A New Statistical Model to Select Target Sequences Bound by Transcription Factors

Utz J. Pape[1,2] utz.pape@molgen.mpg.de
Steffen Grossmann[1] grossman@molgen.mpg.de
Stefanie Hammer[3] hammer@molgen.mpg.de
Silke Sperling[3] sperling@molgen.mpg.de
Martin Vingron[1] martin.vingron@molgen.mpg.de

[1] Computational Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany
[2] Mathematics and Computer Science, Free University of Berlin, Berlin, Germany
[3] Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin, Germany

Abstract

Transcription factors (TFs) play a key role in gene regulation by binding to target sequences. *In silico* prediction of potential binding to a sequence is a main task in computational biology. Although many methods have been proposed to tackle this problem, the statistical significance of the prediction is still not solved. We propose an approach to give a good approximation for the potential of a sequence to be bound by a TF. Instead of assessing distinct binding sites, we motivate to focus on the number of binding sites. Based on a suitable statistical model, probabilities for scoring are approximated for a TF to bind to a sequence. Two examples show the necessity of such a model as well as the superiority of the proposed method compared to standard approaches.

Keywords: position weight matrix, binding site clusters, count statistics, number of occurrences, overlapping occurrences

1 Introduction

Prediction of transcription factor binding sites (TFBSs) is a crucial task in computational biology [11]. The main problem to be addressed is the statistical significance of the prediction. Searching in a long genomic sequence for a small binding site rises the problem of a high number of both false positives and false negatives. Using a statistical approach, it is possible to assess the significance of a predicted site. Position Weight Matrices (PWMs) are widely used because they have been successfully applied in many cases [1] and the retrieved scores are proportional to physical binding energy [10]. So far, most statistics deal with the detection of single binding sites although a higher number of binding sites increase the probability of binding. In addition, many problems require a score indicating whether a TF binds anywhere on a sequence instead of a distinct binding site. In such a setting, it is necessary to construct a score based on the number of significant occurrences on a sequence.

We develop a statistical model to compute the probability to find as many hits as observed by chance. We use this probability as a score. A main issue is the overlapping structure of the PWM. For example, a PWM with the consensus 'CTAACT' has a higher probability to find two hits overlapping in two positions than to find two independent hits. This problem has been discussed in broad range for word counting problems [6]. One solution is the computation of correction terms for the overlapping structure of words because the overlapping structure is discrete and can be explicitly captured by enumeration. Since the PWM is a probabilistic description of the consensus word, overlaps can occur at any position. Therefore, we use the discrete nature of the score to compute the probabilities of overlapping hits. Based on these probabilities, we can give a good approximation for the probabilities of any number of overlapping hits. We present such a score based on a statistical model and show the advantages over naive scores.

2 Methods

We extend a statistical approach to compute the score distribution on a random sequence as well as on a binding site model [5]. We use this approach and extend it such that we can compute the score distribution over a mixture of a random sequence and a binding site model. This distribution approximates the score distribution on a random sequence where we already detected a binding site.

The PWM is the representation of the binding site. It contains the probabilities for each nucleotide at every position. We assume that the binding sites of each TF is described by only one PWM. An extension to more than one PWM is not trivial. The scoring scheme for every position is retrieved by the log-likelihood ratios of the nucleotide distribution of the PWM and the background model. The background model is an i.i.d. model incorporating the average GC content of the upstream sequence. The resulting scoring matrix is called position-specific scoring matrix (PSSM). The PSSM assigns a score to every position of the potential binding site depending on the observed nucleotide. Sliding a window over the sequence of the same length as the PWM and summing up the scores in each window, yields a score for every position of the sequence. At first, we focus on one strand of the sequence, only. Subsequently, we extend the approach to deal with the complementary strand, as well.

2.1 Statistics for Binding Site Detection

We call a position a hit if the corresponding window yields a score s higher than a certain threshold t. The threshold controls the probabilities α and β given by $\alpha = P_{H_0}(s \geq t)$ and $\beta = P_{H_1}(s < t)$ where H_0 is the null model having a random sequence and H_1 the model for the binding site. α has to be very small as the expected number of false positives on a sequence of length n is $n \cdot \alpha$. For long sequences the number becomes very large. Therefore, we control $\alpha_m = 1-(1-\alpha)^m \approx 1-\exp(-m\alpha)$ which is the probability to find at least one false positive on a sequence of length m. We do not set m equal to the sequence length n as this leads to such a high threshold t that we hardly get any hits. Instead, we use the heuristic $m = 500$. This choice is arbitrary but detects a reasonable number of both true and false positives. In general, the approach is robust against the actual threshold because α is incorporated into the computation of the probability as long as neither all positions are hits nor none.

