Journal of Bioinformatics and Computational Biology

RECOGNITION OF CIS-REGULATORY ELEMENTS WITH VOMBAT

Stefan Posch

Jan Grau

Andre Gohr

Institute of Computer Science, University Halle
06099 Halle (Saale), Germany
Jan.Grau@informatik.uni-halle.de

Irad Ben-Gal

Department of Industrial Engineering, Tel-Aviv University
Tel-Aviv, 69978, Israel
bengal@eng.tau.ac.il

Alexander Kel

Biobase GmbH
38304 Wolfenbüttel, Germany
alexander.kel@biobase-international.com

Ivo Grosse

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)
06466 Gatersleben, Germany
grosse@ipk-gatersleben.de

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Variable order Markov models and variable order Bayesian trees have been proposed for the recognition of cis-regulatory elements, and it has been demonstrated that they outperform traditional models such as position weight matrices, Markov models, and Bayesian trees for the recognition of binding sites in prokaryotes. Here, we study to which degree variable order models can improve the recognition of eukaryotic cis-regulatory elements. We find that variable order models can improve the recognition of binding sites of all of the studied transcription factors. To ease a systematic evaluation of different model combinations based on problem-specific data sets and allow genomic scans of cis-regulatory elements based on fixed and variable order Markov models and Bayesian trees, we provide the VOMBAT server to the public community.

Keywords: DNA motifs; variable order models; Markov models; Bayesian networks.

1. Introduction

The in silico prediction of cis-regulatory elements in DNA is an interesting and important problem in genome research. Binding of transcription factors to their binding sites in the promoter of their target gene is a prerequisite for their activation or repression. Hence, knowledge of the location of DNA binding sites is of importance to elucidate the underlying regulatory mechanisms. As wet-lab experiments are expensive to conduct, computational techniques are attractive to tackle this question despite being less accurate in general.

A wide range of techniques for predicting cis-regulatory elements build on statistical models, one for the cis-regulatory elements under consideration and one for the background of non-regulatory sequences. These models are trained using a labeled dataset and subsequently employed in a supervised classifier for predicting new locations of the cis-regulatory element scrutinized. Within such an approach, the choice of appropriate model families is of importance for the accuracy of predictions.

A widely used model class for predicting DNA motifs are Markov models[1,2,3] used also in many other classification problems. One well-known example is the position weight matrix (PWM) model[4], which is an inhomogeneous Markov model of order 0. It assumes each position statistically independent of all other positions. Although it is an open question whether this independence assumption is reasonable[5,6,7], PWM models may outperform Markov models of higher order like the weight array matrix (WAM) model[8], which is an inhomogeneous Markov model of order 1. One possible explanation is the limited amount of experimentally verified binding sites available for the learning phase resulting in the problem of overfitting for models with larger numbers of parameters.

From a statistical point of view, Markov models assume for each nucleotide of the DNA sequence the statistical dependence on a fixed number of directly preceding nucleotides, called the context, and independence otherwise. With the goal of improving the performance of Markov models, two sources of potential shortcoming of this assumption can be identified, namely (i) the strictly sequential order of statistical dependencies with respect to the position of the nucleotides in the sequence and (ii) the exponential growth of parameters for a growing context length. Bayesian networks have been considered as an alternative to Markov models to allow for statistical dependencies among non-adjacent positions[7]. To cope with the exponential growth of parameters, variable order Markov models[10] were proposed. The power of variable order Markov models stems from the freedom to include only those contexts into the model for which there are strong statistical dependencies. The resulting variable order Markov models and Bayesian trees have been shown to outperform traditional models for the recognition of prokaryotic binding sites[9].

We define these statistical models in the next section, and we evaluate them in a case study using binding sites of six eukaryotic transcription factors in Section 3. The web server VOMBAT implementing training, prediction, and cross validation with any of these models on user-supplied data is shortly presented in Section 4.

2. Statistical models

To predict if a DNA sequence $x_1, \ldots, x_L$ of L nucleotides is a cis-regulatory element or not, we use likelihood ratio classifiers as the basic methodology. These require one statistical model $P_{\text{motif}}(x_1, \ldots, x_L)$ for cis-regulatory elements and one model $P_{\text{bg}}(x_1, \ldots, x_L)$ for non-regulatory elements constituting the background.

2.1. *Markov models*

In many classification algorithms for DNA cis-regulatory elements as well as splice sites or nucleosome binding sites the underlying family of distributions are Markov models. Starting from the standard factorization of an arbitrary distribution

$$P(x_1, \ldots, x_L) = P_1(x_1) \prod_{l=2}^{L} P_l(x_l | x_1, \ldots, x_{l-1}), \tag{1}$$

Markov models of order M assume that the conditional probabilities do not depend on all previous nucleotides, but only on the M previous nucleotides $x_{l-M}, \ldots, x_{l-1}$:

$$P(x_1, \ldots, x_L) = P_1(x_1) \prod_{l=2}^{L} P_l(x_l | x_{l-M}, \ldots, x_{l-1}), \tag{2}$$

where $x_a, \ldots, x_b = x_1, \ldots, x_b$ for $a < 1$ and $x_a, \ldots, x_b$ is the empty string for $a > b$.

If the conditional probabilities $P_l(x_l | x_{l-M}, \ldots, x_{l-1})$ are identical at all positions l, the Markov model of order M is called homogeneous (hMM(M)), otherwise it is called inhomogeneous (iMM(M)). In the following we focus on the description of inhomogeneous models, which are used to model the cis-regulatory elements, and comment only briefly on homogeneous models used as background models.

