Using Polygrams and Hidden Markov Models to Recognise Eukaryotic RNA Polymerase II Promoters

Uwe Ohler and Günther Görz
email: goerz@informatik.uni-erlangen.de
IMMD (Computer Science) VIII
University of Erlangen-Nuremberg
Am Weichselgarten 9, D-91058 Erlangen, Germany

A well-known problem in computer based DNA sequence analysis is the recognition of RNA polymerase II promoters because of the inherent structural variety of these sequences. This work deals with the application of two statistical approaches - which have been successfully used in the field of automatic speech recognition - to discriminate between promoter and non-promoter sequences.

In the first approach, a modular Hidden Markov Model was constructed. The model consists of several submodels which represent particular Polymerase II promoter elements such as the well-known TATA-, CAP-, GC-, and CAAT regions, but also the intervening sequences between those prominent elements. At first, these submodels were trained separately using the corresponding parts of primate promoter sequences which were contained in the Eukaryotic Promoter Database (EPD) rel. 40. For each submodel, the best model was chosen from among several others according to its average Z-score (a measure indicating the extent of deviation from arbitrary sequences) and using a disjoint validation sample. An interesting result consists of the Z-score for the -35 region when no TATA box is present: The best model trained with these sequences gained an average Z-score of 1.47, indicating that this region contains statistically significant sequences even when there is no visible element like the TATA box (for instance, the average Z-scores for the CAP region and the TATA box were 1.71 resp. 2.93). Finally, the submodels were combined in an appropriate manner, and the resulting whole-promoter-model was used to discriminate between promoter sequences (a disjoint test set was used) and arbitrary intron and exon sequences extracted from the same database entries as the promoters. The best results were achieved by using a reduced model which contained only the promoter front-end (the upstream region from the TATA box up to the transcription start site) and was trained again after combining the sub-elements. This model yielded a result of 78.25 % correctly classified sequences and outperformed significantly a whole-promoter-model which contained also the submodels for GC- and CAAT elements. The recognition rate for the latter model was only 60.47 %.

With the second approach, different polygram models were examined. A polygram model consists essentially of interpolated Markov chains of different order using optimal weights which are computed by the Expectation-Maximization algorithm.

In a first step, two different polygram models for promoter and non-promoter (exon and intron) sequences were constructed and different interpolation techniques were compared. The best models have a maximum Markov order of 6 (the model therefore takes into account the frequency of all small base sequences of length 1, 2, ..., up to 6 which occur in the examined sequence) and are able to discriminate between promoter and non-promoter sequences with a rate of 74.62 %. An improvement was achieved by dividing the promoter and non-promoter models into two submodels each which were trained unsupervised by a clustering algorithm. The clusters were initialized with *TATA-containing promoters – promoters without a TATA box* resp. *exon sequences – intron sequences.* Using these models, the recognition rate could be increased to 79.74 %, which is even better than the best Hidden Markov Model. This is surprising: Although the polygrams make use of statistics of higher order than the Hidden Markov Models (a standard Hidden Markov Model is always of Markov order 1), they have no knowledge of the particular order of the promoter elements.

It is very likely that further improvement is possible by combining the two approaches.

Acknowledgements:
This work was done in cooperation between IMMD VIII (Artificial Intelligence), IMMD V (pattern recognition) and the Institute of Clinical and Molecular Virolgy at the University of Erlangen-Nuremberg. We would like to thank everybody who was involved in the project, namely Dr. Frank Neipel from the Institute of Clinical and Molecular Virology for the help concerning genetics and DNA databases, and Prof. Dr. Ernst Günter Schukat-Talamazzini, Friedrich-Schiller-Univ. Jena, for the disposition of the POLYGRAM software.