



## Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition

Uwe Ohler<sup>1</sup>, Heinrich Niemann<sup>1</sup>, Guo-chun Liao<sup>2</sup> and Gerald M. Rubin<sup>2</sup>

<sup>1</sup>Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen-Nürnberg, Martensstr. 3, D-91058 Erlangen, and <sup>2</sup>Department of Molecular and Cell Biology, University of California at Berkeley, LSA Rm 539, Berkeley, CA 94708, USA

Received on February 5, 2001; revised and accepted on April 2, 2001

### ABSTRACT

We present an approach to integrate physical properties of DNA, such as DNA bendability or GC content, into our probabilistic promoter recognition system MCPROMOTER. In the new model, a promoter is represented as a sequence of consecutive segments represented by joint likelihoods for DNA sequence and profiles of physical properties. Sequence likelihoods are modeled with interpolated Markov chains, physical properties with Gaussian distributions. The background uses two joint sequence/profile models for coding and non-coding sequences, each consisting of a mixture of a sense and an anti-sense submodel. On a large *Drosophila* test set, we achieved a reduction of about 30% of false positives when compared with a model solely based on sequence likelihoods.

**Contact:** Uwe.Ohler@informatik.uni-erlangen.de

### INTRODUCTION

The notoriously difficult problem of computational promoter recognition in eukaryotic DNA (Fickett & Hatzigeorgiou, 1997) has up to now been largely based on specific features of the DNA promoter sequence: Binding sites of transcription factors, or the base composition in general. But a regulatory region such as a eukaryotic promoter does not only contain specific sequence elements that serve as targets for interacting proteins, but also exhibits distinct physical properties reflected in its sequence (Latchman, 1998). For example, the DNA of an actively transcribed promoter has to be accessible and must not be wrapped up in nucleosomes. CpG islands which hint at regions of generally low methylation and therefore low likelihood to attract nucleosomes are characteristic for many vertebrate promoters (Antequera & Bird, 1993). A recent approach used CpG island features — GC content, ratio of expected to observed CG dinucleotides, and length — to find regulatory regions and aimed at the distinction of promoter-

associated from non-associated CpG islands (Ioshikhes & Zhang, 2000).

Because of the non-existing or very weak methylation in non-vertebrate eukaryotes such as *D. melanogaster* (Gowher *et al.*, 2000), CpG islands features cannot be exploited for eukaryotic promoter finding in general. Studies on human and *E. coli* promoter sequences (Pedersen *et al.*, 1998; Babenko *et al.*, 1999; Pedersen *et al.*, 2000), though, showed that the DNA sequence in promoters causes characteristic profiles in base composition or other DNA properties such as bendability, nucleosome positioning preference and propeller twist (see Figures 1 and 2). These profiles are based on experimentally derived parameter tables for di- or trinucleotides and can easily be calculated from the sequence. Earlier, Lisser & Margalit (1994) investigated such DNA structural profiles of *E. coli* promoters and used the mean value of selected properties within five promoter segments as feature variables. By doing so, they were able to distinguish between promoter and coding sequences by means of linear discriminant analysis, but they did not integrate these features in a system for promoter recognition.

An example where the target of a DNA interacting protein is largely defined by its physical properties is the P transposable element insertion site in *Drosophila*. Here, no clear sequence consensus can be seen, but for a large variety of properties, the average profile at the insertion site shows distinct peaks (Liao *et al.*, 2000).

One reason why property profiles have not yet been taken into account in existing promoter prediction systems might be that individual promoter profiles are extremely noisy and not straightforward to use. Apart from the noise, the profiles are calculated from the primary sequence itself, and it was thus not clear if the properties might already be implicitly captured in a model of the sequence alone, or if they could lead to an improvement in promoter recognition at all.