For any position l and observed nucleotide x_l at this position, we call the nucleotides $x_{l-t}, \ldots, x_{l-1}$ its context of length t in the following. The number of occurrences in a given dataset of any $(t+1)$-mer at positions $l-t$ up to l is denoted by $n_l(x_{l-t}, \ldots, x_l)$. For estimation of parameters we employ a maximum likelihood approach, which leads to

$$\hat{P}_l(x_l | x_{l-t}, \ldots, x_{l-1}) = \frac{n_l(x_{l-t}, \ldots, x_l) + p(t+1)}{\sum_{x \in \Sigma} (n_l(x_{l-t}, \ldots, x_{l-1}, x) + p(t+1))} \tag{3}$$

for any $t \geq 1$ and alphabet Σ. For DNA sequences we have $\Sigma = \{A, C, G, T\}$. Pseudo counts $p(t) = \frac{\epsilon}{d^t}, d = |\Sigma|$, are added to compensate for zero occurrences of some t-mers with an equivalent sample size ϵ[11]. These pseudo-counts are equivalent to using Dirichlet priors for the parameters in Bayesian learning. For hMMs the estimation is done analogously with the only difference that the summation runs over all positions l to obtain one single estimate $\hat{P}(x_l | x_{l-t}, \ldots, x_{l-1})$.

2.2. *Bayesian networks*

One shortcoming of Markov models is the sequential order of statistical dependencies. Generally this is appropriate for time series, but not obviously for DNA

binding sites[7,12]. Hence, Bayesian networks (BNs)[13,14,11] were considered as an alternative by several research groups. For example, first-order BNs were shown to outperform PWM models and WAM models in the prediction of splice sites [15,16] and cis-regulatory elements[7,9]. BNs allow for each position l statistical dependencies on an arbitrary set of other positions – called the parents $Pa(l)$ – as long as no cycles of statistical dependencies are induced. The probability distribution defined by a BN decomposes as

$$P(x_1, \ldots, x_L) = \prod_{l=1}^{L} P_l(x_l | \vec{x}_{Pa(l)}), \tag{4}$$

where the vector of parents may be empty for some of the positions. Bayesian trees (BTs) are special cases of BNs where each position must not have more than one parent. This restriction makes BTs well suited for modeling statistical dependencies in cases where only a limited amount of training data is available. To determine the structure of the BT, i.e., the directed tree encoding the statistical dependencies, we use a maximum likelihood approach. In case of a BT, finding the maximum likelihood structure is equivalent to finding a maximum spanning tree for the set of L positions using the mutual information between positions as edge weights[17].

2.3. *Variable order models*

For Markov models and BNs of fixed order, the number of model parameters grows exponentially with the model order or the number of parents, respectively. For parameter estimation, especially based on limited data, this often results in a sharp transition from under-fitted to over-fitted models. To circumvent this problem, variable order Markov models, originally introduced by Rissanen et al.[10] for data compression and later studied by Ron et al.[18] and Buhlmann et al.[19], were applied to various problems in bioinformatics[20,21,12,9]. Intuitively, the idea is to shorten the context at each position in those cases where the full context of length M does not contain "stronger" statistical dependencies than a shortened one. Hence, the fixed order M of a Markov model becomes a function of the context, i.e., $M_l(x_{l-M}, \ldots, x_{l-1})$, and can be different for each position in general. The joint probability distribution of the resulting inhomogeneous variable order Markov model with initial order M (iVOMM(M)) is derived from (2) as:

$$P(x_1, \ldots, x_L) = P_1(x_1) \prod_{l=2}^{L} P_l(x_l | x_{l-M_l(x_{l-M}, \ldots, x_{l-1})}, \ldots, x_{l-1}). \tag{5}$$

The variable order idea initially applied to Markov models can be applied to BNs as well[9]. For this class of models, called variable order Bayesian networks (VOBNs) or variable order Bayesian trees (VOBTs), the context $x_{l-M}, \ldots, x_{l-1}$ is substituted by $\vec{x}_{Pa(l)}$, where the parents $Pa(l)$ have to be ordered in contrast to ordinary BNs. The joint distribution of VOBNs is derived from (4) analogously to iVOMMs.

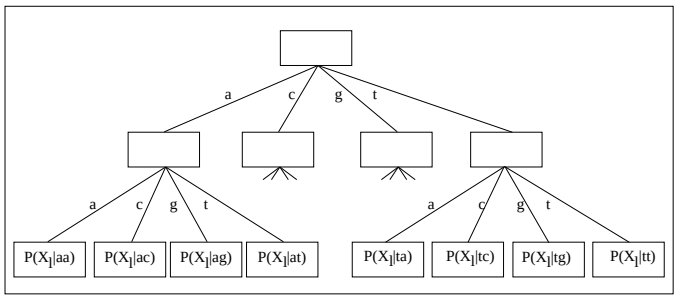


Fig. 1. Full context tree for a Markov model of model order 2

To formalize how to determine M_l, we first describe context trees as a convenient way to represent the set of conditional probabilities at each position for a Markov model or BN. A full context tree (Fig. 1) is a rooted tree of degree d. Each edge is labeled with one nucleotide, where the edges to sibling nodes are required to be pairwise distinct. In the full context tree for position l of a Markov model, each leave has depth M and represents the conditional probabilities $P_l(x_l|x_{l-M},\ldots,x_{l-1}), x_l \in \Sigma$, where $x_{l-M},\ldots,x_{l-1}$ are the concatenated labels on the unique path from that leave to the root.

