Interpolated Markov Chains for Eukaryotic Promoter Recognition

Uwe Ohler[1,2], Stefan Harbeck[1], Heinrich Niemann[1], Elmar Nöth[1], and Martin G. Reese[2]

[1]Chair for Pattern Recognition (Computer Science V)
University of Erlangen-Nuremberg
Martensstraße 3, D-91058 Erlangen

[2]Department of Molecular and Cell Biology
University of California at Berkeley
539 Life Sciences Addition, Berkeley, CA 94720-3200

PREPRINT from *Bioinformatics*, 1999

1 Abstract

Motivation: *We describe a new content based approach for the detection of promoter regions of eukaryotic protein encoding genes. Our system is based on three interpolated Markov chains (IMCs) of different order which are trained on coding, non-coding, and promoter sequences. It was recently shown that the interpolation of Markov chains leads to stable parameters and improves on the results in microbial gene finding (Salzberg et al., 1998). Here, we present new methods for an automated estimation of optimal interpolation parameters and show how the IMCs can be applied to detect promoters in contiguous DNA sequences. Our interpolation approach can also be employed to obtain a reliable scoring function for human coding DNA regions, and the trained models can easily be incorporated in the general framework for gene recognition systems.*

Results: *A fivefold cross-validation evaluation of our IMC approach on a representative sequence set yielded a mean correlation coefficient of 0.84 (promoter vs. coding sequences) respectively 0.53 (promoter vs. non-coding sequences). Applied on the task of eukaryotic promoter region identification in genomic DNA sequences, our classifier identifies 50% of the promoter regions in the sequences used in the most recent review and comparison by Fickett and Hatzigeorgiou (1997), while having a false positive rate of 1/849 bp.*

Contact: *ohler@informatik.uni-erlangen.de*

2 Introduction

Today's state-of-the-art eukaryotic gene finding algorithms (such as Kulp et al., 1996; Burge and Karlin, 1998; Krogh, 1997) are based on a statistical framework which is in many cases a generalization of a hidden Markov model, also called hidden semi-Markov model. Within this framework, several scoring functions for signals such as splice sites and for regions such as exons, introns, or promoters are combined. After the search for possible signals and the judgement of the segments in between, the standard HMM decoding algorithm then provides the best path through the graph of all possible segmentations of the whole sequence. Although much progress has been made with this approach, there still is a considerable need for robust algorithms to classify the individual signals and segments, as the accuracy of the system output depends on the accuracy of its components. In the following we will present new models for the classification of individual DNA segments, and we will mainly focus on the recognition of eukaryotic promoter regions.

Popular content based measures for primary DNA sequences make use of Markov chains (MC) of a fixed order (closely related to oligomer measures) and have been employed for example in the widespread GeneMark and GeneMark.hmm prokaryotic gene finders (Lukashin and Borodovsky, 1998). Recently, the linear interpolation of Markov chains of different order has been described for microbial gene recognition (Salzberg et al., 1998). An interpolation provides a better parameter estimation, as, with increasing order of the Markov chain, the training algorithms lack a suitable amount of data because the number of model parameters increases exponentially.

Here we present a new interpolation scheme which has been successfully applied by our group for various speech recognition tasks (see Schukat-Talamazzini et al., 1997). In the context of speech recognition, interpolated Markov chains (IMCs) to judge the likelihood of symbol sequences are commonly referred to as *stochastic language models.* In contrast to the method described by Salzberg et al. (1998), who provided a function including a χ^2 test on statistical significance to calculate parameters for a linear interpolation, we use a disjoint part of the training sample to automatically estimate *optimal* interpolation parameters with respect to a statistical objective function.

We will show how this kind of IMC can improve the detection of eukaryotic promoter sequences in unknown genomic DNA. Recent progress in the understanding of the structure and function of these polymerase II promoters is reviewed in detail by (Kornberg, 1996, and other articles in the same issue) or (Nikolov and Burley, 1997).

The survey of Fickett and Hatzigeorgiou (1997) provides an excellent introduction to the topic of automated recognition of eukaryotic promoters and a comparison of the available systems for general-purpose Pol II promoter prediction. Among these are linear discriminative (Solovyev and Salamov, 1997) as well as neural network (Reese and Eeckman, 1998) or content based (Audic and Claverie, 1997; Hutchinson, 1996) methods. Content based measures were up to now either plagued by too large a number of false positives, or imposed restrictions on the number of predictions. The results obtained by our interpolated Markov chains will demonstrate the improvement of the recognition rate compared to the best methods available. Our goal was to build a general-purpose promoter recognition system that can be applied to the general task of promoter recognition; computer models constructed for specific tissue types as in (Frech et al., 1998) have a much lower false positive recognition rate. On the other hand, there is an apparent need to add a general promoter recognition module to a gene recognition system. This should help to split contiguous stretches of DNA into the right number of genes and detect the correct transcription start site which might be far upstream from the translated region.