**Table 1.** Parameter table for protein-DNA-twist profile calculation. The entries represent the mean twist angle in degrees.

first base	second base			
	A	C	G	T
A	35.1	31.5	31.9	29.3
C	37.3	32.9	36.1	31.9
G	36.3	33.6	32.9	31.5
T	37.8	36.3	37.3	35.1

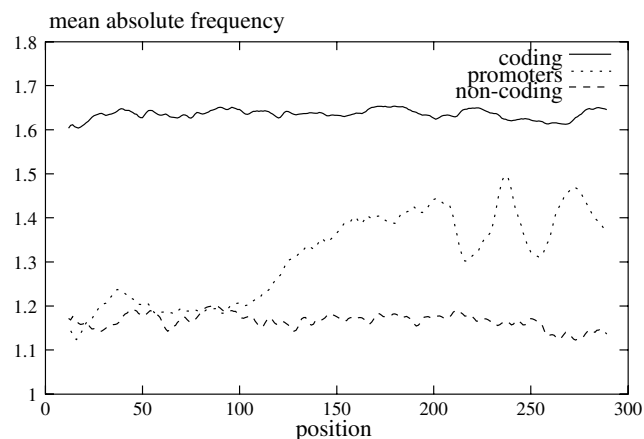
The very nature of the profile data demands a probabilistic modeling. This enables us not only to cope with noisy data, but also allows for a clean integration into existing probabilistic frameworks for promoter prediction. In this paper, we explain how we did this for our probabilistic promoter finding system MCPROMOTER (Ohler *et al.*, 2000; Ohler, 2000). To our knowledge, this is the first time that sequence and profile models are put together for (eukaryotic) promoter recognition, and we are able to show that this considerably improves the performance.

The rest of the paper is organized as follows: After showing some examples of profiles, we describe our promoter prediction system and how to extend it to model physical properties. Next, we explain how features are calculated from a profile, and how we select among the large number of possible properties. We close by showing the results on a large data set of *Drosophila* promoter sequences and discussing the approach.

## DATA SETS AND PROFILE CALCULATION

We explored the 14 different parameter sets of physical DNA properties compiled by Liao *et al.* (2000) (see Table 2). Apart from the GC content based on trinucleotides which simply counts how many Guanines and Cytosines are present in the trinucleotide centered at the actual sequence position (see Figure 1), the parameter tables are all experimentally derived and represent values related to physical properties of di- or tri-nucleotides such as bendability, DNA conformation or protein-DNA-interaction. Table 1 shows the values for protein-DNA-twist as example.

A profile simply consists of the corresponding values from the chosen parameter set along a given DNA sequence. Because the parameters refer to di- or trinucleotides only, those profiles are generally very noisy. Therefore, they are smoothed with a mean value filter of a certain fixed width, usually 20–30 bp (Pedersen *et al.*, 1998; Liao *et al.*, 2000). In our experiments, we used a mean filter of 21 bases.



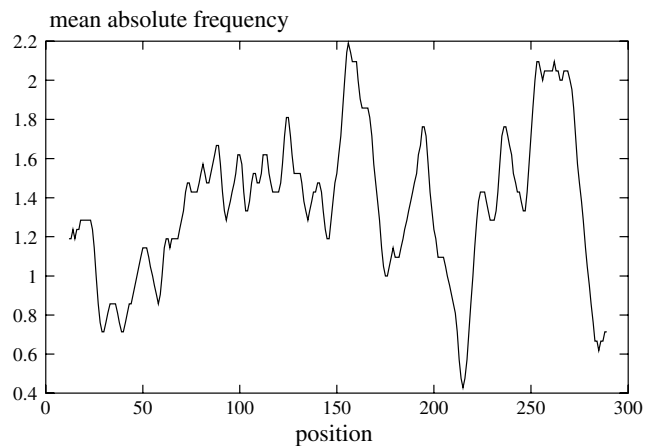
**Fig. 1.** GC content of promoters, coding, and non-coding sequences in *Drosophila*. The transcription start sites of the promoters are aligned at position 250. The profiles were smoothed with a mean filter of width 21 and averaged over all sequences. As the profile is based on tri-nucleotides, the mean absolute frequency may range from 0 to 3.

## Data sets

Our training set consists of non-redundant *Drosophila* promoters, coding, and non-coding sequences. All sequences are 300 bp long; the promoters contain 250 bp upstream and 50 bp downstream of the annotated start site. The training set contains a total of 247 promoters, 240 non-coding and 711 coding sequences. The promoters are a joint set of the non-redundant *Drosophila* promoters from the Eukaryotic Promoter Database (EPD version 63, Perier *et al.* (2000)) and the *Drosophila* promoter database of Arkhipova (1995) following the guidelines of EPD for non-redundant sets; the non-promoters were taken from the training set of the GENIE gene finder for *Drosophila* (Reese *et al.*, 2000b). It is essentially the same set that we compiled for the Genome Annotation Assessment Project (GASP, Reese *et al.* (2000a)), with the slight difference that we extracted the full 300 bp promoter sequences after the *Drosophila* genome became public. A careful check afterwards resulted in the elimination of eight of the previously 255 promoter sequences. To test our approach, we compared the predictors with and without using structural information on the GASP set of 92 potential *Drosophila* transcription start sites.

