# Core promoter identification using ensembles of classifiers

Thomas Abeel, Yvan Saeys and Yves Van de Peer

Department of Plant Systems Biology, VIB, Ghent University, Technologiepark 927, 9052 Gent, Belgium

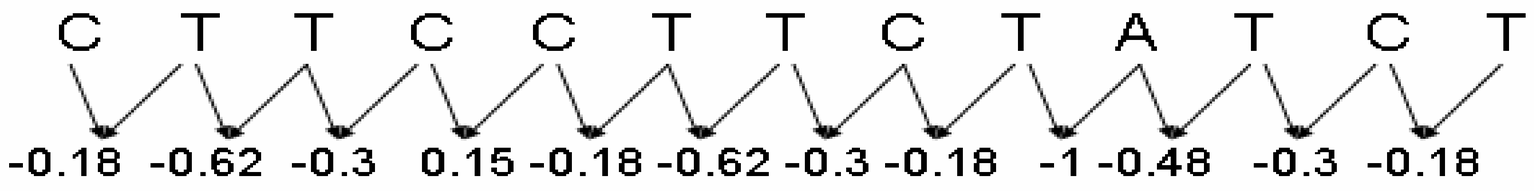
**In the context of automatic genome annotation, a number of typical problems for machine learning algorithms arise, including huge and highly skewed datasets. We propose the use of an ensemble of classifiers to construct a reliable, robust core promoter prediction program that works in a genomic context.**

## Introduction

With genomes being sequenced at an ever increasing pace, there is a need for computational approaches to help with the processing of the vast amounts of generated data. In particular, the automatic annotation of genome sequences is of much interest to genome researchers. One of the more complex tasks of genome annotation is the correct identification and delineation of the transcription start site (TSS) and the core promoter. These regions are of high interest due to their important role in transcription initiation and transcriptional regulation. It has been shown that the region around the TSS differs significantly in terms of structural make-up from other regions in the genome. In this study, we used the DNA denaturation value to convert the nucleotide sequence into a numerical profile. The DNA denaturation value indicates the energy needed to melt the DNA; high values denote rather stable regions, while low values indicate regions that melt easily. Several machine learning techniques exist that can be used for the purpose of classification and prediction. We selected support vector machines with two different kernels (RBF and polynomial) and a classification tree algorithm (C4.5). The grouping of several weak classifiers into one ensemble of classifiers has yielded promising results in many domains, but was hitherto not applied to the classification of core promoters.

## Converting DNA to physico-chemical properties

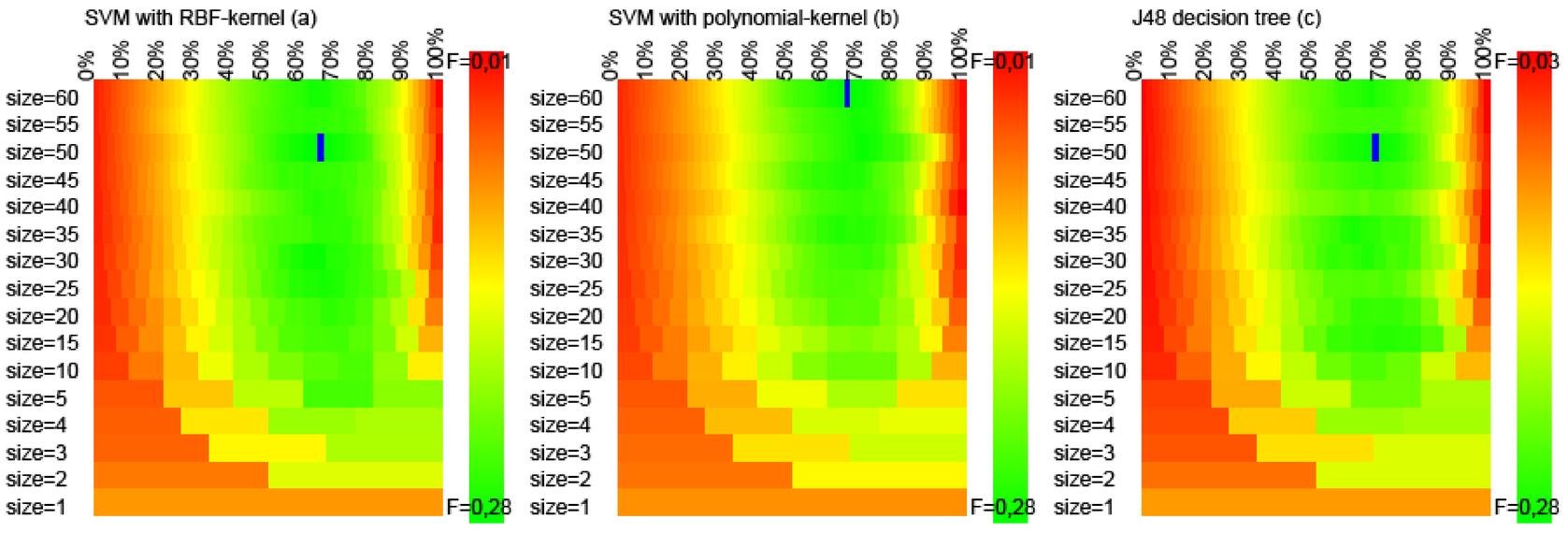
- DNA has physical and chemical properties which depend on the distribution of nucleotides A, T, G, C.
- Some of the structural properties are: Stacking energy, Propeller twist, DNA denaturation value (the one we used here), ...
- Experimentally calculated parameters allow the computation of a structural profile for any given DNA sequence
  - Computed on di- or trinucleotide scales
  - Using experimental conversion tables (see table on the right)
  - Replace every di- trinucleotide with the corresponding value. (see figure below)
  - Convert the DNA sequence into a numeric sequence




