Supervised posteriors for DNA-motif classification

Jan Grau[1], Jens Keilwagen[2], Alexander Kel[3], Ivo Grosse[1,2], and Stefan Posch[1]
[1]*Institute of Computer Science, University Halle, Germany*
[2]*Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany*
[3]*BIOBASE GmbH, Wolfenbüttel, Germany*

Abstract: Markov models have been proposed for the classification of DNA-motifs using generative approaches for parameter learning. Here, we propose to apply the discriminative paradigm for this problem and study two different priors to facilitate parameter estimation using the maximum supervised posterior. Considering seven sets of eukaryotic transcription factor binding sites we find this approach to be superior employing area under the ROC curve and false positive rate as performance criterion, and better in general using sensitivity. In addition, we discuss potential reasons for the improved performance.

1 Introduction

The elucidation of gene regulation is one of the main challenges in functional genomics. One fundamental prerequisite for a gene to be transcribed, or its transcription to be repressed, is the binding of transcription factors (TFs) to their bindings sites (TFBSs) in the promoter region of the gene. Binding of TFs is facilitated by short DNA motifs of typically 10-20 bp length, which show a considerable degree of variation between different TFBSs of the same TF. The detection of TFBSs within a promoter region may be re-formulated as the problem of classifying each subsequence of fixed length.

A wide range of techniques for predicting TFBSs employ statistical models. A successful application of these models requires a problem-specific choice of (i) an appropriate model family for motifs and non-motifs, called background, and (ii) an appropriate training procedure for estimating the model parameters from data sets of known TFBSs and background sequences. Markov models (MMs) have been successfully used for predicting and discovering TFBSs [KGR$^+$03, T$^+$05], cis-regulatory modules [BNP$^+$02], and other DNA motifs [ZM93, Sal97], and so we use Markov models in this paper for predicting TFBSs for seven eukaryotic TFs.

The *generative* approach ([Bis06]) including maximum likelihood (ML) and maximum a posteriori (MAP) is commonly used for parameter estimation. Generally speaking, it aims at an accurate description of the distribution of nucleotides within the TFBSs and within the background. Technically, this results in a separate estimation of parameters for both classes of DNA sequences. This approach is called generative because the resulting distributions allow, amongst others, to generate TFBSs and background sequences from a probabilistic model. In contrast, the *discriminative* approach focuses on the problem of discriminating between sequences of both classes. The resulting distributions are not intended to be accurate descriptions of the true distributions within each class. However, the discriminative approach has often shown a superior classification performance. One example is the maximum conditional likelihood (MCL) principle, which has been applied successfully to Bayesian network classifiers and Markov models for a wide range of data

In this paper, we use inhomogeneous Markov models [ZM93, Sal97] for modeling the class-conditional likelihood of DNA sequences. For a Markov model of order d ($\mathrm{MM}(d)$) each observation at position l may depend only on its $d_l = \min\{d, l-1\}$ predecessors, resulting in

$$P_{\mathrm{MM}(d)}(\boldsymbol{x}|c,\boldsymbol{\theta}) = \prod_{l=1}^{L} P_l(x_l|x_{l-d_l},\ldots,x_{l-1},c,\boldsymbol{\theta}) = \prod_{l=1}^{L} \theta_{l,x_l|c,x_{l-d_l},\ldots,x_{l-1}}. \quad (2)$$

The observations $x_{l-d_l},\ldots,x_{l-1}$ are called the context of position l, which is empty for $l = 1$. In addition to the conditional probabilities $\theta_{l,x_l|c,x_{l-d_l},\ldots,x_{l-1}}$, which constitute the parameters of the Markov model, we denote the prior probability of class c by $\theta_c = P(c|\boldsymbol{\theta})$. A Markov model of order $d = 0$ is equivalent to a position weight matrix (PWM) model [SSGE82, Sta84], which assumes all L positions to be conditionally independent given the class.

The ML estimates of the parameters of a Markov model are the relative frequencies observed in the data set, i.e. $\hat{\theta}^{\mathrm{ML}}_{l,a|c,\boldsymbol{b}} = \frac{n_{l,a|c,\boldsymbol{b}}}{n_{l,\cdot|c,\boldsymbol{b}}}$, $a \in A, \boldsymbol{b} \in A^{d_l}$ where A is the alphabet and $n_{l,a|c,\boldsymbol{b}}$ is the observed absolute frequency of symbol a at position l given context $\boldsymbol{b}$ of the predecessors and class c. In addition we have $\hat{\theta}^{\mathrm{ML}}_{c} = \frac{n_c}{N}$, where n_c is the number of sequences of class c.

