Extended Sunflower Hidden Markov Models for the recognition of homotypic *cis*-regulatory modules

Ioana M. Lemnian[1], Ralf Eggeling[1], and Ivo Grosse[1,2]

1 **Martin Luther University Halle-Wittenberg**
Institute of Computer Science
Von-Seckendorff-Platz 1, 06120 Halle, Germany
{lemnian, eggeling, grosse}@informatik.uni-halle.de

2 **German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig**
Deutscher Platz 5e, 04103 Leipzig, Germany

Abstract

The transcription of genes is often regulated not only by transcription factors binding at single sites per promoter, but by the interplay of multiple copies of one or more transcription factors binding at multiple sites forming a *cis*-regulatory module. The computational recognition of *cis*-regulatory modules from ChIP-seq or other high-throughput data is crucial in modern life and medical sciences. A common type of *cis*-regulatory modules are homotypic clusters of binding sites, i.e., clusters of binding sites of one transcription factor. For their recognition the homotypic Sunflower Hidden Markov Model is a promising statistical model. However, this model neglects statistical dependences among nucleotides within binding sites and flanking regions, which makes it not well suited for *de-novo* motif discovery. Here, we propose an extension of this model that allows statistical dependences within binding sites, their reverse complements, and flanking regions. We study the efficacy of this extended homotypic Sunflower Hidden Markov Model based on ChIP-seq data from the Human ENCODE Project and find that it often outperforms the traditional homotypic Sunflower Hidden Markov Model.

1998 ACM Subject Classification J.3 Life and medical sciences

Keywords and phrases Hidden Markov Models, *cis*-regulatory modules, *de-novo* motif discovery

Digital Object Identifier 10.4230/OASIcs.GCB.2013.101

1 Introduction

The computational recognition of *cis*-regulatory modules (CRMs) is an important task in DNA sequence analysis. If the sequence motifs of the transcription factor binding sites (TFBSs) involved in putative CRMs are known, CRM recognition reduces to finding the composition of TFBS occurrences in a set of promoters or other unaligned sequences, and many methods exist for this task [13]. However, if the sequence motifs are unknown, CRM recognition becomes challenging, and reliable methods are still missing.

A promising model for CRMs is the Sunflower Hidden Markov Model (Sunflower HMM) proposed by Hoffmann and Birney [6], which allows multiple occurrences of TFBSs per sequence. However, the Sunflower HMM assumes statistical independence of the nucleotides within TFBSs and flanking regions, which limits its applicability to *de-novo* motif discovery. There is evidence about the presence of statistical dependences among adjacent nucleotides within TFBSs [9, 2, 1], and neglecting the dependences within flanking regions often leads to

German Conference on Bioinformatics 2013 (GCB'13).
Editors: T. Beißbarth, M. Kollmar, A. Leha, B. Morgenstern, A.-K. Schultz, S. Waack, E. Wingender; pp. 101–109

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

the erroneous identification of repeats instead of putative TFBSs or to poor performance in the recognition of TFBSs [12].

Multiple TFBSs of the same transcription factor often build homotypic clusters, which we call homotypic CRMs. Such homotypic CRMs are frequent not only in invertebrates [8], but also in humans [4]. In this paper we focus on their recognition by a homotypic version of the Sunflower HMM, and by its extension that allows statistical dependences among adjacent nucleotides both within TFBSs and flanking regions.

The rest of the paper is structured as follows: in Section 2 we present the extended homotypic Sunflower HMM and corrresponding learning algorithms, and in Section 3 we study the efficacy of the extended homotypic Sunflower HMM in comparison to the traditional homotypic Sunflower HMM based on ChIP-seq data from the ENCODE project [11].

2 Extended homotypic Sunflower Hidden Markov Models

In the following two subsections, we introduce the extended homotypic Sunflower HMM and the Baum-Welch algorithm for estimating its model parameters. For the sake of convenience, we call the (traditional or extended) homotypic Sunflower HMM simply (traditional or extended) Sunflower HMM from now on.

