Statistical Learning for Detecting Protein-DNA-Binding Sites

Thomas Martinetz, Jan E. Gewehr and Jan T. Kim
Institute for Neuro- and Bioinformatics
University of Lübeck
Seelandstraße 1a, 23569 Lübeck, Germany
Email: martinetz@inb.uni-luebeck.de

***Abstract*— Detecting the sites on genomic DNA at which DNA binding proteins bind is a highly relevant task in bioinformatics. For example, the bindings sites of transcription factors are key elements of regulatory networks and determine the location of genes on a genome. Usually, for a given DNA binding protein, only a few DNA-subsequences at which the protein binds are known experimentally. The task then is to deduce the global binding characteristics of the protein based on these few positive examples. A widespread approach is the so-called Profile-Matrix (PM). The PM-approach can be interpreted as a linear classifier (binding word class/non-binding word class) within the space of sequence words, with the profile of the experimentally verified binding sites determining its parameters. In this paper a novel approach called Binding-Matrix (BM) is introduced. Like the PM, the BM realizes a linear classification, but in contrast to the Profile-Matrix approach the parameters (matrix) of the classifier is now determined by maximum likelihood estimation. Tested on data from the TRANSFAC database, the maximum likelihood estimation leads to an increase in classification performance by about an order of magnitude.**

I. Introduction

A number of different proteins bind to a genome. Some of these DNA-binding proteins only bind to specific locations where they have to execute certain functions. These proteins are called sequence specific DNA-binding proteins. To execute its function at specific sites, a sequence specific DNA-binding protein has to bind at these sites with a binding energy that exceeds a certain threshold. Besides the sites where it has to perform its function there might be sites where the execution of this function might be detrimental. There, this protein may not bind, i.e. the binding energy has to be below the threshold. Then there is a vast majority of sites where the binding of the protein may take place without any consequences.

The DNA-binding domain of the protein interacts with a DNA-segment consisting of L bases $b \in \{A, C, G, T\}$. Such a DNA-segment we call a sequence word. The sequence word $\mathbf{w}_i$ the protein "sees" at site i determines whether i is a binding site for the protein or not. The word length L varies from protein to protein and typically ranges between $L = 10$ and $L = 20$. There are 4^L different words $\mathbf{w}$ or "patterns" of length L (this, however, does not mean necessarily that each of these words appears on a genome, in particular if $4^L > N$ with N as the number of base pairs of the genome). The set of these words we denote by $\mathcal{W}$. Each word of this set is classified by the protein either as being a binding word belonging to the binding word class $\mathcal{W}_B$ or a non-binding word belonging to the non-binding word class $\mathcal{W}_{NB}$.

Specific DNA-binding proteins play a crucial role in the transcription of genes. The sequence specific DNA-binding proteins involved in the transcription of genes are called transcription factors. Knowledge about their binding characteristics, i.e. to which words on the genome they will bind, is crucial for understanding and modeling the transcription machinery which reads the information stored in the genome. Further, transcription factors are integral parts of the so-called genetic regulatory networks. Genetic regulatory networks control the expression of genes which can now be measured on large scales due to the success of the microarray technology. For understanding and modeling these genetic regulatory networks knowledge about the binding characteristics of the transcription factor is crucial. These binding characteristics, however, cannot (yet) be determined from first principles but must be deduced from experimental observations. Usually, for a transcription factor only a few binding words are known experimentally. These few binding words then have to be used as training sample for a classifier which tries to mimic the classification behavior of the protein and to predict whether a word is seen by the protein as binding word or not.

The basis of most state-of-the-art methods for predicting whether a word at a certain site leads to binding of the protein or not is the so-called Profile-Matrix approach (for an overview of the different methods see [1], [2]). If we have $\kappa = |\mathcal{W}_B^E|$ experimentally verified binding words with $\mathcal{W}_B^E \subseteq \mathcal{W}_B$ as the set of these experimentally verified binding words, and if κ_{bl}/κ denotes the occurence frequency of base b at position l in these experimentally verified binding words, then the elements p_{bl} of the Profile-Matrix $\mathbf{p}$ are given by $p_{bl} = \kappa_{bl}/\kappa$. To judge whether a query word $\mathbf{w}$ is a binding word or not, one calculates a score. For this purpose for $\mathbf{w}$ one chooses the so-called orthogonal coding. In this coding $\mathbf{w}$ is a $4L$-dimensional vector (or $4 \times L$-matrix) with component $w_{bl} = 1$ if and only if the word $\mathbf{w}$ contains base b at position l, otherwise $w_{bl} = 0$. Then the score is given by

$$S(\mathbf{w}) = \sum_{b \in \{A,C,G,T\}} \sum_{l=1}^{L} p_{bl} w_{bl} = \mathbf{p}^T \mathbf{w}. \tag{1}$$

