BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 14 2005, pages 3082–3088
doi:10.1093/bioinformatics/bti477

Sequence analysis

A multiple-feature framework for modelling and predicting transcription factor binding sites

Rainer Pudimat, Ernst-Günter Schukat-Talamazzini and Rolf Backofen*

Institut für Informatik, Friedrich-Schiller-Universität, Ernst-Abbe-Platz 3, D-07743 Jena, Germany

Received on January 11, 2005; revised on April 5, 2005; accepted on April 27, 2005
Advance Access publication May 19, 2005

ABSTRACT

Motivation: The identification of transcription factor binding sites in promoter sequences is an important problem, since it reveals information about the transcriptional regulation of genes. For analysing transcriptional regulation, computational approaches for predicting putative binding sites are applied. Commonly used stochastic models for binding sites are position-specific score matrices, which show weak predictive power.

Results: We have developed a probabilistic modelling approach, which allows to consider diverse characteristic binding site properties to obtain more accurate representations of binding sites. These properties are modelled as random variables in Bayesian networks, which are capable of dealing with dependencies among binding site properties. Cross-validation on several datasets shows improvements in the false positive error rate and the significance (*P*-value) of true binding sites.

Supplementary information: A more extensive description of validation results are available at http://www.bio.inf.uni-jena.de/Software/promapper/

Contact: backofen@inf.uni-jena.de

INTRODUCTION

A fundamental challenge of recent biological research has been to understand the transcriptional regulation of gene expression. A major fraction of this regulation is exerted by the binding of transcription factors to regulatory DNA elements (called transcription factor binding sites) in the upstream region of a gene (Alberts *et al*., 2002). It is a standard procedure to determine these binding sites via biological assays. The TRANSFAC database version 8.3 (Wingender *et al*., 2001) contains ∼14 400 entries of experimentally determined sites.

Despite the quite strong sequence similarity among the binding sites of a certain transcription factor, the relatively short sequence motifs often show a certain degree of variability, and matches of one such motif could be present by chance anywhere in a genome without having regulatory functions. Hence, the development of highly specific and accurate computer-aided detection approaches is still an unsolved problem (Levy and Hannenhalli, 2002).

Besides experimental approaches, such as expression analyses and ChIP on chip, there are two widely used *in silico* strategies. The first one, which is known as phylogenetic footprinting (Dieterich *et al*., 2003; Brudno *et al*., 2003), is based on sequence comparison of upstream sequences of orthologous genes from diverse species. Highly conserved regions within these sequences are assumed to have a functional meaning, because their similarity is likely to result from a higher selective pressure. While phylogenetic footprinting is very successful in predicting conserved regulatory patterns, it cannot detect binding sites that only occur in one of the species.

The second widely accepted strategy, which is considered in this paper, is to use stochastic models for predicting transcription factor binding sites. Among these methods, the majority uses position-specific score matrices (PSSMs) (Aerts *et al*., 2003; Boardman *et al*., 2003; Kel *et al*., 2003). Each entry of such a matrix stands for the frequency of certain nucleotides (matrix rows) in certain positions (matrix columns) within the binding site motif (Stormo, 2000).

Albeit their predominant role, PSSMs have only weak predictive power for several reasons. Besides the problems shared by most current *in silico* approaches, namely the inability to model biologically important circumstances, such as cooperativity between factors and the positioning of nucleosomes (Wasserman and Sandelin, 2004), at least two of them are symptomatical for PSSMs.

First, PSSMs assume statistical independence among the motif positions. Recent literature shows that this is too strong an assumption (Bulyk *et al*., 2002; Man and Stormo, 2001; Benos *et al*., 2002).

Second, PSSMs are restricted to motif column distributions. Therefore, they are not suited to describe sequence properties of higher order. Examples of properties that cannot be modelled using PSSMs are the sequence-dependent major-groove width (Ponomarenko *et al*., 1999) around a binding site, the GC-content of its flanking region or the presence of a co-acting factor's binding site in its neighbourhood (Grabe, 2002). Even if higher order properties are representable, it will be problematic to integrate them since this requires the learning of an enormous number of parameters. Usually, there are not enough data available to learn these properties. Any such property in a motif's flanking region has to be modelled in PSSMs with a number of parameters that is linear in the length of the flanking region sequence.

We approach this problem by directly considering higher-order sequence properties (we here call model features). Thus, we have to learn only all possible values for the selected features instead of their underlying sequence. To give a concrete example, it has been shown that the binding sites for HFN1 show a significantly different melting temperature of the surrounding region (Ponomarenko *et al*., 1999). To model this observation, we need only one parameter, namely whether the melting temperature is above or below a given threshold.