2.1 Model

Consider a data set of N sequences $\underline{x}_1, \ldots, \underline{x}_N$, and denote the i-th sequence of length L_i by $\underline{x}_i = (x_{i,1}, \ldots, x_{i,L_i})$, where $i \in \{1, \ldots, N\}$. In analogy to the traditional Sunflower HMM, we define the probability of sequence $\underline{x}_i$ given model parameters π and $\underline{\phi}$ by

$$P(\underline{x}_i|\pi, \underline{\phi}) = \sum_{\underline{u}_i} P(\underline{u}_i|\pi) P(\underline{x}_i|\underline{u}_i, \underline{\phi}), \tag{1}$$

where $\underline{u}_i$ denotes a hidden path consisting of states $u_{i,j} \in \{m_1, \ldots, m_M, m_{\overline{1}}, \ldots, m_{\overline{M}}, f_1, f_2\}$, with M denoting the width of a putative TFBS. Here, π denotes the probability of a transition from a flanking region to a TFBS or its reverse complement, and $\underline{\phi}$ denotes all emission parameters of the model.

The states $u_{i,j}$ are indicator variables for TFBS occurrences in the following manner: $u_{i,j} = m_k$ indicates that $x_{i,j}$ is the k-th nucleotide of a TFBS on the forward strand, $u_{i,j} = m_{\overline{k}}$ indicates that $x_{i,j}$ is the k-th nucleotide (read in $3' \to 5'$ direction) of a TFBS on the reverse complementary strand, and $u_{i,j} = f_1$ and $u_{i,j} = f_2$ indicate that $x_{i,j}$ is part of the flanking region. State f_1 indicates the start position of a flanking region, while state f_2 indicates subsequent positions of a flanking region.

In analogy to the traditional Sunflower HMM, we define the probability of path $\underline{u}_i$ given model parameter π by

$$P(\underline{u}_i|\pi) = P(u_{i,1}|\pi) \cdot \prod_{j=2}^{L_i} P(u_{i,j}|u_{i,j-1}, \pi), \tag{2}$$

which states that the hidden path $\underline{u}_i$ is a realization of a homogeneous first-order Markov model.

The transition of one state to another is parameterized by a sparse transition matrix. There are three possible transitions from states f_1 and f_2: to m_1 with probability $\pi/2$, to $m_{\overline{M}}$ with probability $\pi/2$, and to f_2 with probability $1-\pi$. Here, we assume that the probabilities for the occurrence of a TFBS on the forward strand and on the reverse complementary strand

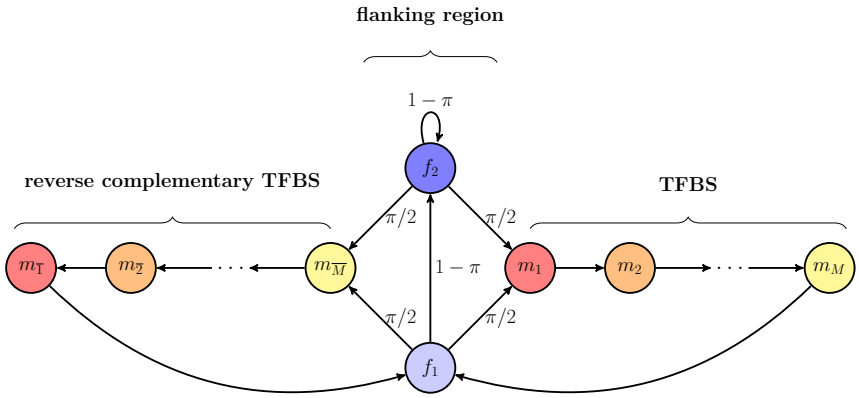

Figure 1 Graphical representation of the transition matrix of the extended Sunflower HMM. Circles denote states of the HMM. States f_1 and f_2 emit the flanking region, $m_1, m_2, \ldots, m_M$ emit the TFBS, and $m_{\overline{M}}, m_{\overline{M-1}}, \ldots, m_{\overline{1}}$ emit the reverse complementary TFBS. States in the same color correspond to the same position in the motif. Arrows represent transition probabilities between states with a probability greater than zero. The transition probability is either marked as a label of the corresponding arrow or is 1 in case of unlabeled arrows.

are equal. There are deterministic transitions from m_k to m_{k+1}, from $m_{\overline{k}}$ to $m_{\overline{k-1}}$, from m_M to f_1, and from $m_{\overline{M}}$ to f_1. All other transitions are forbidden, i.e., their propabilities are zero. The graphical representation of this transition matrix is shown in Figure 1.