The conditional probabilities of variable order models may be represented by shortened or pruned context trees (Fig. 2). An arbitrary leaf at depth t with labels $x_{l-t},\ldots,x_{l-1}$ at the path to the root represents the conditional probabilities $P_l(x_l|x_{l-t},\ldots,x_{l-1}), x_l \in \Sigma$. If an inner node does not have the full degree d, because some but not all of its children have been pruned, it represents the probabilities of the contexts for the pruned children:

$$\hat{P}_l\ (x_l|x_{l-t-1} \in R_l(x_{l-t},\ldots,x_{l-1}), x_{l-t},\ldots,x_{l-1}) = \\ \beta \sum_{x_{l-t-1}\in R_l(x_{l-t},\ldots,x_{l-1})} \hat{P}_l(x_{l-t-1}|x_{l-t},\ldots,x_{l-1})\ \hat{P}_l(x_l|x_{l-t-1},\ldots,x_{l-1}) \quad (6)$$

The set $R_l(x_{l-t},\ldots,x_{l-1})$ denotes all nucleotides that have been pruned from the context tree at the inner node considered. In the example of Fig. 2, these are $\{c,g\}$ for the single inner node with degree 2. The normalizing constant β enforces the conditional probabilities represented at these nodes to sum up to 1 for $x_l \in \Sigma$.

The function M_l is determined by a bottom-up traversal of the full context tree, where we initialize $M_l(x_{l-M},\ldots,x_{l-1}) = M$ for all contexts. At each leaf of the tree, the Kullback-Leibler divergence[22] between the conditional probabilities represented at this leaf and at its parent node is computed as a measure of statistical significance of the last symbol of the context at the leaf. If this divergence is smaller than a given threshold, the last symbol is considered unimportant, the leaf is pruned from the context tree, and we set $M_l(x_{l-M},\ldots,x_{l-1}) = t-1$. More formally we have

$$\Delta_l^{\mathrm{KL}}(x_{l-t},\ldots,x_{l-1}) = \sum_{x\in\Sigma} \hat{P}_l(x|x_{l-t-1},\ldots,x_{l-1}) \log_2 \left(\frac{\hat{P}_l(x|x_{l-t-1},\ldots,x_{l-1})}{\hat{P}_l(x|x_{l-t},\ldots,x_{l-1})} \right), \quad (7)$$

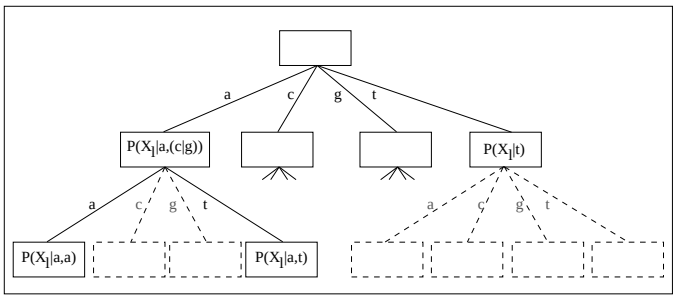


Fig. 2. Example for a pruned context tree with initial order 2. Pruned nodes and edges are depicted with dashed lines.

and a leaf is pruned iff

$$\Delta_l^{\mathrm{KL}}(x_{l-t},\ldots,x_{l-1}) < \begin{cases} c & \text{for a model of regulatory elements} \\ c \cdot \frac{d^{t+1}}{n_l(x_{l-t-1},\ldots,x_{l-1})} & \text{for background models.} \end{cases} \quad (8)$$

This procedure is closely related to the log-loss scores used to derive the ideal compression rate in the field of data compression. The real-valued parameter c is called pruning constant. It controls to which extent a difference between the conditional likelihoods is considered important. For $c = 0$, any extension of the context is considered important, and the resulting VOMM becomes a traditional Markov model of order M. The additional term for background models increases the pruning constant as the number of examples for the corresponding context in the training set decreases. This encourages pruning of these leaves and takes into account the inaccuracies of parameter estimates for small samples. If during traversal all children of an inner node are pruned, it becomes a leaf itself, and the same procedure applies recursively. The recursion terminates if no leaves satisfy (8). Subsequently the conditional probabilities represented at inner nodes with pruned children, i.e., degree less than d, are estimated according to (6). More details are given in[19,9].

For a homogeneous VOMM (hVOMM(M)), the pruning is carried out jointly for all positions computing one single pruned context tree.

3. Case study

In this section we present a case study in which we evaluate the accuracy of the recognition of cis-regulatory elements of fixed order models and variable order models for six different eukaryotic transcription factors.

3.1. *Data*

We analyze six sets of binding sites for mammalian transcription factor families: AP-1, CEBP, GATA, NF-1, SP-1 and thyroid hormone receptor-like factors ("Thyroid"). These sets were built using the TRANSFAC® database (rel. 8.1, 2004)

and comprise experimentally confirmed binding sites for transcription factors collected from scientific literature. For building these sets, we choose those collections containing more than 100 binding sites. The six sets represents three out of four major superfamilies of eukaryotic transcription factors: AP-1 and CEBP belong to the "Basic Domains" factors; NF-1 belongs to "beta-Scaffold Factors with Minor Groove Contacts"; GATA, SP-1, and "Thyroid" belong to the factors with "Zinc-coordinating DNA- binding domains." Each set contains binding sites for one definite family of transcription factors (including factors of different mammalian species) according to the transcription factor classification[23], corresponding to the third family level in the factor hierarchy. All of these sets consist of transcription factor binding sites from various vertebrate species, with the majority of binding sites from human, mouse, and rat.