2.3 Maximum conditional likelihood

The discriminative analogue of the ML principle is the maximum conditional likelihood (MCL) principle,

$$\hat{\boldsymbol{\theta}}^{\mathrm{MCL}} = \operatorname*{argmax}_{\boldsymbol{\theta}} P(\boldsymbol{c}|\boldsymbol{D},\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{n=1}^{N} P(c_n|\boldsymbol{x}_n,\boldsymbol{\theta}) \quad (3)$$

$$= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \log P(c_n|\boldsymbol{x}_n,\boldsymbol{\theta}), \quad (4)$$

which has been successfully applied to Bayesian network classifiers [WGR+02, GSSZ05, GD04] and Markov models [YSH05]. The MCL principle is more directly linked to the classification rule (1) than the ML principle because it focuses on the posterior probabilities $P(c_n|\boldsymbol{x}_n,\boldsymbol{\theta})$. For maximizing the conditional likelihood, the posterior probabilities are expressed in terms of the class-conditional and prior probabilites,

$$P(c|\boldsymbol{x},\boldsymbol{\theta}) = \frac{P(\boldsymbol{x}|c,\boldsymbol{\theta})P(c|\boldsymbol{\theta})}{\sum_{\tilde{c}\in C} P(\boldsymbol{x}|\tilde{c},\boldsymbol{\theta})P(\tilde{c}|\boldsymbol{\theta})}. \quad (5)$$

In contrast to ML estimators, MCL estimators cannot be obtained analytically for several popular models including Markov models. Hence, numerical optimization techniques, such as gradient ascent, are used for the MCL estimation of $\hat{\boldsymbol{\theta}}^{\mathrm{MCL}}$. Unfortunately, neither the conditional likelihood nor the log conditional likelihood are concave functions of $\boldsymbol{\theta}$ [WGR+02]. Hence, numerical optimization techniques often converge only to local maxima or saddle points. To solve this problem, an alternative parameterization is proposed

in [WGR+02] which has also been used for general Bayesian networks [GSSZ05]: Using new parameters $\boldsymbol{\beta}$ the following functions Q are defined:

$$Q_{\mathrm{MM}(d)}(c, \boldsymbol{x}|\boldsymbol{\beta}) \quad = \quad \exp\left(\beta_c + \sum_{l=1}^{L} \beta_{l,x_l|c,x_{l-d_{l,c}},\ldots,x_{l-1}}\right) \tag{6}$$

where $d_{l,c}$ is the order of the Markov model of class c at position l.
Choosing $\beta_{l,x_l|c,x_{l-d_{l,c}},\ldots,x_{l-1}} = \log \theta_{l,x_l|c,x_{l-d_{l,c}},\ldots,x_{l-1}}$ it is easy to verify, that $Q_{\mathrm{MM}(d)}(c, \boldsymbol{x}|\boldsymbol{\beta}) = P_{\mathrm{MM}(d)}(\boldsymbol{x}|c, \boldsymbol{\beta})P(c|\boldsymbol{\beta})$.
Inserting (6) into (4) the log conditional likelihood in the $\boldsymbol{\beta}$-parameterization is given as:

$$\log P(\boldsymbol{c}|\boldsymbol{D}, \boldsymbol{\beta}) = \sum_{n=1}^{N} \left[\log Q_{\mathrm{MM}(d_{c_n})}(c_n, \boldsymbol{x}_n|\boldsymbol{\beta}) - \log\left(\sum_{\tilde{c}\in C} Q_{\mathrm{MM}(d_{\tilde{c}})}(\tilde{c}, \boldsymbol{x}_n|\boldsymbol{\beta})\right)\right] \tag{7}$$

As [WGR+02] prove, the log conditional likelihood is a concave function of $\boldsymbol{\beta} \in \mathrm{R}^{|\boldsymbol{\beta}|}$ for chordal graphs, which are a subclass of Bayesian networks, and which include Markov models. For a two class problem, this property also follows from the relation to logistic regression [WGR+02, NJ02, GSSZ05, FI06], because logistic regression results in a concave objective function [Min03].

We reduce the number of parameters by using a modification of the $\boldsymbol{\beta}$-parameterization proposed by [MP99]. This modification exploits that only $A-1$ of the A parameters at any position possibly given one or more predecessors are free parameters. Without loss of generality we choose the last parameter $\beta_{l,|A||c,\boldsymbol{b}}$ not to be free. In the parameterization of [MP99] this corresponds to fixing this parameter to 0. This reduction of the number of parameters does not affect the concavity of the conditional likelihood, because we consider linear sub-spaces of the full space of parameters $\boldsymbol{\beta}$. Additionally, we can show that for any admissible parameter $\boldsymbol{\theta}$ we find corresponding parameters in the reduced $\boldsymbol{\beta}$ space defining

$$\beta_{l,a|c,\boldsymbol{b}} = \log \frac{\theta_{l,a|c,\boldsymbol{b}}}{\theta_{l,|A||c,\boldsymbol{b}}}, \quad \beta_c = \log \frac{\theta_c}{\theta_{|C|}}. \tag{8}$$

We use the parameterization of [MP99] for all of the models and training approaches in the rest of the paper. It can be shown that the ML estimates of both parameterizations coincide.

2.4 Maximum a posteriori

The maximum a-posteriori (MAP) principle is another common principle for generative parameter learning. In this case, the objective is to choose those parameters $\boldsymbol{\beta}$ that maximize the posterior $P(\boldsymbol{\beta}|\boldsymbol{D}, \boldsymbol{c})$. Decomposing the posterior yields

$$\hat{\boldsymbol{\beta}}^{\mathrm{MAP}} = \operatorname*{argmax}_{\boldsymbol{\beta}} P(\boldsymbol{\beta}|\boldsymbol{D}, \boldsymbol{c}, \boldsymbol{\alpha}) = \operatorname*{argmax}_{\boldsymbol{\beta}} P(\boldsymbol{D}, \boldsymbol{c}|\boldsymbol{\beta})P(\boldsymbol{\beta}|\boldsymbol{\alpha}),$$

where $\boldsymbol{\alpha}$ denotes the hyperparameters of the prior $P(\boldsymbol{\beta}|\boldsymbol{\alpha})$.

sets, e.g. data sets from the UCI machine learning repository [NJ02, RWG+05, GSSZ05], text categorization and protein sequences [YSH05].