If $S(\mathbf{w})$ exceeds a prespecified threshold $S_{\min}$, one assumes $\mathbf{w}$ to be a binding word [3]–[5]. The Profile-Matrix $\mathbf{p}$ can be regarded as the average experimentally verified binding word, i.e.

$$\mathbf{p} = \frac{1}{\kappa} \sum_{\mathbf{w} \in \mathcal{W}_B^E} \mathbf{w} = \langle \mathbf{w} \rangle_E$$

with $[\langle \mathbf{w} \rangle_E]_{bl} = \kappa_{bl}/\kappa$ is valid. All the words whose overlap to this average binding word is large enough one assumes to be binding words.

II. The Structure of the Pattern Space

To approach the classification task and its learning in a more principled way, we first analyze the pattern space, i.e. the space of sequence words the protein ”sees”.

In the orthogonal coding the 4^L possible words $\mathbf{w}$ can be interpreted as points in a $4L$-dimensional space. These points are arranged in an interesting structure. They all lie on a $4L$-dimensional hypersphere, since $||\mathbf{w}||^2 = L$ is valid. At the same time they lie on a $3L$-dimensional linear subspace, since $\sum_b (w_{bl} - 1/4) = 0$ is valid for each $l = 1, \ldots, L$. This $3L$-dimensional linear subspace is the so-called continuous sequence space [6]. With $\bar{\mathbf{w}} = (1/4, 1/4, \ldots, 1/4)^T$ as the center of gravity of all the words $\mathbf{w} \in \mathcal{W}$ one obtains $\bar{\mathbf{w}}^T(\mathbf{w} - \mathbf{w}') = 0$ for all pairs $\mathbf{w}, \mathbf{w}'$. Hence, the $3L$-dimensional linear subspace is orthogonal to $\bar{\mathbf{w}}$. Further, all the words have the same distance from $\bar{\mathbf{w}}$. This means, in addition to lying on a $4L$-dimensional hypersphere around the origin and lying on a $3L$-dimensional linear subspace, the words $\mathbf{w}$ are arranged on a $3L$-dimensional hypersphere around their own center of gravity. For symmetry reasons the words are distributed homogeneously over the surface of this $3L$-dimensional hypersphere. This is illustrated schematically in Fig. 1.

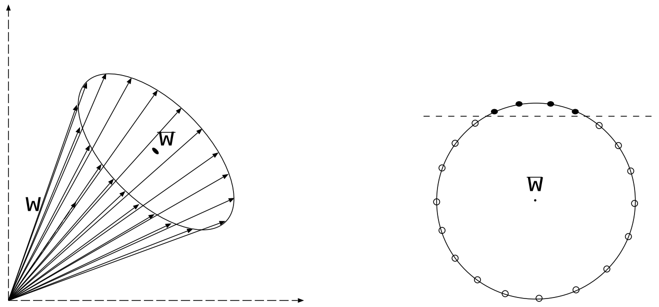

Fig. 1. Left: A low-dimensional sketch of the distribution of the words $\mathbf{w}$ within the $4L$-dimensional orthogonal coding space. Within this space the words lie on a $3L$-dimensional hypersphere (circle) around their center of gravity $\bar{\mathbf{w}}$. Right: On this $3L$-dimensional hypersphere (circle) the words (dots) are distributed homogeneously. Binding words are shown as filled dots. The dotted line indicates the hyperplane which divides the set of words into binding words and non-binding words.

III. Binding Energy Determines Classification Behavior

Elaborate experimental studies have shown that each base at a binding site i contributes to the binding energy E_i independently of the other bases in the word $\mathbf{w}_i$, at least in a very good approximation [7]–[10]. With e_{bl} as the binding energy contribution of base b at position l, and under the assumption of independent contributions from each base, one obtains

$$E(\mathbf{w}) = \sum_{b \in \{A,C,G,T\}} \sum_{l=1}^{L} e_{bl} w_{bl} = \mathbf{e}^T \mathbf{w} \qquad (2)$$

as the binding energy if the protein ”sees” the word $\mathbf{w}$ (see, e.g. [11]). If this binding energy $E(\mathbf{w})$ exceeds a threshold $E_{\min}$, one assumes that the protein binds to this word strongly and long enough to execute its function. Then $\mathbf{w}$ is a binding word, and the locations on the genome where this word occurs are binding sites.

Obviously, with independent binding energy contributions from each base the protein linearly divides the space of sequence words into binding and non-binding words. The vector with the binding energy contributions $\mathbf{e}$ determines the orientation of the dividing hyperplane. The protein acts as a linear classifier which divides the pattern space into two classes. The parameters which determine its classification behavior are the e_{bl} and $E_{\min}$.