To learn these features together with the binding sites themself, we employ Bayesian networks (BNs) (Mitchell, 1997) for several

*To whom correspondence should be addressed.

reasons. First, they provide the necessary flexibility for choosing the most predictive properties of the sites. Second, BNs overcome the second obstacle of PSSMs, in allowing expression of dependencies among these properties. It has been shown previously that dependencies between the positions of a motif are important (Bulyk *et al*., 2002). Barash *et al*. (2003) first applied BNs to binding site prediction. In contrast to our work, their application of BNs is an extension of the PSSM by considering dependencies between sequence positions, without modelling more complex sequence properties. In a related problem of predicting splice sites, Bayesian belief networks have already been used for modelling dependencies between positions (Cai *et al*., 2000; Castelo and Guigo, 2004).

Here, we show that the classification error rates compared with PSSMs can be reduced by our BN approach. We demonstrate this approach by performing cross-validations on three datasets of mammalian transcription factor binding sites.

MODELLING APPROACH

The goal of learning a model is to detect common properties between different samples given in the datasets (e.g. a set of known binding sites for a given transcription factor). For this purpose, it is common to describe these properties by a vector $(F_1, \ldots, F_K)$ of features, whose values $f_1, \ldots, f_K$ can be extracted directly from sample sequences. For PSSMs, these features are simply the nucleotides at the different positions. In our case, we have more complex features like the leftmost starting position of a given consensus within binding sites. Using more complex features, we are also able to characterize important properties of the flanking regions such as structural attributes. In addition, some of the more complex features are also used as a technique for parameter reduction.

Now these features $F_1, \ldots, F_K$ are modelled as discrete random variables, and the problem of learning is to estimate the joint probability distribution $P(F_1, F_2, \ldots, F_K)$ from a set of training samples. In the following, these random variables associated with the features in our model are called model features. We currently distinguish between five main classes of features (called feature types). The first three are based on observations given in the literature.

Motif column feature (PSSM). Corresponds to column distributions of PSSMs. Clearly, it is possible to emulate usual PSSMs with these features. Remarkably, even this feature alone is more expressive than PSSMs since we can model dependencies among the different columns.

Structural property feature. Approximation of the sequence-dependent contribution to a physical property of DNA. Among the 38 parameters provided, there are conformational parameters such as helical twist, helical slide or minor groove width (El Hassan and Calladine, 1997) and physicochemical parameters like free energy change or melting temperature. These parameters are defined for dinucleotides. According to Ponomarenko *et al*. (1999), we compute the average of a parameter for a given subsequence by summing up the values of the dinucleotide steps (Fig. 1). This average is then compared with the average calculated from a null model. Oshchepkov *et al*. (2004) have developed a software tool for extracting such structural patterns from a set of binding sites.

PSSM hits for co-acting factor's site. Since the biological meaning of a sequence as a binding site often depends on the presence of a binding site for a co-acting transcription factor, features of this type evaluate whether or not there is a PSSM hit of a co-actor's site in the neighbourhood of the current position.

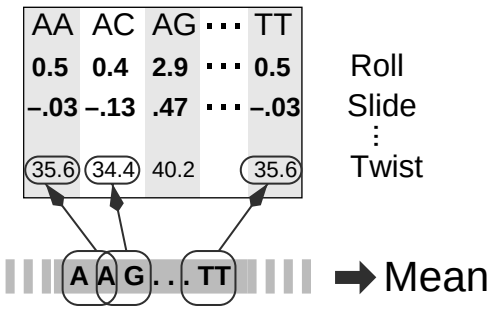

Fig. 1. The structural feature. Here, the helical twist feature is selected.

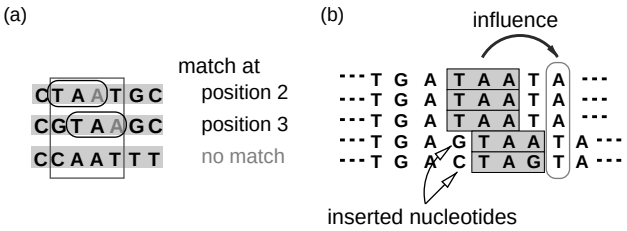

Fig. 2. The feature for the consensus **TAR** (R stands for A or G); (**a**) calculation of the feature values and (**b**) application of the feature in the case of an inserted nucleotides.

In addition, we add the following features that enhance the descriptive power and can be used to reduce the number of parameters:

Consensus match start position. Represent distributions over possible start positions of leftmost or rightmost matches of a given consensus pattern within a given range. Features of that type can help to detect deleted nucleotides in a subset of all training sites, and hence influence the distribution of dependent motif columns (Fig. 2).

Subsequence nucleotide profiles. Measure the coarse base composition of a given subsequence (e.g. a lower or higher A + T content of the flanking region). This type of feature allows consideration of sequence properties of high order with a minimum amount of parameters.

The different features are summarized in Figure 3. To give a concrete example of how these features can be used to reduce the number of parameters, consider the binding sites for the transcription factor Sp1. They usually contain a high proportion of C and G nucleotides. Even though this can be observed for the flanking regions of these sites as well, the exact positions of C and G nucleotides are only weakly conserved. Modelling these flanking regions of length k with PSSM-like motif column features would require the estimation of $4k$ parameters without benefitting from it owing to the weak motif in the flanking regions. In our approach, the observation of high C and G fractions could be considered by including a suitable profile feature which only needs two parameters (above or below a suitable threshold).