In [NJ02] it has been shown for a range of data sets that the performance of MCL classifiers diminishes as the size of the training data available decreases. This demands for approaches employing priors on the parameters in a similar manner as e.g. the MAP approach does for generative learning. Such an approach, called maximum supervised posterior (MSP) approach, has been proposed by [WGR+02, GKM+02, CdM05]. To the best of our knowledge we are among the first who apply MSP to bioinformatical problems. Here, we study if this approach could possibly be useful for the recognition of eukaryotic TFBSs.

2 Methods

In this section we introduce the statistical background and the different principles for learning the parameters of the models.

2.1 Classification

The well-known Bayes classifier assigns a sequence $\boldsymbol{x} = x_1 x_2 \ldots x_L$ of length L to class $c^* \in C$ using

$$c^* = \operatorname*{argmax}_{c \in C} P(c|\boldsymbol{x}) = \operatorname*{argmax}_{c \in C} P(c, \boldsymbol{x}), \tag{1}$$

where $P(c|\boldsymbol{x})$ denotes the posterior probability of class c given sequence $\boldsymbol{x}$, and $P(c, \boldsymbol{x})$ denotes the joint probability.

To apply this classification rule, either the posterior or the joint distribution must be determined. Typically, an appropriate family of distributions is chosen, and its parameters $\boldsymbol{\theta}$ are inferred from the data. We assume a data set of N independent and identically distributed (i.i.d.) data points $(\boldsymbol{x}_n, c_n)$, and we denote $\boldsymbol{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ and $\boldsymbol{c} = (c_1, \ldots, c_N)$. In the remainder of this section, we consider generative and discriminative approaches for the training of parameters and present their application to Markov models.

2.2 Maximum likelihood

Using the *generative* approach, the popular maximum likelihood (ML) principle suggests to choose those parameters $\boldsymbol{\theta}$ that maximize the likelihood $P(\boldsymbol{D}, \boldsymbol{c}|\boldsymbol{\theta})$ of the complete data set $(\boldsymbol{D}, \boldsymbol{c})$,

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{\mathrm{ML}} &= \operatorname*{argmax}_{\boldsymbol{\theta}} P(\boldsymbol{D}, \boldsymbol{c}|\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{n=1}^{N} P(\boldsymbol{x}_n, c_n|\boldsymbol{\theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{c \in C} P(c|\boldsymbol{\theta}) \prod_{n, \text{where } c_n = c} P(\boldsymbol{x}_n|c, \boldsymbol{\theta}). \end{aligned}$$

Assuming the parameters of the class-conditional likelihoods $P(\boldsymbol{x}|c, \boldsymbol{\theta})$ to be pairwise independent allows to determine the ML estimate for each class separately. This approach is called generative because it aims at an accurate estimation of the underlying probabilities $P(\boldsymbol{x}|c, \boldsymbol{\theta})$.

For an inhomogeneous Markov model, we choose a transformed Dirichlet prior, because it is conjugate to the likelihood represented by Markov models. This results in the MAP estimates $\hat{\beta}_{l,a|c,\boldsymbol{b}} = \log \frac{n_{l,a|c,\boldsymbol{b}}+\alpha_{l,a|c,\boldsymbol{b}}}{n_{l,|A||c,\boldsymbol{b}}+\alpha_{l,|A||c,\boldsymbol{b}}}$, and $\hat{\beta}_c = \log \frac{n_c+\alpha_c}{n_{|C|}+\alpha_{|C|}}$ [MP99], We choose the hyperparameters $\alpha_{l,a|c,\boldsymbol{b}} = \frac{\text{ess}_c}{A^{d_l+1}}$, $\alpha_c = \text{ess}_c$, where ess_c denotes the *equivalent sample size* of class c [Bun91]. We choose $\text{ess}_{\text{fg}} = 16$ and $\text{ess}_{\text{bg}} = 256$. The hyperparameters $\boldsymbol{\alpha}$ can be interpreted as pseudo counts stemming from uniformly distributed pseudo data within each class. Another view on pseudo counts is that they compensate for zero frequencies. These are often encountered when only a limited amount of training data is available, such as in the case of TFBSs.

2.5 Maximum supervised posterior

The maximum supervised posterior (MSP) principle [CdM05, WGR$^+$02, GKM$^+$02] suggests using a prior for discriminative learning in the same way as the MAP principle suggests using a prior for generative learning. The MSP principle closely resembles the transition from the ML principle to the MAP principle, multiplying the conditional likelihood (4) by a prior $P(\boldsymbol{\beta}|\boldsymbol{\alpha})$,

$$\hat{\boldsymbol{\beta}}^{\text{MSP}} = \underset{\boldsymbol{\beta}}{\text{argmax}}\, P(\boldsymbol{c}|\boldsymbol{D},\boldsymbol{\beta},\boldsymbol{\alpha})P(\boldsymbol{\beta}|\boldsymbol{\alpha}). \tag{9}$$

One technical advantage of MSP estimators over MCL estimators is that they compensate for zero frequencies. In the $\boldsymbol{\beta}$-parameterization, zero frequencies result in parameters approaching $\pm\infty$, which also causes numerical problems.