If we compare equation (1) and equation (2), both equations have the same structure. The heuristically derived PM-approach implements a linear classification and, in this respect, is the correct ansatz. However, one can question whether the base occurence frequencies $\mathbf{p}$ are the best estimation for the right hyperplane orientation $\mathbf{e}$. Is this approach optimal from a machine learning point of view?

IV. The Estimation Task

As we have pointed out, the $K = 4^L$ words $\mathbf{w}$ are homogeneously distributed on a $3L$-dimensional hypersphere, the center of which is given by $\bar{\mathbf{w}} = (1/4, 1/4, \ldots, 1/4)^T$. The linear classification of the protein according to (2) now divides the hypersphere into two parts. The binding words $\mathbf{w} \in \mathcal{W}_B$ lie on one segment, while the non-binding words $\mathbf{w} \in \mathcal{W}_{NB}$ are located on the other one. This is illustrated in Fig. 1. Usually, the part with the binding words is much smaller than the part with the non-binding words. For example, for the human splice acceptor sites one observes for the relative size of these two classes $K/k \approx 1000$, with $k = |\mathcal{W}_B|$ as the number of binding words. This is the general order of magnitude for K/k, which can be connected to the binding sequence information content, as introduced in [12] and analyzed in detail in [13].

From experiments only binding words are known. Non-binding words, for several reasons, can hardly be determined experimentally. Hence, only positive examples are given and can be used for estimating the right classification parameters $\mathbf{e}$. In addition, the number κ of experimentally verified binding words is quite small, of the order of 10 to 100. The task is to estimate the dividing hyperplane $(\mathbf{e}, E_{\min})$ based on these few training words.

The Profile-Matrix simply takes the vector $\mathbf{p} = \langle \mathbf{w} \rangle_E$ pointing onto the center of gravity of the experimentally verified binding words as an estimation for the orientation of the hyperplane which divides the hypersphere into the binding word segment and the non-binding word segment. This is equivalent to taking the vector pointing from the center of gravity $\bar{\mathbf{w}}$ of all words (the center of the $3L$-dimensional hypersphere) onto $\langle\mathbf{w}\rangle_E$. This can be seen if we calculate the scalar product of an arbitrary word $\mathbf{w}$ with the vector $\langle\mathbf{w}\rangle_E - \bar{\mathbf{w}}$. One obtains

$$\begin{aligned}\left(\langle\mathbf{w}\rangle_E - \bar{\mathbf{w}}\right)^T \mathbf{w} &= \langle\mathbf{w}\rangle_E^T \mathbf{w} - \bar{\mathbf{w}}^T \mathbf{w} \\ &= \mathbf{p}^T \mathbf{w} - L/4,\end{aligned}$$

since $\bar{\mathbf{w}} = (1/4, 1/4, \ldots, 1/4)^T$, and since $\mathbf{w}$ contains L ones and $3L$ zeros. Except for an offset in the threshold $S_{\min}$ this is equivalent to (1). Additional technical details about transformations of the normal vector which leave the classification invariant are provided in the Appendix.

V. Binding-Matrix as Maximum-Likelihood-Estimation

The Profile-Matrix approach does not yield the maximum likelihood estimation of the orientation of the dividing hyperplane $\mathbf{e}$. For estimating $\mathbf{e}$ by maximum likelihood, we have to determine those $(\mathbf{e}, E_{\min})$ which maximize the likelihood

$$P(\mathcal{W}_B^E|\mathbf{e}, E_{\min}) = \prod_{\mathbf{w}\in\mathcal{W}_B^E} P(\mathbf{w}|\mathbf{e}, E_{\min}) \qquad (3)$$

of obtaining $\mathcal{W}_B^E$ as the set of experimentally verified binding words. $P(\mathbf{w}|\mathbf{e}, E_{\min})$ is the likelihood of finding word $\mathbf{w}$ at a binding site, given the binding parameters $(\mathbf{e}, E_{\min})$.

$P(\mathbf{w}|\mathbf{e}, E_{\min})$ is only non-zero for words which lie on the binding word hypersphere segment, i.e. for which $\mathbf{e}^T\mathbf{w} \geq E_{\min}$ is valid. Words which fulfill this constraint are equally likely to be found at binding sites. Hence, for words on the binding word hypersphere segment $P(\mathbf{w}|\mathbf{e}, E_{\min}) = 1/k$ is valid.