3 Algorithm

Let us assume that we have K classes $\Omega_1 \dots \Omega_K$ and wish to classifiy a sequence $\boldsymbol{w} = w_1 \dots w_T$ with symbols w_i, taken from a finite vocabulary $\mathcal{V}$, into one class. In the case of molecular genetics, the alphabet might consist of amino acids or nucleotides. We can make use of the chain rule to compute the likelihood of a particular sequence $\boldsymbol{w}$ for each class:

$$P_k := P(\boldsymbol{w}|\Omega_k) = \prod_{t=1}^{T} P(w_i|\underbrace{w_1 \dots w_{i-1}}_{\text{context}}, \Omega_k). \qquad (1)$$

This equation shows that one symbol in a sequence is dependent on all its predecessors, i. e. on the *context* of preceding symbols. Using Bayes' rule, we are able to classify the sequence into sequence class $\hat{k}$ according to the largest *a posteriori* probability:

$$\hat{k} = \underset{k}{\operatorname{argmax}} P(\Omega_k|\boldsymbol{w}) = \underset{k}{\operatorname{argmax}} (p_k \cdot P_k) \qquad (2)$$

If we have no exact knowledge about the *a priori* probabilities p_k of our sequence classes, the values p_k are assumed to be uniformly distributed and can be neglected. We therefore need to assign a likelihood to the symbol sequence $\boldsymbol{w}$. If we can establish a model which computes this probability, we have the means to determine how likely a sequence will occur in a specific class.

3.1 Maximum Likelihood parameter estimation

In the following we will drop the condition on class Ω_k for simplicity. The right hand side of equation 1 contains a context of arbitrary length which cannot be handled; therefore, an approximation is made by imposing a restriction. A possible approximation of the probability $P(\boldsymbol{w})$ is thus made by limiting the context length to $N-1$:

$$P(\boldsymbol{w}) \approx \prod_{i=1}^{T} P(w_i|w_{i-N+1} \dots w_{i-1}) \qquad (3)$$

The resulting model is called a *Markov chain* of order $N-1$.

Our goal is to obtain parameters — in our case values for the conditional probabilities $P(w_i|w_{i-N+1} \dots w_{i-1})$ — which lead to the best possible recognition rate on the K classes under consideration. As we cannot optimize the recognition rate directly, we have to use objective functions which show the desired behaviour and for which a solution can be found. One well-known objective function is *Maximum Likelihood* (ML). If Λ_k denotes the set of the parameters of model M_k for class Ω_k, we optimize the following function $R(\Lambda_k)$:

$$R(\Lambda_k) = \prod_{i=1}^{n_k} P(\boldsymbol{w}_{ki}|M_k), \qquad (4)$$

where n_k is the number of training sequences for class k. Each class is regarded as independent of the others, and ML estimation tries to maximize the probability that the given training sample was generated, knowing to which class each sequence belongs.

Using a training sample, the ML estimation of the conditional probabilities $\tilde{P}(w_i|w_{i-N+1} \dots w_{i-1})$ can be performed simply by counting the oligomers of length N and $N-1$ in a set of training sequences:

$$\tilde{P}(w_i|w_{i-N+1}^{i-1}) = \frac{\#(w_{i-N+1}^{i})}{\#(w_{i-N+1}^{i-1})}, \qquad (5)$$

where w_x^y is an abbreviation for the partial sequence from position x to position y, and $\#$ denotes the frequency of its argument in the training sample. Here, we have to meet two problems:

1. The approximation by a large context gets closer to the real probability as denoted in equation 1. Unfortunately, the number of parameters which have to be estimated increases exponentially with the number of N, and thus the ML estimates become far from being reliable because of the limited training sample size.

2. With increasing length, some N-mers might not occur at all in the training sample. This has the consequence that the likelihood of the whole sequence $\boldsymbol{w}$ is set to zero if it contains any unseen N-mer. This

might be justified if it really is not a part of the considered class. On the other hand, the sample size might simply be too small to contain every single N-mer. As we do not know which case is true, we must not set any likelihood to zero.

A solution to these problems — the trade-off between the model context and the training sample size, and the problem of unseen N-mers — can be found by introducing a weighted interpolation scheme.