Though model features differ in their value range and their rules for mapping sequences to feature values, we simply assume that

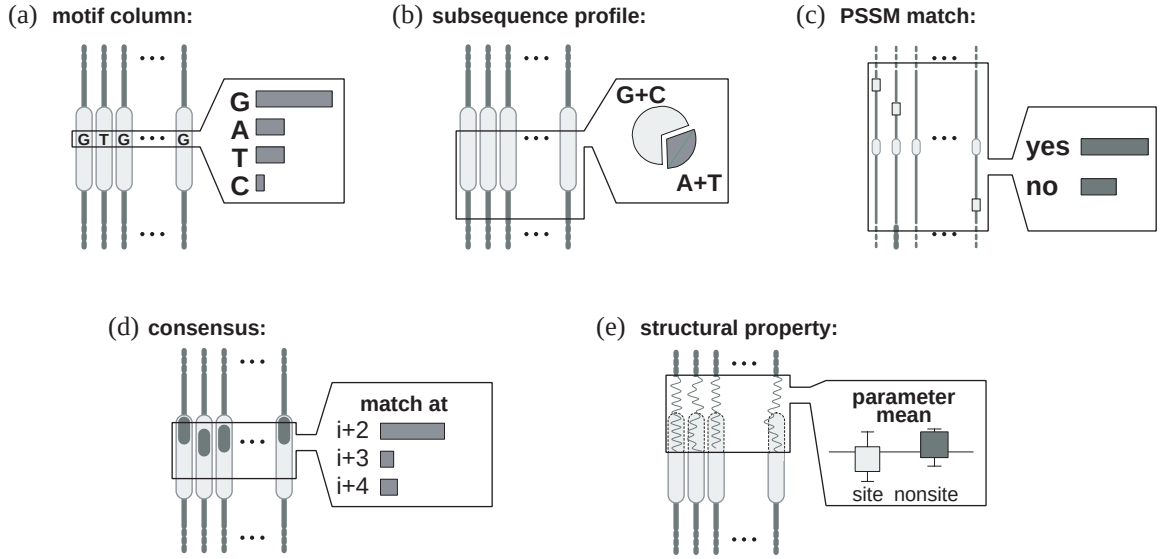

Fig. 3. Currently implemented model features: (**a**) PSSM-like base distribution of one motif column, (**b**) base composition in a subsequence of the binding site, (**c**) PSSM matches for co-acting factor's sites in neighbourhood, (**d**) distribution over start positions of a particular consensus and (**e**) sequence-dependent structural or physicochemical feature.

they are discrete functions $F_k : S \times \mathbb{N} \mapsto \mathrm{ran}(F_k)$ from the Cartesian product of the set S of DNA sequences and integers to the feature range. The additional integer input determines a reference position of the sequence. Since we cannot deal with continuous random variables, the continuous ranges of structural property features and subsequence profile features are discretized. The interval borders for discretization are determined by an entropy-based algorithm developed by Fayyad and Irani (1993).

Bayesian belief networks

Constructing a model for binding sites of a transcription factor requires the choice of model features $F_1, F_2, \ldots, F_K$. The next step is the estimation of the joint probability distribution $P(F_1, F_2, \ldots, F_K)$. If we assumed independence of the different features as in the case of PSSMs, the joint distribution would be calculated as the product of single probabilities of the feature values, i.e. by $\prod_{i=1}^{k} P(F_i)$. However, we cannot assume statistical independence between features as we have mentioned above. Even if there were no possible overlapping features (such as consensus features) in the model, we still would have to model the dependencies between the columns of the binding sites (Bulyk *et al*., 2002). This implies that the joint probability has to be calculated based on conditional probabilities modelling the dependencies.

However, it is not practicable to model all possible pairwise dependencies, since the number of parameters to be estimated would grow exponentially in the number of model features considered in this case. Concomitantly, the available amount of training data is often rather small for transcription factor binding sites. Bayesian belief networks (BNs) are a good trade-off between these two extrema. Formally defined, a BN is a pair $B = (G, \mathcal{P})$. Its first component G is an annotated directed acyclic graph whose vertices correspond to random variables $X_1, X_2, \ldots, X_K$, and whose edges determine direct dependencies between connected variables. The direction of each edge denotes that the value of the parent node influences the value of the child node. The network also encodes implicit independence assumptions in the sense that a random variable is independent of its non-descendants, given its parents in G. The second component $\mathcal{P}$ is a parameter set which quantifies the network. It contains probability parameters $p_{x_k|\pi_{x_k}} = P_B(X_k = x_k | \prod_{x_k} = \pi_{x_k})$ for each possible value x_k of random variable X_k and each configuration π_{x_k} of the set of parent variables $\prod_{x_k}$ (Friedman *et al*., 1997). So, a BN B defines a unique joint probability distribution over all concerned random variables $\boldsymbol{X} = \{X_1, X_2, \ldots, X_K\}$ given by

$$P_B(x_1, x_2, \ldots, x_K) = \prod_{k=1}^{K} P_B(x_k|\pi_{x_k}). \quad (1)$$

It is clear that our model features play the role of the random variables in the BNs. Figure 4 shows what a BN constructed from diverse model features and learned from a set of binding sites could look like.