## Example profiles

Figure 1 shows the GC content profiles of the three sequence classes within our training set, namely coding and intron sequences as well as promoters whose transcription start site is aligned at position 250. The profiles are averaged over all sequences in the set. Coding and non-coding sequences, as expected, show rather uniform values with



**Fig. 2.** Example GC content profile of one *Drosophila* promoter from the training set. Even though one can see the distinct drop at the position of the TATA box, the overall picture is considerably different from the average profile (see Figure 1).

no positional preferences. In contrast, the promoters have a distinct profile with drops in the areas of TATA box and initiator sites.

Figure 1 gives an inaccurate impression because the profile shown is averaged over a large set of sequences and does not reflect that some promoters lack distinct profile features such as the TATA box valley. Moreover, even in the case where a TATA box is present, individual profiles show a high degree of variation from the average profile (Figure 2), resulting from the unique underlying sequence.

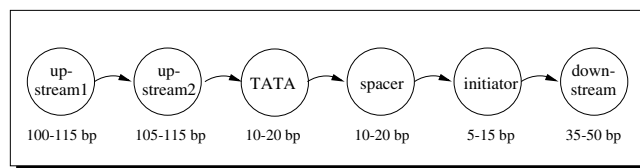
## JOINT MODELING OF PROMOTER SEQUENCE AND STRUCTURE LIKELIHOODS

To illustrate our approach to integrating structural features, we briefly review our existing promoter sequence model and then show how we extended it by adding likelihoods for structural properties.

### Stochastic segment models of promoter regions

Our approach (see Ohler *et al.* (2000) for a description of the algorithms) is based on the observation that a eukaryotic promoter can generally be divided into segments: the region upstream from the transcription start site, the core promoter where the main initiation complex binds, and a region downstream from the start site. The core promoter can be further split into the TATA box and the initiator region (Inr), separated by a spacer of approximately 15 bp. The upstream region is divided into two segments (see Figure 3).

As one can see, this segmental structure of the promoter model corresponds well with the GC profile in Figure 1: The first 100 or so bases have a low average GC content



**Fig. 3.** Structure of the promoter segment model.

(state 1), followed by a gradual increase (state 2), the decline at the TATA box (state 3), an uprise at the spacer (state 4), another decline at the initiator (state 5) and finally a distinct peak at the position of the GC-rich downstream promoter element DPE (state 6, see Kutach & Kadonaga (2000)).

This so-called *stochastic segment model* (SSM) is a generalization of a hidden Markov model. Like an HMM, it consists of a set  $\mathcal{Q}$  of connected states which can be characterized by an initial state distribution  $\pi$  and state transition distribution  $A$  with entries  $a_{ij}$ . Each state  $q_j$  contains an output distribution for the production of symbols which can be observed from the outside. While the output distribution of an HMM state can only emit a single symbol per state (Rabiner & Juang, 1993), each SSM state incorporates a joint distribution  $b_j$  which generates a sequence of symbols (a whole segment). The length of the generated segment underlies a duration distribution  $d_j$  associated with the state. Thus, the probability  $P_j(\mathbf{w}_i)$  that a state produces a partial sequence  $\mathbf{w}_i$  of length  $\tau_i$  is given by

$$P_j(\mathbf{w}_i) = d_j(\tau_i) \cdot b_j(\mathbf{w}_i | \tau_i). \quad (1)$$

With a given valid segmentation  $(s, \tau) = ((q_{s_1}, \tau_1) \dots (q_{s_m}, \tau_m))$  of sequence  $\mathbf{w}$  into segments  $\mathbf{w}_j$ , ( $\sum_j \tau_j = |\mathbf{w}|$ ), the probability of the sequence can be expressed as

$$P(\mathbf{w}, s, \tau) = \pi_{s_1} \prod_{i=1}^{m-1} P_{s_i}(\mathbf{w}_i) a_{s_i s_{i+1}} \cdot P_{s_m}(\mathbf{w}_m) \quad (2)$$