3.2 Interpolation techniques

The basic idea of applying interpolation methods is to fall back on the probability estimation of subsequences shorter than N if the frequencies of an N-mer $\boldsymbol{v} = v_1 \ldots v_N$ cannot be reliably estimated. In principle, interpolation leads us to a re-estimation of the initial parameter values (equation 5). Here, we will consider two different interpolation techniques. The first one is the *linear interpolation* between all conditional probabilities with increasing context length up to $N-1$:

$$\begin{aligned} \hat{P}(v_N|v_1^{N-1}) := & \rho_0 \frac{1}{L} \\ + & \rho_1 \tilde{P}(v_N) \\ & + \rho_2 \tilde{P}(v_N|v_{N-1}) \\ & \vdots \\ & + \rho_N \tilde{P}(v_N|v_1^{N-1}) \end{aligned} \tag{6}$$

The fraction $(1/L)$ accounts for unseen events and ensures that no probability is set to zero. The coefficients ρ_i are non-negative values which sum up to one to guarantee that the new parameter values $\hat{P}(\cdot|\cdot)$ again form a probability distribution.

Setting all the weights $\rho_0 \ldots \rho_{N-1}$ to zero and ρ_N to one is very similar to the well-known oligomer approach, with the only difference that in a Markov chain the parameters are normalized with respect to the context (see equation 5). The models with linear interpolation are thus a straightforward generalization combining oligomers of different length. The advantage of interpolation is that the model can take into account statistics of a higher order without running into the danger of overfitting the model to the training data.

Equation 6 contains only *one* vector of interpolation coefficients, whether all the subsequences up to length N really occured in the training data or not. Additionally, all parameters are treated equally, whereas the interpolation coefficient assigned to a parameter with a frequently occuring context should be larger than the coefficient for a rare event. By introducing an additional function $g_i(\boldsymbol{v}')$ which scores the reliability of the context $\boldsymbol{v}' = v_1^{N-1}$ monotonically, the linear interpolation can be extended to handle this problem accurately:

$$\hat{P}(v_N|\boldsymbol{v}') := \frac{\sum_{i=0}^{N} \rho_i \cdot g_i(\boldsymbol{v}') \cdot \tilde{P}_i(v_N|\boldsymbol{v}')}{\sum_{i=0}^{N} \rho_i \cdot g_i(\boldsymbol{v}')}, \tag{7}$$

where $\tilde{P}_i(v_N|\boldsymbol{v}')$ serves as an abbreviation for the estimates of different context lengths i, as it was shown in detail in equation 6. This interpolation scheme is called *rational interpolation.* It overcomes the problems of linear interpolation by using the function $g_i(\boldsymbol{v}')$, which we chose to be a sigmoid funtion dependent of the frequency of the last i symbols of $\boldsymbol{v}'$:

$$g_i(\boldsymbol{v}') = \frac{\#_i(\boldsymbol{v}')}{\#_i(\boldsymbol{v}') + C} \tag{8}$$

The shape of the sigmoid function is dependent on the constant bias C. In the case of $C = 0$, the function g_i is always equal to one and equation 7 becomes equivalent to linear interpolation. Also, with an increasing amount of training data, the bias C becomes less and less important; the rational interpolation thus has the largest impact if the training sample size is small.

3.3 Maximum Likelihood estimation of interpolation coefficients

We still lack the means to specify appropriate coefficients ρ_i for both linear and rational interpolation. In our approach, optimal coefficients according to the ML objective function are calculated using a second disjoint part of the training sample. This step is called *validation* and is carried out after the initial estimation of the conditional probabilities (section 3.1).

There is no closed solution for a maximum of the ML objective function in the case of interpolated Markov chains, but for the coefficients used in linear interpolation a local optimum can be found with the iterative Expectation Maximization (EM) algorithm (Dempster et al., 1977): we regard the coefficients as *hidden variables* in a double stochastic process. Afterwards, a large weight will be assigned to those contexts for which we can obtain reliable estimations; if only sparse data are at hand, the weights belonging to short contexts will be increased.

For rational interpolation, the EM algorithm cannot be applied and the computation of locally optimal interpolation weights is carried out with a gradient descent algorithm instead. The detailed re-estimation formulas are omitted at this point and can be found in (Schukat-Talamazzini et al., 1997). This automated estimation of optimal parameters is the main difference of our interpolation methods to those described for parsing microbial sequences (Salzberg et al., 1998), where the coefficients are calculated using a predefined function based on the χ^2 statistical test.