Now, we can set $t = t_{bal}$ such that $\alpha_{500} = \beta$ and call it balanced threshold. We compute this threshold by simulating the score distribution according to the background model and the PWM model. Unfortunately, a minor fraction of PWMs contain nucleotide distributions very similar to the background distribution. In these cases, we retrieve very poor probabilities for α and β using the balanced threshold. This results in many false positives diminishing the performance of the prediction. In these special cases, we set the threshold $t = \tilde{t}$ such that $\alpha_{500} = 0.1$.

Next, we show how to compute a score for each upstream sequence reflecting its potential to be bound by a TF. Since the threshold cannot be lowered unlimited, we expect a certain amount of false positives for long sequences. Hence, we cannot use the occurrence of at least one hit as an indicator. In addition, the different lengths of the upstream sequences require a model including the sequence lengths. Therefore, we compute the probability $p = P_{H_0}(X \geq x)$ where X is the random variable for the number of occurrences on a random sequence of length n and x is the observed number of hits on the real sequence.

2.2 An Improved Background Model

In order to compute the probability p, we need an appropriate background model H_0. We assume the positions of the sequence to be i.i.d. The nucleotide distribution at each position is only determined by the GC content of the potential target sequence. We restrict ourselves to the GC content instead of base pair composition due to the extension for both strands we introduce later. In contrast to coding sequence, there is no motivation to handle both strand in the upstream region differently. A simple approach would approximate p by $\hat{p}_1 = 1 - B(x; n, \alpha)$ where $B(\cdot)$ is the binomial distribution with parameters x for number of successes, n number of trials and α probability of success. The number of successes corresponds to the number of detected hits. This model also assumes that a hit is independent of a previous overlapping hit. We propose a more sophisticated model considering the strongest dependencies of overlapping hits because the independence assumption is arguable.

We achieve this by shifting the focus from hits to clusters where a cluster is a collection of overlapping hits. Hence, we exchange the random variable X for the number of hits by Z_I which is the number of clusters with I overlapping hits. We say the cluster has size I. In the special case of $I = 1$, Z_1 is the number of clusters of size 1. This number is equal to the number of hits without any overlaps. Again, we use the binomial distribution to compute $P_{H_0}(Z_I = z_I) = B(z_I; n, \alpha_I)$ where α_I is the probability to find a cluster of size I. The computation of α_I is only tractable for small I. α_I becomes smaller with increasing I because the probability of further hits is always smaller than the probability for no hits. Thus, it is not necessary to compute α_I for large I. In general, α_I becomes negligibly small for $i > 2$ although it is possible to construct rare artificial matrices where this is not true. In the following, we show how to compute α_i.

Let Y_j be an indicator variable which is equal to one if we have a hit starting at position j otherwise zero. The probability to have a hit in a random sequence is $P_{H_0}(Y_j = 1)$. Then, α_1 corresponds to the event that there is exactly one hit at a certain position while no overlapping hits occur. Using the fact that an overlapping hit is a hit within the range of the length of the PWM minus 1 denoted by l, we get:

$$\begin{aligned} \alpha_1 &= P_{H_0}(Y_{j-l} = 0, \ldots, Y_{j-1} = 0, Y_j = 1, Y_{j+1} = 0, \ldots, Y_{j+l} = 0) \\ &= P_{H_0}(Y_{j-l} = 0, \ldots, Y_{j-1} = 0, Y_{j+1} = 0, \ldots, Y_{j+l} = 0 | Y_j = 1) \cdot P_{H_0}(Y_j = 1)\,. \end{aligned} \tag{1}$$

The conditional probability in the last part of (1) is hard to compute because the events in the collection $(\{Y_{j+k} = 0\})_{-l \leq k \leq l, k \neq 0}$ are not independent, given $\{Y_j = 1\}$. However, in a first order approximation we do as if independence would hold here and compute

$$\begin{aligned} \alpha_1 &\approx P_{H_0}(Y_j = 1) \prod_{k=-l, k \neq j}^{l} P_{H_0}(Y_{j+k} = 0 | Y_j = 1) \\ &= P_{H_0}(Y_j = 1) \prod_{k=-l, k \neq j}^{l} (1 - P_{H_0}(Y_{j+k} = 1 | Y_j = 1)). \end{aligned} \tag{2}$$

As described above $P_{H_0}(Y_j = 1) = \alpha$, we only have to compute $P_{H_0}(Y_{j+k} = 1 | Y_j = 1)$ for k as given above. Thus, we need to compute the probability that the score at position $j + k$ exceeds the threshold given there is a hit at position j: $P_{H_0}(S^{(j+k)} \geq t | Y_j = 1)$. Again, this conditional probability is difficult to calculate, because conditioning on $\{Y_j = 1\}$ destroys the position independence of the nucleotide distributions at the positions covered by the hit at position j. Therefore, we make a second approximation by replacing the true conditional distribution by the position specific (and position independent) nucleotide distribution specified by the PWM. This enables us to compute the relevant quantities as convolutions.