In contrast to the traditional Sunflower HMM, we define the likelihood of sequence $\underline{x}_i$ given path $\underline{u}_i$ and model parameter $\underline{\phi}$ by

$$P(\underline{x}_i|\underline{u}_i, \underline{\phi}) = P(x_{i,1}|u_{i,1}, \underline{\phi}) \cdot \prod_{j=2}^{L_i} P(x_{i,j}|u_{i,j}, x_{i,j-1}, \underline{\phi}). \tag{3}$$

In the traditional Sunflower HMM, the conditional probabilities $P(x_{i,j}|u_{i,j}, x_{i,j-1}, \underline{\phi})$ from equation 3 are replaced by $P(x_{i,j}|u_{i,j}, \underline{\phi})$, which states that $x_{i,j}$ and $x_{i,j-1}$ are conditionally independent given $u_{i,j}$ and model parameter $\underline{\phi}$. This conditional independence of the traditional Sunflower HMM is responsible for the occasionally erroneous identification of repeats instead of putative TFBSs, and it is this conditional independence assumption that we drop in the extended Sunflower HMM.

Figure 2 shows the additional conditional dependences among adjacent nucleotides by red and blue arrows. Red arrows represent conditional dependences within TFBSs and within reverse complementary TFBSs, and blue arrows represent conditional dependences within flanking regions. We assume that nucleotides in TFBSs or reverse complementary TFBSs are independent of nucleotides in flanking regions and vice versa, so there are no arrows between TFBSs and flanking regions or between reverse complementary TFBSs and flanking regions.

We denote the probability of emitting symbol a in state f_1 by $\lambda_{1,a}$, the conditional probability of emitting nucleotide b in state f_2 given that the previous nucleotide is a by $\lambda_{2,a,b}$, and all of these model parameters by $\underline{\lambda}$. In analogy to $\underline{\lambda}$, we denote the probability of emitting nucleotide a in state m_1 by $\theta_{1,a}$, the conditional probability of emitting nucleotide b in state m_k given nucleotide a emitted by state m_{k-1} by $\theta_{k,a,b}$, for $k \in \{2, \ldots, M\}$, and all of these model parameters by $\underline{\theta}$. These parameters are equivalent to the parameters of the weight array model of [14], i.e., to those of an inhomogeneous first-order Markov model.

The emission probabilities of states $m_{\overline{1}}, \ldots, m_{\overline{M}}$ corresponding to the reverse complemen-

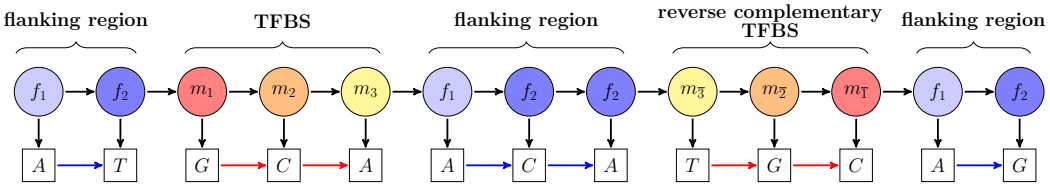

Figure 2 Extended Sunflower HMM for an example sequence of length 13 bp. The sequence contains a 3 bp long TFBS at positions 3-5 and its reverse complement at positions 9-11. The circles denote the values of the hidden states and the boxes the emitted nucleotides. Black arrows encode the dependences present in traditional Sunflower HMMs. Colored arrows encode the additional dependences modeled by the extended Sunflower HMMs: red arrows represent dependences within TFBSs, and blue arrows represent conditional dependences within flanking regions.

tary TFBS can be computed as a function of $\underline{\theta}$ by

$$\begin{aligned} P(x_{i,j} = a | u_{i,j} = m_{\overline{M}}, \underline{\theta}) &= \psi_{M,\overline{a}} \\ P(x_{i,j} = b | u_{i,j} = m_{\overline{k}}, x_{i,j-1} = a, \underline{\theta}) &= \frac{\theta_{k+1,\overline{b},\overline{a}} \cdot \psi_{k,\overline{b}}}{\psi_{k+1,\overline{a}}}, \end{aligned} \tag{4}$$

where $\overline{a}$ denotes the complementary nucleotide to a, and the auxiliary variables $\psi_{k,a}$ are given by the recursion

$$\begin{aligned} \psi_{1,a} &= \theta_{1,a} \\ \psi_{k,a} &= \sum_{b \in \{A,C,G,T\}} \theta_{k,b,a} \cdot \psi_{k-1,b}, \end{aligned} \tag{5}$$

where $k \in \{2, \ldots, M\}$.

