

# A Hybrid Markov Chain – Neural Network System for the Exact Prediction of Eukaryotic Transcription Start Sites

Uwe Ohler, Georg Stemmer and Heinrich Niemann  
Lehrstuhl für Mustererkennung (Informatik 5)  
Universität Erlangen-Nürnberg  
Martensstraße 3, D-91058 Erlangen  
eMail: Uwe.Ohler@cs.fau.de

With an increasing number of completely sequenced eukaryotic organisms such as *Drosophila melanogaster*, the need of a general promoter recognition system to refine the existing and forthcoming genome annotations is more urgent than ever. A reliable transcription start site annotation is *the* prerequisite to start with an analysis of the regulatory regions such as the detection of common binding sites in the promoters of co-expressed genes.

We recently developed a promoter recognition system called MCPROMOTER [1, 2] which uses stochastic segment models (SSMs) to identify promoter sequences. SSMs are a generalization of the widely used hidden Markov models: A state in an SSM contains a probabilistic submodel which generates a whole segment of a sequence at once and can therefore model dependencies among the nucleotides within a segment. This is an advantage compared to the hidden Markov model where every state emits a single symbol. In our model, we use interpolated Markov chains up to 5th order as submodels. The currently best promoter model contains six states: two for the region upstream from the core promoter, three for the core promoter including the TATA box and the initiator, and one for the region downstream from the start site. The competing non-promoter model consists of states for coding and non-coding DNA.

A drawback of the SSM formalism is that it cannot take dependencies between whole segments into account — the mutual dependency between the quality of the TATA box and the initiator is a well-known example. Therefore we added a two-layer neural network to MCPROMOTER to accomplish a nonlinear weighting of the promoter segments. The network is trained on a disjoint part of the training set, and takes the output of the SSMs as input: the likelihoods of the best paths through both promoter and non-promoter models, along with the likelihoods of each state of the promoter model. On the classification of our five-fold cross-validation set of vertebrate promoter and non-promoter sequences (<http://www.fruitfly.org/sequence/human-datasets.html>), we could achieve a sensitivity of > 60% at a false positive rate of 1%, which is a gain in sensitivity of 12 percent points when compared to the SSM alone. This is clearly an affirmation that there are indeed nonlinear dependencies, and that incorporating them leads to an increased performance of prediction methods.

We will show how the system improves on the recognition of TSSs in genomic DNA, using the set of putative transcription start sites that we compiled for the Genome Annotation Assessment Project (described in detail in [3], see also <http://www.fruitfly.org/GASP1>). Finally, we will compare MCPROMOTER's stand-alone performance with the one obtained by the combination with gene finding algorithms. An interface to MCPROMOTER can be found at <http://www5.informatik.uni-erlangen.de/HTML/English/Research/Promoter>.

## References

- [1] U. Ohler. Promoter prediction on a genomic scale — the Adh experience. *Genome Res.*, 10(4):539–542, 2000.
- [2] U. Ohler, S. Harbeck, G. Stemmer, and H. Niemann. Stochastic segment models of eukaryotic promoter regions. In R. B. Altman, K. Lauderdale, A. K. Dunker, L. Hunter, and T. E. Klein, editors, *Pacific Symposium on Biocomputing*, volume 5, pages 377–388, Singapore, 2000. World Scientific.
- [3] M. G. Reese, G. Hartzell, N. L. Harris, U. Ohler, J. F. Abril, and S. E. Lewis. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.*, 10(4):483–501, 2000.