Let S_κ be the random variable for the score at the PWM position κ and $f_{\kappa,\tau}(s)$ its probability mass function with $P_{H_1}(S_\kappa = s) = f_{\kappa,\tau}(s)$ for $\tau = \kappa$. For the mixture of PWM and background model, we have to differentiate between the position κ the score is retrieved from and the position τ the nucleotide distribution is sampled from. Similarly, $P_{H_0}(S_\kappa = s) = g_\kappa(s)$ is the probability for getting a score s at PWM position κ while sampling from the background distribution which is assumed to be i.i.d. The distribution of the overall score S for a sequence of length $l + 1$ is the convolution $f(s) = (f_{1,1} * \ldots * f_{l+1,l+1})(s)$ under H_1 respectively $g(s) = (g_1 * \ldots * g_{l+1})(s)$ under H_0. Now, we can compute $P_{H_0}(S^{(j+k)} = s|Y_j = 1)$ using the local approximation $P_{H_1}(S^{(j+k)} = s)$ implying the nucleotides of the hit at sequence position j to $j + l + 1$ to be distributed according to the PWM:

$$\begin{aligned} & P_{H_0}(S^{(j+k)} = s|Y_j = 1) \\ = & \begin{cases} \left(f_{1,k} * f_{2,k+1} * \ldots * f_{l+1-(k-1),l+1} * g_{l+1-(k-1)+1} * \ldots * g_{l+1}\right)(s) & : k > 0 \\ \left(g_1 * \ldots * g_{-k} * f_{1-k,1} * f_{2-k,2} * \ldots * f_{l+1,l+1+k}\right)(s) & : k < 0 \end{cases} . \end{aligned} \tag{3}$$

Based on the explicit distribution of $P_{H_0}(S^{(j+k)} = s|Y_j = 1)$, we can compute $P_{H_0}(S^{(j+k)} \geq t|Y_j = 1) = P_{H_0}(Y_{j+k} = 1|Y_j = 1)$ given t. With these results, we can compute α_1 using (2). As α_i is the probability to have a cluster with exact i hits, we can approximate α_{i+1} for $i > 0$ by

$$\alpha_{i+1} \approx \sum_{k_1=-l}^{l} \cdot \ldots \cdot \sum_{k_i=-l}^{l} \alpha_1 \cdot \prod_{k \in \{k_1,\ldots,k_i\}} \frac{P_{H_0}(Y_{j+k} = 1|Y_j = 1)}{1 - P_{H_0}(Y_{j+k} = 1|Y_j = 1)} , \tag{4}$$

where $k_1 \neq j$, $k_2 \neq j \wedge k_2 \neq k_1$, and so on. Thus, we neglect dependencies between the additional hits. As the probabilities of these additional hits are very small, the error introduced can be neglected. As already mentioned, the α_i become smaller with increasing i. Thus, it is possible to skip the computation either if α_i has reached a certain threshold or index ε. The probability for clusters with bigger sizes can be approximated by

$$\alpha_{\varepsilon+} = \alpha - \sum_{i=1}^{\varepsilon} \alpha_i .$$

Now, the probability can be computed for observed number of clusters $\mathbf{z} = (z_1, \ldots, z_\varepsilon, z_{\varepsilon+})$ where z_i is the number of observed clusters with size i and $z_{\varepsilon+}$ denotes the number of clusters with size larger than ε:

$$\begin{aligned} p &= P_{H_0}(\mathbf{Z} \geq \mathbf{z}) \\ &= \prod_{i=1}^{\varepsilon} (1 - B(z_i; n, \alpha_i)) \cdot (1 - B(z_{\varepsilon+}; n, \alpha_{\varepsilon+})) . \end{aligned}$$

