Preprint from Genome Res. 10:539-542, 2000.

Promoter Prediction on a Genomic Scale – the Adh Experience

Uwe Ohler

Lehrstuhl für Mustererkennung

(Chair for Pattern Recognition, Computer Science V)

University of Erlangen-Nuremberg

Martensstr. 3, D-91058 Erlangen, Germany

and

Berkeley Drosophila Genome Project

Rubin Lab, Rm 539 LSA MC 3200

University of California at Berkeley, Berkeley, CA 94720

ohler@informatik.uni-erlangen.de

Abstract

We describe our statistical system for promoter recognition in genomic DNA with which we took part in the Genome Annotation Assessment Project (GASP1). We applied two versions of the system; the first uses a region based approach towards transcription start site identification, namely interpolated Markov chains, the second a hybrid approach combining regions and signals within a stochastic segment model. We compare the results of both versions with each other and examine how well the application on a genomic scale compares to the results we previously obtained on smaller data sets.

Introduction

Within the next year, the complete genomes of several eukaryotic organisms will be stored in the data bases, and we have to face the challenge that the annotation process is getting more and more complicated for the genomic sequence of higher eukaryotes such *as D. melanogaster.* The first draft of the annotation of a newly sequenced genome is usually limited to the coding part of a gene, but a complete annotation should also contain the positions of the transcription start sites (TSS), as most of the regulatory elements involved in gene expression are located in the promoter region upstream or close to the TSS.

The untranslated region between transcription and translation start site, the 5’ UTR region, can span up to several kilobases in higher eukaryotes -- it is an average of almost 2,000 bases for the transcription start site set compiled in the paper by Reese *et al.* (2000). Therefore, we cannot simply take the sequence upstream from the start codon. Methods that aim at the identification of regulatory elements in the upstream regions of co-expressed genes such as van Helden *et al.* (1998) have been shown to deliver promising results for the yeast genome which has very short UTRs, but they will be hard to apply when the annotation only consists of the coding part of a gene. Of course, TSS identification is alleviated by full-length cDNA sequencing projects; but as the sequencing always starts at the 3’ end of a gene, we need additional methods to confirm the 5’ end of the sequences, or to hunt for rarely expressed genes that are not contained in the libraries at all. We are in a desperate need to at least get a good guess where the TSS (and thus the promoter region) is located, or we will start looking for the needle in the wrong haystack.

The only available comparison of promoter prediction in genomic DNA was carried out by Fickett and Hatzigeorgiou (1997). At this time, no extensive unstudied genomic sequences were available for complex eukaryotic organisms, and the authors performed their evaluation on a set of 18 newly released vertebrate sequences, the longest of which comprised less than 6

KB. It was therefore a great challenge to see how well a recently developed promoter recognition program performs on a genomic scale, and what we can conclude for the annotation of complex eukaryotic genomes. We will briefly review the two versions of our promoter recognition system that we applied, discuss in detail the results that were described in the paper of Reese *et al.* (2000), and finally draw conclusions on the state of promoter prediction in general.

Methods and Data

McPromoter (Ohler *et al*., 1999a) is a statistical method to look for eukaryotic polymerase II transcription start sites in genomic DNA. It consists of a model for promoter sequences, and a mixture model for non-promoter sequences for coding and non-coding sequences. To localize transcription start sites, a window of 300 bases is shifted over the sequence in steps of 10 bases (see figure 1). At every position, the difference between the log-likelihood of the promoter and the non-promoter model is computed. The resulting plot describes the regulatory potential over the sequence, and is smoothed by a median and hysteresis filter (see Niemann, 1990). The program then makes a prediction for each local minimum below a pre-specified threshold (see figure 2 for an example).