Here, we propose to use two different priors in conjunction with the MSP principle for Markov models of different orders, namely a Gaussian prior and a Laplace prior, which are used for logistic regression [MGL$^+$05, CTG07, GLM05] and maximum entropy models [CR99].

We assume all parameters of $\boldsymbol{\beta}$ to be statistically independent, i.e., we choose as prior a product of univariate densities for each parameter. For the Gaussian prior, we denote the vector of the means by $\boldsymbol{\mu}$ and the vector of the variances by $\boldsymbol{\sigma}^2$, resulting in

$$\begin{aligned} P(\boldsymbol{\beta}|\boldsymbol{\mu},\boldsymbol{\sigma}^2) \quad = \quad & \prod_{c=1}^{|C|-1} \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{1}{2}\frac{(\beta_c-\mu_c)^2}{\sigma_c^2}\right) \\ & \cdot \prod_{c=1}^{|C|} \prod_{l=1}^{L} \prod_{\boldsymbol{b}\in A^{d_l}} \prod_{a=1}^{|A|-1} \frac{1}{\sqrt{2\pi\sigma_{l,a|c,\boldsymbol{b}}^2}} \exp\left(-\frac{1}{2}\frac{(\beta_{l,a|c,\boldsymbol{b}}-\mu_{l,a|c,\boldsymbol{b}})^2}{\sigma_{l,a|c,\boldsymbol{b}}^2}\right). \end{aligned}$$

The parameters $\beta_{l,|A||c,\boldsymbol{b}}$ and $\beta_{|C|}$ do not need to be considered here, because they are fixed to 0.

We determine the hyper-parameters σ_c^2 and μ_c for the classes from prior knowledge about the occurrence of the DNA-motifs of interest. This will be explained in detail in section 3. The means $\mu_{l,a|c,\boldsymbol{b}}$ for the parameters of observational random variables are set to 0. This corresponds to the a-priori assumption that all symbols at every position occur with the same probability. The same assumption was employed for the Dirichlet prior for MAP estimation.

We define the variances as $\sigma^2_{l,a|c,\boldsymbol{b}} = \kappa_c A^{|\boldsymbol{b}|+1} = \kappa_c A^{d_l+1}$. The rationale behind this heuristic is the assumption that the variance of the parameter prior increases exponentially with the (local) order d_l. This assumption stems from the intuition that, on average, the number of samples per parameter decreases exponentially with the order of the model. Consequently, the effect of the prior on the parameters increases with increasing order, which may be balanced by a higher variance. This is again in analogy to the abovementioned choice of the transformed Dirichlet prior. Additionally, we assume that a deviation from the mean of $\mathbf{0}$ becomes more likely with increasing order. Since we do not have a-priori knowledge about the values of κ_{fg} and κ_{bg}, we will choose their values in a prestudy described in section 3.

The Laplace prior is defined as

$$\begin{aligned} P(\boldsymbol{\beta}|\boldsymbol{\mu},\boldsymbol{b}) &= \prod_{c=1}^{|C|-1} \frac{1}{2b_c} \exp\left(-\frac{|\beta_c - \mu_c|}{b_c}\right) \\ &\quad \cdot \prod_{c=1}^{|C|} \prod_{l=1}^{L} \prod_{\boldsymbol{b}\in A^{d_l}} \prod_{a=1}^{|A|-1} \frac{1}{2b_{l,a|c,\boldsymbol{b}}} \exp\left(-\frac{|\beta_{l,a|c,\boldsymbol{b}} - \mu_{l,a|c,\boldsymbol{b}}|}{b_{l,a|c,\boldsymbol{b}}}\right). \end{aligned}$$

We choose $\boldsymbol{\mu}$, b_c and $b_{l,a|c,\boldsymbol{b}}$ such that the Laplace prior has the same mean vector and the same vector of variances as the Gaussian prior, resulting in $b_c = \sqrt{\sigma_c^2/2}$ and $b_{l,a|c,\boldsymbol{b}} = \sqrt{\sigma^2_{l,a|c,\boldsymbol{b}}/2}$. The Laplace prior entails two properties that are disadvantageous from a theoretical point of view: its logarithm is not strictly concave, but only concave, and its derivative with respect to any of the βs is discontinuous at its maximum. We consider both disadvantages relatively mild for numerical optimization, because the first at worst results in a slower convergence, and the second is only relevant if we exactly hit the maximum, which will almost never be the case.

3 Results and Discussion

In this section we compare the classification accuracy of generatively and discriminatively trained models for the TFBSs of seven eukaryotic TFs.

3.1 Data

We consider seven sets of vertebrate TFBSs of length $L = 16$ collected from the TRANSFAC® database (rel. 8.1, 2004), namely AP1 (112 sequences), AR/GR/PR (104 sequences), C/EBP (149 sequences), GATA (110 sequences), NF1 (96 sequences), Sp1 (257 sequences), and thyroid hormone receptor-like factors (Thyroid, 127 sequences). All sets consist of experimentally verified TFBSs collected from the scientific literature. The majority of the TFBSs stems from human, mouse, and rat and cover three of the four superclasses of TFs: AP1 and C/EBP belong to the class of *basic domain* factors, where the latter contains at least two subfamilies; NF1 belongs to the *beta-scaffold factors with minor grove contacts*; GATA, Sp1, and Thyroid are factors with *zinc-coordinating DNA-binding domains*, and AR/GR/PR comprises three *steroid hormone receptors* from the same class of factors. The background data set consists of 267 sequences from second exons of human genes with $68,141$ bp in total.