2.3 Extension for Both Strands

We only consider overlapping hits on the same strand in (1). Searching on both strands introduces the dependency of hits between the strands. Especially PWMs with a palindrome structure containing mainly G and C or A and T, e.g. with a consensus like 'CCCGGG', have a substantially increased probability to find an overlapping hit on the complementary strand. Considering the background model, we modify the definition of α_1 and accordingly α_i by adding a correction term for hits on the complementary strand:

$$\alpha_1 \approx P_{H_0}(Y_j = 1) \prod_{k=-l,k\neq j}^{l} (1 - P_{H_0}(Y_{j+k} = 1|Y_j = 1)) \prod_{k=-l}^{l} (1 - P_{H_0}(Y'_{j+k} = 1|Y_j = 1)) .$$

Here, Y'_k corresponds to Y_k on the complementary strand. We define the position indices on the complementary strand such that two hits Y_k and Y'_k completely overlap. There, $k = j$ is not excluded in the second product because a hit on the complementary strand at the same position as the first hit is counted as a second hit. Furthermore, we change (3) accordingly by substituting f and g with probability mass functions based on the complementary PWM nucleotide distribution f' respectively background nucleotide distributions g'. For the complementary strand, we have to cover the case $j = k$, as well. Then, we have with S' denoting the score on the complementary strand

$$P_{H_0}(S'^{(j+k)} = s|Y_j = 1) = \left(f'_{1,l+1} * \ldots * f'_{l+1,1}\right)(s)\,.$$

2.4 Data

The two main databases holding PWMs are Jaspar [7] and Transfac [4]. It is known that many PWMs have a high self similarity or a palindrome structure. To underline the practical relevance of our approach, we use two real PWMs throughout the results section as well as real sequences. We select the Transfac PWMs 'M00184' for the TF myoblast determining factor (MYOD) and 'M00186' for the serum response factor (SRF) to illustrate the behaviour of our approach in comparison to naive models. Both TFs play a central role in heart malformations [3]. The PWMs are preprocessed by adding position-specific pseudo-counts depending on the information content per position [5].

We choose the two real sequences from the genes MEF2C and IRX4. The human (ENSG00000081189 and ENSG00000113430) and mouse (ENSMUSG00000005583 and ENSMUSG00000021604) sequences are retrieved from Ensembl [2]. We apply the Smith-Waterman algorithm [9] to find local alignments between the human and the mouse sequences. The aligned parts of the sequences are called conserved. We mask the non-conserved parts of the upstream sequences as conservation increases the possibility of a functional site [11]. In addition, we cut sequences if a new gene on either strand appears and restrict the length of the upstream sequence to 10k bp. As the gene has multiple transcription start sites (TSS), we use the upstream sequence of each TSS with a maximum length of 10kb if neither a gene nor another TSS occurs within this range.

3 Results

We compare our method with two naive approaches for the significance of an observed number of hits in a sequence. The first alternative has already been defined in the method section. Let $\hat{p}_1$ denote the probability for the observed number of hits modeled by a binomial distribution. The second approach $\hat{p}_2$ only counts the number of clusters neglecting the number of overlapping hits while the model is still based on the binomial distribution. Our approach is denoted by $\hat{p}_3$. Neither in the example nor usually in practice occur clusters with greater size than 2. Thus, we only compute the probabilities for those clusters (α_1 and α_2).

The conserved upstream region of MEF2C contains 3681 bp. We detect two clusters of size 1 and one cluster of size 2, thus, $\mathbf{z} = (2, 1, 0)$. Looking at the sequence logo [8] in Figure 1 shows that the matrix has a self similarity due to the palindromic structure at positions 3-4 and 11-12 of the PWM. Therefore, the probability to have two overlapping hits is higher than the probability of two single hits. The computation of the probabilities retrieves $\hat{p}_1 = 0.042$, $\hat{p}_2 = 0.0076$, and $\hat{p}_3 = 0.018$. Due to the self similarity, it is clear that $\hat{p}_2$ overestimates the probability. Furthermore, $\hat{p}_1$ ignores the overlapping hit. Therefore, the probability is underestimated as the probability for one single hit is higher than the probability for two overlapping hits. Thus, the application of $\hat{p}_1$ and $\hat{p}_2$ should be avoided. Hence, in this example the real probability should be between $\hat{p}_1$ and $\hat{p}_2$. This holds for $\hat{p}_3$ supporting our approach.

Next, we focus on a matrix without any self similarity. In this case, the probability for two overlapping hits is smaller than the probability for two single hits. Using the not self similar TF MYOD with PWM M00184 (see Figure 2), we search for hits in the sequence IRX4 (5698 bp conserved). There are two clusters of size 1 and two clusters of size 2. The computed probabilities are $\hat{p}_1 = 0.029$, $\hat{p}_2 = 0.0012$, and $\hat{p}_3 = 0.00074$. In fact, $\hat{p}_3$ is the smallest probability. Again, this shows that $\hat{p}_3$ is the more appropriate model.