The likelihood $P(\mathcal{W}_B^E|\mathbf{e}, E_{\min})$ of finding the whole set $\mathcal{W}_B^E$ as binding words is only non-zero for hyperplanes $(\mathbf{e}, E_{\min})$ for which *all* the $\mathbf{w} \in \mathcal{W}_B^E$ are on the binding word segment. If this is the case, $P(\mathcal{W}_B^E|\mathbf{e}, E_{\min}) = k^{-\kappa}$ is valid. Obviously, this expression increases with smaller values for k, i.e. for smaller binding word segments. Hence, we obtain as maximum likelihood estimation for $(\mathbf{e}, E_{\min})$ the hyperplane which cuts off the smallest binding word hypersphere segment by leaving all $\mathbf{w} \in \mathcal{W}_B^E$ on this segment.

This is equivalent to the plane $\mathbf{q}$ which separates the training data points $\mathbf{w} \in \mathcal{W}_B^E$ from the center of gravity $\bar{\mathbf{w}}$ with the maximum distance d. This is illustrated in Fig. 2. The $4 \times L$-matrix (or $4L$-vector) $\mathbf{q}$ which determines the orientation of this plane we call Binding-Matrix (BM). For determining the Binding-Matrix we have to find the distance d and normal vector $\mathbf{q}$ for which

$$d \overset{!}{=} \max \qquad (4)$$

under the constraints

$$||\mathbf{q}||^2 = 1 \qquad \text{and} \qquad \mathbf{q}^T(\mathbf{w} - \bar{\mathbf{w}}) \geq d \quad \forall\, \mathbf{w} \in \mathcal{W}_B^E.$$

This is a constrained optimization problem with a linear target function, a quadratic constrained (normalization of the normal vector of the plane) and κ linear constraints. We will show that, under certain conditions, this can be transformed into a Quadratic-Programming-Problem.

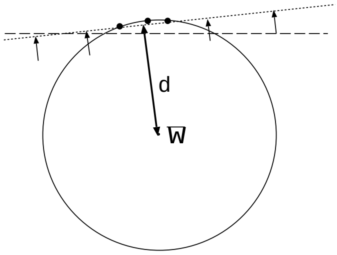

Fig. 2. The maximum likelihood estimation of the binding word distribution is given by the smallest sphere segment which still carries all training data points. Hence, one has to look for the plane which ”cuts off” the smallest segment while keeping all data points on this segment. This is the plane which has the largest distance d to the center of the sphere.

VI. Quadratic-Programming-Problem

We assume that the binding word segment is smaller than the non-binding word segment. This is the biologically relevant case. Then, the maximal distance $d_{\max}$ is positive and we can restrict our search on positive d. Under these conditions we can obtain the Binding-Matrix by solving the slightly modified constrained optimization problem

$$d^2 \overset{!}{=} \max \qquad (5)$$

under the constraints

$$||\mathbf{q}||^2 = 1,\ d > 0, \text{ and } \quad \mathbf{q}^T(\mathbf{w} - \bar{\mathbf{w}}) \geq d \quad \forall\, \mathbf{w} \in \mathcal{W}_B^E.$$

Similar to the calculation of the maximum margin in the Support-Vector-Machine framework [14], we now can define $\mathbf{q}' = \mathbf{q}/d$ and obtain $d^2 = ||\mathbf{q}'||^{-2}$. This transforms our constrained optimization problem above into the Quadratic-Programming-Problem

$$||\mathbf{q}'||^2 \overset{!}{=} \min \qquad \mathbf{q}'^T(\mathbf{w} - \bar{\mathbf{w}}) \geq 1 \quad \forall\, \mathbf{w} \in \mathcal{W}_B^E. \qquad (6)$$

This Quadratic-Programming-Problem can easily be solved by a number of standard procedures. After having obtained $\mathbf{q}'$, there are a number of different ways to normalize $\mathbf{q}'$, all of which leave the classification result invariant. In the Appendix we describe *canonical forms* of the classifier matrix.

The BM as the solution of this Quadratic-Programming-Problem is determined only by a subset of the training data, i.e., only by those training data points which mark the boundaries of the smallest sphere segment. In some respects these ”support binding words” correspond to the support vectors of the SVM.

VII. Results and Comparison

For predicting whether the word $\mathbf{w}_i$ at position i on the genome is a binding word and, therefore, i is a binding site, the BM is employed in the same way as the PM. First, the so-called score value $S(\mathbf{w}_i) = \mathbf{q}^T\mathbf{w}_i$ is calculated, and if this score exceeds a prespecified minimum value $S_{\min}$, the word $\mathbf{w}_i$ is considered to be a binding word and i to be a binding

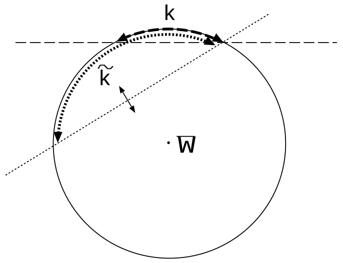

Fig. 3. The ”true” hyperplane ”cuts off” a segment containing k binding words. The estimated hyperplane can be shifted by choosing different threshold values. The threshold value for which the estimated hyperplane ”cuts off” the smallest segment including the ”true” segment is the threshold value for 100% sensitivity. The lower the number $\tilde{k}$ of words which are then recognizes as binding words, the better the estimation.

site. This implements the linear classification with the BM determining the dividing hyperplane.