We applied two versions of *McPromoter* on the *Adh* sequence (see Ashburner *et al.* (1999) for a comprehensive description of the annotated genes). The difference between the two versions lies in the structure of the promoter model, and we wanted to explore how well our more recent modeling approach improved on the recognition of TSSs. Version 1.1 of *McPromoter* is a content based approach and uses a single interpolated Markov chain (IMC) of 5th order to model promoter sequences. As such, the model does not rely on a priori knowledge about the structure of the promoters, but judges the overall composition of the sequence. For the two non-promoter components for coding and non-coding sequences, we also chose interpolated Markov chains. Related methods were described by Audic and Claverie (1997) and

Hutchinson (1996). In the figures of the GASP paper by Reese *et al*. (2000), version 1.1 is denoted by LMEIMC (Lehrstuhl für Mustererkenung – Interpolated Markov Chains). The submodels are trained using the discriminative Maximum Mutual Information (MMI) approach. In contrast to the standard Maximum Likelihood parameter estimation, MMI maximizes the probability of the decision for the correct sequence class, and therefore also takes negative samples into account (Ohler *et al*., 1999b).

In version 2.0, we replaced the single Markov chain promoter model by a more sophisticated *stochastic segment model* which consists of five states for a simplified upstream-TATA-spacer-initiator-downstream structure of eukaryotic promoters (Ohler *et al*., 2000). With this approach, we obtain more accurate statistics for the states, combining region specific states such as the one for the upstream region with states specific for individual signals such as the one for the TATA box. Hybrid approaches that exploit statistics for several regions were previously described by Solovyev and Salamov (1997) and Zhang (1998). Version 2.0 of *McPromoter* is denoted by LMESSM in the GASP overview paper (Reese *et al*., 2000).

Both versions were trained on the same representative data set consisting of *D. melanogaster* promoter and non-promoter sequences of 300 bases length, obtained at http://www.fruitfly.org/sequence/drosophila-datasets.html. Cross-validation classification experiments on this data (described in Ohler *et al*., 2000) gave a recognition rate of 27.9 % for version 1.1 and 58.8 % for version 2.0 at the very low false positive rate of 1 %. We used the system at this threshold for the evaluation of the *Adh* region.

Results

According to the results described by Reese *et al*. (2000), version 1.1 of *McPromoter* could identify 26 (28.2 %) transcription start sites with a false positive rate of 1 / 2,633 bases, and version 2.0 successfully located 31 promoters (33.6%) with the slightly higher false positive rate of 1 / 2,437 bases. This compares well with the results described in the comparison of promoter recognition algorithms in vertebrate DNA (Fickett and Hatzigeorgiou, 1997), especially considered the smaller amount of available training data for the organism of *D. melanogaster*.

A negatively surprising fact for us was the small improvement of the performance that version 2.0 achieved in comparison with the earlier version. With the results from cross-validation experiments on the representative set of promoters and non-promoters in mind, we expected the new version to localize approximately 20-30% more TSSs at the same rate of false predictions. 16 of the 26 predictions made by version 1.1 are contained in the set of 31 predictions from version 2.0. Considered that the methods are closely related, this number is somewhat small, and could be due to the different training algorithms (MMI versus ML parameter estimation).

9 predictions from version 1.1 are located within +/-40 bases of the start site (mean distance 202 bases), as opposed to 13 close predictions and a mean distance of 166 bases of the predictions obtained by version 2.0. As we do not know exactly how far the true TSS differs from our current annotation, this number is encouraging to us. Version 2.0 is clearly more successful to identify the exact position of the start sites.

Discussion

To get a better understanding why the performance of version 1.1 and version 2.0 did not differ very much from each other, we looked at the system performance without the smoothing post-processing steps (table 1). When we look at the results without the smoothing post-processing operations, it becomes obvious that the new version indeed makes great improvement, and that mainly the post-processing is responsible that version 2.0 works less well than expected. The smoothing was designed specifically for a region-based approach like the Markov chains applied in version 1.1, and works less well on a hybrid approach like version 2.0 where the promoter region is divided in several distinct segments.

A rough extrapolation of the cross-validation results at the currently used threshold (1 % false positives) leads to a worst-case rate of 1 / 2,000 bases false predictions. From the non-smoothed results it becomes clear now that this is obviously not met by reality. A possible explanation is that the available training data is still not representative enough. It certainly contains too little non-coding data, and the available promoter set has a bias towards house-keeping genes.