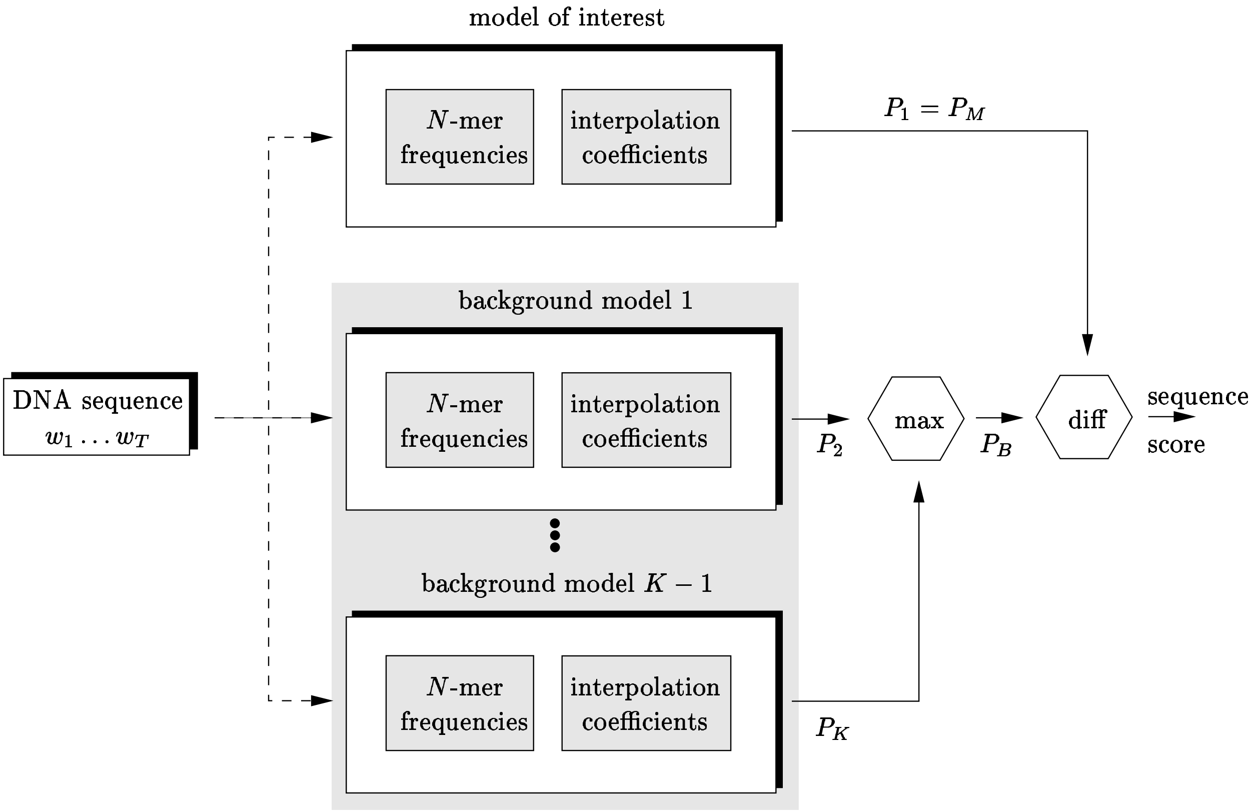

Figure 1: Content based classification with interpolated Markov chains (IMC). The output is the difference between scores of the best background model and the model of interest; this score is then classified with a suitable threshold.

3.4 Sequence classification using interpolated Markov chains

After an IMC has been trained according to sections 3.1–3.3 for each of the considered sequence classes, the IMCs can be used in parallel to classify a sequence using equation 2. Sometimes though the focus is put on the right classification of only one class. In this case, we have one class of interest and one or more "background" classes, and it is not important which particular class a sequence from the backgorund is assigned to, as long as it is not classified into the class of interest. This situation occurs in promoter recognition, where we want to distinguish promoters ("class of interest") from non-promoters (the "background" consists of several models for exonic, intronic and intergenic sequences). We can tune the IMCs with respect to sensitivity and specificity for the class of interest using the following approach: First, we compute the likelihood P_k for each class Ω_k, and then we determine the difference between the score for the model of interest P_M and the best of the background models P_B. Including a length normalization, we obtain the following equation for the total score S:

$$S(\boldsymbol{w}) = \frac{P_B(\boldsymbol{w}) - P_M(\boldsymbol{w})}{len(\boldsymbol{w})} \tag{9}$$

In practice, the logarithms of the probabilities are used because of the more efficient computation and the prevention of numerically unstable values when regarding long sequences. In figure 1 an overview of the resulting algorithm is given.

Choosing a suitable threshold value on the total score S, we can select any percentage of false positives (i. e. patterns out of one of the background classes which were classified into the class of interest). The curve of false positive rate vs. recognition rate over the whole range is called *receiver operating characteristics* (ROC) and will be used to compare the performance of different classifiers. Additionally we will provide the correlation coefficent (CC) which is defined as follows:

$$\mathrm{CC} = \frac{(\mathrm{TP} \cdot \mathrm{TN}) - (\mathrm{FN} \cdot \mathrm{FP})}{\sqrt{(\mathrm{TP}+\mathrm{FN}) \cdot (\mathrm{TN}+\mathrm{FP}) \cdot (\mathrm{TP}+\mathrm{FP}) \cdot (\mathrm{TN}+\mathrm{FN})}} \tag{10}$$

Herein, TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives; these numbers denote the absolute numbers of correctly and wrongly classified sequences.