3.2 Analyses

We use three measures for the accuracy of a classifier, namely the area under the ROC curve (AUC), the sensitivity ($\mathrm{Sn} = \frac{TP}{TP+FN}$) for a fixed specificity ($\mathrm{Sp} = \frac{TN}{TN+FP}$) of 99.9%, and the false positive rate ($\mathrm{FPR} = 1 - \mathrm{Sp} = \frac{FP}{TN+FP}$) for a fixed sensitivity of 95%. AUC indicates the overall performance of a classifier. Sn measures the fraction of correctly classified foreground sequences if a classifier erroneously predicts one out of 1000 background sequences to be a TFBS. FPR measures the fraction of incorrectly classified background sequences if a classifier correctly predicts 95 out of 100 TFBSs. We use a *k-fold stratified holdout sampling* procedure [BGSG$^+$05] for obtaining these measures in a robust way together with estimates of their standard errors.

In the following analyses, we consider only the MAP and the MSP principle, since the number of binding sites is small for all TFs, and zero frequencies occur even for lower-order MMs, resulting in a low classification accuracy for ML and MCL (data not shown).

3.2.1 Choice of hyper-parameters

To determine appropriate values of the hyperparameters μ_{fg} and σ^2_{fg}, we exploit prior knowledge from a study by Stepanova *et al.* [STSB05], who estimate the relative frequencies of occurrence of 184 different TFs in mammalian genomes. We transform these 184 relative frequencies to the $\boldsymbol{\beta}$-parameter space using (8). Assuming the 184 β-values to be statistically independent realizations of a normal density, we estimate $\mu_{\mathrm{fg}} = -8.634$ and $\sigma^2_{\mathrm{fg}} = 5.082$.

To determine appropriate values of κ_c, we perform a pre-study using the data set of Sp1, which is the largest of the seven data sets of TFBSs. For this set, we perform a grid search on κ_{fg} (0.001 to 5, 12 values) and κ_{bg} (0.0005 to 0.5, 10 values), where we fix the order of the TFBS (foreground) model to $d_{\mathrm{fg}} = 0$ and vary the background order from $d_{\mathrm{bg}} = 0$ to $d_{\mathrm{bg}} = 3$. For each combination we use a 100-fold stratified holdout sampling procedure to determine the resulting AUC. For each pair $(\kappa_{\mathrm{fg}}, \kappa_{\mathrm{bg}})$, we then compute the mean AUC over all background orders and choose that $(\kappa^*_{\mathrm{fg}}, \kappa^*_{\mathrm{bg}})$ which yields the maximum AUC. We choose AUC as the measure of accuracy, expecting AUC to be more stable than Sn or FPR, as it integrates over the complete ROC curve. This results in $\kappa^*_{\mathrm{fg}} = 2$ and $\kappa^*_{\mathrm{bg}} = 0.005$ for the Gaussian prior and $\kappa^*_{\mathrm{fg}} = 0.005$ and $\kappa^*_{\mathrm{bg}} = 0.002$ for the Laplace prior. We use these values of the κ_{fg} and κ_{bg} in all further analyses, which implicates that the results for Sp1 and the results for AUC are biased by the pre-study.

3.2.2 Comparison of MAP and MSP

Based on the results of the pre-study, we compare the accuracy of MAP, MSP with Gaussian prior (MSP-G), and MSP with Laplace prior (MSP-L) for each of the seven TFs. We employ MMs of order $d_{\mathrm{fg}} = 0$ and $d_{\mathrm{fg}} = 1$ as foreground models combined with MMs of order $d_{\mathrm{bg}} = 0$ to $d_{\mathrm{bg}} = 4$ as background models. For each of the seven data sets, each of the ten model combinations, and each of the three principles, we record the mean values of the accuracy measures AUC, FPR, and Sn together with their standard errors as obtained from a 1000-fold stratified holdout sampling procedure. We regard a difference of performance as significant if it exceeds twice the standard error.