We denote $\underline{\phi} = (\underline{\theta}, \underline{\lambda})$, and by plugging equations 2 and 3 into equation 1, we obtain the definition of the extended Sunflower HMM with model parameters π and $\underline{\phi}$.

2.2 Learning

In this section we describe how the model parameters π, $\underline{\theta}$, and $\underline{\lambda}$ of the extended Sunflower HMM can be learned and derive the corresponding Baum-Welch algorithm.

Model parameter π encodes the expected frequency with which TFBSs occur in a data set, and we allow the user to externally set this intuitive model parameter. Likewise, we treat the model parameter of the flanking regions $\underline{\lambda}$ as fixed, and we set it as maximum-likelihood estimator of a homogeneous first-order Markov model estimated from the entire data set. The reason for not learning $\underline{\lambda}$ via the Baum-Welch algorithm is that dynamically learning $\underline{\lambda}$ requires computing the sufficient statistics over almost the entire data set, which is unnecessarily time consuming given that the difference to the sufficient statistics of the full data set is only small, since the number of nucleotides in flanking regions and the number of nucleotides in the entire data set differ only slightly.

In analogy to [7] and many other *de-novo* motif discovery algorithms, we estimate the model parameter $\underline{\theta}$ by a maximum-likelihood approach. To this end, we derive a Baum-Welch algorithm [10] for the extended Sunflower HMM, which is a special case of the EM algorithm [3]. Formally, the Baum-Welch algorithm consists of two steps, which we will call E step and M step in analogy to the EM algorithm. The algorithm iterates between computing the expected sufficient statistics from the current values of the model parameters in the E step and computing the model parameters that maximize the log-likelihood of these expected values in the M step.

Here, we denote the conditional probability that nucleotide a is emitted in state m_1 or the complementary nucleotide $\overline{a}$ is emitted in state $m_{\overline{M}}$ by $\gamma_{1,a}$, and we denote the conditional probability that nucleotide b is emitted in state m_k given the previous nucleotide a or the reverse complementary nucleotide $\overline{b}$ is emitted in state $m_{\overline{k}}$ given the following nucleotide $\overline{a}$ by $\gamma_{k,a,b}$ for $k \in \{2, \ldots, M\}$. In the E step we compute $\gamma_{1,a}$ and $\gamma_{k,a,b}$ by using the current estimate of model parameter $\underline{\theta}^{(t)}$ by

$$\begin{aligned}
\gamma_{1,a}^{(t)} &= \sum_{i=1}^{N}\sum_{j=2}^{L_i} P(u_{i,j} = m_1|\underline{x}_i, \underline{\theta}^{(t)}, \pi, \underline{\lambda})\delta_{x_{i,j},a} \\
&\quad + \sum_{i=1}^{N}\sum_{j=2}^{L_i} P(u_{i,j} = m_{\overline{M}}|\underline{x}_i, \underline{\theta}^{(t)}, \pi, \underline{\lambda})\delta_{x_{i,j},\overline{a}} \\
\gamma_{k,a,b}^{(t)} &= \sum_{i=1}^{N}\sum_{j=2}^{L_i} P(u_{i,j} = m_k|\underline{x}_i, \underline{\theta}^{(t)}, \pi, \underline{\lambda})\delta_{x_{i,j-1},a}\delta_{x_{i,j},b} \\
&\quad + \sum_{i=1}^{N}\sum_{j=2}^{L_i} P(u_{i,j-1} = m_{\overline{k}}|\underline{x}_i, \underline{\theta}^{(t)}, \pi, \underline{\lambda})\delta_{x_{i,j-1},\overline{b}}\delta_{x_{i,j},\overline{a}}.
\end{aligned} \tag{6}$$