3.5 Application of IMCs to search for regulatory regions

We will now briefly describe our system for the detection of eukaryotic polymerase II promoters in contiguous DNA sequences. The system consists of one IMC model for promoter sequences and two background IMC models for coding and non-coding sequences. To search for promoters in contiguous sequences, we use a sliding window of 300 bases (motivated by the size of the training sequences; see section 4). Every 10 bases, the current sequence in the window is classified as promoter or non-promoter using a scoring threshold that has previously been selected empirically on the training data (see figure 1). Because a whole promoter region is very likely to cause multiple predictions of several overlapping windows, a prediction is only made for each local minimum of the difference between background and promoter score which lies below the chosen threshold. The transcription start site is then assumed to be located at position 250 within the window.

To eliminate single false predictions, a post processing operation is applied on the graph of the score function S. By a smoothing algorithm, single false promoter predictions as well as single non-promoter predictions within a promoter region are filtered out. We chose to apply the hysteresis threshold algorithm, where a smoothing cursor of a chosen height is shifted over the curve from left to right. As the local minima within the smoothed graph usually comprise several positions with the same value, the prediction is then made at the position with the lowest value in the original graph. More detailed information can be found in (Ohler and Reese, 1998).

4 Data sets

We have built strongly needed representative training and test sets for eukaryotic promoter recognition which allows for a thorough comparison of different methods. These data sets are suited for algorithms aiming at human and *D. melanogaster* promoter prediction.

The data do not contain only promoter sequences which can be retrieved quite easily from the Eukaryotic Promoter Database EPD, but also carefully chosen coding and non-coding sequences. For the human promoter set, we extracted all non-related vertebrate sequences except retroviruses from EPD rel. 50 (Perier et al., 1998). Retrieving only human promoter sequences would result in a too small dataset to fit the parameters of our models; EPD release 50 contained only 181 independent human sequences. Sequences with less than 40 bases upstream or 5 bases downstream from the annotated transcription start site were discarded to assure that at least the possible TATA-box and the initiator site were contained in each entry. This resulted in 565 entries, from which sequences of 300 bases (250 upstream and 50 downstream) were extracted.

For the coding and noncoding sequences, we used the exon and intron sequences of human genes contained in the data set of 1998 for the GENIE genefinding system (Kulp et al., 1996; Reese et al., 1997). The exons were concatenated to form long coding sequences. Then, 300 bases long non-overlapping sequences were extracted. Due to the still limited amount of data, we divided the human data in five sets containing 113 promoter, 180 coding, and 869 non-coding sequences each. On these sets, reliable results can now be obtained by carrying out a fivefold cross-validation: In each experiment, the model is trained on four parts of the sequence data, leaving one part out at a time and testing the performance on the part not used for training. Then the average over all five experiments is computed and used as a result for comparison. All the data sets and more detailed information are publicly available and can be retrievd via the URL http://www-hgc.lbl.gov/inf/human.html; this site also contains a link to the similar set of *D. melanogaster* data. We encourage researchers working in the field of promoter recognition to compare their algorithms on these representative sets.

To evaluate the performance of the system on long contiguous sequences, we made use of the data set in (Fickett and Hatzigeorgiou, 1997). Using this data, we evaluated our IMC based system on a more realistic problem of recognizing transcription start sites and the corresponding promoters in DNA stretches of genomic DNA, and were able to compare our results with other programs. The set consists of 18 vertebrate sequences containing 24 annotated and experimentally proven promoters with a total of 33,120 bp. The evaluation on the contiguous sequences was carried out on both strands; recognition results are therefore given in base pairs instead of single bases.

5 Results and discussion

To get a first impression, we compared different context lengths (4–6 bases) and interpolation methods (none, linear, and rational) on the classification of human promoters and coding sequences from the fixed length sequence set (see section 4). Figure 2 shows a part of the receiver operating characteristics using IMCs of sixth order and pure simple hexamer frequencies, for which the best results could be obtained. The figure shows clearly that rational interpolation outperforms drastically the oligomer approach without interpolation; it is also superior to the simpler linear approach, thus confirming that interpolation helps us to avoid the effect of overfitting the models to the sparse training data.

