

Identification of Transcription Factor Binding Sites with Variable-order Bayesian Networks

Ben-Gal I.^{§†}, Shani A.[†], Gohr A.^{+*}, Grau J.^{+*}, Arviv S.[†], Shmilovici A.[^], Posch S.⁺, and Grosse I.^{*}

[†]Department of Industrial Engineering, Tel-Aviv University,
Tel-Aviv, 69978, Israel

[^]Department of Information Systems Engineering, Ben-Gurion University
P.O.Box 653, Beer-Sheva, 84105, Israel

^{*}Institute of Plant Genetics and Crop Plant Research, 06466 Gatersleben, Germany

⁺Institute of Computer Science, University Halle, 06099 Halle (Saale), Germany

ABSTRACT

Motivation: We propose a new class of variable order Bayesian network (VOBN) models for the identification of transcription factor binding sites (TFBSs). The proposed models generalize the widely-used position weight matrix (PWM) models, Markov models and Bayesian network (BN) models. In contrast to these models, where for each position a fixed subset of the remaining positions is used to model dependencies, in VOBN models these subsets may vary based on the specific nucleotides observed, which are called the context.

This flexibility turns out to be of advantage for the classification and analysis of TFBSs, as statistical dependencies between nucleotides in different TFBS positions (not necessarily adjacent) may be taken into account efficiently – in a position-specific and context-specific manner.

Results: We apply the VOBN model to a set of 238 experimentally verified sigma-70 binding sites in E.coli. We find that the VOBN model can distinguish those 238 sites from a set of 472 intergenic ‘non-promoter’ sequences with higher accuracy than fixed-order Markov models or Bayesian trees (BT). We use a replicated stratified-holdout experiments having a fixed true-negative rate of 99.9%. We find that for a foreground inhomogeneous VOBN model of order 1 and a background homogeneous variable-order Markov (VOM) model of order 5 the obtained mean true-positive (TP) rate is 47.56%. In comparison, the best TP rate for the conventional models is 44.39%, obtained from a foreground PWM model and a background 2nd-order Markov model. As the standard deviation of the estimated TP rate is ~ 0.01%, this improvement is highly significant.

[§] Corresponding author, email: bengal@eng.tau.ac.il.

Availability: All datasets are available upon request from the authors at bengal@eng.tau.ac.il. A web server for utilizing VOBN and VOM models is available at <http://www.eng.tau.ac.il/~bengal/>.

1. INTRODUCTION

One problem in the analysis of DNA sequences is the identification of transcription factor binding sites (TFBSs). The importance of this problem stems from the fact that the combinatorial presence and absence of TFBSs is – to a large degree – responsible for the complexity of gene regulation in virtually every living organism (Wingender *et al.*, 2000, Wingender *et al.*, 2001, Pickert *et al.*, 1998, Kel-Margoulis *et al.*, 2003). The interest in TFBS analysis has dramatically grown with the arrival of microarray gene-expression data which heralds an important advance in the identification of co-expressed genes (Fickett and Hatzigeorgiou, 1997, Chu *et al.*, 1998, Spellman *et al.*, 1998, Thijs *et al.*, 2001, Ohler and Niemann, 2001, Hanisch *et al.*, 2002). While experimental techniques, such as footprinting experiments or chromatin immunoprecipitation experiments, allow the identification of TFBSs, these experiments are expensive and time consuming, and require the availability of a well-equipped lab together with professionally trained personnel. Today, the availability of computer hardware installed with cheap or often free bioinformatics software has enabled bench scientists in almost every research lab to run programs such as MatchTM (Kel *et al.*, 2003) or HMMgene and HMMPro (e.g., see Baldi and Brunak, 2001) to predict the location of TFBSs. While bioinformatics identifications are probabilistic in nature and cannot achieve the accuracy of "wet" experimental data, their value lies in the low-cost and high-speed with which these identifications can be obtained. To expand on the above-mentioned example, MatchTM, a publicly available and free-of-charge software, can scan several megabytes of DNA for the presence of putative TFBSs within minutes. Hence, TFBS identification programs are extremely popular, despite their limited accuracy.

Most TFBS classification algorithms compute some numerical score reflecting the degree to which a given sequence site matches a given motif. In many TFBS classification algorithms the underlying scoring model is either a fixed-order Markov model or simply a position weight matrix (PWM) model with no context dependencies at all (e.g., MatchTM). The latter can be regarded as a fixed zeroth-order model. In the following, we note on the main differences among the proposed

Variable-Order Bayesian Network (VOBN) model, the PWM model, the fixed order Markov model and the Bayesian Network (BN) model.

Presumably, the most common context-independent model is the PWM (or the Position Specific Score Matrix – PSSM). The PWM model has been successfully applied not only to the problem of TFBS classification but also to other diverse problems in DNA and protein sequence analysis (e.g., see Salzberg, 1997). Although other scoring models have been able to improve the accuracy of the PWM model in certain cases, they are not as prevalent as the simple PWM model in the classification of TFBS (Fickett and Hatzigeorgiou, 1997). Hence, many TFBS classification algorithms that are developed today still rely on the PWM models that obtain a relatively good performance, as will be seen below.

The basic assumption of the PWM model is that the basepairs at different positions are statistically independent, hence, the joint probability of finding a multiple-position site factorizes into the product of single-position probabilities (e.g., Djordjevic *et al.* 2003, Ewens and Grant, 2001, Stormo and Fields, 1998). As indicated in Barash *et al.* (2003) it is an open question whether the “strong” independence assumption of the PWM model is reasonable in view of recent results that point to the dependence between positions (e.g., see Benos *et al.*, 2001, Bulyk *et al.*, 2002). This dependence is used by fixed-order models, such as Markov models, Hidden Markov models (HMMs) and Interpolated Markov models to detect motifs in upstream regions of co-regulated genes (Liu *et al.*, 1995, Ohler *et al.*, 1999, Hughes *et al.*, 2000, Thijs *et al.*, 2001, Ohler and Niemann, 2001, Liu *et al.*, 2001, Salzberg *et al.*, 1998, Salzberg *et al.*, 1999). Although it is well known (and intuitively clear) that the PWM assumption is violated in almost every TFBS studied to date, this violation, nevertheless, does not prevent the PWM model from being a leading model in the classification of putative TFBSs. In fact, the PWM model, which is based on the (unsupported) independence assumption, is often found to outperform fixed-order Markov models of higher order that are based on the (reasonable and supported) dependence assumption.

In this paper, we show that the above-mentioned contradiction is due to the unbalanced comparison between the PWM model and a high-order Markov model with respect to their number of parameters. Although the PWM independence assumption is unsound and results in an under-fitted model with a smaller-than-necessary number of parameters, it often outperforms fixed-order Markov models that tend to be over-fitted due to their large dimensionality, given the limited amount of

training data. The solution presented here is based on the development of a variable-order model that, in terms of its order, stands in between these two types of models. We show that the variable order models do not ignore the statistical dependencies between basepairs, yet, they take into account only those dependencies that are found statistically significant.

As an extension to fixed-order models, we suggest to use the inhomogeneous VOBN model. The VOBN model is a generalization of the variable order Markov (VOM) model, which was originally proposed for data-compression (Rissanen, 1983) and later applied in various forms to prediction and identification (Weinberger *et al.*, 1995), statistical process control for finite-state processes (Ben-Gal *et al.*, 2000, Ben-Gal *et al.*, 2003), text clustering (Vert, 2001); modelling of genetic texts, including TFBS and protein coding regions (Ron *et al.*, 1996, Buhlmann and Wyner, 1999, Orlov and Potapov, 2000, Orlov *et al.* 2002, Bilu *et al.* 2002); and modelling of protein families (Bejerano and Yona, 2001). In contrast to fixed-order Markov models, where the order is the same for all positions and for all contexts, in VOM and VOBN models the order may vary for each position based on its contexts. There are three main differences between the above VOM models and the proposed VOBN model. The first is the variations in the construction of the models as seen in section 2. The second is the use of *inhomogeneous* (position-dependent) VOBN models vs. the homogeneous VOM model – a crucial property for the classification of E. Coli TFBSs. The third is the VOBN generalization to contexts from non-adjacent positions in a manner similar to BN models, which are discussed next.

