Available online at www.sciencedirect.com

ScienceDirect

Journal of Biomedical Informatics 40 (2007) 139–149

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

A new approach to the assessment of the quality of predictions of transcription factor binding sites

Szymon Nowakowski *, Jerzy Tiuryn

Institute of Informatics, Warsaw University, ul. Banacha 2, 02-097 Warszawa, Poland

Received 20 January 2006
Available online 21 July 2006

Abstract

In this paper, we describe a novel method called *Secondary Verification* which assesses the quality of predictions of transcription factor binding sites. This method incorporates a distribution of prediction scores over positive examples (i.e. the actual binding sites) and is shown to be superior to *p*-value, routinely used statistical significance assessment, which uses only a distribution of prediction scores over background sequences. We also discuss how to integrate both distributions into a framework called *Secondary Verification Assessment* method which evaluates the quality of a model of a transcription factor. Based on that we create a hybrid representation of a transcription factor: we select the description (with or without dependencies) which is best for the transcription factor considered.

Keywords: Modeling dependencies; Mixture of PSSMs; Statistical significance; Motif finding; Model assessment; Predicting binding sites

1. Introduction

Motif discovery in the case of DNA or protein sequences has a wide range of applications in modern molecular biology: from modeling mechanisms of transcriptional regulation [1,2] to local protein structure prediction [3]. In this work, we address the former problem.

Generalizing alignments as a method to discover motifs has been extensively studied during the last few years. It is well known that the simplest estimator of observing a nucleotide in an alignment, called *Maximum Likelihood* estimator [4], which only counts occurrences of nucleotides, is of no use for alignments consisting of too few sequences—it may happen that a nucleotide *is not observed* in the alignment at all, but it *could be observed*, if the number of aligned sequences were greater. To this end various methods introducing prior knowledge were proposed. These techniques range from the *zero-offset method* [5] which simply adds 1 (or another constant) to every nucleotide count, through methods based on substitution matrices [6] or feature alphabets [7], to the most advanced *Dirichlet mixture method* [8,9] for which a mixture of Dirichlet distributions is supplied as prior knowledge. The *pseudocount method* [10] is a special kind of the Dirichlet mixture method, where only one Dirichlet distribution is used. It is shown in [5] that, when the columns of an alignment are independent, the Dirichlet mixture method is close to the theoretical optimum.

Models of probability distributions that generalize alignments can be classified by their ability to model column dependencies into two classes. PSSM (Position Specific Score Matrix) is a model not taking into account the dependencies that may exist between columns of an alignment. On the other hand, modeling dependencies between columns in DNA sequence alignments has been recently studied [1,11–16]. In [12] dependencies are modeled only between adjacent columns with the use of Hidden Markov Models (HMMs).

Methods exist to model dependencies between nonadjacent columns in an alignment. Barash et al. [1] analyze a

[*] Corresponding author. Fax: +48 22 5544400.
E-mail address: s.nowakowski@mimuw.edu.pl (S. Nowakowski).

1532-0464/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2006.07.001

number of such techniques, one of them being the *mixture of PSSMs* method. They also use *tree models*, which are Bayesian networks with the restriction that the dependency graph is a tree. In their work the prior knowledge is introduced with the pseudocount method (modeled by a single Dirichlet distribution). King and Roth [13] introduce another model called NONPAR which is able to model arbitrary dependencies between positions. As two parameters are varied, it smoothly interpolates from a single PSSM model to the full-dependency empirical distribution of binding sites. Although HMMs can model dependencies only between adjacent columns in an alignment, it has been shown [14] that it is possible to rearrange the columns in the way which increases the number of adjacent dependencies and in fact allows to model nonadjacent dependencies. There are also approaches in which the frequency of nucleotide tuples (pairs, triples, etc.) is analyzed [15,16] in order to describe dependencies.

This paper addresses the question of statistical significance of a prediction obtained by an application of a probability model to a query sequence. We propose a novel method called Secondary Verification (SV) which assesses the quality of predictions of transcription factor binding sites. This method, in addition to the distribution of prediction scores in the background (negative) sequences, incorporates the distribution of prediction scores over independent positive examples (i.e., the actual binding sites). It is shown to be superior to *p*-value, other statistical significance assessment which uses only the distribution of prediction scores in the background sequences. The statistical significance of a prediction was previously analyzed in [17], but the authors did not use the independent example set. Their conclusions were derived under the assumption (which is not necessarily correct) that the motif model gives the true distribution of the motif sequences. The independent example set used in our method makes it possible to verify this assumption.

In this paper, we discuss also the question of reliability of including dependencies for alignment models. As it was shown in [18], the inclusion of dependencies can actually decrease the quality of a model for alignments comprising of too few sequences. A novel technique based on the SV score, called Secondary Verification Assessment (SVA), is proposed to evaluate the quality of models and consequently to decide whether inclusion of dependencies leads to improvement of a model. The SVA technique uses both distributions of prediction scores (i.e. over the actual binding sites and over the background sequences).

We use PSSMs as independent models and mixtures of PSSMs as models with dependencies. The rationale behind choosing a mixture of PSSMs to model dependencies was to allow dependencies between nonadjacent columns (as nonadjacent sites and the dependencies between them can play a major role in the protein–DNA binding process [14]). A mixture of PSSMs has relatively few parameters in the class of such models [1]. There is also no need to estimate the dependency structure. Finally, following the discussion in [1], we point out one more reason that a mixture of PSSMs is a very suitable representation of the transcription factor binding site motifs. It can model different types of transcription factor binding, each related to a different physical configuration of the protein. The result is that the transcription factor can bind to more than one class of sequences, each related to one type of its configuration. Each of these sequence classes can be described by one of the PSSMs in the mixture.

We show that the PSSM representation with carefully chosen prior knowledge can give a description of an alignment as good as the ones obtained with the use of dependencies and we compare our results to the results of Barash et al. [1] and King and Roth [13]. We also show how to choose the best model for an alignment. It is either the PSSM model or a mixture of PSSMs (for the cases in which modeling dependencies improves the alignment description). We call this representation *a hybrid model*, since it is based on more than one model and chooses the best model available for the alignment. We use the SVA value to select the best representation for each alignment in the construction of a hybrid model. The hybrid model is shown to be superior to any other model (with or without dependencies) considered in our experiment and in experiments described in Barash et al. [1] and King and Roth [13].