For all linear classifiers investigated here, the threshold value $S_{\min}$ determines how many words are classified as binding words (see Fig. 3). Thus, the threshold value $S_{\min}$ can be used to control the sensitivity/specificity of the binding word recognition procedure. Lowering the threshold value increases sensitivity for a decrease in specificity and vice versa. As one reduces $S_{\min}$, more and more words are classified as binding words. Consequently, sensitivity is increased. At a certain setting of $S_{\min}$, all binding words are correctly classified, i.e. sensitivity reaches 100%. By lowering the threshold beyond this point, only more and more false positives are introduced, resulting in reduced specificity.

The number of words which are predicted to be binding words at the threshold setting which yields 100% sensitivity is denoted by $\tilde{k}$. This is illustrated in Fig. 3. The $\tilde{k}$ words include all k ”true” binding words (by definition), and a number of false positives, i.e. non-binding words which are incorrectly predicted to be binding words. Thus, the smaller $\tilde{k}$, i.e. the smaller the difference $\tilde{k}-k$, the larger the specificity $1-(\tilde{k}-k)/(K-k)$ at 100% sensitivity and the better the estimation approach. This fact is used for testing and comparing the different approaches.

The true number of binding words, k, is unknown in practice, and, therefore, specificity cannot be measured directly. However, comparative specificity assessment is possible. Consider two normal vectors $\mathbf{u}$ and $\mathbf{v}$, with values $\tilde{k}_{\mathbf{u}}$ and $\tilde{k}_{\mathbf{v}}$, and assume that $\mathbf{u}$ achieves a higher specificity than $\mathbf{v}$ at 100% sensitivity. This implies that $1-(\tilde{k}_{\mathbf{u}}-k)/(K-k) > 1-(\tilde{k}_{\mathbf{v}}-k)/(K-k)$. From this, one obtains $\tilde{k}_{\mathbf{u}}/K < \tilde{k}_{\mathbf{v}}/K$. Thus, a smaller $\tilde{k}/K$ ratio indicates higher specificity.

The problem of testing and comparing the classification performance of the PM and the BM is the very low number of binding words which are typically known for a transcription factor, i.e., the low number of data points. The most comprehensive database for transcription factors is the TRANSFAC database [15]. Even there, the sets of experimentally verified binding words only have a size ranging between $\kappa = 1$ and $\kappa = 73$. A division of such small data sets into training and test sets reduces the number of training data even further. Therefore, we perform two different tests. The first test includes almost all transcription factors, also the ones with small sets of binding words, and performs a leave-one-out test (test set of size 1 only). The second test includes only transcription factors with large sets of binding words, which then allow larger test sets. In these tests 2/3 of the data set were used for training and 1/3 for testing.

After having calculated the Profile-Matrix and Binding-Matrix from the training set, one has to test whether the unseen test words will be recognized as binding words. This, however, depends on the threshold value $S_{\min}$ one chooses. For our performance tests, we ask how far one has to reduce the threshold value $S_{\min}$ to reach a sensitivity which recognizes all known binding words, including those in the test set. This is taken as an approximation of the point of 100% sensitivity and used to estimate $\tilde{k}/K$, the fraction of of words which have to be classified as binding words for this sensitivity. The estimation is performed by generating and scoring $100\,000$ random words and determining the fraction of words which satisfy $S(\mathbf{w}) \geq S_{\min}$. As discussed above, the lower the fraction $\tilde{k}/K$, the higher the specificity at 100% sensitivity and the better the real hyperplane is estimated.

From the TRANSFAC data base we were able to select 95 binding word sets with a size of at least 5. For these sets, the leave-one-out-test was performed. One binding word of the binding word set is removed for testing, the rest is used for training. This is repeated until each binding word has been used for testing. Over all data sets we obtained 1604 $\tilde{k}/K$ values. The average binding word set size was $\kappa = 17$ and the maximum number was $\kappa = 73$. In Fig. 4 the result is shown for the PM, the BM and a simple consensus sequence. The consensus sequence in its orthogonal coding is a matrix (vector) $\mathbf{c}$ which in each column $l = 1, \ldots, L$ contains a one ($c_{bl} = 1$) for the base b which occurs most frequently and three zeros ($c_{bl} = 0$). As for the PM and BM a score $S(\mathbf{w}) = \mathbf{c}^T\mathbf{w}$ is calculated, and if this score exceeds the chosen threshold $S_{\min}$, the word $\mathbf{w}$ is classified as a binding word.