The BN model is a graphical representation of probabilistic dependency knowledge (Pearl, 1988) that was applied to analysis of gene expression data (Friedman *et al.*, 2000), genetic linkage analysis (Fishelson and Geiger, 2002) and identification of TFBSs and other functional DNA regions (Cai *et al.*, 2000, Barash *et al.*, 2003, Castelo and Guigo, 2004). In the BN graph an edge is directed from an influencing position (parent) into an influenced position (child). In contrast to fixed-order Markov models, in BNs it is not assumed that dependencies are necessarily between adjacent positions. The difference between BNs and the proposed VOBN is that, in general, the order of the model in BNs depends only on the size of the parents' subset, while in VOBNs it also depends on the specific nucleotides observed in each parent subset. As a result, the number of parameters that need to be estimated from the data is substantially smaller, yielding a smaller chance for over-fitting of the model to the training dataset. A class of models which is closely related to VOBNs is the context-

specific Bayesian networks (CSBN) (e.g., Boutilier *et al.*, 1996, Friedman and Goldszmidt 1996). The main differences between CSBNs and the proposed VOBN are in the method of encoding and constructing the context-specific dependencies (e.g., complete vs. non-complete trees), parameter estimation and refinement methods, and the manner in which the context-dependencies are integrated in the model-learning phase (e.g, starting with an over-fitted model), as described below.

The proposed VOBN model is a true generalization (rather than a replacement) of PWM, fixed-order Markov and BN models in the sense that these models are special cases of the VOBN model. This means that in cases where the statistical dependencies are insignificant, the VOBN model “automatically” degenerates to the PWM model. If dependencies exist only amongst adjacent positions and the ideal memory length for a position is identical for all basepairs, the VOBN generalizes to a fixed-order Markov model. Similarly, if the ideal memory length for a given position is identical for all basepairs and depends only on the number of parents, the VOBN generalizes to a BN model (see also figure 1).

In the remainder of the paper we study the degree to which the VOBN yields useful generalizations of the PWM and the fixed-order models in the context of TFBS classification.

[Insert figure 1 about here]

2. METHODS

In this section we introduce the homogeneous and inhomogeneous models, which we will apply, respectively, for modelling the DNA background not containing TFBSs and for the TFBSs. We start with a homogeneous zeroth-order Markov model, continue with homogeneous fixed-order Markov models, and end with homogeneous VOM models as generalizations of fixed-order Markov models. We then introduce the VOM tree and outline the VOM construction algorithm. Following, we discuss different inhomogeneous models, starting with the PWM and inhomogeneous VOM models and ending with the proposed VOBN models. Finally, we introduce the used classification rule, the used stratified-holdout procedure, and the performance measure by which we quantify the classification accuracy.

2.1 Homogeneous models

In the following, we adopt some definitions and notations from Buhlmann and Wyner (1999), Ohler *et al.* (1999), and Ben-Gal *et al.* (2000, 2003). Let $x_l^n = x_l, x_{l+1}, \dots, x_{n-1}, x_n$ define a sequence with $n-l+1$ symbols over a finite alphabet X

of cardinality $|X| = d$. In case of the TFBS classification problem $d=4$, $X=\{A,C,G,T\}$, and x_j is the nucleotide at position j , with $1 \leq j \leq N$ in the DNA sequence x_1^N of length N . The likelihood of sequence x_1^N is computed by the multiplication chain rule (e.g., Ohler *et al.*, 1999):

$$P(x_1^N) \equiv \prod_{j=1}^N P(x_j | x_1^{j-1}), \quad (1)$$

where $P(\cdot)$ stands for probability, X_j is the random variable representing the nucleotide at position j with x_j as its realization, X_1^{j-1} is a sequence of random variables for the *context* with x_1^{j-1} as the actually observed context of the symbol x_j containing its preceding $j-1$ symbols, x_1^0 is the empty string.

Traditional models often consider only a small part of the available context (or no context) in order to minimize the number of parameters to be estimated and to avoid over-fitting the training dataset. Many papers suggest a homogeneous zeroth-order Markov model (a Bernoulli model) as a simple background model (e.g., see Liu *et al.*, 1995, Hughes *et al.*, 2000). The likelihood for this model, which we abbreviate by Markov(0), is computed by multiplying the probabilities of the symbols, i.e.,

$$P(x_1^N) = \prod_{j=1}^N P(X_j = x_j), \quad (2)$$

because this model considers no context at all. Using a zeroth-order Homogeneous model implies that $P(\cdot)$ is identical for all j . Other studies indicate that such a model poorly reflects the complex structure of genome sequences (Thijs *et al.*, 2001, Liu *et al.*, 2001) and suggest higher-order models. Accordingly, in L^{th} -order Markov models, denoted here by Markov(L), the likelihood of the sequence depends on the sequence of predecessors of a **fixed** length $L < N$, i.e.,

$$P(x_1^N) = \prod_{j=1}^N P(X_j = x_j | X_{j-L}^{j-1} = x_{j-L}^{j-1}), \quad (3)$$

where $x_{j-L}^{j-1} = x_1^{j-1}$ if $j-L \leq 0$, i.e., the memory length cannot exceed the number of preceding symbols and the subscript $j-L \equiv \max(j-L, 1)$.

As indicated in Buhlmann and Wyner (1999) and Orlov *et al.* (2002), a main problem of fixed-order Markov models is that the number of model parameters grows exponentially with the model order L , resulting in a very sharp and discontinuous transition from under-fitted models (that do not capture enough statistical dependencies in the data) to over-fitted models (that contains a redundant number of

parameters). For example, the number of free parameters in fixed-order Markov models with $d=4$ and $L = 2,3,4,5$ is equal to 63, 255, 1023, and 4095 respectively.

Some approaches to solve the above-mentioned problems look for the optimal L , which maximizes the likelihood of the training dataset, or apply the interpolation of different model orders, as suggested in Salzberg *et al.* (1998), Salzberg *et al.* (1999) and Ohler *et al.* (1999). The difficulty with these approaches is that the model order is averaged or weighted over different sub-sequences in the training set, and thus might be either too short or too long for different symbols in the set. Symbols along the sequence might depend on contexts that are shorter than an averaged L , even if it has a relatively small value. We suggest to allow a variable model order L_j that depends on the preceding symbols to position j , thus, the order of the Markov model becomes a function of the context at each position,

$$P(x_1^N) = \prod_{j=1}^N P\left(X_j = x_j \mid X_{j-L_j}^{j-1} = x_{j-L_j}^{j-1}\right), \quad (4)$$

where the variable order $L_j = L(x_{j-1}, x_{j-2}, \dots)$ is itself a function of preceding symbols. An optimal value for L_j defines the shortest context for which the transition probability of symbol x_j is practically equal to the transition probability of that symbol given the context of maximal order L , i.e.,

$$L_j = \min\left\{\tilde{L} \mid P\left(X_j = x_j \mid X_{j-\tilde{L}}^{j-1} = x_{j-\tilde{L}}^{j-1}\right) = P\left(X_j = x_j \mid X_{j-L}^{j-1} = x_{j-L}^{j-1}\right)\right\}. \quad (5)$$

Note from eq. (3) that for the fixed-order Markov chain $L(x_{j-1}, x_{j-2}, \dots) = L$ for all x_j , whereas, for the suggested variable-order Markov model, $L_j \leq L$, implying that some transition probabilities of the Markov chains can be lumped together (e.g., Buhlmann and Wyner, 1999, Orlov and Potapov, 2000).

2.2. The context tree representation

VOM models, including fixed-order Markov models, can be represented by a tree, which Rissanen (1983) calls *context tree* (called also VOM tree in this paper).

For illustration purposes, let us start with simple examples of homogenous VOM trees that will then be defined more formally. The trees were constructed from the Intergenic Background dataset described in section 3. Figure 2a represents a (degenerated) tree that consist of a single root that is equivalent to a Markov(0) model. The root contains four probabilities for nucleotides A, C, G, and T,