2. Materials and methods

2.1. Data

We use the same dataset which was used by Barash et al. [1] and King and Roth [13]. It consists of the binding sites of 95 transcription factors from TRANSFAC database [19] and can be downloaded as described in Barash et al. [1]. These binding sites form 95 alignments which we use in our analysis. The alignments contain from 20 to 88 sequences. Thirty-nine of the alignments are gapped. Eleven of these alignments are associated to human transcription factors and contain at least 40 sequences and we use this subset in two experiments: we assess techniques scoring prediction quality in the first experiment and we assess model quality in the second experiment. The third experiment, i.e. the construction and evaluation of a hybrid model, is conducted with the use of all 95 alignments.

2.2. PSSMs and mixtures of PSSMs

In what follows we assume that the alphabet of nucleotides consists of four letters $\mathcal{L} = \{A, C, T, G\}$. Let us assume that $\mathcal{A}$ is the alignment over $\mathcal{L}$ with N columns created from a number of transcription factor binding sites of length N. We treat $\mathcal{A}$ as training data and we want to construct a predictor scoring every sequence of length N. The score should indicate whether the sequence is the binding site. A common first step is constructing a probability distribution over all sequences of length N, with $\mathcal{A}$ as a sample from that distribution.

A widely used model of a probability distribution over all sequences of length N, which can be reliably estimated from a given alignment, is called a *Position Specific Score Matrix* (*PSSM*). It assumes that the columns of the alignment are independent. PSSM is represented by a matrix $P = (p_{li})$ of size $|\mathcal{L}| \times N$, where N is the number of columns in the alignment. For $l \in \mathcal{L}$ and $i = 1, \ldots, N$ the value of p_{li} is the probability of seeing the letter l in the ith column of the alignment, for which the PSSM was estimated. Then the probability of the sequence $S = s_1 s_2 s_3 \cdots s_N$ in the PSSM model is

$$Pr(S) = p_{s_1 1} \cdot p_{s_2 2} \cdot p_{s_3 3} \cdot \cdots \cdot p_{s_N N}.$$

The mixture of PSSMs is a distribution which introduces dependencies into alignment analysis. The alignment is described by K PSSMs $P^{(1)}, \ldots, P^{(K)}$, and by K positive weights of these PSSMs, $q_1, \ldots, q_K$, such that $\sum_{k=1}^{K} q_k = 1$. Let $P^{(k)} = (p_{li}^{(k)})$ for $k = 1, \ldots, K$. The probability of the query sequence $S = s_1 s_2 s_3 \cdots s_N$ in this model is then

$$Pr(S) = \sum_{k=1}^{K} q_k \cdot p_{s_1 1}^{(k)} \cdot p_{s_2 2}^{(k)} \cdot \cdots \cdot p_{s_N N}^{(k)}.$$

We use notation $Pr(S|M)$ when we want to make explicit dependence of the probability measure on a model M of an alignment. M can be a PSSM or a mixture of PSSMs or any other probability distribution.

2.3. Estimating a model and prior knowledge

In this paper, we estimate the models using two techniques: an *optimal PSSM method* and *MAP* (*Maximum a Posteriori*) *estimation for two PSSMs*. The first technique results in one PSSM, the second in a mixture of two PSSMs.

The mixture of Dirichlet distributions, which will be used as a prior for the models we consider, is a probability distribution over the space of all probability vectors with four nonnegative coordinates (representing four nucleotides) summing to 1. We do not provide the details of the probability density function of the mixture of Dirichlet distributions, referring the interested reader to [4,8,9]. Observe that methods exist to efficiently estimate the mixture of Dirichlet distributions from the data as well as to use it as a prior distribution for the model estimation. In the remainder of this section we focus on these methods.

Optimal PSSM estimation is described in detail in [8,9]. In short, PME (Posterior Mean Estimator) is used on separate columns of the alignments, resulting in independent probabilistic description of every column. In [5] it is argued that the PSSM obtained in this way is close to theoretical optimum. It requires a prior in the form of a mixture of Dirichlet distributions.

The MAP estimation [18] finds local maximum of likelihood function based on a PSSM mixture and thus finds a PSSM mixture well describing the query alignment. Since the maximization is local, an extensive search of the probability space of the prior distribution is first conducted to identify good starting points. This procedure also requires a prior in the form of a mixture of Dirichlet distributions.

Although the two methods described above require a mixture of Dirichlet distributions as prior knowledge, a pseudocount prior (i.e. a single Dirichlet distribution) may also be used for these methods.

To identify good prior distributions we follow [8,9]. They describe a procedure to estimate the mixture of Dirichlet distributions which provides the best possible prior description of the data.

2.4. Score of a prediction

2.4.1. Log-odds scoring and posterior scoring

For a model M of a binding site (of a given transcription factor), and a model B of a background distribution, we can score the chance that the query sequence S is the binding site rather than it comes from the background using the standard formula $\frac{Pr(S|M)}{Pr(S|B)}$. Logarithm of this value is the well known *log-odds score* [4]

$$\text{log-odds}(S) = \log \frac{Pr(S|M)}{Pr(S|B)}.$$

When we have a prior estimate $Pr(M)$ of the probability of observing an actual binding site of the transcription factor, before we have seen any information about the sequence itself, we can include that estimate and obtain the posterior score that the query sequence S is the binding site. The posterior score is then equal $\frac{Pr(S|M)Pr(M)}{Pr(S|B)Pr(B)}$, where $Pr(B) = 1 - Pr(M)$. Logarithm of this value is usually used in practice, since it is computationally more convenient

$$\text{posterior}(S) = \log \frac{Pr(S|M)Pr(M)}{Pr(S|B)Pr(B)}.$$

Since both the above scores (the log-odds score and the posterior score) are sums of many similar values, we make a simplifying assumption that they are normally distributed [20]. The normality assumption is theoretically sound (it is based on the Central Limit Theorem) and makes our model both simple and efficient. One may extend the methods presented in this paper by estimating the exact distributions of scores as it is described in [17,21,22]. This should lead to the further improvement of the results.

Both the log-odds score and the posterior score will be called *simple scores*.

2.5. Statistical significance of a score

In the discussion below we assume that the posterior or log-odds scoring was used for a given sequence S and we have obtained a score s of that prediction.

2.5.1. Normal distribution

Since we make an assumption of the normal distribution of the score statistics, we define here the probability density function, *pdf*, and the cumulative distribution function, *cdf*, of the normal distribution $N(\mu, \sigma^2)$ with the expected value μ and standard deviation σ.