Learning The input of the learning procedure is a set of aligned sites with their corresponding contexts. The learning process of multiple-feature TFBS models comprises three tasks: (1) the selection of an appropriate set of model features, (2) the calculation of the dependencies between the different features (i.e. the determination of the network structure) and (3) the estimation of the corresponding parameters (probabilities). The first task requires the second and third as a subtask.

Concerning the selection of an appropriate feature subset, this task is a common instance of the feature subset selection problem, a widely researched problem concerning the dimensioning of classifiers in pattern recognition. We have chosen to apply the sequential floating feature selection (SFFS) method described by Pudil *et al*. (1994), which works well with the type of optimization criteria used in our case (namely classification error rate). We start with an initial

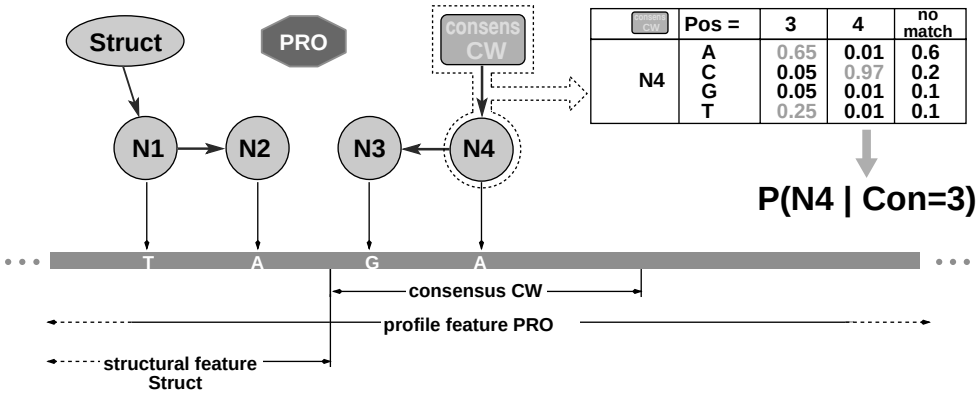

Fig. 4. Multiple-feature BN model for transcription factor binding sites. The BN consists of so-called model features (random variables—the nodes of the graph), stochastic dependencies (edges) among them, and a table of conditional probabilities for each node. An exemplary probability table is given on the right. Different shapes and colours emphasize diverse types of model features. Each feature takes a particular subsequence to compute its values. The corresponding intervals are shown below the input sequence).

feature set consisting of the motif column features for the aligned positions of the binding sites. This implies that the corresponding initial network is similar to a PSSM model, with the exception that existing dependencies among the different positions can be handled beforehand. Then, it successively adds that feature from the entirety of features which leads to the highest improvement of the classification performance. After each add-operation, the search algorithm removes model features from the subset, as long as this improves the performance previously achieved. Allowing the removal of previously chosen features avoids being trapped in a local optimum. The algorithm stops when the previously determined number of model features has been included in the feature subset.

For the second task of determining the best network structure, given some training data, it is known that this an NP-hard problem (Pearl, 1988). Hence, we confine ourselves to considering a subclass of all possible network structures, in which the freedom of setting up edges is constrained. In this subclass, for each model feature, at most, one incoming edge is allowed. Thus, the probability of a certain value of a model feature can depend on, at most, one other model feature. Obviously, BNs that fulfil this property form a set of trees. For that reason, they are called Tree-augmented networks (TANs) (Friedman *et al*., 1997).

In the case of connected TANs (a TAN is connected if there is a path from each node to any other node, not considering edge orientations), there are efficient structure-learning algorithms that reduce the problem of determining the optimal tree structure to finding a maximum-weighted spanning tree (Chow and Liu, 1968). After having obtained such a spanning tree, a direction has to be assigned to each edge of the tree. This is done by randomly choosing a root node and orientating all edges to be directed outward to it. It was shown that this procedure results in an optimal network structure among all possible TAN structures (Friedman *et al*., 1997). The procedure of Chow and Liu (1968) always results in connected graphs with one edge per model feature. To achieve connectivity, some edges are included without seeing any dependence between the according features. Especially in cases in which two model features have nearly constant values in all training samples, the mutual information content (MIC) which is assigned to each edge and indicates the interdependencies between adjoining features is not meaningful. Therefore, the estimation of the conditional probabilities is far from being robust. In order to circumvent this problem, we re-arrange the tree network by removing all edges with an MIC below a given threshold. As a consequence, the tree is split into several smaller trees which finally form our network structure.

To perform supervised learning BNs with transcription factor binding sites, one needs a sample set of known binding sites justified with respect to a reference position. We have chosen the reference position to be the first position according to TRANSFAC (Wingender *et al*., 2001). In addition, we must include as much flanking region relative to the reference position as the included model features demand. Each site in the sample set has to be transformed into a vector of variable assignments by applying the model feature functions. Finally, these vectors are presented to the network learning procedure.