respectively. In this case the model has no memory, and the likelihood is computed by multiplying the probabilities of the nucleotides, as indicated in eq. (2). Figure 2b represents a more developed tree that consists of a single root with four leaves. Each node in the tree contains four parameters – the conditional probabilities of nucleotides – ordered as $P(A/x_{j-1})$, $P(C/x_{j-1})$, $P(G/x_{j-1})$, $P(T/x_{j-1})$, where x_{j-1} is the context corresponding to each of the nodes. The tree is equivalent to a first-order Markov model. The tree root contains the (unconditional) nucleotide probabilities, while the leaves contain 16 conditional probabilities for each nucleotide given a preceding nucleotide. The likelihood is computed by multiplying the transition probabilities of the symbols given the previous symbol, as shown in eq. (3) with $L=1$. For example, $P(TCCGGA) = P(T) \times P(C/T) \times P(C/C) \times P(G/C) \times P(G/G) \times P(A/G) = 0.26 \cdot 0.21 \cdot 0.24 \cdot 0.27 \cdot 0.25 \cdot 0.23$. In a similar manner, a tree that represents a fixed-order Markov(L) model contains d^L leaves. Figure 2c represents a VOM tree that consists of 14 nodes. In this case, the depth of the tree is no longer fixed. All branches of the original tree, which corresponds to a 5th-order Markov model, were pruned by the VOM construction algorithm (see below). Note that the tree branches from the root on top down to the leaves represent the *reversed* contexts. Thus, an extension of a branch by adding a node represents the extension of a context by an *earlier* observed symbol (see Buhlmann and Wyner, 1999, Ben-Gal *et al.*, 2003). For example, the first level node for context A represents a 1st-order Markov model, which is used instead of *the longer contexts for all nucleotides* except for context C. The single node at level three represent a 3rd-order that consist of the context GAT. Although the maximal order allowed in this case is equal to five, all branches are pruned to a lower order. As a result, instead of using a full Markov model of order $L=5$ with $d^{L+1} - 1 = 4,095$ free parameters, the VOM model in this example has only $14 \cdot 3 = 42$ free parameters. The likelihood of a sequence given a VOM for class k depends on the contexts of a varying-order L_j , as seen in eq. (4). Using the above example, $P(TCCGGA) = P(T) \times P(C/T) \times P(C/TC) \times P(G/CC) \times P(G/CG) \times P(A/G) = 0.26 \cdot 0.20 \cdot 0.24 \cdot 0.34 \cdot 0.29 \cdot 0.23$ (The difference in the probabilities in level 1 compared to figure 2b stems from the probability adjustment explained in the meta code below). Thus, from the fourth up to the sixth nucleotide, the order of the model, as represented by the equivalent branches in the tree, is smaller than the number of proceeding symbols. For example, the tree does not contain the full-order branch for $TCCGG$, since for the last nucleotide $P(A/TCCGG) \approx P(A/G)$. Let us now define the VOM tree more formally.

[Insert Figure 2a 2b and 2c about here]

The VOM model assigns a context for each element in the sequence, and defines the transition probability of each symbol x_j given its context. Graphically, the VOM-tree has a root node on top, from which the branches are developed downwards, with the constraint that each internal node has at most d children, with differently labelled edges. The tree is not necessarily balanced (i.e., not all the branches need to be of the same length) nor complete (i.e., not all the nodes need to have d children). Each node contains d transition probabilities of symbols given the context, which is represented by the reversed path from that node to the root (this is why these trees are also called *suffix trees*). *Optimal context* that satisfies eq. (5) are represented by the reversed path $x_{j-L}, x_{j-L+1}, \dots, x_{j-1}$ from the leaves to the root. Sometimes an optimal context is represented by a *partial leaf*, which is an *internal* node whose reversed path satisfies eq. (5) for some (but not all) nucleotides (see Ben-Gal *et al.*, 2003).

2.3. Construction of the VOM tree

In the following, we describe the algorithm that we use for the construction of the VOM trees for sigma-70 sites (foreground set) and non-sigma-70 sites (background set). The algorithm consists of two main stages. First, it constructs a complete and balanced tree of depth L , which corresponds to a fixed-order Markov model of order L . Second, it iteratively prunes the tree by a backward procedure. The initial order L of the models is estimated using the number of samples in the training sets, such that on average each leaf counter contains at least 10 data points (in our case approximately 40 data points per leaf). Once an initial (complete and balanced) tree of order L is constructed, its probability parameters are estimated (as denoted by the tilde sign) by the frequencies of the respective subsequences, i.e.,

$$\tilde{P}(X_j = x_j | X_{j-L}^{j-1} = x_{j-L}^{j-1}, \Omega_k) = \frac{n_k(x_{j-L}^j)}{n_k(x_{j-L}^{j-1})}, \quad (6)$$

where $n_k(\cdot)$ denotes the frequency of its argument in a training set taken from class Ω_k : Ω_1 is the class of TFBSs (the foreground set), and Ω_2 is the class of non-TFBSs (the background set). To compensate for zero occurrences of certain oligonucleotides we use a pseudo count, which is added to all frequency counters. The value of the pseudo count depends on the level of the node in the tree. For the background model we assure that the sum of all pseudo counts in a level is 4096, for the foreground model the sum is 16. This rule corresponds to an equivalent sample size (ESS) of 4096 and 16, respectively, as used in Bayesian networks (Heckerman *et al.*, 1994). It

results in an effective pseudo count of 1024 for each frequency counter in the root and a pseudo count of 1 in the leaves of a 5th-order tree as a background model.

Then, we compute the Kullback-Leibler (KL) divergence (Kullback, 1959) of the conditional probabilities of symbols between each leaf and its parent node. If the KL divergence is smaller than a pre-selected pruning threshold, the leaf is pruned. A small KL divergence implies that there is no significant divergence in the symbol distribution when using the reduced order of the model, or in other words, that the larger model order, which is represented by the leaf, does not add much information and can be pruned without affecting the likelihood measure significantly. The pruning procedure is repeated until the KL divergence is larger than the predefined pruning threshold for all the leaves in the tree.

A simple pseudo-code to construct a VOM tree from a given (supervised) training set is now detailed. In our example we train two independent VOM models, a foreground VOM model on TFBS oligonucleotides and a background VOM model on non-TFBS oligonucleotides (for simplicity of presentation, we omit the class notation, k , since the algorithm is applied to all the classes independently).

1. Construct an initial complete and balanced Markov tree of a maximal fixed-order L such that in the average each leaf counter contains 10 data points. This rule yields $L=5$ for our background set, and $L=1$ for our foreground set.
2. Estimate the conditional probabilities in the tree nodes by using eq. (6) and by adding a pseudo-count $\varepsilon = ESS/(d \cdot d^t)$ to all frequency counters, with t being the depth of the node, starting from the root with $t=0$.
3. Estimate the KL divergence of the distribution of nucleotides between leaves and their parent nodes:

$$\Delta_{leaf} = \sum_{x_j \in X} \tilde{P}(x_j | x_{j-L_j}^{j-1}) \log_2 \left(\frac{\tilde{P}(x_j | x_{j-L_j}^{j-1})}{\tilde{P}(x_j | x_{j-L_j+1}^{j-1})} \right), \text{ for all leaves.}$$

4. Prune the leaves with a small KL divergence. The pruning process is executed bottom-up from each leaf to the root according to the following rule: If $\Delta_{leaf} \leq c \cdot \psi$, where c is a predefined pruning parameter, prune that leaf by setting $L_j = L_j - 1$. For the foreground model we use $\psi=1$ and for the background model we use $\psi = d^{t+1}/n(x_{j-L_j}^{j-1})$. For such a ψ , the deeper a node

is in the tree it is easier to prune it, and the more samples that reached that node it is harder to prune it. Otherwise, set L_j as the optimal order for that leaf.

5. If all leaves are left unpruned – stop. Otherwise, go back to step 3 and repeat for all the pruned leaves.
6. Refine the probability parameters in the obtained VOM tree by subtracting the counts of each symbol in the decedents nodes from the count of that symbol in the parent node respectively, i.e., $n_k(x_j | x_{j-L_j+1}^{j-1}) - n_k(x_j | x_{j-L_j}^{j-1})$ (for further details see Ben-Gal *et al.*, 2003).

Note again that in general L_j can be equal for all positions, as in the case of a fixed-order Markov model, including the zeroth-order Markov model with $L_j = 0$.

There are several variations for the construction of VOM trees (e.g., Rissanen, 1983, Buhlmann and Wyner, 1999, Orlov *et al.*, 2002, and Ben-Gal *et al.*, 2003) that might affect their classification performance significantly. For example, Orlov *et al.* (2002) use homogeneous (rather than inhomogeneous) VOM trees to model (rather than classify) TFBS oligonucleotides. Their initial tree depth is set to ten and does not depend on the size of the dataset. They use different pseudo counts and prune the VOM tree based on a stochastic complexity measure, which is related to KL divergence, which we use, yet penalizes directly the model complexity. Instead, to avoid model overfitting, we use eq. (5) and search for a good value of the pruning constant over a set of cross-validation experiments. Finally, Orlov *et al.* (2002) do not refine the probability parameters as we do in step 6 above.