The background set was built by extracting sequences of second exons of human genes. Second exons were chosen to minimize the chance of having unknown transcription factor binding sites in the background sequences, since there is no recorded evidence on known transcription factor binding sites in any second exon of any gene. This results in six foreground sets containing 112 AP-1 binding sites, 149 CEBP binding sites, 110 GATA binding sites, 96 NF-1 binding sites, 257 SP-1 binding sites, and 127 Thyroid binding sites, respectively, and one background set containing 267 second exons with a total length of 68,141 bp.

3.2. *Stratified holdout sampling*

For the evaluation of the performance of the classifiers we use the following stratified holdout sampling procedure.

(1) Partition the foreground and background set randomly into a training set containing 90% and a test set containing the remaining 10% of the sequences.
(2) Train the foreground and background model on their training sets.
(3) Compute the likelihood ratio of each of the overlapping L-mers of the sequences of the background test set and define the threshold T such that the specificity reaches 99.9%. This ensures that the classifier yields at most one false negative prediction per kb.
(4) Compute the likelihood ratio of each of the sequences of the foreground test set and determine the sensitivity using T.

Repeat these four steps 10^4 times, and record the mean sensitivity and its standard error. The standard error of the mean sensitivity is at most 0.15% for all of the classifiers and all of the datasets studied below.

3.3. *Fixed order models*

In this subsection we study the sensitivity of classifiers based on fixed order models for each of the six transcription factors. We choose an iMM(0), an iMM(1), and a

BT as foreground model, and an hMM(M) for M ranging from 0 to 5 as background model, yielding 3×6 different model combinations. For all of the studies presented in this paper, we set the equivalent sample size $\epsilon = 16$ for any foreground model, and $\epsilon = 4096$ for any background model. This corresponds to a pseudo count of 1 for each leaf in the initially full context trees.

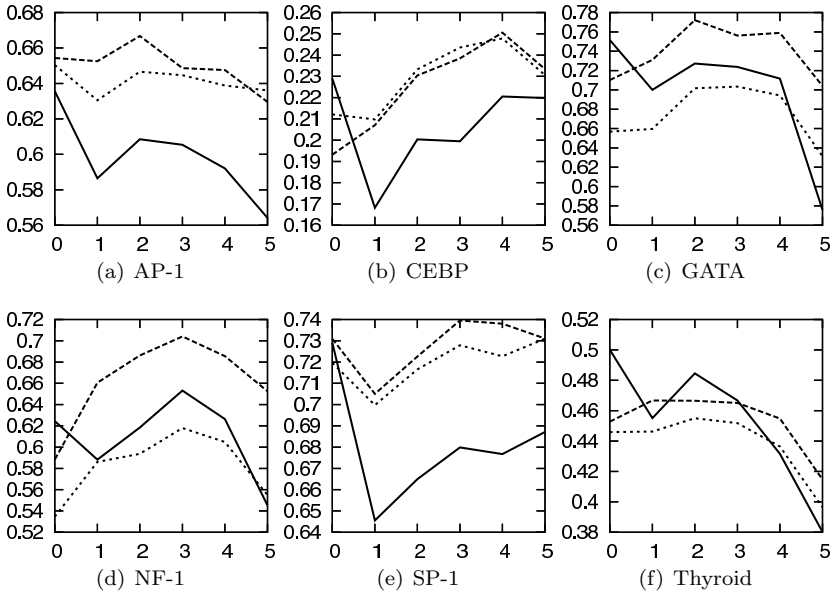


Fig. 3. Mean sensitivity versus M of the hMM(M) for three foreground models applied to the six TFBSs (solid: iMM(0), dashed: iMM(1), dotted: BT).

Fig. 3(a) shows the mean sensitivity of each of the 18 classifiers for the transcription factor AP-1. We find that a traditionally used iMM(0) combined with its optimal background model (an hMM(0) in this case) yields a mean sensitivity of 63.6%, an MM(1) yields 66.7%, and a BT yields 65.0%. This indicates that there are statistical dependencies among nucleotides in the binding sites of AP-1, which can be used to improve their recognition with the statistical models investigated. Surprisingly, the sensitivity of the BT is lower than that of the iMM(1).

One possible explanation is that the statistical dependencies among neighboring nucleotides are stronger than those among non-neighboring nucleotides in the binding sites of AP-1. Another possible explanation is an overfitting of the BT given the limited amount of training data and the additional degrees of freedoms compared to an iMM(1). We observe similar patterns for CEBP and SP-1 (Fig. 3).

Fig. 3(d) gives the results for the transcription factor NF-1. We find that the sensitivity increases by approximately 5% for most of the studied background models when using an iMM(1) instead of an iMM(0).

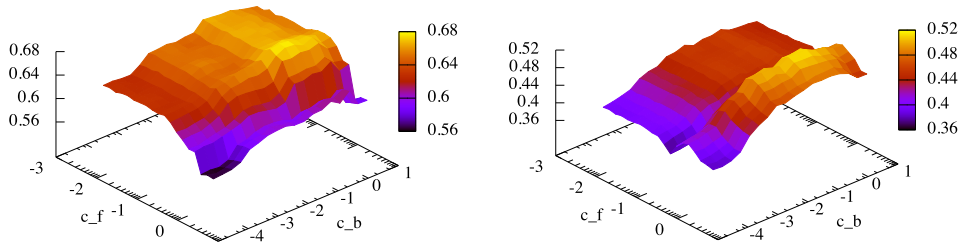
In contrast to AP-1, CEBP, and SP-1, we find that for NF-1 the sensitivity of a

BT is – for most background models – even lower than that of an iMM(0). Again, this is may be due to weak or strongly varying statistical dependencies among non-neighboring nucleotides. Neglecting them improves the recognition of NF-1 binding sites significantly, probably by avoiding overfitting effects. We find a similar pattern for GATA (Fig. 3(c)).