Most gene finding systems which make use of stochastic models fit into the framework of SSMs. The GenScan system (Burge & Karlin, 1997), in particular, uses a model structure similar to ours. The difference is that we cannot expect the training material to be annotated in advance, i. e. how a promoter is divided up into the six segments that our model contains. We therefore adopted the Viterbi training algorithm to include length distributions: First, we determine the most likely state sequence for each training sequence, then we treat this segmentation as the correct annotation. The resulting training material for each state

is used to estimate the output and duration distribution. Of course, the probabilities of the state transitions and initial states are modified as well. The algorithm maximizes the Viterbi score of the model, i. e., the score obtained on the best segmentation is guaranteed to increase after each iteration.

The probability of generating a sequence  $w$  with a segment model is equal to the sum of all possible segmentations over which the sequence can be produced. This can be computed efficiently by an adaption of the forward algorithm that also takes the length distribution into account. The most likely segmentation can be computed using the Viterbi algorithm, in which the sum over all possible segmentations is replaced by its maximum.

As submodels for each state, we use interpolated Markov chains (Ohler *et al.*, 1999, IMCs) of different order, depending on the size of the segment. IMCs can be very efficiently evaluated which allows to apply the model on large genomic sequences. We also use IMCs for the background model for coding and non-coding sequences, each of which consists of two IMCs trained on sense and anti-sense sequences.

### Extending the sequence model with profile likelihoods

The above segment model can easily be extended to handle profile features; instead of calculating the segment probabilities based on the sequence alone, we replace Equation 1 by the joint probability on sequence and profile:

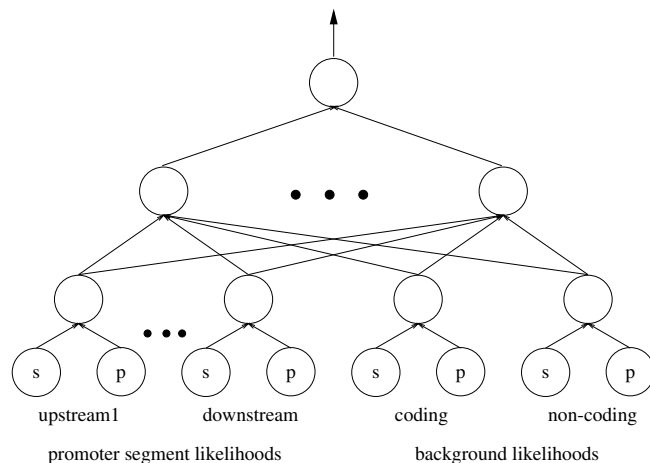
$$P_j(w_i, p_i) = d_j(\tau_i) \cdot b_j(w_i | \tau_i) \cdot c_j(p_i | w_i, \tau_i) \quad (3)$$

We include a probabilistic submodel  $c_j$  in each state that describes the likelihood of a profile  $p_i$ , given the sequence and its length. No other changes are necessary, all algorithms that were applicable to the sequence model can also be used for the joint sequence/profile model. In the section on feature extraction and selection below, we will explain which features we extract from structural profiles and how we model those features in a probabilistic way.

For efficiency, the best segmentation remains solely based on the sequence probabilities instead of running the Viterbi algorithm using the sequence/profile product likelihoods. The profile features are thus calculated based on the best path delivered by application of the sequence model. In the section on the segmental profile features we will show that this approximation makes little difference for the features that we extract from the profiles.

### Allowing non-linear combination of segments

So far, the modeling of promoters allows for dependences within a model state, but conditional independence is



**Fig. 4.** Topology of the neural network component. In the first layer, sequence (s) and profile (p) likelihoods are grouped together. An additional input node with the total promoter sequence likelihood as computed by the Viterbi algorithm is connected with all nodes within the third layer of the net. For clarity, this input node and its connections are not shown.

assumed between the states. This does certainly not reflect the biological reality — studies have shown that there are dependences among the states, namely between TATA box and initiator or TATA box and downstream promoter element (Kutach & Kadonaga, 2000). If one of them is weakly conserved, it is much more likely that the other one is strong and will obtain a good score under the model.