The pdf of the normal distribution $N(\mu, \sigma^2)$ is given by the following equation:

$$f_{N(\mu,\sigma^2)}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

The cdf of the normal distribution $N(\mu, \sigma^2)$ at the point $x \in \mathbb{R}$ is equal to the probability of obtaining a value of at most x by sampling from $N(\mu, \sigma^2)$. It is given by the equation

$$F_{N(\mu,\sigma^2)}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(u-\mu)^2}{2\sigma^2}} du.$$

2.5.2. P-value

When we have s, the score of the prediction, we still do not know how significant it is. To this end p-value is used, which measures the probability that the score this high or higher was obtained by chance, i.e. from the background. Many approaches were proposed to compute p-value of a score. One of them uses the fact, that the distribution of scores is normal. We can estimate mean μ and variance σ^2 of that distribution in a population of random background sequences. Then the p-value is calculated as a right tail of the cdf of estimated distribution:

$$p\text{-value}(s) = 1 - F_{N(\mu,\sigma^2)}(s).$$

In fact, since we want the scoring method to give higher values for better predictions, we use $\mathrm{PV}(s) = 1 - p\text{-value}(s)$ under the name p-value scoring. Observe that the scores is transformed into $\mathrm{PV}(s)$, which can be further used as a new score of that prediction.

2.5.3. Secondary Verification

As it was explained, p-value of the score measures its statistical significance taking into consideration only the background distribution of scores (which we refer to as the *negative distribution*). But there is also a *positive distribution* of scores, i.e. the distribution of scores among the sequences actually being the binding sites of the transcription factor. We propose a method which includes *both* distributions in assessing the statistical significance of the score.

We assume that the query sequence is either a positive sequence, i.e. a binding site, or a negative one, coming from the background. The score s of the prediction carries some information on the probability of each case, but we do not know its statistical significance. It is possible, for instance, that the positive distribution and the negative distribution of scores are very similar, which means, that it is not possible to tell the truth based only on the value of the score of the prediction (and the p-value of the score s is not relevant in this case).

To further discuss the possibility that the p-value is not reliable enough, we consider the situation when the p-value of the score indicates that the hit is very significant, i.e. the background probability of such a score is very low. It is possible, however, that the probability of the binding site with such a score is very low, too. In such a case we have either a rare background sequence or a rare motif sequence. We must take into account both the positive and the negative distributions to be able to choose the better variant. All these aspects make us believe, that it would be very beneficial to consider both distributions of scores.

We use two priors: the prior that a score comes from the negative distribution, denoted Pr^-, and the prior that a score comes from the positive distribution, denoted Pr^+. We have $Pr^- + Pr^+ = 1$. Let us suppose that the score of the prediction is s. We are interested in the value of $Pr(+|s)$, i.e. the probability that the distribution is positive, given we observe the score s. By the use of Bayes Theorem the value of $Pr(+|s)$ can be expressed as

$$Pr(+|s) = \frac{Pr(s|+)Pr^+}{Pr(s|+)Pr^+ + Pr(s|-)Pr^-}.$$

Note, that the value $Pr(+|s)$ has one desirable feature different from the p-value score $PV(s)$: it can decrease for higher scores s if the $Pr(s|-)$ increases and $Pr(s|+)$ decreases with s, which is possible in practice (usually for very poor models).

Both distributions $Pr(\cdot|+)$ (the positive distribution of scores) and $Pr(\cdot|-)$ (the negative distribution) can be estimated from the data. To estimate $Pr(\cdot|-)$ we randomly generate a large number of sequences from the background and estimate the mean and variance of a normal distribution of their scores, as we would do when computing p-value.

The case of $Pr(\cdot|+)$ is a little bit more complex. We need a set I of binding sites independent from the alignment $\mathcal{A}$ (the alignment $\mathcal{A}$ is the set of binding sites on which the basic scoring model M was trained). Then we score every binding site from I and we estimate the mean and variance of a normal distribution of their scores. We must keep in mind, however, that our estimate is less reliable than in the case of negative distribution, as we usually have only a few examples available in I for the estimation.

As before, we treat $s \to Pr(+|s)$ as the transformation of a simple score s into a new score, which we call the *Secondary Verification score* (or the *SV score* in short):

$$\mathrm{SV}(s) = Pr(+|s).$$

Observe, that log-odds(S) and posterior(S) are calculated from the sequence S, from the model M of the transcription factor and from the background model B. The values $\mathrm{PV}(s)$ and $\mathrm{SV}(s)$ are calculated from $s = \mathrm{score}(S)$ and from two families of scores $s_i^+ = \mathrm{score}(S_i^+)$ for $i = 1, \ldots, n^+$ and $s_i^- = \mathrm{score}(S_i^-)$ for $i = 1, \ldots, n^-$. The sequences S_i^+ and S_i^- are positive and negative (background) sequences, respectively, and score($\cdot$) represents the scoring system (i.e. either posterior($\cdot$) or log-odds($\cdot$)). In the case of p-value scoring we have $n^+ = 0$.

Both the *p*-value score and the SV score will be called *significance assessment scores*.

2.6. Assessment of the quality of a model

Before actually applying a motif model and assessing statistical significance of prediction scores one may want to assess the quality of an estimated model. We use *average log-probability* method and we introduce a novel Secondary Verification Assessment method. Below we provide a description of both techniques.

2.6.1. Average log-probability and cross-validation test

To assess the quality of a method modeling the transcription factor by a probability distribution we use a *10-fold cross-validation test*. The dataset for the test consists of the aligned binding sites of the transcription factors. The test itself is performed as follows: sequences from every alignment are randomly divided into 10 subsets of equal size. Each of the subsets is treated as a test set T in one of 10 runs of the cross-validation test. At the same time the remaining 9 subsets are treated as a training set—a model (a PSSM or a mixture of PSSMs) is estimated from the union of these 9 subsets.

Logarithm of a score (the score being a probability of a sequence in a probability model) is computed for every sequence from T and the model of the alignment from which that sequence was removed. This value is called *log-probability* of the sequence. For each alignment we then compute its mean log-probability value for a given model, averaged over all sequences in all runs of cross-validation procedure. Let us fix an alignment $\mathcal{A}$. To address the statistical significance of the difference in mean log-probability values for $\mathcal{A}$ between two models, the paired *t*-test is performed. Following Barash et al. [1], we call one model *better* than the other on $\mathcal{A}$ when the difference in mean log-probability values for $\mathcal{A}$ is positive, and *significantly better* when the associated paired *t*-test *p*-value falls below 0.05 threshold. It is called *worse* or *significantly worse* when the other method is better or significantly better, respectively. We can compare two models judging them by the number of alignments, for which one of them is better, significantly better or significantly worse than the other.