For the transcription factor Thyroid (Fig. 3(f)), the iMM(0) outperforms both the iMM(1) and the BT, which is in contrast to the other five factors. As the amount of data available for Thyroid is comparable to that of the other data sets (Table 1), this may indicate that statistical dependencies between pairs of nucleotides are too weak or too diverse to be of value for the recognition of Thyroid binding sites.

3.4. *Variable order models*

In this subsection we study the sensitivity of classifiers based on variable order models. We choose an iVOMM(1) and a VOBT as foreground model, and an hVOMM(5) as background model. For both model combinations we vary the pruning constants c_{f} (foreground model) from 2^{-8} to 2 and c_{b} (background model) from 2^{-14} to 2^{3}.



(a) (b)

Fig. 4. Mean sensitivity versus foreground pruning constant c_{f} and background pruning constant c_{b} for (a) AP-1 using iVOMM(1)/hVOMM(5) and (b) Thyroid using VOBT/hVOMM(5) classifiers. The axes are marked with $\log_{10}(c_f)$ and $\log_{10}(c_b)$ respectively.

The mean sensitivity of the iVOMM(1)/hVOMM(5) classifier for the transcription factor AP-1 is shown in Fig. 4(a). We find that the sensitivity varies relatively smoothly with c_{f} and c_{b}, and that the highest sensitivity of 67.9% is obtained for $c_{\mathrm{f}} = 2^{-2.5}$ and $c_{\mathrm{b}} = 2^{-2}$. Hence, the maximum is not located in the vicinity of the four corners, where the iVOMM(1)/hVOMM(5) classifier reduces to a fixed order classifier. A comparison of Fig. 3(a) and Fig. 4(a) shows that the iVOMM(1)/hVOMM(5) classifier increases the sensitivity by 1.2% over the optimal fixed order iMM/hMM classifier. This indicates that VOMMs make a more economical use of their model parameters compared to fixed order MMs, resulting in an improved recognition of AP-1 binding sites.

Fig. 4(b) shows the mean sensitivity of the VOBT/hVOMM(5) classifier for the transcription factor Thyroid. Again, the sensitivity varies relatively smoothly with

c_f and c_b, and the highest sensitivity of 52.0% is obtained for $c_\mathrm{f} = 2^{-1}$ and $c_\mathrm{b} = 2^{-2}$, which again yields variable order models that are notably different from their fixed order limits. Comparing Fig. 3(f) and Fig. 4(b), we find that the VOBT/hVOMM(5) classifier can increase the sensitivity by 2.0% over the optimal fixed order classifier. The plots for all factors are available at *http://www2.informatik.uni-halle.de/agprbio/AG/Publication/OnlineMaterial/jbcb07.html*.

3.5. *Discussion*

Even though all six transcription factor families correspond to a similar level of factor hierarchy, they bear quite different degrees of functional heterogeneity as well as structural heterogeneity of the factor heterodimers binding to the corresponding cis-regulatory elements. Especially heterogeneous is the family of thyroid hormone receptor-like factors, which belongs to the superfamily of nuclear hormone receptors. Nuclear hormone receptors are ligand-activated transcription factors that belong to a superfamily consisting of over 150 different members, reviewed in[24,25]. DNA binding sites of nuclear hormone receptors are typically composed of two 6-bp half-sites that may be arranged as direct, inverted, or everted repeats[25,26]. The family of thyroid hormone receptor-like factors show a large variety of modes of DNA binding. Their DNA binding sites are generally direct or inverted repeats with variable spacing between 0 and 5 bp[27,28]. The repeated structure of the binding sites in the Thyroid set may be the reason for the VOBT to slightly outperform VOMMs for this set. The presence of direct and inverted repeats may explain the statistical dependencies between non-neighboring nucleotides. It is interesting to observe that in the case of AP-1 and CEBP binding sites, which are also characterized by a short inverted palindromic structure, the difference between maximal sensitivity obtained by VOMM and VOBN models is minimal, which indicates the presence of dependencies between non-neighboring nucleotides.

The maximal sensitivity level achieved by any of the models is very different for different sets of binding sites. It does not correlate with the membership of the factors to specific DNA binding domains or to any functional groups. CEBP sites yield the lowest sensitivity level (25%). This is not a surprise, since factors of the CEBP family exhibit an extremely relaxed DNA-binding specificity[29], which is still poorly understood. One possible explanation is that CEBP factors (CCAAT/enhancer binding proteins) are ubiquitously involved in regulating a huge number of promoters under many different conditions. Their binding to a large variety of promoters is relatively position specific, which makes the "CCAAT-box" statistically profound and even considered as a part of the basic promoter structure. Hence, the loose DNA-binding specificity of these factors can be the result of the requirement of these factors to be able to bind to the CCAAT-box under many different promoter contexts. One possibility to improve the computational recognition of CEBP binding sites is to build separate models for different specific subgroups of CEBP binding sites, which was explored for the case of the PWMs[30] and can be extended for the

Table 1. Percentage of mean sensitivity averaged over 10^4-fold replicated stratified-holdout experiments for different model combinations and the six data sets. For row 2 (MM) we vary the order M of the fixed order foreground model from 0 to 1. For both rows 2 and 3, we vary M of the fixed order background MM from 0 to 5, and we record the maximum sensitivities in columns 3 through 8. For rows 4 and 5, we use variable order foreground models of initial order 1 and variable order background models of initial order 5. The maximum sensitivities achieved for varying pruning constants c_f and c_b are given in columns 3 through 8.