To account for this, we added a neural network that takes the promoter and background likelihoods and the likelihoods produced by each state as input and is therefore able to respect arbitrary dependencies between the promoter parts. The network is trained on a disjoint part of the training set after the segment models have been established.

In principle, we are able to simply provide the sequence/structure product probability of Equation 3 in place of the sequence probability alone. A more promising way, though, is to replace each likelihood input node by a sequence/profile double node and connect them solely with each other in the first hidden layer. We thus automatically derive a linear weighting for the relative importance of sequence and physical property, similar to the approach to combine acoustic and linguistic evidence in speech recognition. Figure 4 shows the resulting network topology. In the experiments below, we used one output node, and six nodes in the previous layer.

When we look for promoters in genomic sequences, we calculate the score of the model every 10 bp, and smooth the output of the neural network with a median filter of width 3. We then set a threshold on the NN output and

report local maxima above the threshold as hits, along with the exact position where the sequence model entered the initiator state. Figure 5 shows the whole system and flow of information.

## FEATURE EXTRACTION AND SELECTION

### Segmental profile features

Even after smoothing with a mean filter, the profile of a single sequence appears rather noisy. We therefore decided not to use position-based probabilities for the profile values but rather use the mean profile values of whole segments as feature variables. The likelihood of observing a particular profile in state  $q_j$  is assumed to be generated by a Gaussian distribution:

$$c_j(p_i; [w_i, \tau_i]) := \frac{1}{\sqrt{|2\pi \Sigma_j|}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_j)\right) \quad (4)$$

with

$$\mathbf{x}_i = \frac{1}{\tau_i} \sum_{k=1}^{\tau_i} p_i^k \quad (5)$$

being the mean of the profile values within segment  $s_i$  (according to some parameter set), and  $\boldsymbol{\mu}_j$  and  $\boldsymbol{\Sigma}_j$  the parameters of the Gaussian. We therefore have the following setup for profile feature calculation:

1. Compute a profile of a selected physical property over the sequence and smooth it with a mean filter of width 21.
2. Compute the segmentation of the actual 300 bp window with the sequence-based model.
3. Calculate the mean profile values for the six segments.
4. Judge these values with a corresponding Gaussian distribution for each segment.

These features are then fed into the neural network input layer along with the corresponding sequence likelihoods, as described in the previous section. Similar profile distributions are also used for the coding and non-coding background classes. In contrast to the sequence background that consists of mixture models for sense and anti-sense, the profiles are the same on both strands. Because we use a distribution only on the mean value within a segment, small changes in the segmentation result in small changes of the likelihood; it is therefore justified to calculate the optimal segmentation based on the sequence alone (see Equation 3).

**Table 2.** Classification of promoters based on physical properties of DNA.

physical property	ERR	ROC
tri-nucleotide CG content	68.2	7336
DNA bendability	61.5	6519
A-philicity	66.1	7291
protein induced deformability	64.3	6909
B-DNA twist	50.7	5054
protein-DNA twist	71.5	7772
Z-DNA stabilizing energy	67.8	7362
nucleosome positioning	66.8	7252
stacking energy	64.3	7133
propeller twist	70.6	7650
duplex stability (disrupt energy)	61.6	6698
DNA denaturation	63.8	6982
DNA bending stiffness	70.7	7633
duplex stability (free energy)	65.0	7136

### Classification based on structural profile features

To select among the many possible parameter sets, we first studied how well a classification based on every single property could be achieved. We performed three cross-validation experiments on the data, each time leaving aside a different third as an independent test set. Half of the training data was used to train the Gaussian distributions (similar to the outlined algorithm, but with training of the Gaussian distributions in step 4), the other half as an independent data set to train a neural network that combines the likelihoods of the Gaussians. This neural network is a multi-layer perceptron with eight inputs, six hidden and one output node (corresponding to the network in Figure 4 without sequence likelihoods), and is trained with simple back-propagation.

The performance of the individual features can be seen from Table 2. We use the equal recognition rate (ERR) for classification into promoter/non-promoter and the integral over the receiver operating characteristics (ROC) as measures. A ROC curve shows the recognition rate (true positives) for pre-selected values of false positives, in our case 0 to 100 percent in one-percent steps. Then, the trapezoid rule is used to numerically compute the integral. The highest achievable value is 10,000 (100\*100, i.e. perfect recognition for all rates of false positives); a random classification results in a value of 5,000. ERR gives the recognition performance at the point where the rate of true positives equals the rate of true negatives; the ROC integral judges the performance in a more global manner. In our case, ERR and ROC are highly correlated and lead to the same ranking in most cases.