2.6.2. Secondary Verification Assessment

The average log-probability assessment takes into account only the positive distribution of scores. And it completely ignores the negative distribution. But it was already pointed out that in cases when the positive distribution and the negative distribution of scores are similar it is not possible to tell whether the query sequence is the binding site or not. In other words, for similar positive and negative distributions the Secondary Verification score $\mathrm{SV}(s)$ is low for every sequence and its score s. The desirable situation would be, however, that $\mathrm{SV}(s)$ is high for the positive sequences. This is why we are interested in the distribution of $\mathrm{SV}(s)$ for the positive scores. If the SV score is low for a majority of the positive examples, it is the indication that we are unable to distinguish between the background B and the model M for a majority of positive scores, i.e. for a majority of sequences really being the binding sites we are unable to tell if they are the binding sites, no matter what the average log-probability value tells us.

To this end we propose a simple statistic called the Secondary Verification Assessment value (or the SVA value in short) which scores the model for a transcription factor on the additional dataset I which we treat as a set of positive sequences. Contrary to the case of the average log-probability method, we require that I is reasonably large, as we need to estimate the distribution of scores in I and we need a sample large enough.

The positive and negative distributions are first estimated from the set I and from the sequences sampled from the background model, respectively. Then the expected value of $\mathrm{SV}(\cdot)$ with respect to the positive distribution is calculated:

$$\mathrm{SVA}(B,M,I) = E^{+}(\mathrm{SV}) = \int_{-\infty}^{\infty} \mathrm{SV}(s) Pr(s|+)\,\mathrm{d}s = \int_{-\infty}^{\infty} \frac{Pr^{2}(s|+)Pr^{+}\mathrm{d}s}{Pr(s|+)Pr^{+} + Pr(s|-)Pr^{-}}.$$

The value $\mathrm{SVA}(B,M,I)$ tells us what quality of scores we can expect on average for the sequences really being the binding sites. Consequently, it assesses the model M as a description of the transcription factor. Moreover, we have $\mathrm{SVA}(B,M,I) \in (0,1)$ so we can tell if the model M is reliable without a need to compare it to another model. The higher SVA value, the better description of a motif is provided by a model.

It should be stressed that the SVA value is conceptually different and much better suited for model evaluation than methods related to the distances between probability distributions. Widely used *information content* [23], which measures decrease in uncertainty between a model and a background distribution in terms of information theory, is an example of such a method. Unfortunately high information content does not imply that a model is useful. It only proves that a model is different from the background but not necessarily describes the motif we want to describe. The same holds for *relative entropy*, called also Kullback–Leibler divergence [24]. As an example: Maximum Likelihood (ML) estimator tends to have higher information content and relative entropy than the Bayesian estimators incorporating prior knowledge. There is a consensus among the researchers that ML is less suited for modeling motifs than Bayesian estimators. We performed an experiment in which for all 95 alignments two models were computed: the ML model and a Bayesian model with 2.0 pseudocounts evenly distributed between nucleotides. We calculated information content [23] and relative entropy [24] for both models. Let a PSSM model be represented by a matrix $P = (p_{li})$ of size $|\mathcal{L}| \times N$, where N is the number of columns in the alignment and $\mathcal{L}$ is the nucleotide alphabet. Let the background distribution be given by $B = (q_1, \ldots, q_{|\mathcal{L}|})$. The information content of a model versus background is defined as

$$\mathrm{IC}(P,B) = \sum_{i=1}^{N} \sum_{l\in\mathcal{L}} (p_{li}\log_2 p_{li} - q_l \log_2 q_l)$$

and relative entropy is defined as

$$D(P\|B) = \sum_{i=1}^{N} \sum_{l\in\mathcal{L}} p_{li}\log_2 \frac{p_{li}}{q_l}$$

As one can easily see, for the uniform background distribution both methods are equivalent and render the same score. Thus, for simplicity, in our experiment we used the uniform background distribution. In Table 1 we present the results of this experiment: in all 95 cases the better model (i.e. the Bayesian model) has lower information content score and lower relative entropy. It is a side effect of the way Bayesian methods improve the model: they smoothen the distribution, which makes it more similar to the background. Summing up, both distance-based methods are very useful to highlight the columns which play the major role in the model (and are used for such a purpose with great success [25]), but are not very well suited to choose the best model.

In [17] a method to assess the statistical power of a probabilistic model is presented. The main difference between the SVA method proposed in the present paper and the method proposed in [17], as well as the above mentioned methods of information content/relative entropy, is that the former assesses the quality of the model with the use of independent example set, while the latter methods do not. Their conclusions are derived under the assumption that motif sequences are distributed according to the model. This assumption may be incorrect. The SVA method addresses this problem verifying this assumption on an independent example set.

3. Experiments

The data, binding sites of 95 transcription factors taken from TRANSFAC database [19], were randomly divided into 10 sets $S_1, \ldots, S_{10}$ as a preparation for a cross-validation experiment. For every choice of 9 sets out of these 10 possibilities the mixture of Dirichlet distributions was estimated as described in Section 2.3 (with the best prior description possible of these 9 sets). Let us denote the mixture of Dirichlet distributions D_i if it was obtained with the exclusion of S_i, $i = 1, \ldots, 10$. Additionally, the mixture of Dirichlet distributions $D_{1,2,3}$ was estimated from $\sum_{i=4}^{10} S_i$. It was used for Secondary Verification calculation.