In the M step we use the conditional probabilities from the E step to compute the next estimate of model parameter $\underline{\theta}^{(t+1)}$ by

$$\begin{aligned}
\theta_{1,a}^{(t+1)} &= \frac{\gamma_{1,a}^{(t)}}{\sum\limits_{a\in\{A,C,G,T\}} \gamma_{1,a}^{(t)}} \\
\theta_{k,a,b}^{(t+1)} &= \frac{\gamma_{k,a,b}^{(t)}}{\sum\limits_{b\in\{A,C,G,T\}} \gamma_{k,a,b}^{(t)}}.
\end{aligned} \tag{7}$$

For the computation of $P(u_{i,j} = k|\underline{x}_i, \underline{\theta}, \underline{\lambda})$ needed by each E step, we derive a Forward-Backward algorithm for the extended Sunflower HMM. First, we compute the forward variables $\alpha_{j,k}^i = P(x_{i,1}, \ldots, x_{i,j}, u_{i,j} = k|\underline{\theta}, \underline{\lambda})$ for $k \in \{m_1, \ldots, m_M, m_{\overline{1}}, \ldots, m_{\overline{M}}, f_1, f_2\}$ by the recursion

$$\begin{aligned}
\alpha_{1,k}^i &= P(x_{i,1}|u_{i,1} = k, \underline{\theta}, \underline{\lambda})\delta_{k,f_1} \\
\alpha_{j,k}^i &= \sum_l \alpha_{j-1,l}^i P(u_{i,j} = k|u_{i,j-1} = l, \pi) P(x_{i,j}|u_{i,j} = k, x_{i,j-1}, \underline{\theta}, \underline{\lambda}).
\end{aligned} \tag{8}$$

Second, we compute the backward variables $\beta_{j,k}^i = P(x_{i,j+1}, \ldots, x_{i,L_i}|u_{i,j} = k, \underline{\theta}, \underline{\lambda})$ for $k \in \{m_1, \ldots, m_M, m_{\overline{1}}, \ldots, m_{\overline{M}}, f_1, f_2\}$ by the recursion

$$\begin{aligned}
\beta_{L_i,k}^i &= 1 \\
\beta_{j,k}^i &= \sum_l \beta_{j+1,l}^i P(u_{i,j+1} = l|u_{i,j} = k, \pi) P(x_{i,j+1}|u_{i,j+1} = l, x_{i,j}, \underline{\theta}, \underline{\lambda}).
\end{aligned} \tag{9}$$

Finally, we combine the forward and backward variables and obtain

$$P(u_{i,j} = k|\underline{x}_i, \underline{\theta}, \underline{\lambda}) = \frac{\alpha_{j,k}^i \beta_{j,k}^i}{\sum_k \alpha_{j,k}^i \beta_{j,k}^i}. \tag{10}$$

The Baum-Welch algorithm iterates the E step and the M step, yielding monotonically increasing log-likelihoods, and we terminate the algorithm when the difference of two subsequent log-likelihoods falls below $\varepsilon = 10^{-6}$. Typically, the algorithm reaches different local

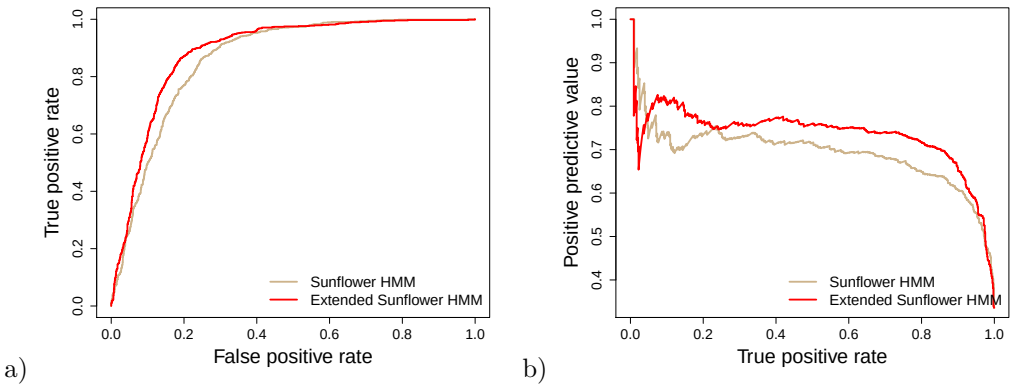

Figure 3 ROC curves (a) and PR curves (b) for the classification on the CREB1 data set. We find that in both cases the curves of classifier B, which uses the extended Sunflower HMM, lie above the curves of classifier A, which uses the traditional Sunflower HMM, except for recalls near zero, where the precisions of the traditional Sunflower HMM are greater than the precisions of the extended Sunflower HMM.

maxima or saddle points for different initializations, so we run it multiple times with different initializations and finally select the model parameter $\underline{\theta}$ with the highest log-likelihood.