We already realized a number of plans to improve the model performance of version 2.0. The first idea was to include reverse sequence models for the non-promoter states, as we scan both directions of the sequence independently. It is well known that the reverse sequences of genes still resemble the true genes on the opposite strand, and that the statistics of reverse exon and intron sequences are close to the forward sequence -- hence the problem of shadow gene predictions. Nevertheless, we added two new states for reverse exon and intron sequences to have a more accurate model for the non-promoters.

In a second step, we increased the amount of training data. For the *Adh* experiment, we took the model that performed best on three cross-validation experiments and left out one third of the available data to see if our predictions on this set were met by reality. Instead, we took the whole set and determined the 1 % false positive threshold by choosing the mean threshold of the three experiments.

Finally, we replaced the median and hysteresis filters by a simple approach to allow only one prediction below the threshold within 300 bases (the model size). A similar smoothing approach is implicitly carried out by the gene finders with integrated promoter predictors; they choose the best prediction in accordance with the model topology which allows for only one prediction before the start codon. But the question remains if some predictions close to the best one might correspond to alternative transcription start sites, and if such a reduction actually filters out useful information.

As a result of these improvements, 20 predictions instead of 13 are now located within +/-40 bases from the putative start site, and we could increase the performance to 34 identified promoters with a false positive rate of 1 / 3,000 bases.

Conclusions and Outlook

The analysis of the *Adh* region showed us clearly that promoter recognition by itself, without context information, still delivers too many false positives to be practically useful on a *genomic* scale. There is still a lot of room for improvement – we think of parallel states for the TATA box region and the downstream region, discriminative training of the segment model, and a non-linear combination of the segment likelihoods. But the overall picture will maybe not change in the near future when we exploit only the primary sequence. We will see if the usage of other features such as DNA bendability (Pedersen *et al.*, 1998) can lead to the necessary improvement.

From a different point of view though, the rate of one false positive in three kilobases seems reasonable if one has already an idea where the coding part of the gene is. This information can be provided in both by alignments of cDNA to genomic sequence and ab initio gene finding. We therefore envision a promoter recognition system used within a gene finder that also incorporates EST and cDNA alignment information to extend the coding region on the 5'end. The accuracy of the TSS localization of *McPromoter* is good enough to then use such a preliminary annotation of the transcription start site for the analysis of upstream regions of co-expressed genes.

Both versions of the McPromoter system can be accessed via the World Wide Web at http://www5.informatik.uni-erlangen.de/HTML/English/Research/Promoter.

Acknowledgements

Uwe Ohler thanks the Boehringer Ingelheim Fonds for financial, and his collegues at the universities of Erlangen and Berkeley for scientific support. Special thanks to Sima Misra, George Hartzell and Martin Reese for the work on the collection and evaluation of putative TSSs in the *Adh* region, and to G. Rubin, the head of the Berkeley Drosophila Genome Project.

References

Ashburner, M., S. Misra, J. Roote, S. E. Lewis, R. Blazej, T. Davis, C. Doyle, R. Galle, R. George, N. Harris, G. Hartzell, D. Harvey, L. Hong, K. Houston, R. Hoskins, G. Johnson, C. Martin, A. Moshrefi, M. Palazzolo, M. G. Reese, A. Spradling, G. Tsang, K. Wan, K. Whitelaw, B. Kimmel, S. Celniker and G. M. Rubin. 1999. An exploration of the sequence of a 2.9-Mb region of the genome of Drosophila melanogaster: The Adh region. Genetics 153(1):179-219.

Audic, S. and J.-M. Claverie. 1997. Detection of eukaryotic promoters using Markov transition matrices. Comput Chem. 21(4):223-7.

Fickett, J. and A. Hatzigeorgiou. 1997. Eukaryotic promoter recognition. Genome Res. 7:861-878.

Hutchinson, G. B. 1996. The prediction of vertebrate promoter regions using differential hexamer frequency analysis. Comput Appl Biosci. 12(5):391-8.

Niemann, H. 1990. Pattern Analysis and Understanding, 2nd edition. Springer, Berlin.