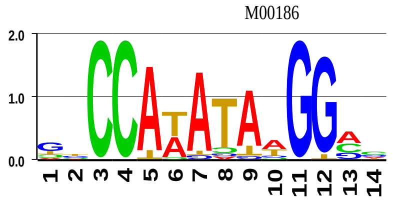

Figure 1: Sequence Logo for PWM M00186 retrieved from Transfac.

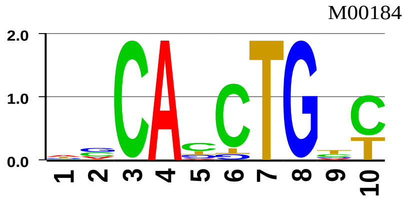

Figure 2: Sequence Logo for PWM M00184 retrieved from Transfac.

4 Discussion

We have presented a reasonable statistical model to improve the computation of a probability as a score for the number of occurrences on a sequence. The results show that an adequate statistical model is necessary for valid statements about statistical significance. Furthermore, the palindromic structure and self similarity have a strong impact on the dependence of overlapping hits. Thus, it is necessary to model these dependencies to get reasonable probabilities. The examples give support for our model such that it is able to capture the strongest dependencies in contrast to the naive approaches. Furthermore, we want to stress that all computations run in less than 1 minute on a standard PC with 1.6 GHz processor and 512 MB RAM.

Our approach can serve as a starting point for more realistic models. For example, TFs often have multiple PWMs which are not independent of each other. Extending this approach to deal with multiple PWMs seems to be possible though not trivial due to the combinatorics for different scoring distributions. Another extension could preserve the restriction of our model to consider dependencies only in an interval of twice the length of the PWM. In case of a long chain of overlapping hits, our approach will underestimate the probability of such an observation leading to a artificial low probability. Although, the impact of this problem on real analyses is small because one hardly observes clusters with size larger than 2. In addition, the occurrence of such large clusters could be used as an indicator for erroneously small probabilities.

We are aware of the fact that we abstained from giving biological evidence for the results. Since the decision whether a sequence can be bound by a TF or not is based on a threshold for the probability, the threshold influences the result substantially. Focusing on only two genes and TFs makes the choice of a threshold very randomized. Thus, a larger set including true positives and true negatives are necessary for a thorough analysis. This can be the goal of further research. In contrast, we aimed at showing the shortcomings of the naive approaches and the superiority of our model from a principle point of view.

References

[1] Benos, P. V., Bulyk, M. L., and Stormo, G. D., Additivity in protein-DNA interactions: How good an approximation is it? *Nucleic Acids Res.*, 30:4442–4451, 2002.

[2] Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey,

R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., and Birney, E., Ensembl 2005, *Nucleic Acids Res.*, 33:D447–D453, 2005.

[3] Kaynak, B., von Heydebreck, A., Mebus, S., Seelow, D., Hennig, S., Vogel, J., Sperling, H.-P., Pregla, R., Alexi-Meskishvili, V., Hetzer, R., Lange, P. E., Vingron, M., Lehrach, H., and Sperling, S., Genome-Wide array analysis of normal and malformed human hearts, *Circulation*, 107(19):2467–2474, 2003.

[4] Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D.-U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E., TRANSFAC(R): Transcriptional regulation, from patterns to profiles, *Nucleic Acids Res.*, 31(1):374–378, 2003.

[5] Rahmann, S., Müller, T., and Vingron, M., On the power of profiles for transcription factor binding site detection, *Stat. Appl. Genet. Mol. Biol.*, 2(7):Article 7, 2003.

[6] Reinert, G., Schbath, S. and Waterman, M. S., Probabilistic and statistical properties of words: An overview, *J. Comput. Biol.*, 7:1–46, 2000.

[7] Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W., and Lenhard, B., JASPAR: An open-access database for eukaryotic transcription factor binding profiles, *Nucleic Acids Res.*, 32:D91–D94, 2004.

[8] Schneider, T. D. and Stephens, R. M., Sequence logos: A new way to display consensus sequences, *Nucleic Acids Res.*, 18:6097–6100, 1990.

[9] Smith, T. F. and Waterman, M. S., Identification of common molecular subsequences, *J. Mol. Biol.*, 147:195–197, 1981.

[10] Stormo, G. D., DNA binding sites: Representation and discovery, *Bioinformatics*, 16:16–23, 2000.

[11] Wasserman, W. W. and Sandelin, A., Applied bioinformatics for the identification of regulatory elements, *Nature Reviews Genetics*, 5:276–287, 2004.