3 Results

We have implemented the traditional and extended Sunflower HMMs including all algorithms in Java using Jstacs [5]. To assess the efficacy of the traditional and extended Sunflower HMM for the recognition of homotypic CRMs, we perform a classification of ChIP-seq positive versus negative regions based on data of human embryonic cells from the ENCODE project. We use the data for the six transcription factors CREB1, SP1, GABP, TEAD4, USF1, and YY1 from the HAIB TFBS track of the UCSC Genome Browser [1]. We select genomic regions covered by peaks with a score above 200 as positive sequences and the adjacent genomic regions of the same length as negative sequences. We split the positive and negative data sets for each transcription factors in two subsets, one for training and one for testing, at a ratio of 2:1.

We build two classifiers as follows: classifier A combines a traditional Sunflower HMM as foreground model for the positive sequences with a homogeneous Markov model of order 0 as background model for the negative sequences, while classifier B combines an extended Sunflower HMM for the positive sequences with a homogeneous Markov model of order 1 for the negative sequences.

For each transcription factor, we estimate the parameters of the homogeneous Markov models of order 0 and 1 by maximum likelihood from the union of positive and negative training data sets. We use these values as parameters of the background models of the classifiers and also as parameter $\underline{\lambda}$ of the Sunflower HMMs.

We learn the model parameters $\underline{\theta}$ on the positive data set by applying the traditional and extended Baum-Welch algorithm of Section 2.2. As a consequence, the only features discriminating between positive and negative sequences in each classifier are TFBSs in the

[1] `http://genome.ucsc.edu/cgi-bin/hgTables`

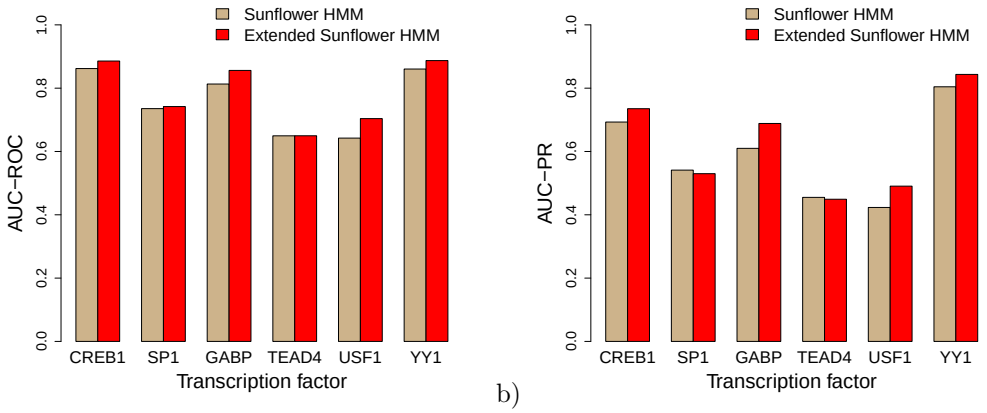

Figure 4 Classification results for six transcription factors. We classify the test data sets for CREB1, SP1, GABP, TEAD4, and USF1 using the two classifiers A and B trained on the training data sets. We show results of classifier A based on the traditional Sunflower HMM in light brown and those of classifier B based on the extended Sunflower HMM in red. Figure (a) shows the results in the area under the receiver operating characteristic curve (AUC-ROC), and Figure (b) shows the area under the precision-recall curve (AUC-PR).

foreground model. We may thus reason that a foreground model with a better classification performance has recognized TFBSs more accurately, so the classification performance may be regarded a measure of accuracy of recognition of homotypic CRMs from ChIP-seq data.