Ohler, U., S. Harbeck, H. Niemann, E. Nöth, and M. G. Reese. 1999a. Interpolated Markov chains for eukaryotic promoter recognition. Bioinformatics 15(5): 362-369.

Ohler, U., S. Harbeck and H. Niemann. 1999b. Discriminative training of language model classifiers. Proc. European Conference on Speech and Signal Processing Technology, p. 1607-1610, Budapest.

Ohler, U., S. Harbeck, G. Stemmer, and H. Niemann. 2000. Stochastic segment models of eukaryotic promoter regions. Pacific Symposium on Biocomputing 5:377-388.

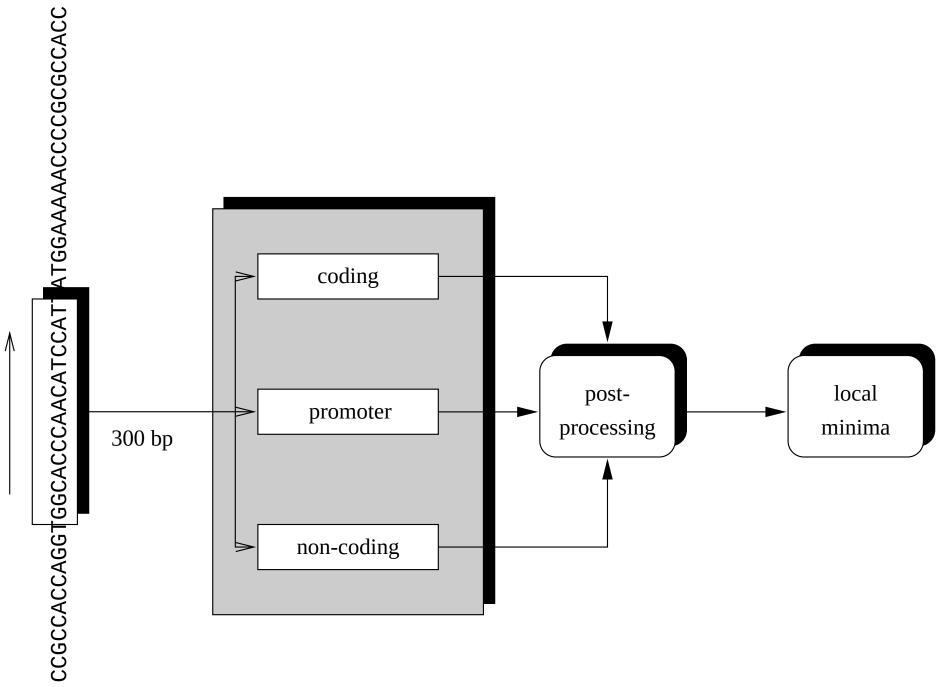
Pedersen, A. G., P. Baldi P, Y. Chauvin, S. Brunak. 1998. DNA structure in human RNA polymerase II promoters. J Mol Biol. 28;281(4):663-73.

Reese, M. G., N. Harris, G. Hartzell, U. Ohler, and S. Lewis. 2000. The genome annotation assessment project. Genome Res. 10, to appear.

Solovyev, V. and A. Salamov. 197. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. Proc ISMB 5:294-302.

Van Helden, J., B. Andre and J. Collado-Vides.1998. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J. Mol. Biol. 281(5):827-842.

Zhang, M. Q. 1998. Identification of human gene core promoters in silico. Genome Res 8(3): 319-326.


CCGCCACCAGGTGGCACCCAACATCCATTATGGAAAAACCCCGCGCCACC
300 bp
coding
promoter
non-coding
post-processing
local minima