TFBS model	**background model**	**AP-1**	**CEBP**	**GATA**	**NF-1**	**SP-1**	**Thyroid**
MM	MM	66.7	25.1	77.2	70.4	74.0	50.0
BT	MM	65.0	24.8	70.3	61.8	73.1	45.5
VOMM	VOMM	**67.9**	**25.9**	**79.0**	**71.0**	**75.1**	51.8
VOBN	VOMM	67.5	25.7	78.1	69.2	73.5	**52.0**
Size of dataset		112	149	110	96	257	127

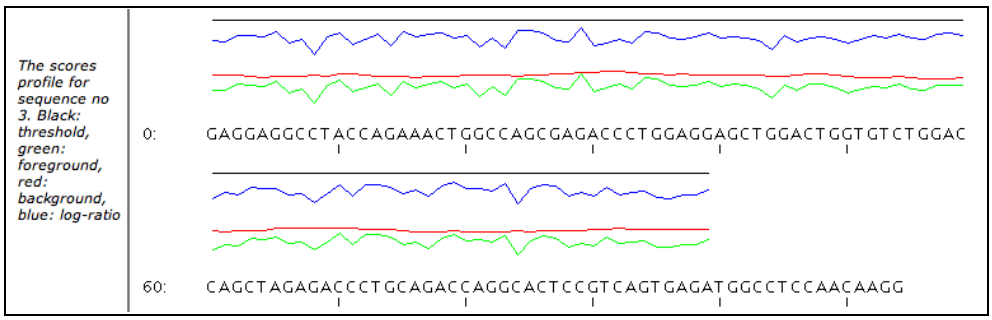
more complex models considered in this paper. The relatively low sensitivity level achieved for the Thyroid set (50%) can be explained by a high heterogeneity of the binding sites in the set with variable spacing between the repeats of "half sites". This demonstrates some limits of the approach considered in this paper, where the "inhomogeneous" models are required to be strictly position specific, and therefore incapable of capturing the variability of spacers between different subparts of the sites. A generalization of this approach can be considered in the future, where inhomogeneous models may be partially homogeneous in some sub-regions of the sites. First attempts towards building such models with sub-regional homogeneity were made for splice sites[31] and some transcription factor binding sites[32,33].

Comparing the performance achieved for the six different sets, we not only find a wide range of sensitivities. Also the optimal model combination (including the choice of the optimal model orders for fixed order models and the choice of the optimal pruning constants for the variable order models) varies strongly from transcription factor to transcription factor. Hence, we recommend a systematic analysis of the performance of different model and parameter combinations based on problem-specific training datasets in advance to a genome-wide analysis of cis-regulatory elements. In order to ease such systematic analyses and genomic scans of cis-regulatory elements, we provide a web server to the public community, which we describe in the next section.

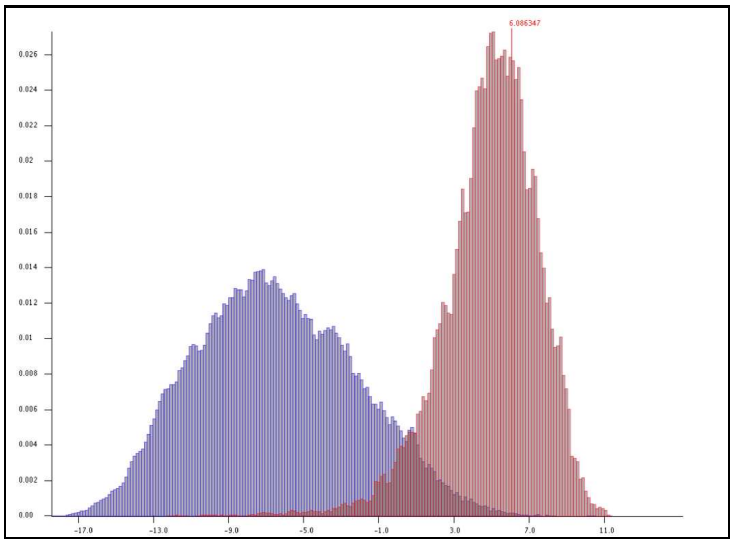
4. VOMBAT server

The VOMBAT web-server[34] implements an easy-to-use web-interface and allows users to apply different combinations of Markov models and Bayesian trees to their data. This includes models with variable and fixed orders as well as homogeneous and inhomogeneous Markov models. Tasks available are learning statistical models from data, predicting putative binding sites, and stratified holdout experiments for different model combinations.

To train a model the user is queried for an input file of training sequences and



(a)



(b)

Fig. 5. (a) Example for a profile of log-likelihood ratios predicting putative TFBSs;
(b) Histogram of the log-likelihood ratios for a TFBS (left) and the background (right)

parameters to specify the type and order of the model. The output is an XML-representation of the model including the parameters learnt. It can be downloaded and used subsequently for prediction and for optional graphical representation.

The trained foreground and background models can be used to predict putative cis-regulatory elements in a new set of input sequences. The output of the prediction is an HTML file containing a textual description of the putative cis-regulatory element. This includes the motif position, its sequence, the likelihood obtained by the two models, and the corresponding log-likelihood ratio. Additionally, profiles of the log-likelihoods and the log-likelihood ratio are plotted for each sequence of the input file (Fig. 5(a)).