2.4. Inhomogeneous models

TFBS oligonucleotides are often represented by inhomogeneous models where a model is built for each position in the sequence. The PWM is possibly the most popular inhomogeneous model for binding sites (Fickett and Hatzigeorgiou, 1997, Ewens and Grant, 2001, Mount, 2001, Barash *et al.*, 2003). The underlying independence assumption for this model implies that the likelihood of a sequence can be computed by eq. (2) with a distinction that the marginal probability $P(X_j = x_j)$ is estimated for each position independently – based on the nucleotides frequency at that position only. Table 1 presents the probability parameters of the PWM model for the two hexamers found respectively in the “-35 box” and the “-10 box”, as estimated from the sigma-70 foreground dataset. The last row in the table presents the consensus sequence. We use this PWM model as a reference model to the proposed inhomogeneous VOM tree, which is described next.

[insert Table 1 around here]

The construction of a foreground inhomogeneous VOM tree is performed by applying the VOM meta code outlined above to each position of the set of aligned binding sites. Thus, we construct twelve independent VOM trees, one for each of the twelve positions of the sigma-70 binding site.

The proposed VOBN model extends the inhomogeneous VOM model by allowing dependencies between non-adjacent positions. To obtain a VOBN we first learn a BN structure. Then, for each position the VOM tree is constructed given the context of its parents in the graph. Subsequent to the trees' construction, we prune and adjust probabilities in every tree as in the original VOM model. Minor modifications are necessary in order to adapt eq. (5) for the VOBN model. The preceding symbols to position j in the sequence X_{j-L}^{j-1} are replaced by the L parents in the BN dependence graph, denoted by $Pa(X_j)_L \equiv (Pa(X_j)_1, Pa(X_j)_2, \dots, Pa(X_j)_L)$, where for every i and given a dependence measure $D(\cdot, \cdot)$: $D(Pa(X_j)_i, X_j) \geq D(Pa(X_j)_{i+1}, X_j)$. In particular, we use *mutual information* as the dependence measure. The equivalent equation to eq. (5) for optimal order of positions in the VOBN model is:

$$L_j = \min \left\{ \tilde{L} \left| P \left(X_j = x_j \mid Pa(X_j)_{\tilde{L}}^1 = pa(X_j)_{\tilde{L}}^1 \right) = P \left(X_j = x_j \mid Pa(X_j)_L^1 = pa(X_j)_L^1 \right) \right. \right\}. \quad (5a)$$

For the case of a Bayesian tree (BT) networks an efficient algorithm exists to learn the maximum likelihood graph structure from the data. This algorithm is equivalent to finding the maximum spanning tree over a fully connected, undirected graph using positions as nodes and mutual information between positions as edge weights (Chow and Liu, 1968). For 1st-order VOBN models we use the dependencies learned with the BT algorithm. Figure 3 shows a first order VOBN model constructed from the foreground dataset of 238 sigma-70 binding sites. This example illustrates an important property of the VOBN model: since insignificant branches are pruned, the tree can serve as a compact and accurate exploratory tool for the dependencies among the basepairs in the sequence. In section 4, we comment on some of the dependencies observed in the sigma-70 data-set. Note that pruning a VOBN model is not equivalent to pruning edges of a BN dependence graph, as the VOBN pruning is context specific. The performance of the VOBN model is further studied in section 4.

Classification Rule

Once the foreground and background models are selected, the parameters are estimated and refined from a training set, and the log-likelihood ratio of these models

is used for classification. In the case of our TFBS classification, a site is declared as TFBS if the log-likelihood ratio is greater than a given threshold T , i.e., if

$$\log_2 \frac{P(x_1^N | \Omega_1)}{P(x_1^N | \Omega_2)} \geq T. \quad (7)$$

Specific biological knowledge, such as the low number of TFBS sequences in the genome, can be taken into account by specifying the threshold value T to guarantee the balance of the specificity and the sensitivity of the classifier. In the experiments presented in section 4 we choose a value of T that guarantees a true negative (TN) rate of 99.9%. The E.coli genome consists of approximately $4 \cdot 10^6$ basepairs (bp) containing approximately 4000 genes. Since we expect one gene per 1000bp, a true negative rate of 99.9% keeps the number of false predictions smaller than the expected number of true sigma-70 sites for the whole-genome analyses.

The performance measure

Our main performance measure is the mean true-positive (TP) rate (also known as *sensitivity*), which is the ratio between true-positives and all positive samples, for replicated stratified-holdout experiments having a fixed true-negative rate of 99.9%. We also compute the standard deviation of the estimated mean TP rate, as a measure for the model robustness and the dependence of its accuracy on the training set.

Stratified-holdout sampling

In order to minimize over-fitting effects, we conduct a 10^6 -fold stratified-holdout experiment by iteratively applying the following procedure. A random 10% of all TFBS sequences are excluded from the foreground dataset and a model is constructed based on all other remaining sequences in the set. A random 10% of all background sequences are excluded from the background dataset and a model is constructed based on all other remaining sequences in the set. Based on these models the likelihood of every sequence in the excluded foreground and background sequences is calculated. Using the model likelihoods the score (7) is computed, according to which each excluded foreground sequence is marked as either true-positive or false-negative.

3. DATASETS

Throughout the experiments we use three datasets: one foreground dataset that contains 238 carefully selected E.coli sigma-70 binding sites of length 12 bp, and two background datasets – one that contains randomly permuted TFBS sequences, and another that contains 12-mers sampled from 472 intergenic “non-promoter”

sequences in E.coli. All datasets are available upon request from bengal@eng.tau.ac.il.

The Sigma-70 foreground dataset. Transcription initiation in E.coli is controlled to a large degree by the binding of the RNA polymerase holoenzyme together with multiple cofactors to the promoter region just upstream of the transcription start sites. One of the cofactors, which is believed to convey a large fraction of the DNA binding specificity, is the sigma-70 factor. Two well-conserved cis elements, the “-35 box” and the “-10 box,” can be found in close proximity – approximately 35 bp and approximately 10 bp upstream – of the transcription start site of many E.coli mRNA genes. In order to obtain a dataset of these cis-element pairs (and likely sigma-70 factor binding sites) we perform the following steps:

- We start with a dataset of 300 binding site pairs from PromEC (Margalit *et al.*, 2000). Each of these binding site pairs consists of two hexamers, so the motif length for all of the models discussed in this paper is $L=12$.
- We remove all binding site pairs that could not be found in the database RegulonDB 3.0 (Salgado *et al.*, 2000) created by Julio Collado-Vides and co-workers, or which are not annotated there as sigma-70 binding sites.
- We map the remaining binding site pairs to the E.coli genome, including the “spacer” sequences between the two hexamers boxes, and remove all binding site pairs that could not be mapped uniquely to the E.coli genome (see Blattner and Schroeder, 1984), or that got mapped to a protein-coding region (NCBI genbank).

Following the above procedure we obtain a set of 238 binding site pairs, which we call the “sigma-70 foreground dataset,” or simply the “foreground dataset.” We choose the above-mentioned very stringent rules to derive the “sigma-70 foreground set” for a simple reason: it is generally true that, for statistical analyses, a small dataset of high quality is more valuable than a larger dataset of lower quality. Hence, we would like to obtain a foreground set with a minimum amount of contamination, and we are willing to sacrifice true-positive binding site pairs in order to guarantee a very low number of false-positive binding site pairs in the foreground set.

The Background Datasets. We generate two background datasets – the “random background set” and the “intergenic background dataset” – for two different studies. We use the “random background dataset” in order to study the degree to which the VOM/VOBN models capture the existing statistical dependencies in the sigma-70 foreground set. For this study it is important (i) to eliminate possible correlations in

the sequences of the background set; and (ii) to eliminate a possible classification success simply due to a different nucleotide composition in the foreground and the background set. We use the “intergenic background dataset” in order to study the degree to which higher-order background models can improve the classification of sigma-70 binding site pairs versus 12-mers sampled from intergenic regions.

The Random Background dataset. The homogeneous zeroth-order Markov model is a popular background model of intergenic sequences (Liu *et al.*, 1995, Neuwald *et al.*, 1995, Hughes *et al.*, 2000, Thijs *et al.*, 2001). In order to generate a dataset without 'built-in' statistical dependencies among different positions, and in order to eliminate any composition difference between the foreground and the background sets, we generate the random background dataset as follows: Learn a zeroth-order Markov model from the foreground dataset. Use this Markov model to generate 427 random sequences of length 182, which gives approximately the same number of 12-mer windows as in the “intergenic background dataset”.