As we can see, a classification based on the profile means of the six promoter segments already results in a surprisingly high classification performance for many of the parameter sets. The physical property leading to the

highest classification rate is protein-DNA-twist (71.5% ERR). Only B-DNA twist leads to a classification that is just slightly above chance (50.7% ERR).

### Combination of several features

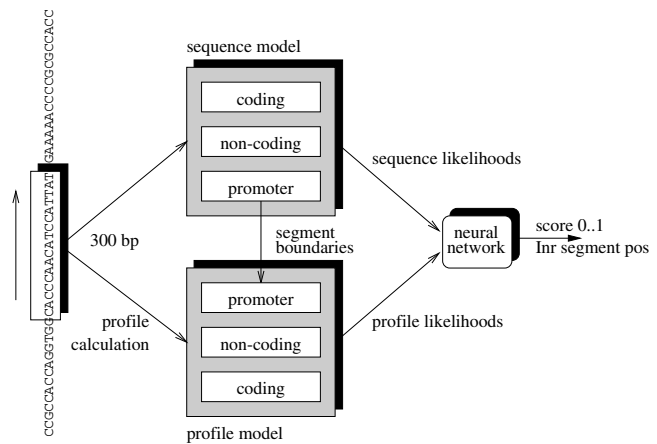
The next step is the combination of several of the features. A feature selection scheme should be based on both classification performance and correlation of the feature under consideration to the already existing feature set (Niemann, 1983). Many of the 14 parameter sets are highly correlated (see Liao *et al.* (2000) and the web supplement at <http://www.fruitfly.org/~guochun/pins.html>), and even if a certain feature delivered a good classification rate when used on its own, it will not improve much on the overall classification when it is correlated too closely to a feature in the current set.

The properties that gave the best results after protein-DNA-twist are DNA bending stiffness and propeller twist. The parameters of propeller twist are correlated with the ones of protein-DNA-twist with a correlation coefficient (CC) of 0.68, the ones of bending stiffness with a CC of -0.11. We therefore used the mean values of both bending stiffness and protein-DNA-twist and trained a two-dimensional Gaussian with full covariance matrix for every segment. Although the parameters are hardly correlated, this did not lead to an improved recognition rate.

Instead, we added the coefficient of the regression line computed for the values of protein-DNA-twist within a segment as additional feature: Even though the individual values do not properly reflect it, a distinct ascent or descent such as the increase in GC content before the TATA box (see Figure 1) might be visible from a regression line. This resulted in a slightly improved recognition rate (ROC integral value of 7812).

For some features, a modeling of the profile with a *mixture* distribution can be advantageous; e. g. for GC content, this should account for different GC isochores, or for TATA-box containing versus TATA-less promoters. In the case of protein-DNA-twist, the training of a mixture with two components by the expectation-maximization algorithm lead to two almost identical distributions in all cases and therefore to equal classification results. We manually checked this result for the TATA box region: For the three cross-validations experiments, we divided the training set into TATA box containing and TATA-less promoters<sup>†</sup>, and trained two Gaussians on the respective subsets. In all three cases, the parameters differed only slightly (Fisher criterion score,  $(\mu_1 - \mu_2)^2 / (\sigma_1^2 + \sigma_2^2)$ , between  $5.3 \cdot 10^{-7}$  and  $1.2 \cdot 10^{-6}$ ). We consequently

<sup>†</sup> MatInspector public domain release (Quandt *et al.*, 1995) with a hit above 0.7 core and 0.8 total similarity for TBP site within position -50 and 0.



**Fig. 5.** Overview of the MCPROMOTER system including profile features.

used only one two-dimensional Gaussian distribution with the mean value and the regression line coefficient of the protein-DNA-twist profile as features.