The top of Fig. 4 shows a box-plot of the distribution of the $\tilde{k}/K$ values one obtains for each approach. The Binding-Matrix can reduce the fraction $\tilde{k}/K$ of words which have to be classified as binding words to recognize all the known binding words by about an order of magnitude compared to the Profile-Matrix. This increase in specificity is about the increase the PM was able to achieve compared to the consensus sequence. The consensus sequence was the common approach before it was improved by the Profile-Matrix.

Figure 4 also shows the result one obtains if one restricts the leave-one-out test to binding word sets of at least 30 words. There are 13 such sets, and one obtains 508 $\tilde{k}/K$ values, i.e. on average each set consisted of $\kappa = 39$ experimentally verified binding words. In this case the performance increase becomes even more significant.

For the 13 sets of binding words from TRANSFAC which contain at least $\kappa = 30$ binding words, in addition to the leave-one-out test, the performance test based on larger test sets

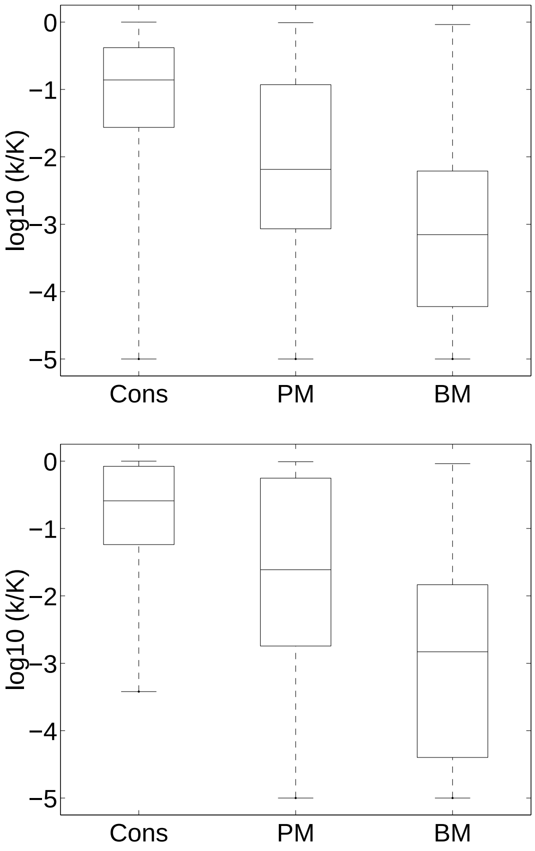

Fig. 4. The results of the leave-one-out-test for transcription factors with at least 5 (top) and at least 30 (bottom) experimentally verified binding words. The box-plots show the distribution of the $\tilde{k}/K$ values. The $\tilde{k}/K$ value is an indicator for the specificity of the classifier at 100% sensitivity. Boxes indicate the quartiles, the horizontal line in the box shows the median. The bars extend to the minimal and the maximal value. Compared to the Profile-Matrix (PM), the Binding-Matrix (BM) achieves an increase in recognition performance of about an order of magnitude, about the same increase the PM was able to achieve compared to the Consensus Sequence (Cons).

was carried out. The test sets were generated by randomly chosing $1/3$ of the binding words of a set. The remaining words were put into the corresponding training set. For each of the 13 binding word sets, $10\,000$ such separations into training and test sets were generated independently, and $\tilde{k}/K$ was estimated as described above.

Figure 5 shows the result of this performance test. For the Consensus Sequence and the Profile-Matrix the $\tilde{k}/K$ distribution is about the same as in the leave-one-out test (bottom of Fig. 4). A reduced training set size does not seem to reduce the estimation quality. This might be due to the fact that the specificity levels achieved by the Consensus Sequence and the Profile-Matrix is already comparatively low. More than 90% of the sites classified as binding sites one has to expect to be false positives at this sensitivity level. This is different for the Binding-Matrix. In both tests only the binding matrix was able to reach $\tilde{k}/K$ values of the order of magnitude of 10^{-3}, which one expects from experimental observations and theoretical analyzes [12], [13]. On this specificity level one can expect that the classifier reacts much more sensitively to withheld training data. Indeed, the Binding-Matrix loses estimation performance by reducing the training data set by 1/3. The improvement by the Binding-Matrix now amounts to approximately half an order of magnitude instead of one order of magnitude as in the leave-one-out test.