The Intergenic Background dataset. Many experiments have been performed to obtain sigma-70 factor binding sites in the E.coli genome, but only little work has been devoted to identify – with experimental rigor – sequences that are free of sigma-70 binding sites. Hence, we adopt the following protocol (Gelfand 2003, Thijs *et al.* 2001) to obtain sequences that are unlikely to contain sigma-70 binding sites. Two neighbouring genes are located either on the same strand or on opposite strands. If they are located on opposite strands, they either overlap or they share a common intergenic region. If they share a common intergenic region, that region is either a common 5' (or upstream) intergenic region of both genes, or a common 3' (or downstream) region of both genes. Provided that the gene annotation is reliable, the common 3' (or downstream) intergenic regions between two neighbouring genes should not contain sigma-70 binding sites. We extract the set of common 3' (or downstream) intergenic regions between two neighbouring genes from the complete E.coli genome (NCBI 2003), and we obtain a dataset consisting of 472 sequences with a total of 77,644 nucleotides, which we call the “intergenic background dataset.”

4. RESULTS AND DISCUSSION

The analysis of the above-mentioned data is performed in three stages. First, we study the degree to which the VOB model is capable of capturing statistically significant (and perhaps biologically relevant) dependencies among the different positions within sigma-70 factor binding sites, compared to inhomogeneous fixed-order Markov

models. Since we focus on the contribution of the foreground set, we use the “random background dataset” and find the zeroth-order Markov model as the best background model when we train the VOM on this dataset. Second, we study the degree to which statistical dependencies present in the “intergenic background dataset” can improve the classification performance, hence, we increase the order of the homogeneous Markov models for the background from $L=0$ to $L=5$. Third, we apply VOBN models constructed with different pruning constants to the “foreground dataset” and VOM models constructed with different pruning constants to the “intergenic background datasets” and we compare the accuracy of these models with the accuracy of fixed-order Markov models including the PWM models and Bayesian trees (BTs).

We briefly describe the notations used in the following figures and discussion.

1. **Inhomogeneous models:** The model constructed from the "sigma-70 foreground dataset". Two types of inhomogeneous models are considered:
 - a. **Markov(L)** – the inhomogeneous Markov model of order L , which includes the PWM model in case of $L=0$.
 - b. **VOBN(L, c)** – the generalized VOM model with L denoting the initial maximal order and c denoting the pruning constant, equivalent to the PWM model in the case of $L=0$ and $c=0$. We frequently use a BT model - equivalent to a VOBN(1,0) model.
2. **Homogeneous models:** the model constructed from the background dataset. Two types of background models are considered as follows.
 - a. **Markov(L)** – the homogeneous Markov model of order L .
 - b. **VOM(L, c)** – the homogeneous VOM model with L denoting the initial maximal order and c denoting the pruning constant, which is equivalent to the Markov(L) model in case of $c=0$.
3. **Mean TP:** The obtained mean true-positive rate for a fixed true-negative level of 99.9%. The standard deviation of the estimated mean TP (in figures 4, 5, and 6) with 1M fold equals $S/\sqrt{10^6} \approx 0.01\%$, where S denotes the sample *standard deviation* that is obtained from the 10^6 replicated experiment. The standard deviations of each experiment are not shown as they are similar.
4. **Number of nodes:** the average number of nodes in the model over the stratified-holdout experiments.

In the first stage of the experiment we study the classification performance of foreground inhomogeneous VOBN models vs. inhomogeneous Markov models, including the widely-used PWM model (e.g., see Fickett and Hatzigeorgiou, 1997,

Ewens and Grant, 2001, and Mount, 2001). We use the “random background dataset” and obtain a homogeneous Markov(0) as the best background model. We find that inhomogeneous Markov(2) and Markov(3) models are over-fitted and that the inhomogeneous Markov(1) model achieves the highest classification performance of the fixed order models with mean TP rate=29.4%. Pruning the inhomogeneous VOM models by different pruning constants does not result in considerable improvements over fixed-order models: the VOBN(1, c) with $c=0.2102241$ achieves a statistically significant improvement with a mean TP rate of 30.7%.

Next, we analyze the performance of higher-order Markov models based on the “Sigma-70 foreground dataset” and the “intergenic background dataset”. We summarize the results in three figures: 4, 5, and 6. Figure 4 focus on fixed-order models. It shows the performance of different combinations of foreground and background models. Figure 5 shows the enhancement from pruning the foreground model to obtain a VOBN, where the background model is fixed to Markov(3). Figure 6 shows the improvement from pruning a 5th-order VOM model in the background with the foreground model fixed to the best VOBN model from Figure 5.

Figure 4 presents the classification accuracy for different combinations of fixed-order foreground (PWM, inhomogeneous Markov (1), and a BT) and background models (homogeneous Markov $L=0$ up to $L=5$). For a PWM model as foreground model, a Markov(2) model is the optimal background model, yielding a mean TP rate of 44.39%. We also find that, if an inhomogeneous Markov(1) model is chosen as foreground model, then a Markov(3) model is the optimal background model with a mean TP rate of 43.5%. Finally, with a BT model as the foreground model, a Markov(3) model is the optimal background model, resulting with the largest mean TP rate over all fixed order models of TP = 45.65%. All findings are in qualitative agreement with previous studies, such as, e.g., Thijs *et al.* (2001) and Barash *et al.* (2003), that indicate that homogeneous Markov(0) models may not be optimal for modelling genomic DNA. Clearly, the optimal Markov model order for the background depends on the foreground model chosen as well as on the measure of classification accuracy. In addition, we caution that one cannot infer from Figure 4 the optimal Markov model order for other datasets or other classification problems. One interesting observation from Figure 4 is that for all background models the mean TP rates of the PWM as foreground models are higher than for the Markov(1) model. Thus, for all of background models tested, the PWM model is superior to the inhomogeneous Markov(1) model. This finding is consistent with the high popularity

of the PWM model for TFBS recognition and explains why the weight array model proposed by Zhang and Marr (1993), which corresponds to a inhomogeneous Markov(1) model and which has been shown to be more accurate than the PWM model for splice site recognition, has not replaced the PWM model for TFBS recognition. The unimodal behaviour of the TP rate as a function of the number of model parameters might reflect the trade-off between over-fitted models and under-fitted models. It can also be seen from Figure 4 that for third and higher order background models the BT always outperforms the PWM. For lower order background models there is no clear dominance.

Although for certain background models the BTs achieve better results than PWMs, it is not clear whether the BTs are overfitted. In the following, we test for such an overfitting scrutinizing VOBN models for different pruning constants since these models have the potential of modelling only a few (significant) statistical dependencies, while neglecting statistically insignificant ones.

[Insert Figure 4 about here]

Figure 5 presents the classification performance of different foreground VOBN models for different choices of the pruning constant c (including the PWM model) for a background Markov(3) model, which was found best in Figure 4. The mean TP rate is given as a function of the number of nodes left in the VOBN(1, c). Note, that increasing the pruning constant decreases the number of nodes left in the VOBN. The behaviour of the mean TP rate is nearly unimodal except for models very similar to PWM. The best foreground model is a VOBN(1,2^{-3.75}) achieving a mean TP rate of 46.46%. This is an improvement of 0.8% over the BT model. As one can see from Figure 5 the best VOBN(1, c) foreground model has approximately half the number of nodes compared to the full model, yet reaches a significantly higher mean TP rate.

To further explore the dependencies between foreground and background model we now fix the foreground model as VOBN(1,2^{-3.75}) and turn to homogeneous VOM models for the background. Figure 6 presents the improvement gained by pruning a background VOM model with maximum order 5. The plot has again an essential unimodal shape as in Figure 5. The best background model found is VOM(5,2^{-5.5}), which has 94 nodes (it is too large to be presented in the paper, but it is available upon request from the authors). This best combination of a VOBN(1,2^{-3.75}) with a VOM(5,2^{-5.5}) achieves a mean TP rate of 47.56%. This is an improvement of 3.17% compared to the combination of PWM/Markov(2) models in Figure 4.

It is interesting to note that this mean TP rate can only be gained, if we use a maximal order of 5 for the VOM. In the VOM(5,2^{-5.5}) model there are still contexts up to the 5th order left (e.g., GCCGG, TCCGG), while others are already pruned to down to order 3. These long contexts seem to be responsible for the high classification accuracy of the background model, as pruning a 4th or 3rd order model reaches significantly lower mean TP rates.

As evident from Figure 6, for low pruning constants (models with more than 300 nodes) the mean TP rate is even below that of the common combination of PWM and Markov(L). The VOM(5,c) models between $c=2^{-10}$ and 2^{-9} still have 400 to 600 nodes, which is more than for a 4th order model (with 341 nodes) and causes a strong over-fitting effect. The best mean TP rate is reached for a pruned model (with 94 nodes) that has only few more parameters to estimate than the fixed Markov(3) model, but utilizes those parameters much more efficiently. If the pruning constant is further increased the classification accuracy decreases again, as now significant and vital contexts are pruned from the model.