3.1. Predicting binding sites

For this experiment 11 human transcription factors were chosen. The requirement was that at least 40 binding sites are aligned in the chosen alignments. This was necessary in order to perform reliable Secondary Verification calcula-

Table 1
The relative entropy $D(P\|B)$ scores for ML model and Bayesian model obtained for all 95 alignments

Transcription factor	$D(P\|B)$	
	ML	Bayes
F$ABAA_01	14.23	11.16
F$GCN4_01	16.85	14.62
F$MCM1_01	14.21	11.31
I$CF1_02	13.59	11.76
I$CF2II_01	10.72	10.01
I$DL_01	14.77	11.65
I$DRI_01	9.20	8.03
I$OVO_01	10.40	8.14
I$SN_01	13.10	11.36
I$UBX_01	9.31	8.67
P$ABF1_01	12.20	9.41
P$ABF_Q2	14.70	13.21
P$ANT_01	14.09	11.91
P$ATHB1_01	15.48	12.35
P$ATHB5_01	12.73	10.59
P$EMBP1_Q2	12.05	9.71
P$GAMYB_01	9.40	7.54
P$LIM1_01	6.29	5.04
P$P_01	9.79	8.41
V$AHRARNT_01	4.59	3.78
V$AHRARNT_02	20.42	16.43
V$AML1_01	9.16	8.29
V$ARNT_01	5.36	4.14
V$AR_01	14.49	12.15
V$ATF6_01	15.06	11.57
V$ATF_01	12.61	10.22
V$BRACH_01	29.64	25.82
V$CART1_01	13.63	11.15
V$CDPCR1_01	9.51	8.17
V$CDPCR3HD_01	10.38	8.78
V$CDPCR3_01	16.52	13.32
V$CEBPA_01	7.08	6.29
V$CEBPB_01	10.70	8.35
V$CEBP_01	6.99	5.53
V$CETS1P54_02	8.72	7.55
V$CHX10_01	10.20	7.97
V$CIZ_01	11.53	9.88
V$E4BP4_01	15.40	12.16
V$ELK1_02	10.16	8.57
V$ERR1_Q2	11.78	9.62
V$EVI1_03	19.26	15.74
V$FAC1_01	8.00	6.64
V$FOXD3_01	11.77	9.95
V$FOXJ2_01	13.90	12.04
V$FOXO1_02	14.34	11.31
V$FOXO4_01	10.77	8.95
V$GATA6_01	6.97	6.36
V$GKLF_01	8.77	7.93
V$GNCF_01	21.51	18.65
V$HAND1E47_01	11.25	9.45
V$HNF1_01	14.54	11.96
V$IRF1_01	13.20	10.35
V$IRF7_01	7.34	6.26
V$LHX3_01	13.83	12.08
V$LUN1_01	27.01	21.58
V$MZF1_01	9.18	7.18
V$NCX_01	4.73	4.25
V$NKX22_01	11.15	8.89
V$NRSF_01	30.18	24.91
V$NFKAPPAB_01	13.36	11.63
V$OCT1_01	17.57	15.82

(continued on next page)

Table 1 (*continued*)

Transcription factor	$D(P\|B)$	
	ML	Bayes
V$OCT1_02	11.83	10.50
V$OCT1_03	6.89	6.20
V$OCT1_04	10.22	9.12
V$PAX2_01	5.91	5.11
V$PAX2_02	6.40	5.52
V$PAX6_01	15.15	13.64
V$PAX8_01	4.85	4.32
V$PAX8_B	6.14	5.15
V$PBX1_02	11.42	10.11
V$PPARG_01	16.39	15.31
V$PPARG_02	19.94	16.59
V$RORA1_01	13.75	11.09
V$RORA2_01	17.48	14.82
V$RSRFC4_01	18.54	15.94
V$RSRFC4_Q2	14.23	11.53
V$R_01	17.96	15.48
V$S8_01	5.89	4.06
V$SOX5_01	11.10	8.88
V$SOX9_B1	11.05	10.21
V$SPZ1_01	5.84	5.02
V$SRY_01	6.86	5.46
V$SRY_02	9.99	8.35
V$STAT5A_01	11.20	9.57
V$STAT5A_02	15.44	13.51
V$STAT5B_01	12.53	11.14
V$TBP_01	9.80	8.71
V$VJUN_01	20.51	16.36
V$VMYB_01	9.35	7.63
V$XBP1_01	13.64	11.62
V$ZIC1_01	5.68	4.94
V$VMYB_02	10.99	9.03
V$ZIC2_01	5.52	4.84
V$ZID_01	14.41	12.37

In all cases the ML model has higher score, although it is well known that Bayesian models have higher quality. Observe that the same results hold for information content: its value is equal to the value of relative entropy because we used the uniform background distribution.

tions as the positive distribution was estimated from 20% of the data and we wanted to have at least 8 scores to estimate their distribution.

For every transcription factor 10% of the binding sites (the ones corresponding to the transcription factor and contained in S_1) were treated as test sequences.

In the case of log-odds, posterior and p-value scoring, the remaining 90% was used as training sequences for the estimation of a PSSM model with D_1 used as a prior. Observe, that all sequences from the test set were excluded from the estimation of D_1.

In the case of Secondary Verification calculations, 70% was used as basic training sequences for the estimation of a PSSM model, and another 20% of the data (sequences from S_2 and S_3) was in an additional training set and was used to estimate positive distribution of scores. In the Secondary Verification test the posterior scoring was used as basic scoring. The PSSM was estimated with $D_{1,2,3}$ used as a prior. Observe, that all sequences from both the test set and the additional training set were excluded from the estimation of $D_{1,2,3}$.

We trained the SV method only on a part of all the training data available (leaving the rest as an additional training set for estimation of a positive distribution of scores). It must be stressed, however, that the rest of the methods to which the SV score was compared, were trained on all the available training data.

A random sequence of 2,500,000 nucleotides was generated from a third order Markov chain trained on human promoter sequences from PromoSer database [26]. The rationale behind choosing a random sequence instead of a real human promoter sequence is that in the case of real promoter sequences one can never be sure whether all binding sites of the transcription factor were identified. It would lead to misclassification of positive examples as negative ones and consequently would severely affect the result of the test.

The test sequences from all 11 transcription factors were planted into the artificial sequence at random positions. Both the p-value score and the SV score were computed based on the posterior score calculated with the models estimated from 90% and 70% of the data, respectively.

We assumed that prior values (for a chosen transcription factor) used in posterior scoring and Secondary Verification scoring, i.e. $Pr(M)$ and Pr^{+} (see Sections 2.4.1 and 2.5.3), are equal $\frac{n}{2{,}500{,}000}$, where n is the number of planted test sequences for the transcription factor considered. Value of n ranged from 4 to 7 and is presented in Table 2. The rest of the parameters were estimated from the data as explained in appropriate subsections of Section 2.