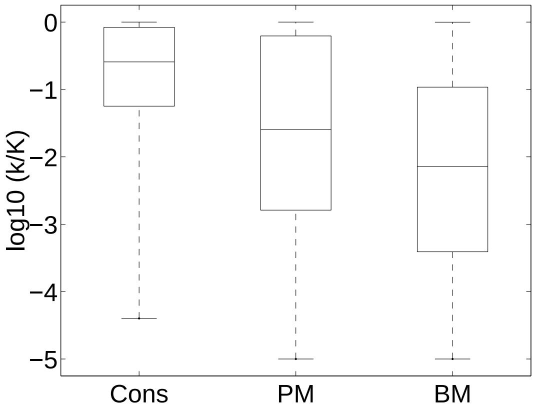

Fig. 5. Results of the performance evaluation based on 1/3 test sets. Boxes indicate the quartiles, the horizontal line in the box shows the median. The bars extend to the minimal and the maximal value.

VIII. Conclusion

We have presented a novel approach to the problem of predicting the binding sites of sequence specific DNA-binding proteins, e.g., transcription factors. The approach determines the distribution of binding words within the space of orthogonally coded words by maximum likelihood estimation. Since there are good reasons to assume that bases contribute independently to the binding energy, at least to a good approximation, and under the assumption that at binding sites a minimum binding energy is required, binding and non-binding words are linearly separated within the word space. The question we posed was how to optimally determine this linear separation from a small set of known binding words. Does the widespread Profile-Matrix provide the optimal solution? This task is well-known from machine learning: how to determine the optimal linear classifier from a small set of positive examples. First, we analyzed the structure of the data distribution. We showed that all the patterns (words) are points on a $3L$-dimensional hypersphere within the $4L$-dimensional word space. From this it followed that the set of binding words is given by a segment ”cut” from this $3L$-dimensional hypersphere by a hyperplane.

The Profile-Matrix estimates this hyperplane by taking as its normal vector the vector which points onto the center of gravity of the given training data points (the usually small set of experimentally verified binding words). This converges against the correct solution for an increasing number of training points, but only suboptimally. Particularly for small sets of training data one expects the maximum likelihood estimation to provide superior results. We have shown that the maximum likelihood estimation of the linear separation can be obtained from solving a Quadratic-Programming-Problem which, in some respects, is similar to the one which has to be solved in the context of the Support-Vector-Machine. This yields what we call Binding-Matrix (BM).

We tested and compared the BM and PM on data from the TRANSFAC database. Because of the small size of most data sets, first a leave-one-out-test was chosen. Then, with data sets of sufficient size, in addition performance tests with test sets of $1/3$ of the known binding words were carried out. After having determined the Profile- and Binding-Matrix for a transcription factor from the training set, we calculated the fraction of words each matrix has to classify as binding words to recognize all known binding words including the ”unseen” test words. This test gives a measure for the specificity of each approach at the same sensitivity level. The larger this fraction to successfully recognize the test words, the lower the specificity and the worse one has to rate the overall recognition performance. This test showed an increase in recognition performance of the Binding-Matrix of about an order of magnitude compared to the Profile-Matrix.

References

[1] K. Frech, K. Quandt, and T. Werner, “Finding protein-binding sites in DNA sequences: The next generation,” *TIBS*, vol. 22, pp. 103–104, 1997.

[2] G. D. Stormo, “DNA binding sites: Representation and discovery,” *Bioinformatics*, vol. 16, pp. 16–23, 2000.

[3] G. Stormo, T. Schneider, L. Gold, and A. Ehrenfeucht, “Use of the perceptron algorithm to distinguish translational initiation sites in *E. coli*.” *Nucl. Acids Res.*, vol. 10, pp. 2997–3011, 1982.

[4] R. Harr, M. Haggstrom, and P. Gustafsson, “Search algorithm for pattern match analysis of nucleic acid sequences,” *Nucl. Acids Res.*, vol. 11, pp. 2943–2957, 1983.

[5] R. Staden, “Computer methods to locate signals in nucleic acid sequences.” *Nucl. Acids Res.*, vol. 12, pp. 505–519, 1984.

[6] M. Vingron and P. R. Sibbald, “Weighting in sequence space: A comparison of methods in terms of generalized sequences,” *Proc. Natl. Acad. Sci. USA*, vol. 90, pp. 8777–8781, 1993.

[7] A. Sarai and Y. Takeda, “Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically,” *Proc. Natl. Acad. Sci. USA*, vol. 86, pp. 6513–6517, 1989.

[8] Y. Takeda, A. Sarai, and V. Rivera, “Analysis of the sequence-specific interactions between *Cro* repressor and operator DNA by systematic base substitution experiments,” *Proc. Natl. Acad. Sci. USA*, vol. 86, pp. 439–443, 1989.

[9] M. Mulligan, D. Hawley, R. Entriken, and W. McClure, “*Escherichia coli* promoter sequences predict *in vitro* RNA polymerase selectivity,” *Nucl. Acids Res.*, vol. 12, pp. 789–800, 1984.