[insert figure 5,6 about here]

To summarize, the variable order concept applied to foreground Bayesian trees (obtaining the VOBN) and to background Markov models (obtaining the VOM) is shown to outperform the common PWM by 3.17% (31 standard deviations higher).

Figure 3 shows the foreground VOBN(1,2^{-3.75}) model which reached the highest TP rate. Note that more than half the edges in the dependencies' graph are between non-adjacent positions, and that two of the edges are between positions from separate boxes (the “-35 box” and “-10 box”). Approximately half of the nodes were found insignificant and were pruned.

5. CONCLUSIONS

VOBN models are one promising generalization of the widely-used position weight matrix (PWM) model, fixed-order Markov models and Bayesian Networks (BNs). In this paper we show that VOBN models are useful for predicting the location of transcription factor binding sites (TFBSs). Specifically, we show in stratified-holdout experiments that a VOBN model can predict the location of sigma-70 binding sites in E.coli with higher accuracy than a PWM model, a fixed-order Markov model and a BT model. We speculate that VOBN models might be useful for predicting the location of TFBSs in other genomes.

Acknowledgements. We are thankful to the Minerva Short-Term Research Grants for supporting mutual visits of students in the research team.

REFERENCES

- Baldi,P. and Brunak,S. (2001) *Bioinformatics: The Machine Learning Approach*, 2nd ed., The MIT Press, Cambridge, MA.
- Barash,Y., Elidan,G., Friedman,N. and Kaplan,T. (2003), Modeling Dependencies in Protein-DNA Binding Sites. *Proc. Seventh Annual Inter. Conf. on Computational Molecular Biology (RECOMB)*.
- Bejerano,G. and Yona,G. (2001) Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics*, **17**, 23-43.
- Ben-Gal, I., Morag, G., Shmilovici, A. (2003) CSPC: A Monitoring Procedure for State Dependent Processes, *Technometrics*, **45**(4), 293-311.
- Ben-Gal,I. and Shmilovici,A. (2001) Promoters recognition by variable-length Markov models. *Workshop on Artificial Intelligence and Heuristic Methods for Bio-informatics*, San-Miniato, Italy.
- Ben-Gal,I., Shmilovici,A. and Morag,G. (2000), Design of control and monitoring rules for state dependent processes. *The International Journal for Manufacturing Science and Production*, **3**(2-4), 85-93.
- Benos,P.V., Lapedes,A.S., Fields,D.S. and Stormo,G.D. (2001), SAMIE: statistical algorithm for modeling interaction energies. In PSB'01.
- Bilu Y., Linial M., Slonim N. Tishby N. (2002), Locating Transcription Factors Binding Sites Using a Variable Memory Markov Model. Leibnitz Center TR 2002-57. Also available at <http://www.cs.huji.ac.il/~johnblue/papers/>
- Blattner F.R., Schroeder, J. L. (1984), A computer package for DNA sequence analysis. *Nucleic Acids Research* 12(1): 615-617
- Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. (1996) Context-specific independence in Bayesian networks. *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, 115-123.
- Buhlmann,P. and Wyner,A. J. (1999), Variable length Markov chains. *Ann. Statist.* **27**(2), 480-513.
- Bulyk,M.L., Johnson,P.L. and Church,G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–61.
- Cai, D., Delcher, A., Kao, B. and Kasif, S. (2000) Modeling splice sites with Bayes Networks. *Bioinformatics*, **16**, 152-158.
- Castelo, R., and Guigo, R. (2004) Splice site identification by idIBNs. *Bioinformatics*, **20**, i69-i76.
- Chow,C.K., Liu,C.N. (1968) Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, **14**, 462-467.
- Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O., Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699-705.
- Djordjevic, M., Sengupta, A.M., and Shraiman, B.I. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res.* **13**(11), 2381-2390.
- Ewens,W. J. and Grant,G.R. (2001) *Statistical Methods in Bioinformatics: An Introduction*, Springer-Verlag New York, Inc.
- Fickett,J.W. and Hatzigeorgiou,A.G. (1997) Eukaryotic Promoter Recognition. *Genome Research*, **7**, 861-878.

- Friedman, N. and Goldszmidt, M. (1996) Learning Bayesian Networks with Local Structure. *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, 252-262.
- Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. *Journal of Computational Biology* **7**, 601-620
- Gelfand, V. (2003), personal communication; a similar protocol is presented in Thijs *et al.*, 2001).
- Hanisch,D., Zien,A., Zimmer,R. and Lengauer,T. (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **1**(1), 1-10.
- Heckerman, D., Geiger, D., Chickering, D. M. (1994) Learning Bayesian Networks. Microsoft Research, MSR-TR-95-02.
- Hughes,J.D., Estep, Preston,W., Tavazoie,S. and Church,G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205-1214.
- Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V., Wingender,E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **31**(13), 3576-9.
- Kel-Margoulis,O.V., Tchekmenev,D., Kel,A.E., Goessling,E., Hornischer,K., Lewicki-Potapov,B. and Wingender,E. (2003) Composition-sensitive analysis of the human genome for regulatory signals. *In Silico Biology* **3**, 13.
- Kullback,S. (1959) *Information Theory and Statistics*, New York: Wiley (reprinted in 1978 by MA: Peter Smith).
- Liu,J.S., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, **90**, 1156-1170.
- Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, **6**, 127-138.
- Margalit,H., Hershberg,R., Bejerano,G. and Santos-Zavaleta,A. (2000) at <http://bioinfo.md.huji.ac.il/marg/promec/index.html>.
- Mount,D.W. (2001) Bioinformatics, sequence and genome analysis. *Cold Spring Harbor Laboratory Press*, 357-365.
- Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.*, **4**, 1618-1632.
- Ohler,U. and Niemann,H. (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.*, **17**, 56-60.
- Ohler,U., Harbeck,S., Neimann,H., Noth,E. and Reese,M.G. (1999) Interpolated Markov chains for eukaryotic promoter recognition. *Bioinformatics*, **15**(5), 362-369.
- Orlov,Y.L., Filippov,V.P., Potapov,V.N., Kolchanov, N.A. (2002), Construction of stochastic context trees for genetic texts. *In Silico Biology* **2**(3), 233-247.
- Orlov,Y.L., Potapov,V.N., (2000) Determining Markov model of genetical texts by stochastic complexity estimation. BGRS, Novosibirsk, 71-73.
- Pickert,L., Reuter,I., Klawonn,F. and Wingender,E. (1998) Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics*, **14**, 244-251.
- Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference*. Morgan Kaufmann, California.
- Rissanen,J. (1983) A Universal Data Compression System. *IEEE Transactions on Information Theory*, **29**(5), 656-664.

- Ron D., Singer Y. and Tishby N. (1996) The power of amnesia: learning probabilistic automata with variable memory length. *Machine Learning*, **25**, 117-149.
- Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millán-Zárate,D., Blattner,F.R. and Collado-Vides,J. (2000) RegulonDB (version 3.0): transcriptional regulation and operon organization in Escherichia coli K-12. *Nucleic Acids Res.* **28** (1), 65–67.
- Salzberg,S.L. (1997) A method for identifying splice sites and translational start sites in Eukaryotic mRNA. *Computer Applications in the Biosciences (CABIOS)*, **13**(4), 365-376.
- Salzberg,S.L., Delcher,A.L., Kasif,S. and White,O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, **26**(2) 544-548.
- Salzberg, S.L., Pertea,M., Delcher,A.L., Gardner,M.J. and Tettelin,H. (1999) Interpolated Markov models for eukaryotic gene finding. *Genomics*, **59**, 24-31.
- Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273-3297.
- Stormo, G.D. and Fields, D.S. (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci* **23**, 109-113.
- Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor,B., Rouzé,P. and Moreau,Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**, 1113-1122.
- Vert,J.P. (2001) Adaptive context trees and text clustering. *IEEE Transactions on Information Theory*, **47**(5), 1884-1901.
- Weinberger,M., Rissanen,J.J. and Feder,M. (1995) A Universal Finite Memory Source. *IEEE Transactions on Information Theory*, **41**, 643-652.
- Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhäuser,R., Prüß,M., Schacherer,F., Thiele,S. and Urbach,S. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.* **29**, 281-283.
- Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* **28**, 316-319.
- Zhang, M.Q. and Marr, T.G. (1993) A weight array method for splicing signal analysis. *Computational Applications for Bioscience*, **9**, 499-509.