As a second step we applied careful five-fold cross-validation experiments on the complete fixed length sequence set (promoters, introns, coding sequences), using IMCs with a context length of six and rational interpolation. To get a better insight, we tested the promoter model not only against both non-promoter models at once, but also individually against one non-promoter class. Table

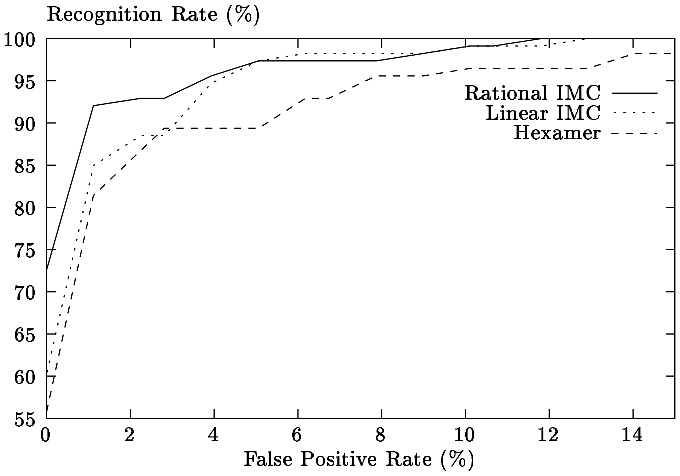

Figure 2: Comparison of the performance of IMC models with oligomer statistics without interpolation for eukaryotic promoter recognition. The results (which are the best for each considered method) were achieved for hexamers resp. IMCs based on 7-mers. The ROC curve for promoter/coding sequence classification in the range of 0–15 % of false positives is shown. The models were trained on a set of 452 promoters and 720 coding sequences of 300 bases length and evaluated on a disjoint test set of 113 promoters and 180 coding sequences.

1 therefore contains the average of the five experiments for three discrimination tasks: promoter vs. coding sequences, promoter vs. intron sequences, and promoters vs. both coding and non-coding sequences. Choosing a threshold for more than five percents of false positives here does not lead to a practically useful number of predictions.

The discrimination performance between promoters and coding regions is stunning; at a false positive rate of 5 % almost 89 % of the promoter sequences were classified correctly (correlation coefficient 0.84). Nevertheless it is also very clear that a classification between promoter and introns is much more difficult — the best CC value obtained was 0.53, at a false positive rate of 3 % and a recognition rate of 49.7 %. Most probably this stems from the much weaker information contained in the introns compared to the strong coding information of the exons. On applying models on the three-part set of promoters, non-coding, and coding sequences, the results are comparable to the two-class problem of promoters and non-coding sequences, resulting from the much larger sample size of intronic sequences. Corresponding results were obtained for the *D. melanogaster* set (Ohler and Reese, 1998).

We applied one model trained on promoters, coding, and non-coding sequences to the task of finding promoter regions in longer vertebrate DNA sequences, following the principles described in section 3.5 and using the set of contiguous sequences from the promoter prediction program survey of Fickett and Hatzigeorgiou (1997). In this survey, a prediction is judged as correct if an annotated transcription start site lies within 200 bases downstream and 100 bases upstream from the predicted site. Using this criterion, and a threshold set at a rate of 4 % false positives (highest CC value), we could detect 12 out of the 24 promoters (50 %) while having one false prediction on average every 849 base pairs. The two programs which achieved the best performance in the survey could detect 54 % and 42 % of the promoters with a false positive rate of 1/460 bp and 1/789 bp, respectively (Reese and Eeckman, 1998; Solovyev and Salamov, 1997). These numbers show that the performance of the IMCs is slightly better than the best available tools for promoter prediction, but the number of test sequences is too small to make a general statement possible.

false posit. (%)	**recognized promoters (%)**		
	promoter vs. CDS	*promoter vs. intron*	*promoter vs. CDS/intron*
0.0	58.6 (0.68)	12.9 (0.33)	3.9 (0.16)
1.0	69.4 (0.74)	32.2 (0.46)	29.9 (0.45)
2.0	78.8 (0.80)	42.5 (0.51)	41.8 (0.50)
3.0	80.5 (0.81)	**49.7 (0.53)**	48.7 (0.52)
4.0	85.7 (0.83)	51.9 (0.52)	**53.6 (0.52)**
5.0	**88.9 (0.84)**	54.7 (0.51)	56.6 (0.51)

Table 1: Promoter classification on vertebrate sequences with Markov chain models using rational interpolation and an order of six. For a certain percentage of false positives, the corresponding cross-validated recognition rate and the correlation coefficient is given. The recognition rate with the highest correlation coefficient is printed in bold (CDS = coding sequence).

An example of the performance on the longest test sequence, the human phenol sulfotransferase gene (5,663 bases, forward strand of GenBank accession code HSU54701), is shown in figure 3. Following the approach described in section 3.5, two predictions are made within this sequence, one of which is located close to one of the two annotated transcription start sites. A complete graph describing the regulatory potential over the sequence positions is calculated. Even if no clear decision is possible at the default threshold, a manually inspection of the graph may still reveal where a sudden change from regulatory (low values) to non-regulatory (high values) takes place.