Application of trained models The procedure of scanning an input sequence $\boldsymbol{s} = s_1 s_2 \cdots s_L$ is quite similar to the learning process. Given a model $M_{\mathcal{F}}$ with a feature set $\mathcal{F}$, a variable assignment vector $\boldsymbol{f}(l) = (f_1(l), f_2(l), \ldots, f_K(l))$ is computed for each position l of the sequence. The network returns the joint probability of each vector. Owing to the fact that a model feature could use base pairs upstream of the reference position, some positions at the 5′ end of a sequence cannot be evaluated (the same holds for the 3′ end).

To decide whether a sequence position is a putative binding site or not, we compare the output probability $P_{M_{\mathcal{F}}}(f_1, f_2, \ldots, f_K)$ of the binding site model with the output probability $P_N(f_1, f_2, \ldots, f_K)$ of a background model. This is, according to the features which were chosen, an equally dimensioned Bayesian belief network trained on arbitrary eukaryotic promoter sequences, and could be done by using the common log-odds scores

$$S(\boldsymbol{f}) = \log_2 \frac{P_{M_{\mathcal{F}}}(f_1, f_2, \ldots, f_K)}{P_N(f_1, f_2, \ldots, f_K)}. \qquad (2)$$

This way of calculating scores does not ensure their comparability, given the fact that models for different transcription factors can differ in their sets of features. The reason is that, it is not possible to compare probabilities produced by features of a different nature (e.g. ‘Is it better to see a T at position 1 with probability 0.9 or to see a helical twist above 34.5° with probability 0.75 at subsequence $s_m \cdots s_n$?’).

Furthermore, the question arises which features should be contained in the background model.

We start to tackle these problems by determining the features that should be integrated in the background model. Let $\mathcal{U}$ be the set of all model features that occur in any model within our classification system. Then the background model N is constructed by including all features $U \in \mathcal{U}$ and learning the probability distribution and network structure, based on the background data described above. The next step is to expand each site model [i.e. the numerator in Equation (2)] in a simple way to include the missing context of the background model. Let $\mathcal{F} \subset \mathcal{U}$ be the set of model features considered in a model $M_{\mathcal{F}}$. All remaining features $\mathcal{G} = \mathcal{U} - \mathcal{F}$ are included in other models and in the background model N. The scores would no longer be comparable, if they were defined on a joint probability distribution $P(\boldsymbol{u})$ of all values $\boldsymbol{u} = (\boldsymbol{f}, \boldsymbol{g})$ of the whole set $\mathcal{U} = \mathcal{F} \uplus \mathcal{G}$. Since the features $G \in \mathcal{G}$ are supposed to contain no information about binding sites modelled in $M_{\mathcal{F}}$ (otherwise they would have been included in the model), we assume that the contribution of $\mathcal{G}$ can be described by the background distribution N. Starting from

$$P(\boldsymbol{u}) = P(\boldsymbol{f}, \boldsymbol{g}) = P(\boldsymbol{f}) \cdot P(\boldsymbol{g}|\boldsymbol{f}), \tag{3}$$

the first part of the product on the right is clearly substituted by the joint probability $P_{M_{\mathcal{F}}}(\cdot)$ of the binding site model $M_{\mathcal{F}}$. The second part, which is not modelled in $M_{\mathcal{F}}$, is approximated by the conditional probability of observing values $\boldsymbol{g}$ of variables in $\mathcal{G}$, given the values $\boldsymbol{f}$ of variables in $\mathcal{F}$ according to the background distribution $P_N(\cdot)$. Fortunately, efficient algorithms to approximate these conditional probabilities with Bayesian belief networks exist (Lauritzen and Spiegelhalter, 1988). For each model $M_{\mathcal{F}}$ with its corresponding feature set $\mathcal{F} \subset \mathcal{U}$, scores are calculated with respect to:

$$\begin{aligned} S(\boldsymbol{u}) = S(\boldsymbol{f}, \boldsymbol{g}) &= \log \frac{P(\boldsymbol{f}, \boldsymbol{g})}{P_N(\boldsymbol{f}, \boldsymbol{g})} \\ &\approx \log \frac{P_{M_{\mathcal{F}}}(\boldsymbol{f}) \cdot P_N(\boldsymbol{g}|\boldsymbol{f})}{P_N(\boldsymbol{f}, \boldsymbol{g})}. \end{aligned} \tag{4}$$

By computing the scores in this way, we achieve comparable scores. Note that the conditional probability in the numerator of Equation (4) differs with respect to the particular decomposition $\mathcal{U} = \mathcal{F} \uplus \mathcal{G}$. Cancelling down to $S(\boldsymbol{u}) = \log[P_{M_{\mathcal{F}}}(\boldsymbol{f})]/[\sum_{\boldsymbol{g}} P_N(\boldsymbol{f}, \boldsymbol{g})]$ is theoretically possible, but requires the computation of the huge marginalization over all assignments of variables in $\mathcal{G}$, which is technically not feasible.