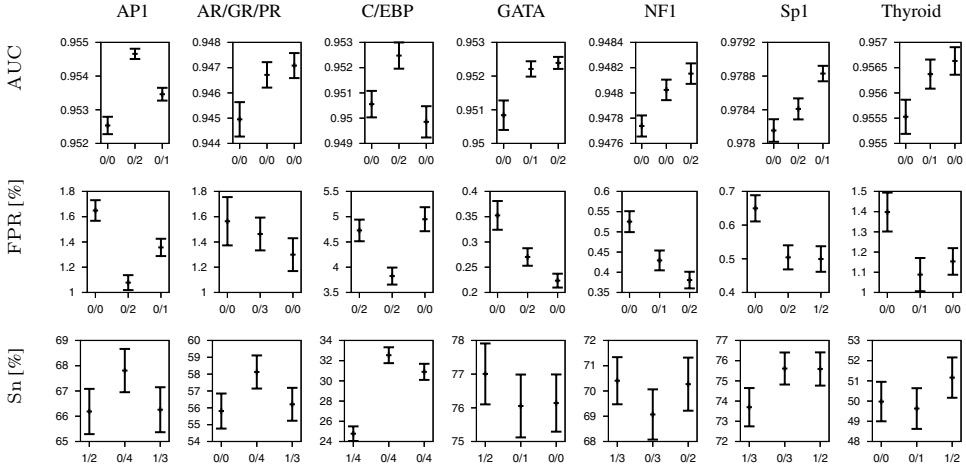

Figure 1: AUC, FPR, and Sn for 7 eukaryotic TFBSs. In each plot, the first column shows the best result for MAP, the second column for MSP-G, and the third column for MSP-L. The optimal orders of the foreground and background model are given on the abscissa as $\mathrm{fg} \,/\, \mathrm{bg}$. The whiskers indicate a deviation by the twofold standard error in each direction.

Figure 1 shows the results for MAP, MSP-G, and MSP-L for each of the seven TFs. Comparing the AUC obtained by MAP and MSP, we observe a significant improvement for MSP for both priors and for each data set, with the exception of MSP-L applied to C/EBP. Comparing the results of MSP-G and MSP-L, we cannot see a clear preference. Out of the seven data sets, MSP-G performs significantly better than MSP-L for AP1 and C/EBP, significantly worse for NF1 and Sp1, and comparable for the remaining three data sets.

For FPR, we see a significant improvement, a decrease in this case, regardless of the prior for five of the seven TFs. For the remaining two data sets, MSP performs better than MAP for one of the priors. Comparing the MSP approaches, again no clear pattern is evident: MSP-G yields a significantly lower FPR compared to MSP-L for AP1 and C/EBP, a significantly higher FPR for AR/GR/PR, GATA, and NF1, and a similar FPR for Sp1 and Thyroid.

Considering Sn, we again see an improvement for many cases, although the pattern is less clear. On the one hand, we observe a significant improvement for both discriminative approaches only for C/EBP and Sp1. On the other hand, MSP-G is superior to MAP for AP1, AR/GR/PR, C/EBP, and Sp1, and MSP-L is superior to MAP for C/EBP, Sp1 and Thyroid. For NF1, only MSP-L performs as well as MAP, whereas, for GATA, MSP-G and MSP-L perform worse than MAP. Interestingly, we see the most impressive improvements in Sn for AR/GR/PR (2.3 %) and C/EBP (7.8 %), which are known to comprise the binding sites of different subfamilies of TFs.

For Sn, and to a minor extend for FPR and AUC, we observe that MSP works especially well for higher model orders for some of the TFBSs. One possible explanation might be that for these TFBSs long-distance dependencies exist, which can be captured by higher-order models, suggesting the use of models that can capture non-adjacent dependencies, such as Bayesian trees, in the future.

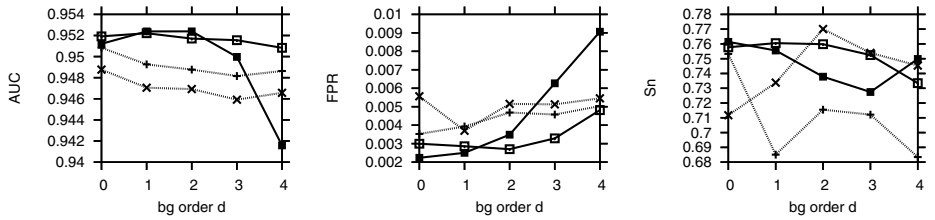

Figure 2: AUC, FPR, and Sn for the GATA set. The models considered are MM(0)/MM(d) (····+····) and MM(1)/MM(d) (····×····) for MAP, and MM(0)/MM(d) for MSP-G (—□—) and MSP-L (—■—).

3.2.3 Priors and orders

In the following we study to which degree the performance of MSP-G and MSP-L may vary from model to model. Although MSP-L yields a higher accuracy than MSP-G on the GATA set considering AUC and FPR using the best combination of models (see figure 1), this could possibly not be the case for all model orders. Figure 2 presents the results of MAP, MSP-G, and MSP-L for different orders of the MMs and this TF. We find that MSP-L is more sensitive to the order of the models employed than MSP-G. Interestingly, the performance of MSP-L significantly decreases for AUC and FPR with increasing order, whereas that of MSP-G stays relatively constant. This observation is in agreement to the observation that MSP-G, in contrast to MSP-L, shows a comparable or a better performance than MAP for any model order considering AUC and FPR. Both observations also hold for the other data sets (data not shown) and suggest the future use of MSP-G for the classification of eukaryotic TFBSs.

3.2.4 Differences between generative and discriminative learning

With the goal to understand to some degree why MSP shows a superior classification performance in many cases, we compare the parameter values obtained by MAP and MSP-G training. We transform the parameter estimates of the $\mathrm{MM}(0)$ into the θ-parametrization and compute the log ratios of the parameters between the foreground and background model. This results in log ratios $\mathrm{lr}(l, a)$ for each position l and symbol a. As we compute these values for the MAP and the MSP-G principle, we obtain two sets of values $\{\mathrm{lr}(l, a)^{\mathrm{MSP-G}}\}$ and $\{\mathrm{lr}(l, a)^{\mathrm{MAP}}\}$. The difference of the corresponding values $\mathrm{d}(l, a) = \mathrm{lr}(l, a)^{\mathrm{MSP-G}} - \mathrm{lr}(l, a)^{\mathrm{MAP}}$ then provides an insight into the reasons of differing classification.