We used receiver operator characteristic (ROC) curves to compare tested scoring methods. A ROC curve shows the ability of a scoring method to separate true positives (correctly identified binding sites) from false positives (background sequences incorrectly identified as binding sites). The ROC curves were obtained with the following procedure:

Table 2
Eleven human transcription factors used in "Predicting binding sites" and "assessment of a predictive model" experiments

Transcription factor	Binding sites	Planted seqs	SVA value	Classification
V$AML1_01	56	5	0.03	Worst
V$CEBPA_01	43	4	0.001	Worst
V$FOXJ2_01	41	4	0.03	Worst
V$NFKAPPAB_01	40	4	0.10	Best
V$OCT1_01	56	5	0.63	Best
V$OCT1_02	44	4	0.33	Best
V$OCT1_03	51	5	0.02	Worst
V$OCT1_04	47	4	0.005	Worst
V$PAX6_01	47	4	0.62	Best
V$PBX1_02	40	4	0.37	Best
V$SOX9_B1	73	7	0.08	Unclassified

The second column presents the number of binding sites in the database of Barash et al. [1]. Next column presents the number of binding sites planted into the artificial sequence (10% of the binding sites available). The SVA value of the transcription factor is presented in one but last column. The last column indicates if the transcription factor was considered one of the *best* or the *worst* five, based on the SVA value. V$SOX9_B1 was left unclassified due to the requirement of equal sizes of both classes.

1. Let us fix a prediction method. For all 11 transcription factors, all contents of a sliding window (of width equal to the length of a motif) through a test sequence were considered as input sequences for the prediction method.
2. For a given threshold t (the sequences with a score higher than t were considered as *binding site predictions*) we plotted a point with coordinates $(\frac{FP}{AB}, \frac{TP}{AP})$, were AB is the number of all background sequences considered in all possible windows (in our experiment AB was approximately equal to $11 \cdot 2{,}500{,}000$), AP is the number of all planted positive sequences, TP is the number of correct predictions (true positives) and FP is the number of incorrect predictions (false positives).
3. The ROC curve was obtained by changing t from $-\infty$ to $+\infty$.

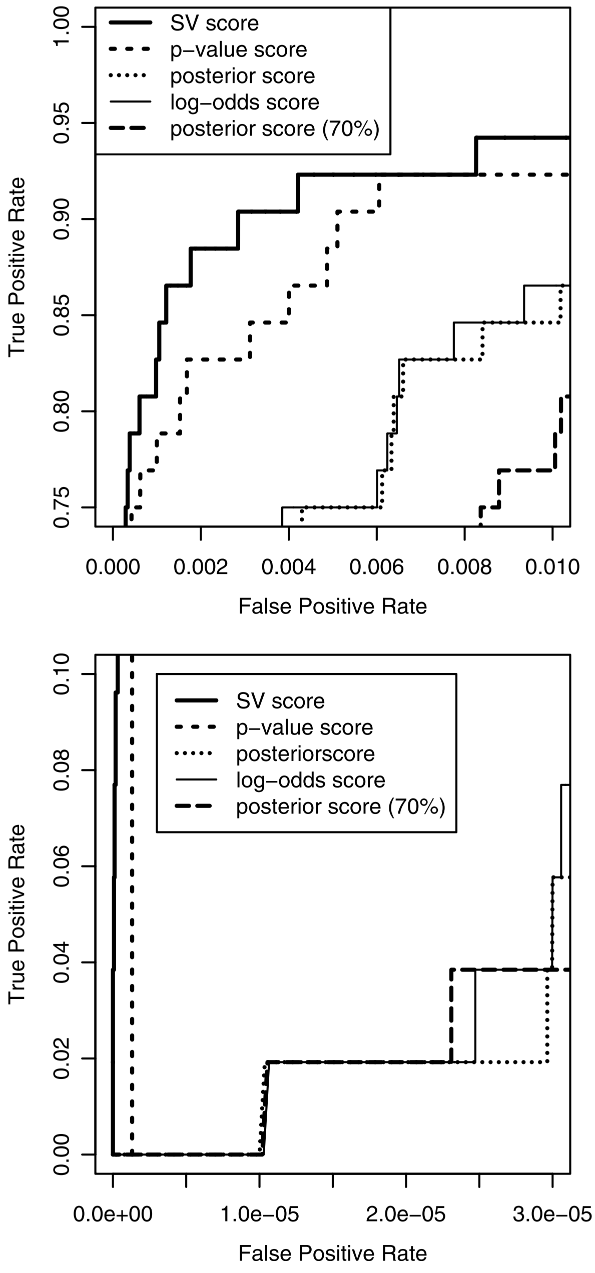

Fig. 1. (Upper) The comparison of ROC curves of 5 scoring methods showing their ability to identify binding sites of 11 human transcription factors. (Lower) ROC curves zoomed close to point (0,0) show the ability of the SV score to identify positive binding sites with false positive rate equal 0.

Fig. 1 presents ROC curves for log-odds scoring, posterior scoring, p-value scoring and Secondary Verification scoring. Additionally, one more ROC curve is presented: it is associated with the posterior scoring with the model estimated from 70% of the data, instead of 90% as is the case with the regular posterior scoring. We can see that it is by far worse than any other. On the other hand, after calculating SV transformation of this poor score we can see that the SV score becomes superior to any other score. The log-odds and posterior scores are very comparable in this experiment. Both the significance assessment scores are much better in separating positive and negative examples than both simple scores. As it was said, the SV score is better than the p-value score. It is caused by the fact that 20% of the data is not used to estimate a model, but instead significance assessment is performed with these data. In fact, as it is shown in Fig. 1 (lower picture), the SV score is able to identify only the real binding sites: there exists a threshold for which only the true positive rate is positive (the real binding sites are identified) and the false positive rate equals 0. No other score can achieve that on these data. One may find it surprising that although there might be as few as 6 positions significantly contributing to the score of some motifs (so binding sites of these motifs should occur by chance many times in a random sequence of 2,500,000 bases), we are still able to identify true binding sites without any incorrect matches in the background. The two highest SV scores found were 0.94 and 0.82, belonging to the transcription factors V$PAX6_01 and V$OCT1_01, respectively. These motifs are both long (21 and 19 nucleotides, respectively). One cannot expect to see them in the background by chance. This very fact is reflected by the high SV score of these findings and it is the reason that they are true positives.

We stress that (contrary to many approaches currently in use) it is worth to estimate a model from only a part of data and use the rest of it as a positive sample, for instance in SV scoring.

3.2. Assessment of a predictive model

To illustrate usefulness of Secondary Verification Assessment we performed the following experiment, based on the data described in Section 3.1.