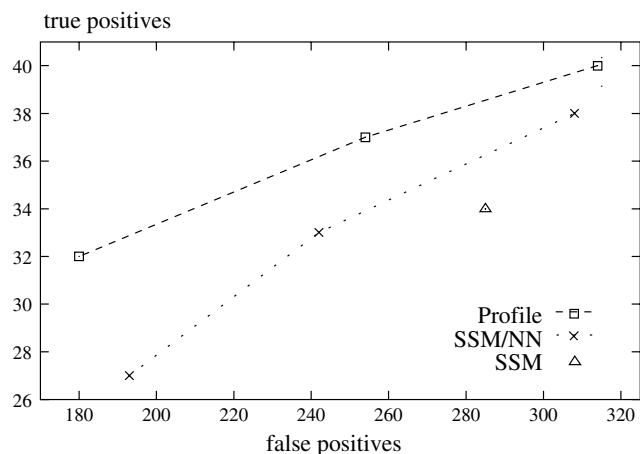
## EXPERIMENTS AND RESULTS

To test our new model on a realistic data set, we trained the integrated model as in Figure 5; half of our data was used to train the distributions (i. e., the Markov chains for the sequence and Gaussians for the structure), then the likelihoods of other half were computed and used as features to train the neural network.

We applied this model on the extensive GASP promoter data test set. The criteria for success were the same as in the GASP experiment: A promoter was considered to be found if there was a hit in the region from -500 to +50 of the annotated 5' end of the full-length cDNA to genomic alignment. The region from +50 up to the end of the annotation of the gene was used as "negative region". The total positive region is 50,600 bases long, and the negative region consists of 802,580 bases.

Figure 6 shows the results we obtained (a) for the stochastic segment models as reported by Ohler (2000), (b) the stochastic segment/neural network model without profile features, and (c) the model with profile features as in Figure 5. The different rates of true and false positives are obtained by using several threshold values of the neural network classifier.

It becomes clear that both the combination of likelihoods with a neural network and especially the integration of profile features significantly reduce the number of false positives. At a slightly lower absolute recognition rate (32 instead of 33 promoters), the number of false predictions is reduced by roughly 30% (from 1/3,530 to 1/4,740 bases)



**Fig. 6.** Results of promoter recognition. Shown are the absolute numbers of true and false positives for different thresholds. We compare the result of our previous segment model (SSM, see Ohler (2000)) with the new results obtained by a segment model/neural network hybrid (SSM/NN) and the model including profile features (Profile).

when taking profile features into account.

## DISCUSSION

We present a first step towards an integration of sequence and structural profile likelihoods in a probabilistic promoter recognition system. Our approach models a promoter as a sequence of consecutive segments, each of which is evaluated for both sequence and profile likelihood and represented by possibly arbitrarily complex sequence and profile submodels. The likelihoods produced by each state are finally combined in a neural network to allow for a non-linear weighting of the segments and for a linear weighting of sequence and profile likelihood contribution.

The idea of modeling the structural profiles solely by a Gaussian distribution on their mean values within a segment is clearly a very simple one. Nevertheless, the classification of promoter and non-promoter sequences using only these simple profile features achieves recognition rates of up to 71.5%. The ranking in Table 2 contains the result if the same single property is taken into account for all segments. It therefore represents an overall picture, and we cannot necessarily conclude from it what is most or least important for promoter function in the organism. In contrast, in their study of *E. coli* promoters, Lisser & Margalit (1994) combined three properties calculated on five segments with linear discriminant analysis. They observed that while one property contributed much to the overall classification within one segment, another one was specific for a different promoter segment. This might be an explanation for the fact that the combination of several

profiles in our model did not improve on the recognition so far: The considered properties are chosen based on the overall classification and included in all of the segments. Instead, one could think of first assessing the quality of the properties for each individual segment independently, and then model only on the suitable ones by the Gaussian distribution.

As the general setup of Equation 3 allows for arbitrary models to represent the profile likelihood, a more exact modeling of the profile slope, e. g. with continuous hidden Markov models, is also worth exploring. It remains to be tested whether this will lead to a further improvement or if the profile data simply is too noisy. Besides, there are other DNA physical properties that we did not consider so far; a more extensive collection can be found in the PROPERTY database (Ponomarenko *et al.*, 1999). We are currently including the profile likelihoods in our vertebrate promoter predictor as well; the non-linear modeling by a neural network and additional CpG island features already delivered promising results. Most important of all, though, is that a clear improvement on our *Drosophila* data set is visible, and that the incorporation of structural features helps us to recognize promoters more reliably.