TABLES

	1	2	3	4	5	6		1	2	3	4	5	6
A	0.10	0.07	0.10	0.54	0.17	0.49		0.07	0.74	0.15	0.57	0.53	0.08
C	0.10	0.09	0.13	0.19	0.54	0.14		0.12	0.07	0.12	0.12	0.22	0.07
G	0.12	0.09	0.57	0.12	0.12	0.17		0.10	0.07	0.14	0.16	0.09	0.05
T	0.68	0.74	0.20	0.15	0.18	0.20		0.72	0.12	0.58	0.15	0.15	0.80
	T	T	G	A	C	A		T	A	T	A	A	T

Table 1. PWM and consensus sequences for the “-35 box” and the “-10 box”, as derived from the "sigma-70 foreground dataset"

FIGURES

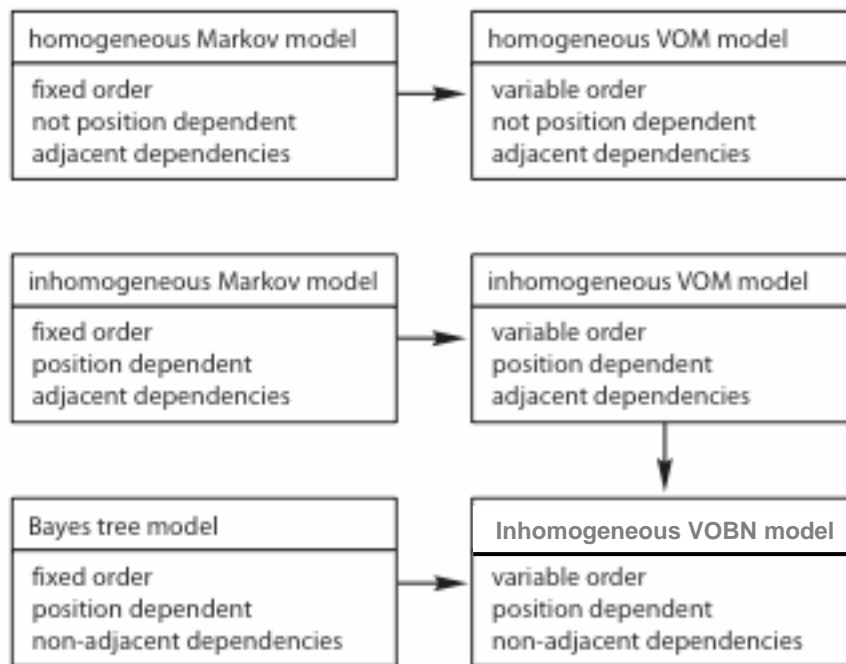


Figure 1. The generalization hierarchy of the variable order models. The generalized Variable-order Markov (VOBN) model can be interpreted as a generalization of the VOM model as well as of the BT model.

A	C	G	T
0.26	0.24	0.24	0.26

Figure 2a. A degenerated VOM tree represented by a root node, which is equivalent to a Markov(0) model. The tree is constructed from the “intergenic background dataset.” The label of the node is the 4-dimensional probability vector of the single-nucleotide probabilities $P(A)$, $P(C)$, $P(G)$, and $P(T)$ in this order. Note that in this and the following figures the sum of probabilities in a node might not sum exactly to 1 due to rounding and pseudo-count effects.

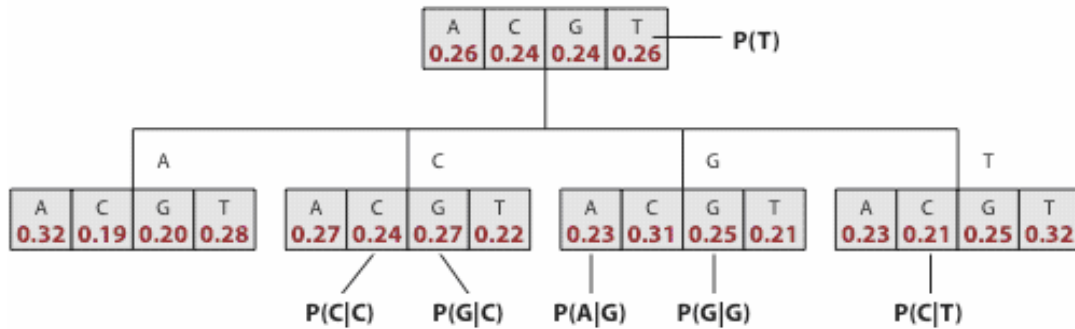


Figure 2b. An homogeneous unpruned VOM tree, which is equivalent to Markov(1) model. The tree is constructed from the “intergenic background dataset.” The leaf nodes are labeled with the transition probability vector given a single-nucleotide context. The root node is labeled with the unconditional probability vector of nucleotides.

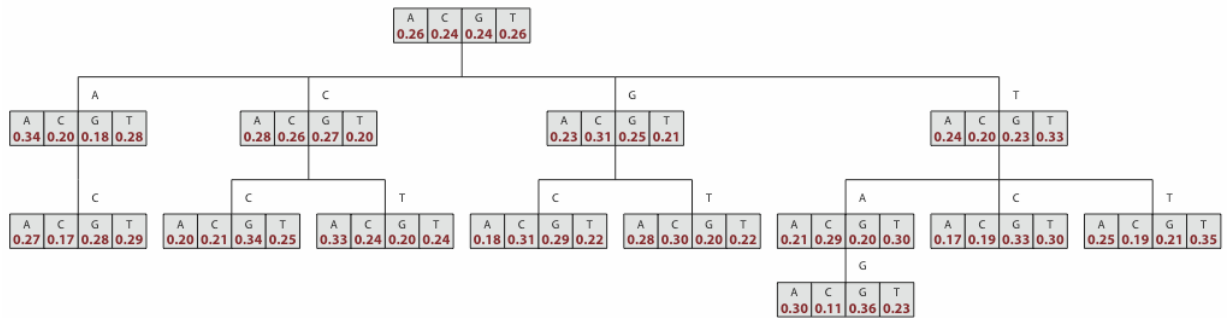
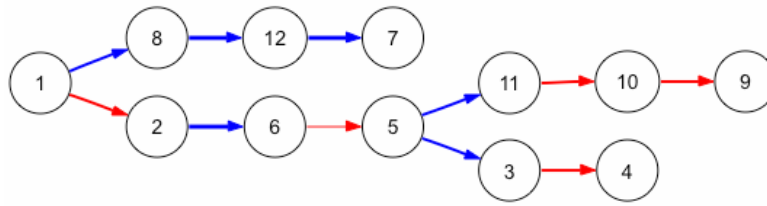
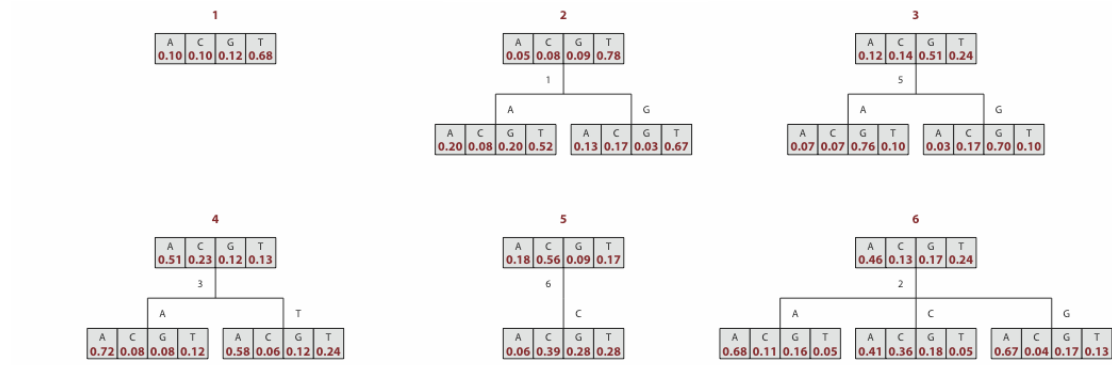


Figure 2c. The VOM(5, 0.65) homogeneous VOM tree. The tree is constructed from the “intergenic background dataset.” The nodes are labelled with the transition probability vector given the context defined by the reversed path from the nodes to the root. The root node is labelled with the unconditional probability vector of nucleotides.