The SVA method was used on 11 human transcription factors and the PSSM models estimated from the training data with $D_{1,2,3}$ used as a prior. The PSSM models were trained on 70% of the data, as it was explained in Section 3.1. Based on the SVA value of the posterior scores we identified 5 best and 5 worst transcription factors. Since we wanted to compare the performance of both groups,

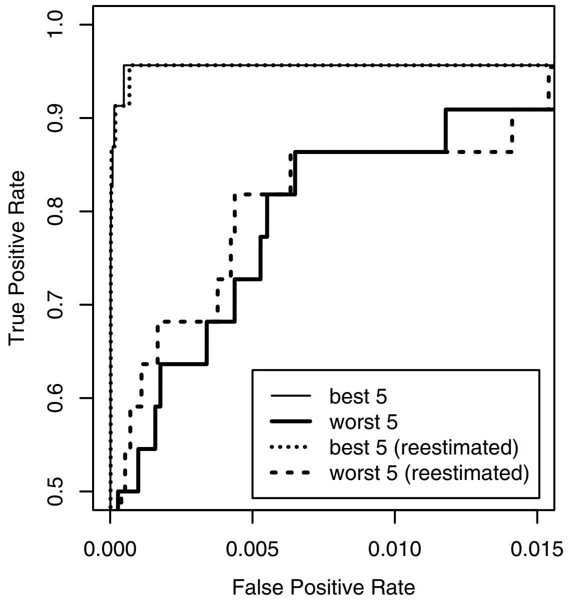

Fig. 2. The ROC curves for p-value scoring obtained for 5 best and 5 worst transcription factors, as decided by the SVA value. Indeed there is a huge difference in the ability to identify test binding sites between these two groups. The difference does not change after the model is reestimated on all data

we required that both groups have equal sizes (5 transcription factors), which left V$SOX9_B1 unclassified. The ROC curves were prepared for both groups with the p-value of a posterior score used for scoring. They are presented in Fig. 2. One can easily see in Fig. 2 that the quality of predictions for the transcription factors considered as bad is much worse than for the ones considered as good. We can conclude, that the SVA value of the transcription factor model really tells much about the chances to predict the real binding site using this model.

Moreover, we prepared ROC curves for the p-value of a posterior score that was obtained from a PSSM model trained on 90% of the data with D_1 used as a prior (see Section 3.1 for details). They are also presented in Fig. 2. Observe that alignments on which we trained the models were changed. But again we note that there is a great difference in prediction quality between the transcription factors we qualified as best and the worst ones. It suggests that the method can be used even in the absence of the additional training set and it is informative to apply the following approach: (1) divide the training data into two sets (basic and additional training sets), (2) perform the SVA analysis on both sets and finally (3) train the model on all training data if the SVA value is above chosen threshold.

3.3. Hybrid representation of a transcription factor

The final experiment was performed in order to compare our results with the results obtained by Barash et al. [1] and King and Roth [13]. To this end we used all 95 transcription factors that they used. We analyzed the following models, on which the cross-validation test was performed:

- the optimal PSSM representations with the mixture of Dirichlet distributions D_i used as the prior in the ith run of cross-validation experiment.
- the optimal PSSM representations with 5 pseudocounts distributed among the data used as the prior.
- the optimal PSSM representations with 1.6 pseudocounts distributed among the data used as the prior.
- the MAP estimation for two PSSMs with the mixture of Dirichlet distributions D_i used as the prior in the ith run of cross-validation experiment.
- the MAP estimation for two PSSMs with 5 pseudocounts distributed among the data used as the prior.
- the hybrid model which was designed to choose the best model for every alignment and was constructed as explained below.

Before we describe the construction of our hybrid model, recall that on the basis of the SVA value we can choose the best model for a transcription factor. Since some of the alignments comprise of as few as 20 sequences we decided not to create additional training sets. We trained and performed the SVA analysis on the same sets of examples of the binding sites that the models were trained. As a result of not using the additional independent training set, the two PSSMs representation had higher SVA value for all cases, since it has more parameters and can be more overfitted. Obviously, the higher SVA value did not indicate the superiority of a model in this case. We wanted to counter this effect. To this end we introduced a *correction factor* c, which was set to the square root of the number of PSSMs in a model. To construct a hybrid representation for every transcription factor in the ith run of cross-validation experiment we chose one of the following models, the one with the highest value of $\frac{1}{c} \cdot$ SVA:

- the optimal PSSM representation with the mixture of Dirichlet distributions D_i used as the prior,
- the MAP estimation for two PSSM with 5 pseudocounts distributed among the data used as the prior.

Following Barash et al. [1] and King and Roth [13] we compare different methods *indirectly*, assessing them relative to a chosen *reference method*. In their work Barash et al. [1] used 5-pseudocount PSSM as a reference method, while King and Roth [13] used 5-pseudocount PSSM and 1.6-pseudocount PSSM. Tables 3 and 4 present our results of that comparison together with the results of these two other groups. The reference method row is presented in bold face.

We can see in Table 3 that one PSSM with carefully tuned prior knowledge is almost as effective in that comparison as any model with column dependencies. It is caused by a small amount of data on which the estimation was performed, which makes estimating column dependencies less reliable, as it is shown in [18]. Moreover, when we estimate column dependencies we get worse results for very complex priors than we get when we use simple priors. For

Table 3
Comparison of different estimation methods with the reference method (5-pseudocount PSSM presented in the first row in bold face)

Estimation method	Parameters	PSSMs	Better	Sig. better	Sig. worse
Optimal PSSM	**5 pseudocounts**	**1**	**0**	**0**	**0**
Hybrid	—	1 or 2	86	62	0
Optimal PSSM	1.6 pseudocounts	1	71	45	8
Optimal PSSM	Dirichlet mixture	1	83	53	1
MAP estimation	5 pseudocounts	2	54	32	22
MAP estimation	Dirichlet mixture	2	29	7	34
2 PSSMs (Barash et al. [1])	5 pseudocounts	2	59	36	No data
Tree (Barash et al. [1])	5 pseudocounts	—	33	22	No data
2 trees (Barash et al. [1])	5 pseudocounts	—	57	35	No data
NONPAR (King and Roth [13])	$b = 1.7$, $\beta = .54$	—	84	59	3
Optimal PSSM (King and Roth [13])	1.6 pseudocounts	1	75	43	5

The third column presents the number of PSSMs used to represent estimated distributions. Last 3 columns present number of alignments with higher mean log-probability value than that of the reference method (i.e. better results), significantly better results and significantly worse results (see the definition of these terms in Section 2.6.1).