We present the results of this analysis for AR/GR/PR in the lower plot of figure 3, while the upper plot shows the sequence logo of the AR/GR/PR foreground data set [SS90]. Interestingly, we find the most noticeable differences $\mathrm{d}(l, a)$ between the MSP and the MAP classifier for those positions l with the greatest nucleotide conservation according to the sequence logo. We might speculate that these positions are the most important for the binding of AR/GR/PR to its TFBSs. Interestingly, it is exactly these conserved positions on which the MSP-G principle focuses even more strongly than the MAP principle. This might explain the superior performance of the MSP-G principle.
For most of the positions with high nucleotide conservation (7, 8, 10, and 11) the parameters of the MSP-G classifier compared to MAP more strongly tend to the consensus nucleotide (G,T,C, and T, respectively). In figure 3 this shows as large negative differ-

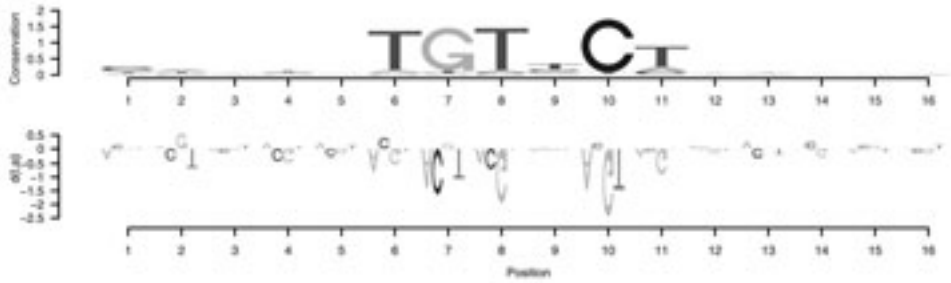

Figure 3: Sequence logos for AR/GR/PR. The upper plot shows the sequence logo (created with `seqLogo` R-package [Bem07]) and the lower plot shows the difference of the log ratios of the parameters of the MSP-G and the MAP classifier at each position.

ences $\mathrm{d}(l, a)$ for the non-consensus nucleotides. As the consensus nucleotides are very conserved at these positions, this can be explained by the general tendency of discriminative principles to concentrate on the differences between the classes. On the other hand, we observe an increased probability for nucleotide C at position 6 compared to the MAP parameters, which cannot be explained by this means, but still might to some extend be responsible for an improved classification performance.

These results are related to those of Mirny and Gelfand [MG02], who find that for protein-DNA interactions the number of contacts between the protein and a nucleotide is highly correlated with its conservation, supporting our speculation. The findings of [MG02] are (amongst others) used by Keleş *et al.* [KvdLD$^+$03] to improve the de-novo identification of TFBSs. As MSP classifiers to some extend seem to focus on positions with nucleotide conservation, it might be worthwhile to apply the MSP approach to the problem of de-novo motif identification.

4 Conclusion

Several approaches exist for improving the computational recognition of TFBSs. One approach, which is very popular as measured by the number of publications over the last decade, focuses on finding more appropriate statistical models for modeling both the motifs and the DNA background. A complementary approach focuses on improving methods for parameter estimation from small data sets. While generative training approaches attempt to model each sequence class (motifs and DNA background) as accurately as possible, discriminative training approaches attempt to choose those parameters that discriminate the sequence classes from each other as accurately as possible.

One discriminative parameter estimation approach proposed in the machine learning community is the maximum supervised posterior principle. In this paper, we study if the recognition of eukaryotic TFBSs can be improved by using the MSP principle for parameter estimation.

We compare the classification accuracy of the MAP and MSP principles applied to seven data sets of eukaryotic TFBSs. As models for motifs and the DNA background, we choose Markov models, which are at the heart of most of the TFBS recognition algorithms. For parameter estimation, we use the MAP principle using transformed Dirichlet priors and the

MSP principle using Gauss (MSP-G) and Laplace (MSP-L) priors. As measures of accuracy, we use the area under the ROC curve, the false positive rate for a fixed sensitivity of 95%, and the sensitivity for a fixed specificity of 99.9%. Performing a 1000-fold stratified holdout sampling procedure, we find that the recognition of TFBSs can be improved significantly for most of the studied data sets and measures of classification accuracy by using the MSP approach in favor of the MAP approach. Although the MSP approach achieves an impressively higher sensitivity for a subset of the studied TFs including AR/GR/PR and C/EBP, we do not see an improvement as measured by the sensitivity for all of the factors. With respect to varying the orders of the Markov models MSP-G is more stable, even though MSP-L yields a higher accuracy than MSP-G in a few cases. In all of the studied cases, MSP-G achieves a comparable or a better classification performance than the MAP approach considering area under curve and false positive rate regardless of the orders of the Markov models. This suggests that the MSP approach using Gaussian priors could be useful for the prediction of other TFBSs or other DNA motifs, such as nucleosomal binding sites, splice sites, or splicing enhancers.