MATERIAL AND RESULTS

In order to evaluate our modelling approach we have performed tests on three datasets. We restricted ourselves to genomic sequences. Artificial sequences, such as those of SELEX experiments, were not taken into account, since they do not contain an appropriate context.

Data

The first dataset consists of 26 experimentally proven MEF-2 binding sites. MEF-2 is a transcription factor, which is involved in the regulation of several genes concerning skeletal, smooth and cardiac muscles. The binding sites of the MEF-2 dataset were taken from a study of Wasserman and Fickett (1998), and from the TRANSFAC database (Wingender *et al*., 2001). The second dataset contains 78 AP-1 boxes. These are binding sites for either heterodimers made of transcription factors JUN and FOS, or for homodimers made of two JUN molecules. The sequences of this dataset were taken completely from TRANSFAC. As a third transcription factor we have chosen, Sp1, which is ubiquitous in most tissues and cell states. Sp1 is a member of the class of zinc-finger transcription factors. Three zinc fingers handle the binding to target DNA. The preferred binding sequence is quite G rich. In all datasets, the binding sequences given in TRANFAC were enriched with their flanking regions obtained via the EMBL links (Stoesser *et al*., 1999) given in TRANSFAC entries.

Procedure

We compared the classification performance of our approach with PSSM models using the 10-fold cross-validation technique. Although PSSM can be based on many different scoring schemes [e.g. free-energy change (Stormo and Fields, 1998)], we follow the approach of Barash *et al*. (2003) and use PSSMs modelling nucleotide distributions for each position of the binding site. Note that this is the type of PSSM used in the TRANSFAC tools (Kel *et al*., 2003).

This means that in each trial, a PSSM model and a multiple-feature TAN model were learned using 90% of the dataset. PSSMs were trained and applied using the weight matrix framework of BioJava. After being learned, both models were used to detect the remaining 10% of the sample in a test set consisting of these samples and 5000 random sequences. We used random sequences to lower the risk of finding unknown, but true, binding sites within promoter sequences which would have been counted as false positives (FPs). These random sequences were sampled using a third-order Markov chain trained on the promoter sequences of the particular dataset.

Thus, a complete cross-validation consists of 10 trials. Each binding site was used exactly one time for testing the models. For each trial, models are tested on positive and negative samples. Positive samples are the 10% test binding sites, negative samples are sequence-windows of random sequences. For each sample in a test set, a score is calculated. Knowing the label (positive or negative) of all samples in a test set, we were able to enumerate a common contingency table with the four statistics: true positives (TPs), true negatives (TNs), FPs and false negatives (FNs) for a given score threshold. Contingency tables were used to create ROC plots in the following manner: Calculate an entry in contingency tables for a TP rate of x%, the x% best scoring positive samples, where taken; the worst score among this set was taken as a threshold to calculate the corresponding FP rate.

Furthermore, we have used exemplary contingency tables to calculate the $F_{0.5}$-measure: $F_{0.5} = (2 \cdot r \cdot p)/(r + p)$, where the recall $r = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ is the part of real sites that were considered as matches, and the precision $p = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ is the fraction of TPs among all matches. Hence, the $F_{0.5}$-measure describes the quality of a model considering both, the ability to detect binding sites and the ability to avoid FPs.

Scores of negative samples were used to estimate a score distribution for the relevant model. Therefore, it was assumed that scores are distributed according to an extreme value distribution (Durbin *et al*., 1998). The parametric approximation of the score distribution provides an easy computation of P-values of scores. The P-value of the mean of all positive sample scores was computed for each model as a measure for the expected number of FPs.

Results

The cross-validation experiments revealed a better performance of our multiple-feature TAN models, compared with PSSMs in all