Table 4
Comparison of different estimation methods with the reference method (1.6-pseudocount PSSM presented in the first row in bold face)

Estimation method	Parameters	PSSMs	Better	Sig. better	Sig. worse
Optimal PSSM	**1.6 pseudocounts**	**1**	**0**	**0**	**0**
Hybrid	—	1 or 2	73	26	4
Optimal PSSM	Dirichlet mixture	1	69	19	9
MAP estimation	5 pseudocounts	2	31	20	30
NONPAR (King and Roth [13])	$b = 1.7, \beta = .54$	—	65	41	7

The third column presents the number of PSSMs used to represent estimated distributions. Last 3 columns present number of alignments with higher mean log-probability value than that of the reference method (i.e. better results), significantly better results and significantly worse results (see the definition of these terms in Section 2.6.1).

the mixture of many Dirichlet distributions used as a prior the results are much worse than for a simple 5-pseudocount prior in the case of MAP estimation for two PSSMs, which is a model with dependencies. It is caused by the fact that more complex priors require more parameters and there is too little data available to estimate these parameters reliably. The MAP estimation for two PSSMs with a simpler prior (5 pseudocounts) is as effective as a 2-PSSM mixture obtained with the same prior in the experiments of Barash et al. [1].

In Table 3 there are a few methods which behave very similarly: 1.6-pseudocount PSSM, the Dirichlet mixture PSSM, NONPAR and our hybrid model. To see that there is really a difference between 1.6-pseudocount PSSM and other models we performed a comparison of best methods from Table 3 and 1.6-pseudocount PSSM used as a reference method. The results are shown in Table 4.

We can see in Tables 3 and 4 that although the MAP estimation for two PSSMs method behaves poorly in that comparison, the hybrid model (based also on the MAP estimation model) outperforms all other. It indicates that before dependencies are included, it is worth to analyze whether they improve the model. The SVA value can be used for such an analysis.

4. Conclusion

In this work, we presented a way of using the distribution of scores in the populations of positive and negative sequences for the construction of the prediction scoring system much more efficient than the systems currently in use. We also showed how to incorporate this scoring system into the framework that can rank the models for a given transcription factor and consequently choose the best model for a transcription factor.

These concepts unfortunately require that a researcher has access to additional data, not used in model estimation. As the number of known binding sites increases constantly, they are going to be applicable to the increasing number of transcription factors. Nevertheless we proposed two methods which can help to overcome the problem of lacking data. The first one permits the use of training data as positive examples but requires introducing a correction factor which compensates overfitting and the fact, that different models can be differently overfitted. The other method requires the division of the training data into two sets: (1) training data for a temporary model and (2) positive examples for significance assessment. Then the final model can be re-estimated from all training data after significance assessment is finished.

We also argued that the simplest solution, that is to divide the training data into two sets and to estimate the model only from the first training set while the other set is used as positive examples, is very efficient. In fact we showed that SV scoring of a model trained on partial data is more efficient than the *p*-value scoring of the model trained on all training data available. Even though it was shown on a dataset of only 11 transcription factors, this is promising for the future as motif datasets are expected to grow constantly.

The code is fully accessible upon request, please email the corresponding author.

Acknowledgments

This work was supported by scientific Grants 3 T11F 006 27 and 3 T11F 021 28 of Polish Ministry of Education and Science.

References

[1] Barash Y, Elidan G, Friedman N, Kaplan T. Modeling dependencies in protein–DNA binding sites. In: RECOMB'03; 2003. p. 28–37.

[2] Liu X, Brutlag DL, Liu JS. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput 2001:127–38.

[3] Aloy P, Stark A, Hadley C, Russell RB. Predictions without templates: new folds, secondary structure, and contacts in CASP5. Proteins 2003;53(Suppl. 6):436–56.

[4] Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis. Cambridge University Press; 1998.

[5] Karplus K. Evaluating regularizers for estimating distributions of amino acids. Proc Int Conf Intell Syst Mol Biol 1995;3:188–96.

[6] Altschul SF. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol 1991;219:555–65.

[7] Smith RF, Smith TF. Automatic generation of primary sequence patterns from sets of related protein sequences. Proc Natl Acad Sci USA 1990;87(1):118–22.

[8] Brown MP, Hughey R, Krogh A, Mian IS, Sjölander K, Haussler D. Using Dirichlet mixture priors to derive hidden Markov models for protein families. In: Hunter L, Searls D, Shavlik J, editors. ISMB-93. Menlo Park, CA: AAAI/MIT Press; 1993. p. 47–55.

[9] Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, et al. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comp Appl Biosci 1996;12:327–45.

[10] Tatusov RL, Altschul SF, Koonin EV. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. Proc Natl Acad Sci USA 1994;91:12091–5.

[11] Agarwal P, Bafna V. Detecting non-adjoining correlations with signals in DNA. In: RECOMB'98; 1998. p. 2–8.

[12] Bulyk ML, Johnson PLF, Church GM. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res 2002;30(5):1255–61.

[13] King OD, Roth FP. A non-parametric model for transcription factor binding sites. Nucleic Acids Res 2003;31(19):e116. Evaluation Studies.

[14] Ellrott K, Yang C, Sladek FM, Jiang T. Identifying transcription factor binding sites through Markov chain optimization. Bioinformatics 2002;18(Suppl. 2):100–9. Evaluation Studies.

[15] Gershenzon NI, Stormo GD, Ioshikhes IP. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. Nucleic Acids Res 2005;33(7):2290–301. Evaluation Studies.

[16] Keich U, Pevzner PA. Finding motifs in the twilight zone. In: RECOMB'02; 2002. p. 195–203.

[17] Rahmann S, Müller T, Vingron M. On the power of profiles for transcription factor binding sites detection. Stat Appl Genet Mol Biol 2003;2(1).

[18] Nowakowski S, Fidelis K, Tiuryn J. Introducing dependencies into alignments analysis and its use for local structure prediction in proteins. LNCS 3911; 2006. p. 1106–13.

[19] Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, et al. The TRANSFAC system on gene expression regulation. Nucleic Acids Res 2001;29(1):281–3.

[20] Goldstein L, Waterman MS. Approximations to profile score distributions. J Comput Biol 1994;1(2):93–104.

[21] Staden R. Methods for calculating the probabilities of finding patterns in sequences. CABIOS 1989;5(2):89–96.

[22] Bailey TL, Gribskov M. Combining evidence using *p*-values: application to sequence homology searches. Bioinformatics 1998;14(1):48–54.

[23] Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. J Mol Biol 1986;188:415–31.

[24] Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat 1951;22(1):79–86.

[25] Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 1990;18:6097–100.

[26] Halees AS, Leyfer D, Weng Z. PromoSer: a large-scale mammalian promoter and transcription start site identification service. Nucleic Acids Res 2003;31(13):3554–9.