For advanced users VOMBAT provides functions for stratified holdout sampling analyses of different model combinations. This allows the user to find the optimal

model combination for his or her classification problem and data sets. The output of the stratified holdout sampling contains the sensitivity obtained for a selected fixed specificity, the maximum correlation coefficient over all possible classification thresholds, and a histogram of the log-likelihood ratios for the classified foreground and background samples (Fig. 5(b)).

The VOMBAT server is based on a three-tier architecture, where the presentation layer, the management layer, and the execution framework are logically separated and physically located on different servers. The web front-end of VOMBAT is based on standard technologies like JavaServer Faces and Servlets. The user-supplied parameters and input files are stored in a mySQL-database, which is queried by the execution framework running on a Linux cluster with 150 processors. VOMBAT is available free of charge at http://bic-gh.de/vombat. The web pages also include use cases and a detailed manual.

5. Summary

We study the recognition of eukaryotic cis-regulatory elements by fixed and variable order Markov models and Bayesian trees. Compared to fixed order models, variable order models improve the recognition of binding sites for all transcription factors studied. The combination of a VOMM(1) for the foreground and a VOMM(5) for the background yields the optimal classifier for AP-1, CEBP, GATA, NF-1, and SP-1 among the model combinations studied. In contrast, for Thyroid the additional freedom to exploit statistical dependencies also among non-neighboring nucleotides when using a BT or VOBT increases the performance compared to MMs or VOMMs. For all of the studied examples, the sensitivity of the variable order classifiers varies relatively smoothly with c_{f} and c_{b}, which allows to robustly determine the optimal pruning constants for each dataset. However, the values of the optimal pruning constants, the choice of the optimal combination of models, and the resulting sensitivities vary strongly from transcription factor to transcription factor. Hence, we recommend a systematic evaluation of different model combinations based on problem-specific data sets before starting genomic scans of cis-regulatory elements. To allow such systematic evaluations, the training of fixed and variable order MMs and BTs based on user-supplied data sets, and genomic scans of cis-regulatory elements based on the trained models, we provide the VOMBAT server to the public community.

Acknowledgements

We thank the German Ministry of Education and Research (BMBF Grant No. 0312706A/D) for financial support.

References

1. Fickett J, Hatzigeorgiou A, Eukaryotic promoter recognition, *Genome Research* **7**(9):861–878, 1997.

2. Salzberg S, A method for identifying splice sites and translational start sites in eukaryotic mrna, *CABIOS* **13**:365–376, 1997.
3. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E, MATCH^{TM}: a tool for searching transcription factor binding sites in DNA sequences, *Nucleic Acids Research* **31**(13):3576–3579, 2003.
4. Staden R, Computer methods to locate signals in nucleic acid sequences, *Nucleic Acids Research* **12**:505–519, 1984.
5. Benos PV, Lapedes AS, Fields DS, Stormo GD, Samie: Statistical algorithm for modeling interaction energies., *Pacific Symposium on Biocomputing*, pp. 115–126, 2001.
6. Bulyk ML, Johnson PLF, Church GM, Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors, *Nucleic Acids Research* **30**(5):1255–1261, 2002.
7. Barash Y, Elidan G, Friedman N, Kaplan T (eds.), *Modeling Dependencies in Protein-DNA Binding Sites*, Proc. Seventh Annual Inter. Conf. on Computational Molecular Biology (RECOMB), Vol. 7, ACM, ACM, New York, 2003.
8. Zhang MQ, Marr TG, A weight array method for splicing signal analysis, *Computational Applications for Bioscience* **9**:499–509, 1993.
9. Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, Posch S, Grosse I, Identification of transcription factor binding sites with variable-order bayesian networks, *Bioinformatics* **21**:2657–2666, 2005.
10. Rissanen J, A universal data compression system, *IEEE Transactions on Information Theory* **29**(5):656–664, 1983.
11. Heckerman D, Geiger D, Chickering DM, Learning bayesian networks: The combination of knowledge and statistical data, *Machine Learning* **20**:197–243, 1995.
12. Zhao X, Huang H, Speed TP, Finding short dna motifs using permuted markov models, *Proceedings of the 8th Annual International Conference on Computational Molecular Biology*, ACM, San Diego, California, USA, pp. 68–75, 2004.
13. Pearl J, *Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference*, Morgan Kaufmann, California, 1988.
14. Buntine W, Theory refinement on bayesian networks, *Proceedings of Seventh Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, Los Angeles, CA, pp. 52–60, 1991.
15. Cai D, Delcher A, Kao B, Kasif S, Modeling splice sites with bayes networks, *Bioinformatics* **16**:152–158, 2000.
16. Castelo R, Guigo R, Splice site identification by idlbns, *Bioinformatics* **20**:i69–i76, 2004.
17. Chow CK, Liu CN, Approximating discrete probability distributions with dependence trees, *IEEE Transactions on Information Theory* **14**:462–467, 1968.
18. Ron D, Singer Y, Tishby N, The power of amnesia: learning probabilistic automata with variable memory length, *Machine Learning* **25**:117–149, 1996.
19. Buhlmann P, Wyner AJ, Variable length markov chains, *Ann Statist* **27**(2):480–513, 1999.
20. Bejerano G, Yona G, Variations on probabilistic suffix trees: statistical modeling and prediction of protein families, *Bioinformatics* **17**:23–43, 2001.
21. Orlov YL, Filippov VP, Potapov VN, Kolchanov NA, Construction of stochastic context trees for genetic texts, *In Silico Biology* **2**(3):233–247, 2002.
22. Kullback S, Leibler RA, On information and sufficiency, *Ann Math Statist* **22**:79–86, 1951.
23. Wingender E, Classification scheme of eukaryotic transcription factors, *Mol Biol Engl Tr* **31**:483–497, 1997.