Acknowledgements

We thank André Gohr and Yvonne Pöschl for valuable discussions and the German Ministry of Education and Research (BMBF Grant No. 0312706A/D) for financial support.

References

[Bem07] O. Bembom. seqLogo: An R package for plotting DNA sequence logos. `http://cosmoweb.berkeley.edu/software.html`, January 2007.

[BGSG+05] I. Ben-Gal, A. Shani, A. Gohr, J. Grau, S. Arviv, A. Shmilovici, S. Posch, and I. Grosse. Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics*, 21(11):2657–2666, 2005.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[BNP+02] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci USA*, 99(2):757–762, Jan 2002.

[Bun91] W. L. Buntine. Theory Refinement of Bayesian Networks. In *Uncertainty in Artificial Intelligence*, pages 52–62. Morgan Kaufmann, 1991.

[CdM05] J. Cerquides and R. López de Mántaras. Robust Bayesian Linear Classifier Ensembles. In *ECML*, pages 72–83, 2005.

[CR99] S. Chen and R. Rosenfeld. A Gaussion Prior for Smoothing Maximum Entropy Models. Technical report, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, February 1999.

[CTG07] G. Cawley, N. Talbot, and M. Girolami. Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

[FI06] A. Feelders and J. Ivanovs. Discriminative Scoring of Bayesian Network Classifiers: a Comparative Study. In *Proceedings of the third European workshop on probabilistic graphical models*, pages 75–82, 2006.

[GD04] D. Grossman and P. Domingos. Learning Bayesian network classifiers by maximizing conditional likelihood. In *ICML*, pages 361–368. ACM Press, 2004.

[GKM+02] P. Grünwald, P. Kontkanen, P. Myllymäki, T. Roos, H. Tirri, and H. Wettig. Supervised posterior distributions. Presented at the Seventh Valencia International Meeting on Bayesian Statistics, 2002.

[GLM05] A. Genkin, D. D. Lewis, and D. Madigan. Sparse Logistic Regression for Text Categorization. Project Report, Center for Discrete Mathematics & Theoretical Computer Science, April 2005.

[GSSZ05] R. Greiner, X. Su, B. Shen, and W. Zhou. Structural Extension to Logistic Regression: Discriminative Parameter Learning of Belief Net Classifiers. *Machine Learning Journal*, 59(3):297–322, 2005.

[KGR+03] A. E. Kel, E. Gössling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*, 31(13):3576–3579, July 2003.

[KvdLD+03] S. Keles, M. J. van der Laan, S. Dudoit, B. Xing, and M. B. Eisen. Supervised detection of regulatory motifs in DNA sequences. *Stat Appl Genet Mol Biol*, 2(1), 2003.

[MG02] L. A. Mirny and M. S. Gelfand. Structural analysis of conserved base pairs in protein-DNA complexes. *Nucl. Acids Res.*, 30(7):1704–1711, 2002.

[MGL+05] D. Madigan, A. Genkin, D. D. Lewis, S. Argamon, D. Fradkin, and L. Ye. Author Identification on the Large Scale. In *Joint Annual Meeting of the Interface and the Classification Society of North America*, June 2005.

[Min03] T. P. Minka. Algorithms for maximum-likelihood logistic regression. Technical Report 758, Carnegie Mellon University, Department of Statistics, 2001, revised Sept. 2003.

[MP99] M. Meila-Predoviciu. *Learning with Mixtures of Trees*. PhD thesis, Massachusetts Institute of Technology, 1999.

[NJ02] A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 605–610. MIT Press, Cambridge, MA, 2002.

[RWG+05] T. Roos, H. Wettig, P. Grünwald, P. Myllymäki, and H. Tirri. On Discriminative Bayesian Network Classifiers and Logistic Regression. *Machine Learning*, 59(3):267–296, June 2005.

[Sal97] S. L. Salzberg. A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Comput. Appl. Biosci.*, 13(4):365–376, 1997.

[SS90] T. D. Schneider and R. M. Stephens. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res.*, 18:6097–6100, 1990.

[SSGE82] G. D. Stormo, T. D. Schneider, L. M. Gold, and A. Ehrenfeucht. Use of the 'perceptron' algorithm to distinguish translational initiation sites. *NAR*, 10:2997–3010, 1982.

[Sta84] R. Staden. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research*, 12:505–519, 1984.

[STSB05] M. Stepanova, T. Tiazhelova, M. Skoblov, and A. Baranova. A comparative analysis of relative occurrence of transcription factor binding sites in vertebrate genomes and gene promoter areas. *Bioinformatics*, 21(9):1789–1796, 2005.

[T+05] M. Tompa et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1):137 – 144, 2005.

[WGR+02] H. Wettig, P. Grünwald, T. Roos, P. Myllymäki, and H. Tirri. On Supervised Learning of Bayesian Network Parameters. Technical Report HIIT Technical Report 2002-1, Helsinki Institute for Information Technology HIIT, 2002.

[YSH05] O. Yakhnenko, A. Silvescu, and V. Honavar. Discriminatively Trained Markov Model for Sequence Classification. In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 498–505, Washington, DC, USA, 2005. IEEE Computer Society.

[ZM93] M.O. Zhang and T.G. Marr. A weight array method for splicing signal analysis. *Comput. Appl. Biosci.*, 9(5):499–509, 1993.