We compute the receiver operating characteristic (ROC) curves and the precision-recall (PR) curves for all six pairs of data sets. The ROC curve shows the true positive rate, also known as recall, as a function of false negative rate. The true positive rate is defined as the ratio of true positives and all positives. Analogously, the false positive rate is the ratio of false positives and all negatives. The PR curve shows the positive predictive value, also known as precision, as a function of the true positive rate. The positive predictive value is the ratio of true positives and the sum of true positives and false positives. Figure 3 shows the ROC and PR curves for the CREB1 data set. Both the ROC curves and the PR curves indicate that taking into account dependences among adjacent nucleotides leads to an improved recognition of CREB1 binding sites.

For calculating the overall classification performance we use the area under the ROC curve (AUC-ROC) and the area under the PR curve (AUC-PR). Figure 4 shows the AUC-ROC and AUC-PR values for both classifiers and each of the six transcription factors. The first pair of columns in each figure corresponds to the area under the curves shown in Figure 3. For CREB1, we observe that classifier B increases the AUC-ROC by 0.02 and the AUC-PR by 0.04 over classifier A. For the remaining transcription factors, we observe that the extended Sunflower HMM achieves higher AUC-ROC values and higher AUC-PR values than the traditional Sunflower HMM also for GABP, USF1, and YY1, whereas we do not observe an improved recognition of homotypic CRMs by taking into account dependences for SP1 and TEAD4.

4 Conclusions

In this work, we have extended the Sunflower HMM for homotypic CRMs by allowing statistical dependences among adjacent nucleotides within TFBSs and flanking regions. We

have derived a modified Baum-Welch algorithm including modified forward and backward algorithms, and we have found by case studies on ChIP-seq data that this extension improves the recognition of TFBSs for four out of six studied transcription factors.

However, this work is limited in several aspects. First, we have considered only one motif type and only first-order dependences. Second, the learning algorithm is based on the maximum likelihood principle, which neglects prior knowledge. Despite these limitations, the extended homotypic Sunflower HMM presented here might possibly be a useful starting point for the reliable recognition of CRMs. Further promising extensions could involve Bayesian or discriminative learning approaches or a generalization of the model to heterotypic CRMs, to higher-order nucleotide dependences.

Acknowledgements We thank Jesús Cerquides Bueno, Andreas Gogol-Döring, and Jan Grau for valuable discussions and DFG (grant no. GR 3526/2-1) and *Reisestipendium des allg. Stiftungsfonds der MLU Halle-Wittenberg* for financial support.

References

1 Yoseph Barash, Gal Elidan, Nir Friedman, and Tommy Kaplan. Modeling dependencies in protein-DNA binding sites. pages 28–37. ACM Press, 2003.

2 Martha L. Bulyk, Philip L. F. Johnson, and George M. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Research*, 30(5):1255–1261, March 2002.

3 Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

4 Valer Gotea, Axel Visel, John M. Westlund, Marcelo A. Nobrega, Len A. Pennacchio, and Ivan Ovcharenko. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Research*, 20(5):565–577, May 2010.

5 Jan Grau, Jens Keilwagen, André Gohr, Berit Haldemann, Stefan Posch, and Ivo Grosse. Jstacs: A java framework for statistical analysis and classification of biological sequences. *Journal of Machine Learning Research*, (13):1967–1971, 2012.

6 Michael M. Hoffman and Ewan Birney. An effective model for natural selection in promoters. *Genome research*, 20(5):685–692, May 2010.

7 Charles E. Lawrence and Andrew A. Reilly. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7(1):41–51, 1990.

8 Alexander P. Lifanov, Vsevolod J. Makeev, Anna G. Nazina, and Dmitri A. Papatsenko. Homotypic regulatory clusters in Drosophila. *Genome Res*, 13(4):579–588, 2003.

9 Tsz-Kwong Man and Gary D. Stormo. Non-independence of Mnt repressor–operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Research*, 29(12):2471–2478, June 2001.

10 Lawrence R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16, Jan 1986.

11 The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, June 2007.

12 Gert Thijs, Magali Lescot, Kathleen Marchal, Stephane Rombauts, Bart De Moor, Pierre Rouzé, and Yves Moreau. A higher-order background model improves the detection of

promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 17(12):1113–1122, December 2001.

13 Peter Van Loo and Peter Marynen. Computational methods for the detection of *cis*-regulatory modules. *Briefings in Bioinformatics*, 10(5):509–524, 2009.

14 M. Q. Zhang and T. G. Marr. A weight array method for splicing signal analysis. *Comput Appl Biosci*, 9(5):499–509, October 1993.