[10] G. Stormo, S. Strobl, M. Yoshioka, and J. Lee, “Specificity of the *Mnt* protein. independent effects of mutations at different positions in the operator,” *J. Mol. Biol.*, vol. 229, pp. 821–826, 1993.

[11] G. D. Stormo and D. S. Fields, “Specificity, free energy and information content in protein-DNA-interactions,” *Trends in Biochemical Sciences*, vol. 23, pp. 109–113, 1998.

[12] T. D. Schneider, G. D. Stormo, and L. Gold, “Information content of binding sites on nucleotide sequences,” *J.Mol.Biol.*, vol. 188, pp. 415–431, 1986.

[13] J. T. Kim, T. Martinetz, and D. Polani, “Bioinformatic principles underlying the information content of transcription factor binding sites,” p. to appear, 2003, journal of Theoretical Biology, accepted in revised form on 21. August 2002.

[14] V. Vapnik, *Statistical Leraning Theory*. New York: John Wiley & Sons, 1998.

[15] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Prüß, I. Reuter, and F. Schacherer, “TRANSFAC: An integrated system for gene expression regulation,” *Nucl. Acids Res.*, vol. 28, pp. 316–319, 2000.

Appendix: Canonical Normal Form for Matrix-Based Classifiers

With a normal vector $\mathbf{v}$ and a threshold score V given as parameters, a linear classifier is unambiguously specified by the inequality

$$\mathbf{v}^T\mathbf{w} \geq V \tag{7}$$

However, in our case the reverse is not true: Because all words lie in a $3L$-dimensional linear subspace of $\mathbb{R}^{4L}$, the plane of separation can be rotated such that its intersection with the subspace populated by the words remains unchanged. There exist L rotational degrees of freedom of this kind, and one additional degree of freedom due to scaling. This appendix presents a straightforward procedure for mapping all inequalities (7) that describe the same classifier to one unique form, which is the *canonical normal form*, denoted by $(\mathbf{v}_c, V_c)$.

The degrees of freedom are reflected by the fact that the classification implied by a hyperplane remains unchanged when a constant c is added to V and to all four components in $\mathbf{v}$ belonging to one position l. Formally, one can construct a matrix $\mathbf{u}$ from $\mathbf{v}$ by letting $u_{bl} = v_{bl} + c$ for $b \in \{A, C, G, T\}$ and a fixed l, and a corresponding threshold $U = V + c$. Then, the hyperplanes specified by $\mathbf{u}^T\mathbf{w} \geq U$ and by $\mathbf{v}^T\mathbf{w} \geq V$ implement identical classifiers on $\mathcal{W}$, since $\mathbf{u}^T\mathbf{w} = \mathbf{v}^T\mathbf{w} + c$ is valid for each $\mathbf{w} \in \mathcal{W}$.

It is convenient to translate the origin of the $4L$-dimensional coordinate system to $\bar{\mathbf{w}}$. After translation, words are represented by $\mathbf{w}' = \mathbf{w} - \bar{\mathbf{w}}$, with

$$\sum_{b\in\{A,C,G,T\}} w'_{bl} = 0, \quad 1 \leq l \leq L \tag{8}$$

being valid. With $V' = V - \mathbf{v}^T\bar{\mathbf{w}}$, Equation 7 can be rewritten as $\mathbf{v}^T\mathbf{w}' \geq V'$.

The rotational degrees of freedom are disambiguated by requiring that the normal vector $\mathbf{v}'$ must be in the $3L$-dimensional subspace occupied by the words, i.e that the v'_{bl} must satisfy Equation (8). The parameters meeting this requirement can be obtained by computing $c_l = -\sum_b v_{bl}/4$. Then one constructs $\mathbf{v}'$ by letting $v'_{bl} = v_{bl} + c_l$ and $V'' = V' + \sum_{l=1}^{L} c_l$.

The scaling degree of freedom is removed by requiring the canonical normal vector to have unit length, i.e. $||\mathbf{v}_c|| = 1$. This is achieved by simply multiplying the normal vector and the threshold with $1/||\mathbf{v}'||$. Let $\mathbf{v}_c = \mathbf{v}'/||\mathbf{v}'||$ and $V_c = V''/||\mathbf{v}'||$. The inequality $\mathbf{v}_c^T\mathbf{w}' \geq V_c$ is the canonical normal form that represents the classifier described by the original inequality.

A canonical normal vector can be transformed into a *canonical profile matrix*. Let $m = -\min_{b,l}[\mathbf{v}_c]_{bl}$. Then, the canonical profile matrix is defined by $\mathbf{v}_p = \mathbf{v}_c/(4m) + \bar{\mathbf{w}}$. With the corresponding threshold $V_p = V_c/(4m) + L/4$, the canonical profile matrix provides an alternative to the canonical normal form, e.g. if an application cannot process the negative values in the canonical normal vector correctly.