24. Mangelsdorf DJ, Thummel C, Beato M, Herrlich P, Schutz G, Umesono K, Blumberg B, Kastner P, Mark M, Chambon P, Evans RM, The nuclear receptor superfamily: The second decade, *Cell* **83**:835–839, 1995.
25. Mangelsdorf DJ, Evans RM, The RXR heterodimers and orphan receptors, *Cell* **83**:841–850, 1995.
26. Glass C, Differential recognition of target genes by nuclear receptor monomers, dimers, and heterod imers, *Endocr Rev* **15**(3):391–407, 1994.
27. Umesono K, Murakami KK, Thompson CC, Evans RM, Direct repeats as selective response elements for the thyroid hormone, retinoic acid, and vitamin D3 receptors, *Cell* **65**:1255–1266, 1991.
28. Mader S, Chen J, Chen Z, White J, Chambon P, Gronemeyer H, The patterns of binding of RAR, RXR and TR homo- and heterodimers to direct repeats are dictated by the binding specificites of the dna binding domains, *EMBO J* **12**:5029–5041, 1993.
29. Costa RH, Grayson DR, Xanthopoulos KG, Darnell JE, A Liver-Specific DNA-Binding Protein Recognizes Multiple Nucleotide Sites in Regulatory Regions of Transthyretin, alpha 1-antitrypsin, Albumin, and Simian Virus 40 Genes, *PNAS* **85**(11):3840–3844, 1988.
30. Shelest E, Kel AE, Gößling E, Wingender E, Prediction of potential C/EBP/NF-kappaB composite elements using matrix-based search methods., *In Silico Biology* **3**:7, 2003.
31. Kel AE, Ponomarenko MP, Likhachev EA, Orlov YL, Ischenko IV, Milanesi L, Kolchanov NA, SITEVIDEO: a computer system for functional site analysis and recognition. investigation of the human splice sites., *Computer Applications in the Biosciences* **9**(6):617–627, 1993.
32. Kel A, Reymann S, Matys V, Nettesheim P, Wingender E, Borlak J, A novel computational approach for the prediction of networked transcription factors of Aryl hydrocarbon-receptor-regulated genes, *Mol Pharmacol* **66**(6):1557–1572, 2004.
33. Kel AE, Kel-Margoulis OV, Farnham PJ, Bartley SM, Wingender E, Zhang MQ, Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors, *Journal of Molecular Biology* **309**(1):99–120, 2001.
34. Grau J, Ben-Gal I, Posch S, Grosse I, VOMBAT: prediction of transcription factor binding sites using variable order Bayesian trees, *Nucl Acids Res* **34**(suppl_2):W529–533, 2006.

Stefan Posch received his Diploma and Doctoral degrees in Computer Science from the University of Erlangen-Nuremberg, Germany in 1985 and 1989. From 1990 to 1991, he held a postdoctoral position at the International Computer Science Institute, Berkeley, California. In 1991, he joined the Applied Computer Science Group at the University of Bielefeld, Germany. Since 1999, he is Professor at the Institute of Computer Science at the Martin Luther University of Halle-Wittenberg. Research interests include bioinformatics and computer vision and scene understanding.

Jan Grau received his M.Sc. degree in Bioinformatics from Martin-Luther-University Halle (MLU), Germany in 2006. Currently he is working as a scientific assistant at the Institute of Computer Science of MLU in the Bioinformatics group.

André Gohr received his M.Sc. degree in Bioinformatics from Martin-Luther-University Halle (MLU), Germany in 2006. Currently he is working as a scientific assistant at the Institute of Computer Science of MLU in the Bioinformatics group.

Irad Ben-Gal is a Senior Lecturer and the Head of the Division of Industrial Engineering at Tel-Aviv University, Israel. Irad holds a B.Sc. (1992) degree from Tel-Aviv University, M.Sc. (1996) and Ph.D. (1998) degrees from Boston University. Irad has worked for several years in various industrial organizations. His research interests include Statistical methods for control and analysis of stochastic processes; Applications of Information Theory to industrial problems; Design of Experiment; and Machine Learning.

Alexander E. Kel has received both Master of Science and Ph.D. degrees both in genetics and mathematical biology from the Institute of Cytology and Genetics, Novosibirsk, Russia in 1985 and 1991 respectively. Dr. Kel was a Group Leader and Principal Investigator at the Institute of Cytology and Genetics, Novosibirsk, Russia. He joined BIOBASE in 2000 and currently holds the position of Senior Vice President Research and Development. His research interests include gene regulation, molecular evolution, and structural biology, database development, application of machine learning techniques.

Ivo Grosse is Assistant Professor of Bioinformatics at the Institute of Computer Science at Martin Luther University Halle-Wittenberg and head of the Plant Data Warehouse Group at the Leibniz Institute of Plant Genetics and Crop Plant Research at Gatersleben. He holds a Diploma degree from Humboldt University Berlin (1995) and a Ph.D. degree from Boston University (1999). Ivo worked on gene finding and promoter recognition at the Institute of Molecular Biology and Biochemistry at Free University Berlin and at Cold Spring Harbor Laboratory. His research interests include computational biology, machine learning, and data integration.