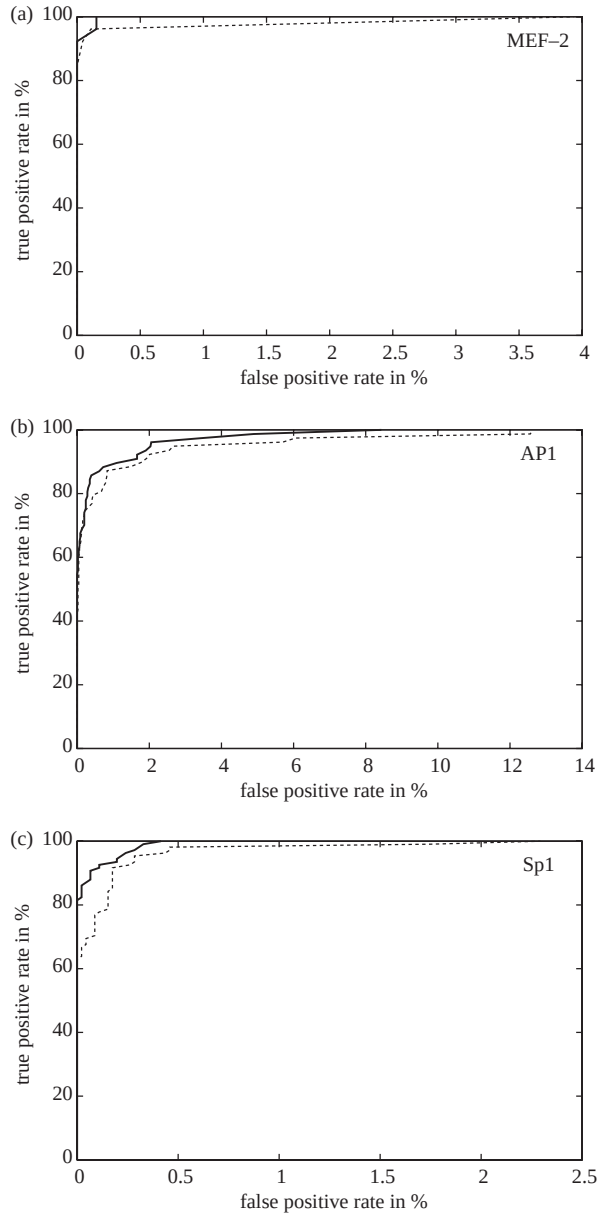

Fig. 5. ROC plots of multiple-feature TAN models (solid line) and PSSMs (dashed line) for (**a**) MEF-2 dataset, (**b**) AP-1 dataset and (**c**) Sp1 dataset. They were calculated by successively setting thresholds so that a defined TP rate was established and by determining the FP rate for each of these thresholds.

calculated quality measures. The ROC-plots of Figure 5a–c illustrate that TAN models achieved lower FP rates for almost every fixed TP rate. Two properties of these plots deserve closer attention: first, the minimal fixed TP rate that leads to a vanishing FP rate is higher for TAN models than for PSSMs. Second, when adjusting the score threshold to a TP rate of 100%, the FP rates of TAN models are lower by a multiple compared with PSSMs. One reason for these reduced error rates is shown for the AP1 dataset in Figure 6 (analogous plots for the other datasets are given in the supplements). In the figure, estimated score distributions for random sequences on the one hand

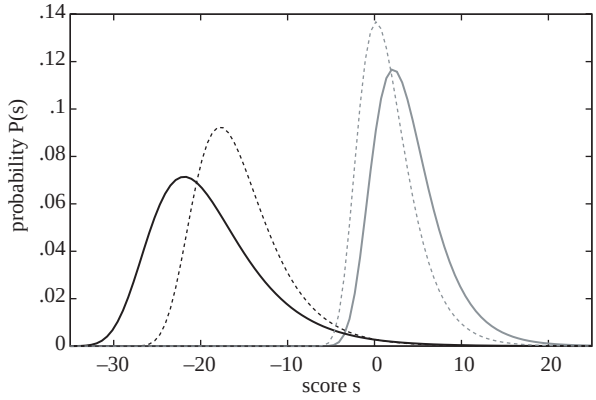

Fig. 6. Score distributions of multiple-feature TAN models (solid line) and PSSMs (dashed line) for the AP1 dataset. The two leftmost curves denote the probability distribution of random sequence scores assigned by both models, the two rightmost curves denote the distribution of scores assigned to true binding sites. The larger distance between the means of both TAN curves compared with that of both PSSM curves promises improvement in classification error rates.

Table 1. Overview of calculated measures for all datasets

Dataset	Model type	Mean TP P-value	Best $F_{0.5}$-value
MEF-2	PSSM	0.00406	0.90
	TAN	0.00054	0.94
AP1	PSSM	0.01856	0.79
	TAN	0.00659	0.81
Sp1	PSSM	0.00966	0.91
	TAN	0.00513	0.93

In all cases our TAN approach shows lower P-values for real binding sites and higher $F_{0.5}$-value.

and true binding sites on the other hand were drawn for both, TAN models and PSSMs. While a better separation of random sequences and true sites is achieved by higher scores for true binding sites and lower scores for random sequences, this alone does not ensure lower error rates. What is crucial is a reduction of the cutting area under the two curves of TAN models compared with those of PSSMs. The features-selection algorithm (SFFS) described in the Method section prefers model feature sets with these properties.

Another commonly used measure for the quality of prediction is the P-value of a true binding site, according to the estimated random sequence score distribution. The P-value for the average true-site score is lower for TAN models than for PSSMs in all datasets. At the same time, the additionally calculated $F_{0.5}$-measure, which reflects the overall performance of a model, is higher for TAN models in all cases. The comparisons according to these two measurements are listed in Table 1.

MEF-2. In contrast to the other datasets, the number of model features was fixed at 10 owing to the small number of samples. The SFFS algorithm found consensus features (e.g. for the consensus `TNWWW` between positions 1 and 5), profile features (e.g. the fraction of adenine nucleotides between positions 1 and 10) and structural

features (e.g. the free-energy change between positions 3 and 5 to be lower/higher than −1.3). A complete description of the model feature set is included in the Supplementary Material. The PSSMs, which were used for comparison with our TAN models, were designed to model positions −4 to 13 (same positions as in the PSSM given in TRANSFAC).