A closer look at the contiguous sequences in the Fickett et al. data set and the behaviour of the system concludes this section and helps to reveal some advantages and shortcomings of the current approach:

- The overall results are certainly influenced by the fact that our system was established as a promoter predictor for human sequences, whereas seven of the 18 sequences were of non-human origin.

- One start site missed was located only a few bases downstream of the sequence start. As we score a window which is assumed to contain 250 bases upstream and 50 bases downstream, no predictions are made

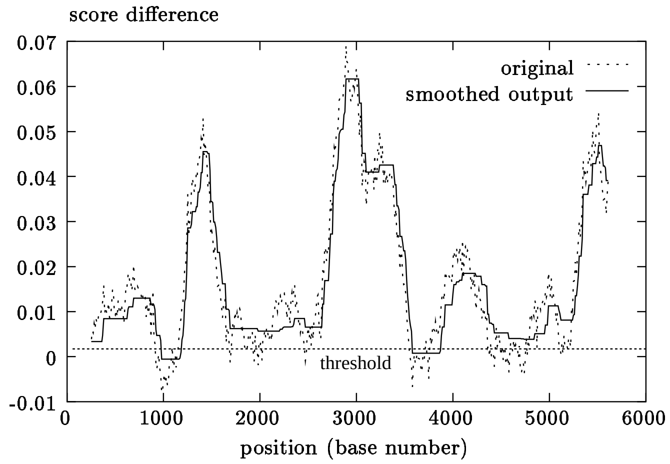

Figure 3: Output of the system on a large contiguous sequence (GenBank accession HSU54701), before and after applying the automated smoothing step. The original graph depicts the difference of the the best non-promoter and the promoter model score on the forward strand of the sequence. Two predictions are made, one for each local minimum below a predefined threshold on the smoothed output. Here, the predictions are located at positions 980 and 3580; the identified annotated transcription start site is located at position 935. Another start site located at position 2002 is not revealed.

before position 250.

- In sequence MMG67PRO, three annotated start sites were located within 300 bp, and our program made only one detection. This is not unexpected since a whole 300 bp region is scored at once, and the post processing smoothes out the small local maxima that might help to seperate the individual start sites.

- The prediction accuracy of the TSS location was quite good despite the fact that Markov chains do not use location specific information: Seven of the 12 correct predictions were made within 30 bases from the annotated start site.

- Only two of the missed promoters were also not detected by any of the nine programs evaluated in (Fickett and Hatzigeorgiou, 1997). On the other hand, one promoter detected by our Markov chains could not be identified by any other program. This means that much improvement could be achieved by a combination of several systems.

At the moment, we do not have a non-promoter model for intergenic sequences; if a reliable training sample for this sequence class can be obtained, the performance is likely to improve because of the more accurate sequence modeling. Obtaining such a sample though is difficult; most database entries contain only single genes, and for large sequences generated in the genome projects, the genes and especially the transcription start site annotations are mostly computational and not experimentally verified and therefore not reliable.

The probably most widespread application of MCs so far is found in gene recognition systems, where they serve as a classifier for coding versus non-coding parts of a DNA sequence. Thus, we also compared the performance of our interpolated models to the standard Markov chains, following the guidelines of the coding measure survey of Fickett and Tung (1992). On the GENIE data set of human exons and introns, the average recognition rate on 108 bp long sequences is 85% which is an improvement of 2.2 percent points (frame independent classification) compared to the best reviewed method, a non-interpolated Markov chain. Detailed results will be presented elsewhere.

6 Conclusions

In this paper, we describe the application of interpolated Markov chains to content based DNA classification problems. The performance of our models on two different applications, the recognition of promoter regions and the discrimination of coding and non-coding sequences, is consistently better than the one of oligomer models which realize Markov chains of a fixed order. We therefore recommend the use of interpolated models in any case, even if enough data is at hand — due to the estimation of optimal interpolation parameters, the interpolated model will in the "worst" case again result in a conventional non-interpolated Markov chain.

For the classification of promoter regions, we could demonstrate on the test set of Fickett and Hatzigeorgiou (1997) that our method performs equally or better than any signal or content based method in the survey. Signal based approaches rely on the application of position specific models, e. g. neural networks or weight matrices trained on frequently occuring pattern such as the TATA box or the initiator site. In the case of general purpose promoter prediction where no certain combination of transcription factor binding sites is expected in advance, the judgement of the overall sequence proves to be equally suitable. Further research towards the integration of content and signal based approaches therefore seems appropriate; a first step in this direction was described by (Solovyev and Salamov, 1997).