“-35 box”



“-10 box”

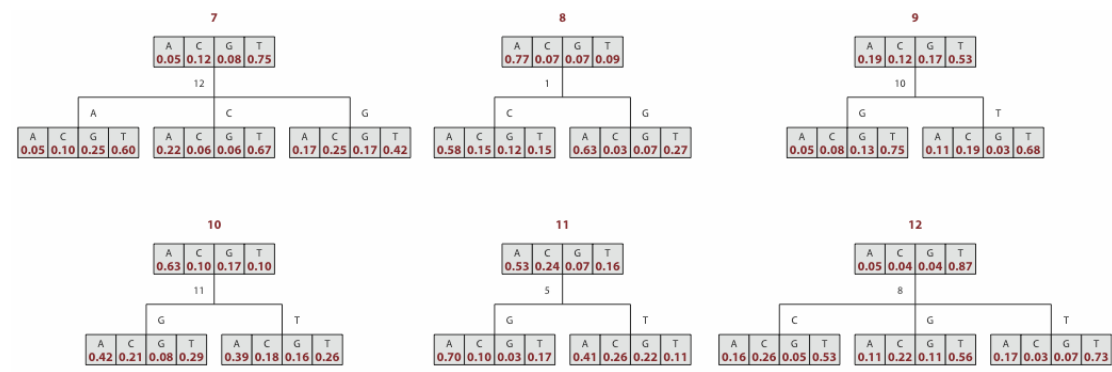


Figure 3. The BT dependence graph and the corresponding $\text{VOBN}(1, 2^{-3.75})$ forest, as constructed from the “Sigma-70 foreground dataset.” In the VOBN forest the nodes are labelled with the unconditional probability vector of nucleotides at that position in the box. The position is denoted by the upper numerical label (above the root). The position on which’s context we split is labelled by the numerical label near the split.

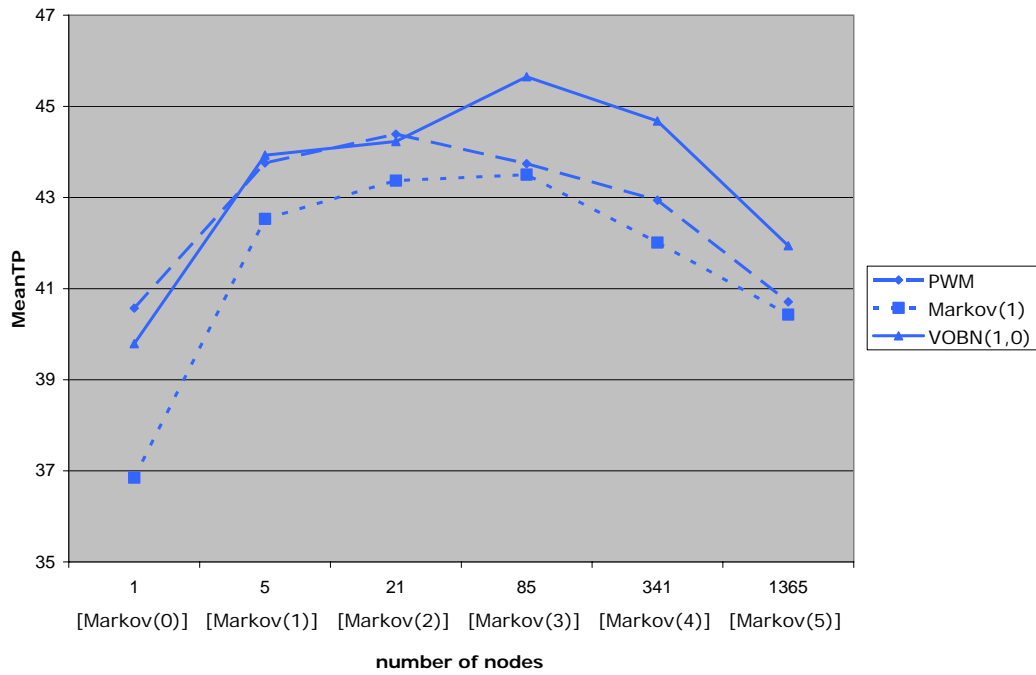


Figure 4. Mean TP values for different fixed-order model combinations. The Markov model order and their number of nodes in the background model are presented in the x axis. The three different curves signify the different foreground models: PWM, Markov(1) and a VOB(1,0) which is equivalent to a BT.

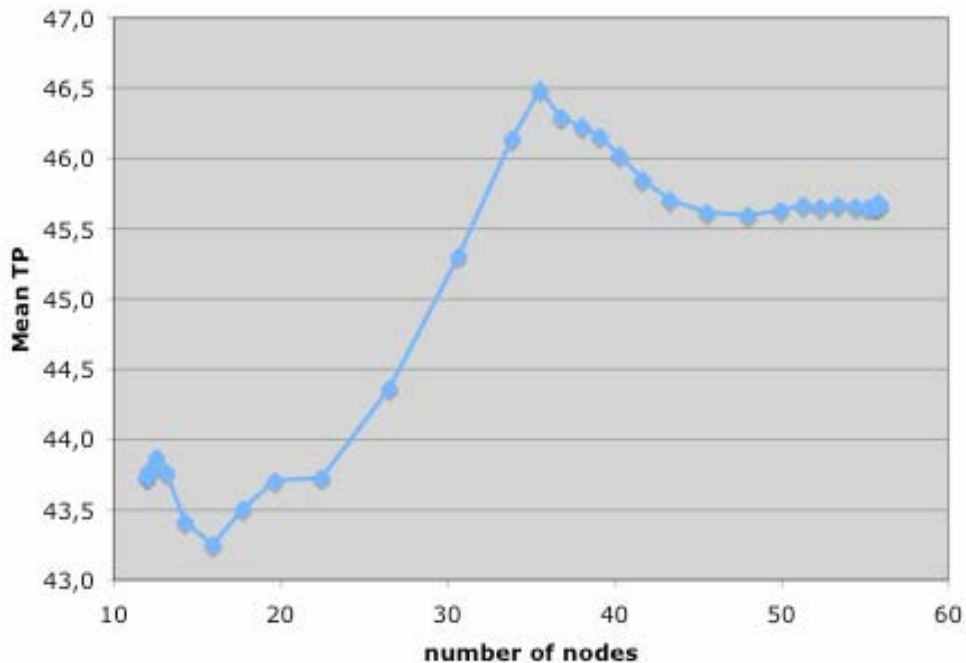


Figure 5. Pruning of the foreground VOB(1,c) model for different values of pruning constant c for a Markov(3) background model. The number of nodes in the foreground model is presented on the x-axis to show the size of the model.

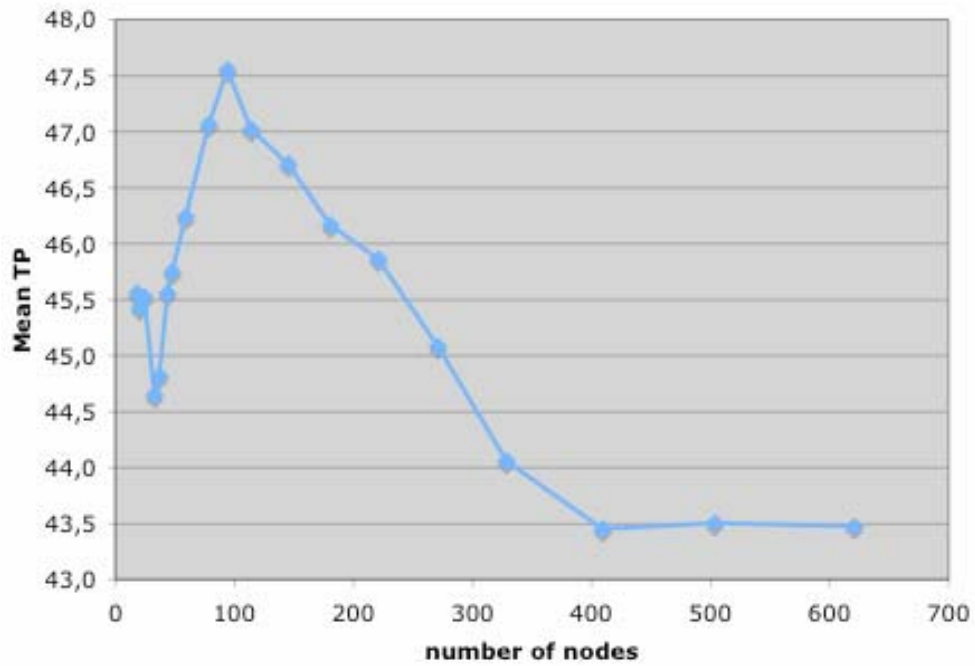


Figure 6. Pruning the background VOM(5,c) model for different values of the pruning constant c . The foreground model is the VOBN(1,2^{-3.75}) model, which was found best in Figure 5 and has about 36 nodes. The number of nodes in the background model is presented on the x -axis to show the size of the model.