AP1 and Sp1. Since these datasets have a more comfortable size, we adjusted the number of model features to 15. Besides some of the nucleotide features of the initial feature set, the SFFS algorithm included several additional features of nearly all types. In the case of the AP1 model, these were mainly structural features. In the case of the Sp1 model, the SFFS algorithm obtained profile features and structural features to be predicative for Sp1-binding sites. As for MEF-2, the PSSMs, which were used for comparison, were dimensioned according to the positions that are considered in TRANSFAC.

CONCLUSIONS

We have developed a flexible modelling approach for transcription factor binding sites that allows the consideration of arbitrary sequence-related or measurable biological properties of binding sites. The binding site properties, which are called model features, are modelled together in Bayesian belief networks.

We have presented a study on this framework, where we have implemented five possible binding site properties. We have used structural features (such as helical twist) in combination with standard sequence features (such as consensus start-positions or motif-column distribution). We have presented an approach to automatically learn the set of features that are predicative for the modelled binding site.

To validate our approach, we compared its predictive power to that of PSSM models. Therefore, we performed cross-validation tests on two datasets, namely MEF-2 binding sites, AP-1 boxes and Sp1 binding sites. In all cases we find considerable improvement of classification error rates using TAN models.

ACKNOWLEDGEMENTS

We thank Michael Hiller for proofreading the manuscript. This work was supported by the German Ministry of Education and Research under grant no.

REFERENCES

Aerts,S. *et al*. (2003) Toucan: deciphering the *cis*-regulatory logic of coregulated genes. *Nucleic Acids Res.*, **31**, 1753–1764.

Alberts,B., Johnson,A., Lewis,J., Raff,M., Roberts,K. and Walter,P. (2002) *Molecular Biology of the Cell*, 4th edn. Garland Science, NY.

Barash,Y., Elidan,G., Friedman,N. and Kaplan,T. (2003) Modeling dependencies in protein–DNA binding sites. In *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology (RECOMB)*, pp. 28–37.

Benos,P.V. *et al*. (2002) Additivity in protein–DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.

Boardman,P.E. *et al*. (2003) SiteSeer: Visualisation and analysis of transcription factor binding sites in nucleotide sequences. *Nucleic Acids Res.*, **31**, 3572–3575.

Brudno,M. *et al*. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.*, **13**, 721–731.

Bulyk,M.L. *et al*. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.

Cai,D. *et al*. (2000) Modeling splice sites with Bayes networks. *Bioinformatics*, **16**, 152–158.

Castelo,R. and Guigo,R. (2004) Splice site identification by idlBNs. *Bioinformatics*, **20**, I69–I76.

Chow,C.K. and Liu,C.N. (1968) Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inform. Theory*, **14**, 462–467.

Dieterich,C. *et al*. (2003) Corg: a database for comparative regulatory genomics. *Nucleic Acids Res.*, **31**, 55–57.

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

El Hassan,M.A. and Calladine,C.R. (1997) Conformational characteristics of DNA: empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. *Phil. Trans. R. Soc. Lond. A*, **355**, 43–100.

Fayyad,U. and Irani,K. (1993) Multi-interval discretization of continous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*. Morgan Kauffmann, Chambery, France, pp. 1022–1027.

Friedman,N. *et al*. (1997) Bayesian network classifiers. *Machine Learning*, **29**, 131–163.

Grabe,N. (2002) AliBaba2: context specific identification of transcription factor binding sites. *In Silico Biol.*, **2**, S1–S15.

Kel,A.E. *et al*. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.

Lauritzen,S.L. and Spiegelhalter,D.J. (1988) Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc.*, **50**, 157–224.

Levy,S. and Hannenhalli,S. (2002) Identification of transcription factor binding sites in the human genome sequence. *Mamm. Genome*, **13**, 510–514.

Man,T.K. and Stormo,G.D. (2001) Non-independence of Mnt repressor–operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.*, **29**, 2471–2478.

Mitchell,T.M. (1997) *Machine Learning*. WCB/McGraw-Hill.

Oshchepkov,D.Y. *et al*. (2004) SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. *Nucleic Acids Res.*, **32**(Web Server issue), W208–W212.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems*, 2nd edn. Morgan Kauffmann.

Ponomarenko,J.V. *et al*. (1999) Conformational and physicochemical DNA features specific for transcription factor binding sites. *Bioinformatics*, **15**, 654–668.

Pudil,P. *et al*. (1994) Floating search methods in feature-selection. *Pattern Recog. Lett.*, **15**, 1119–1125.

Stoesser,G. *et al*. (1999) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **27**, 18–24.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Stormo,G.D. and Fields,D.S. (1998) Specificity, free energy and information content in protein–DNA interactions. *Trends Biochem. Sci.*, **23**, 109–113.

Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.

Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.

Wingender,E. *et al*. (2001) The transfac system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.