In our opinion, another important factor for the success of our promoter recognizer is the competition of several models. Promoter predictors which only consist of a model for promoter sequences and rely on a certain fixed threshold have to meet the problem that it often depends not only on the sequence itself, but also on the particular context whether a region is functionally active. Because we use several models and judge the *difference* of the particular likelihoods, this is implicitely captured.

The integration of a promoter recognition module into gene parsers like GENIE (Kulp et al., 1996) or GenScan

(Burge and Karlin, 1997), where the different sensors are trained seperately and can be easily exchanged, is in principle straightforward. But up to now, the only system incorporating a promoter module is GenScan, and this is a fairly simple model incorporating weight matrices for the TATA and the initiator region, coupled with a null model to cope with promoters with a weakly conserved core region. According to (Burge and Karlin, 1997), this approach is due to the lack of sensitivity of current predictors. The performance of promoter prediction algorithms is still much worse than those for coding regions or signals involved in the transcription process such as splice sites, and therefore a cautionless employment of a promoter module may lead to an overall deterioration of the system. Nevertheless, especially the good classification results for promoters vs. exons leads us to the expectation that a future integration of our promoter recognizer into a gene parsing framework will be successful.

7 Acknowledgments

The work of Uwe Ohler is funded by the Boehringer Ingelheim Fonds. The authors wish to thank G. M. Rubin who kindly hosted U. O. during a research sabbatical.

References

S. Audic and J.-M. Claverie. Detection of eukaryotic promoters using Markov transition matrices. *Computers and Chemistry*, 21(4):223–227, 1997.

C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268:78–94, 1997.

C. Burge and S. Karlin. Finding the genes in genomic DNA. *Current Opinion in Structural Biology*, 8:346–354, 1998.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Royal Statistical Society*, 39(1):1–38, 1977.

J. W. Fickett and A. G. Hatzigeorgiou. Eukaryotic promoter recognition. *Genome Research*, 7:861–878, 1997.

J. W. Fickett and C.-S. Tung. Assessment of protein coding measures. *Nuc. Ac. Res.*, 20(24):6441–6450, 1992.

K. Frech, K. Quandt, and T. Werner. Muscle actin genes: a first step towards computational classification of tissue specific promoters. *In Silico Biology*, 1(0005), 1998. http://www.bioinfo.de/isb/1998/01/0005/.

G. B. Hutchinson. The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comp. Appl. Biosc.*, 12(5):391–398, 1996.

R. D. Kornberg. RNA polymerase II transcription control. *Trends in Biochemical Sciences*, 21:325–326, 1996.

A. Krogh. Two methods for improving performance of an HMM and their application for gene finding. In *Proc. Fifth Int. Conf. Intelligent Systems in Molecular Biology*, pages 179–186. AAAI Press, 1997.

D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman. A generalized hidden Markov model for the recognition of human genes in DNA. In *Proc. Fourth Int. Conf. Intelligent Systems in Molecular Biology*, St. Louis, 1996.

A. V. Lukashin and M. Borodovsky. GeneMark.hmm: new solutions for gene finding. *Nuc. Ac. Res.*, 26(4):1107–1115, 1998.

D. B. Nikolov and S. K. Burley. RNA polymerase II transcription initiation: A structural view. *Proc. Natl. Acad. Sci*, 94:15–22, 1997.

U. Ohler and M. G. Reese. Detection of eukaryotic promoter regions using stochastic language models. In R. Hofestädt, editor, *Molekulare Bioinformatik*, pages 89–100, Aachen, 1998. Shaker.

R. C. Perier, T. Junier, and P. Bucher. The Eukaryotic Promoter Database EPD. *Nuc. Ac. Res.*, 26(1):353–357, 1998.

M. G. Reese and F. H. Eeckman. Time-delay neural networks for eukaryotic promoter prediction, 1998. submitted.

M. G. Reese, F. H. Eeckman, D. Kulp, and D. Haussler. Improved splice site detection in GENIE. *J. Comp. Biol.*, 4(3):311–323, 1997.

S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White. Microbial gene identification using interpolated Markov models. *Nuc. Ac. Res.*, 26(2):544–548, 1998.

E. G. Schukat-Talamazzini, F. Gallwitz, S. Harbeck, and V. Warnke. Rational interpolation of Maximum Likelihood predictors in stochastic language modeling. In *Proc. European Conf. on Speech Communication and Technology*, pages 2731–2734, Rhodes, Greece, 1997.

V. Solovyev and A. Salamov. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. In *Proc. Fifth Int. Conf. Intelligent Systems in Molecular Biology*, pages 294–302. AAAI Press, 1997.