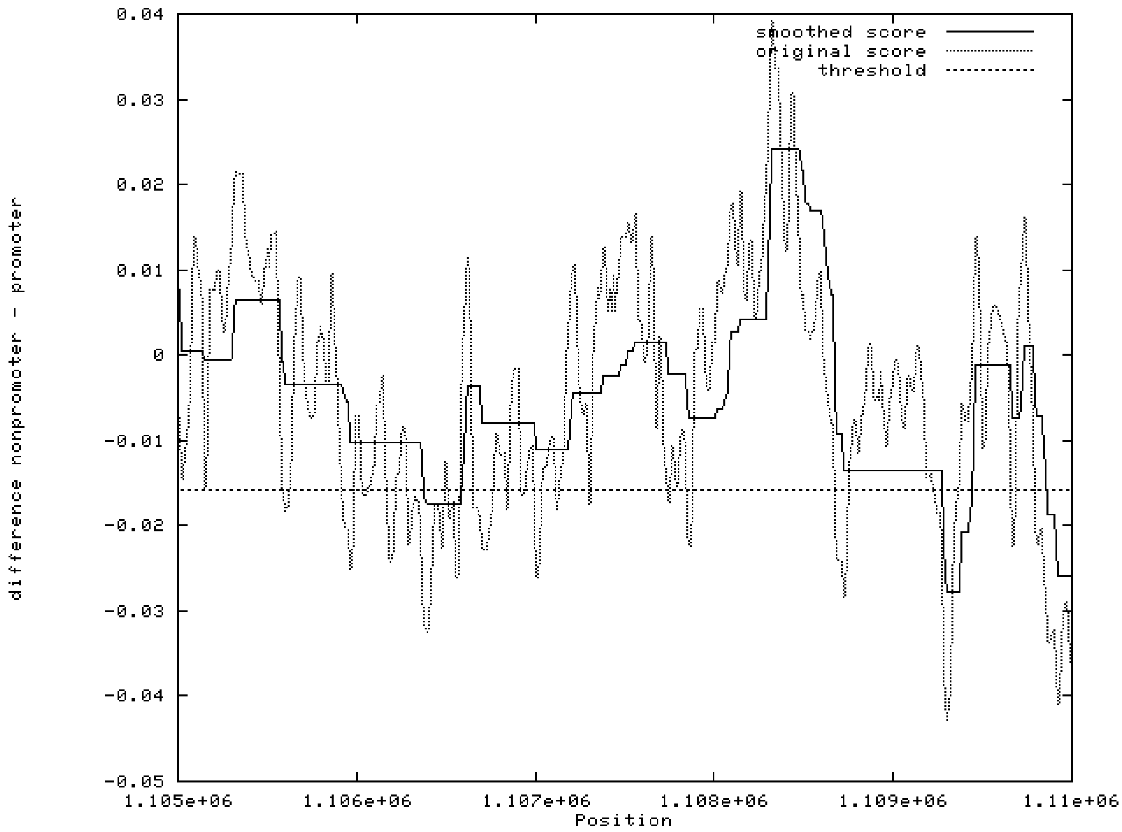

smoothed score
original score
threshold
0.04
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
-0.04
-0.05
difference nonpromoter - promoter
1.105e+06
1.106e+06
1.107e+06
1.108e+06
1.109e+06
1.11e+06
Position

Figure 1. Structure of the McPromoter system. A window of 300 bases is shifted over the sequence in steps of 10 bases, and the content is evaluated with the promoter and non-promoter models. The difference between the promoter and the non-promoter log likelihood is stored. After post-processing, the local minima are reported as transcription start site predictions.

Figure 2. Application of McPromoter version 2.0 on a 5 kB part of the *Adh* region containing the transcription start site for the *Adh* gene. We show the non-smoothed as well as the smoothed output of the system. The strongest local minimm corresponds to the annotated transcription start site of *Adh*.

	Version 1.1		**Version 2.0**	
Post-processing	*Recognized promoters*	*False positive rate*	*Recognized promoters*	*False positive rate*
None	47	1 / 450	57	1 / 719
Hysteresis	33	1 / 1,833	43	1 / 1,653
Median & Hysteresis	26	1 / 2,633	31	1 / 2,437

Table 1. Comparison of the influence of post-processing on the performance of the promoter predictors. Shown are the results without any post-processing (i.e., every local minimum is used as prediction), after hysteresis smoothing, and after both median and hysteresis smoothing. The post-processing operations reduce the number of false positives for both versions, but it becomes clear that the effect is much better for the pure region-based approach of version 1.1.