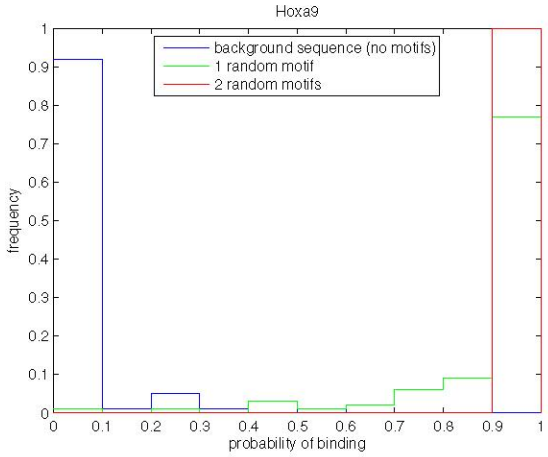

Figure 2. Histograms of binding probabilities for `Hoxa9`.

directly reflect our belief of having (at least one) binding site.

4. DISCUSSION

We have developed a probabilistic formulation for predicting TF binding. The method uses the standard higher order Markovian background model and the PSFM matrix as its' building blocks. An artificial simulation data set (for which we know the ground truth) clearly demonstrates the potential of the method.

PSFM is currently the most commonly used motif model although it is incapable of representing dependencies between nucleotides within binding site, which have been observed in real sequences [13]. Thus, a natural, and more or less straightforward, extension of the proposed framework would include the use of, e.g., generalized weight matrices that incorporates pair-wise dependencies [14] or Bayesian networks [15] to model binding sites.

5. REFERENCES

[1] E. Wingender, *et al.*, "TRANSFAC: an integrated system for gene expression regulation," *Nucleic Acids Research*, vol. 28, no. 1, pp. 316–319, 2000.

[2] R. Staden, "Computer methods to locate signals in nucleic acid sequences," *Nucleic Acids Research*, vol. 12, no. 1, pp. 505–519, 1984.

[3] K. Quandt, *et al.*, "MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data," *Nucleic Acids Research*, vol. 23, no. 23, pp. 4878–4884, 1995.

[4] J.-M. Claverie and S. Audic, "The statistical significance of nucleotide position-weight matrix matches," *Bioinformatics*, vol. 12, no. 5, pp. 431–439, 1996.

[5] L. Hertzberg, *et al.*, "Finding motifs in promoter regions," *Journal of Computational Biology*, vol. 12, no. 3, pp. 314–330, 2005.

[6] I. V. Bajić, "Detection-theoretic analysis of MatInspector," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2388–2393, 2006.

[7] G. Thijs, *et al.*, "A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes," *Journal of Computational Biology*, vol. 9, no. 2, pp. 447–464, 2002.

[8] D. J. Reiss and B. Schwikowski, "Predicting protein-peptide interactions via a network-based motif sampler," *Bioinformatics*, vol. 20, no. Suppl. 1, pp. i274–i282, 2004.

[9] S. Sinha, "On counting position weight matrix matches in a sequence, with application to discriminative motif finding," *Bioinformatics*, vol. 22, no. 14, pp. e454–e463, 2006.

[10] W. P. Lehrach, *et al.*, "A regularized discriminative model for the prediction of protein-peptide interactions," *Bioinformatics*, vol. 22, no. 5, pp. 532–540, 2006.

[11] G. Thijs, *et al.*, "A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling," *Bioinformatics*, vol. 17, no. 12, pp. 1113–1122, 2001.

[12] M. Tompa, *et al.*, "Assessing computational tools for the discovery of transcription factor binding sites," *Nature Biotechnology*, vol. 23, no. 1, pp. 137–144, 2005.

[13] M. L. Bulyk, *et al.*, "Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors," *Nucleic Acids Research*, vol. 30, no. 5, pp. 1255–1261, 2002.

[14] Q. Zhou and J. S. Liu, "Modeling within-motif dependence for transcription factor binding site predictions," *Bioinformatics*, vol. 20, no. 6, pp. 909–916, 2004.

[15] Y. Barash, *et al.*, "Modeling dependencies in protein-DNA binding sites," in *Proceedings of the Seventh Annual International Conference on Computational Biology (RECOMB 2003)*. 2003, ACM.

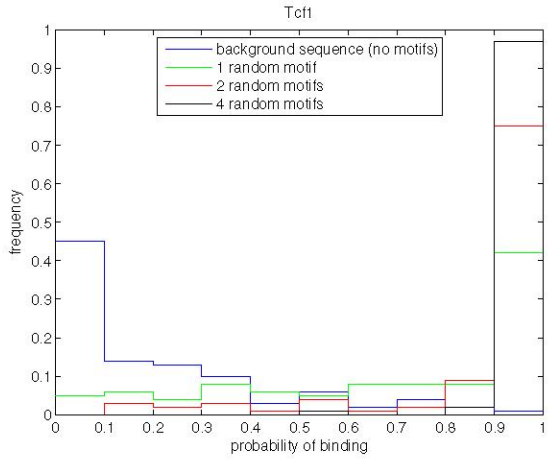

Figure 3. Histograms of binding probabilities for `Tcf1`.