The MCPROMOTER system can be accessed at <http://promoter.informatik.uni-erlangen.de>.

## ACKNOWLEDGEMENTS.

The authors wish to thank Georg Stemmer for helpful comments on the paper. Uwe Ohler is a fellow of the Boehringer Ingelheim Fonds. Gerald Rubin was supported by the Howard Hughes Medical Institute.

## REFERENCES

- Antequera, F. & Bird, A. (1993). Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. U.S.A.*, **90**, 11995–11999.
- Arkhipova, I. (1995). Promoter elements in *Drosophila melanogaster* revealed by sequence analysis. *Genetics*, **139**, 1359–1369.
- Babenko, V. N., Kosarev, P. S., Vishnevsky, O. V., Levitsky, V. G., Basin, V. V. & Frolov, A. S. (1999). Investigating extended regulatory regions of genomic DNA sequence. *Bioinformatics*, **15**, 644–653.
- Burge, C. & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol.*, **268**, 78–94.
- Fickett, J. W. & Hatzigeorgiou, A. G. (1997). Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.
- Gowher, H., Leismann, O. & Jeltsch, A. (2000). DNA of *Drosophila melanogaster* contains 5-methylcytosine. *EMBO J.*, **19**, 6918–6923.
- Ioshikhes, I. P. & Zhang, M. Q. (2000). Large-scale human promoter mapping using CpG islands. *Nat. Genet.*, **26**, 61–63.
- Kutach, A. K. & Kadonaga, J. T. (2000). The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. *Mol Cell Biol.*, **20**, 4754–4764.

- Latchman, D. S. (1998). Gene Regulation — A Eukaryotic Perspective. Stanley Thornes Ltd, third edition.
- Liao, G.-C., Rehm, E. J. & Rubin, G. M. (2000). Insertion site preferences of the P transposable element in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. U.S.A.*, **97**, 3347–3351.
- Lisser, S. & Margalit, H. (1994). Determination of common structural features in *Escherichia coli* promoters by computer analysis. *Eur. J. Biochem.*, **223**, 823–830.
- Niemann, H. (1983). Klassifikation von Mustern. Springer, Berlin.
- Ohler, U. (2000). Promoter prediction on a genomic scale — the Adh experience. *Genome Res.*, **10**, 539–542.
- Ohler, U., Harbeck, S., Niemann, H., Nöth, E. & Reese, M. G. (1999). Interpolated Markov chains for eukaryotic promoter recognition. *Bioinformatics*, **15**, 362–369.
- Ohler, U., Harbeck, S., Stemmer, G. & Niemann, H. (2000). Stochastic segment models of eukaryotic promoter regions. *Pac. Symp. Biocomput.*, **5**, 380–391.
- Pedersen, A. G., Baldi, P., Chauvin, Y. & Brunak, S. (1998). DNA structure in human RNA polymerase II promoters. *J Mol Biol.*, **281**, 663–673.
- Pedersen, A. G., Jensen, L. J., Brunak, S., H.-H. Staerfeldt & Ussery, D. W. (2000). A DNA structural atlas for *Escherichia coli*. *J Mol Biol.*, **299**, 907–930.
- Perier, R. C., Praz, V., Junier, T., Bonnard, C. & Bucher, P. (2000). The Eukaryotic Promoter Database (EPD). *Nucleic Acids Res.*, **28**, 302–303.
- Ponomarenko, J. V., Ponomarenko, M. P., Frolov, A. S., Vorobyev, D. G., Overton, G. C. & Kolchanov, N. A. (1999). Conformational and physicochemical DNA features specific for transcription factor binding sites. *Bioinformatics*, **15**, 654–668.
- Quandt, K., Frech, K., Karas, H., Wingender, E. & Werner, T. (1995). MatInd and MatInspector – New fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.
- Rabiner, L. & Juang, B.-H. (1993). Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs.
- Reese, M. G., Hartzell, G., Harris, N. L., Ohler, U., Abril, J. F. & Lewis, S. E. (2000a). Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.*, **10**, 483–501.
- Reese, M. G., Kulp, D., Tammana, H. & Haussler, D. (2000b). Genie — gene finding in *Drosophila melanogaster*. *Genome Res.*, **10**, 529–538.