| | | | |
|---|---|---|---|
| **aa** | 66.51 | **ga** | 80.03 |
| **ac** | 108.8 | **gc** | 135.83 |
| **ag** | 85.12 | **gg** | 99.31 |
| **at** | 72.29 | **gt** | 108.8 |
| **ca** | 64.92 | **ta** | 50.11 |
| **cc** | 99.31 | **tc** | 80.03 |
| **cg** | 88.84 | **tg** | 64.93 |
| **ct** | 85.12 | **tt** | 66.51 |

The conversion table for DNA denaturation value. Each dinucleotide corresponds to a numeric value.

## Results and conclusion




**Tuning SVM parameters**

The heat map of the results of the tuning of ensemble size and threshold for a collection of support vector machines with a RBF kernel (left), a polynomial kernel (degree=2) (center) and a C4.5 tree (right). The scale goes from red to green. The parameter combination that yields the highest F-measure is shown in blue. The X-axis shows the number of agreeing models and the Y-axis shows the size of the ensemble.

### Results

The results for the RBF ensemble when applied to an assembly of the whole human genome and compared with a database of experimentally verified transcription start sites. We compared the predictions made by the ensemble to a database of experimentally verified transcription start sites that was compiled using the CAGE technique (Carninci, et al., 2006).

| Chromosome | TP | FP | FN | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|
| 1 | 27766 | 41669 | 8139 | 39.99% | 77.33% | 52.72% |
| 2 | 20659 | 28375 | 12630 | 42.13% | 62.06% | 50.19% |
| 3 | 13228 | 14199 | 13033 | 48.23% | 50.37% | 49.28% |
| 4 | 10334 | 58098 | 9789 | 15.10% | 51.35% | 23.34% |
| 5 | 16861 | 37083 | 5211 | 31.26% | 76.39% | 44.36% |
| 6 | 11872 | 12529 | 10632 | 48.65% | 52.76% | 50.62% |
| 7 | 10112 | 48161 | 11240 | 17.35% | 47.36% | 25.40% |
| 8 | 11449 | 16462 | 6575 | 41.02% | 63.52% | 49.85% |
| 9 | 9978 | 9099 | 7222 | 52.30% | 58.01% | 55.01% |
| 10 | 11559 | 14258 | 7199 | 44.77% | 61.62% | 51.86% |
| 11 | 13565 | 11778 | 8037 | 53.53% | 62.80% | 57.79% |
| 12 | 9274 | 29220 | 10772 | 24.09% | 46.26% | 31.68% |
| 13 | 6202 | 10075 | 4735 | 38.10% | 56.71% | 45.58% |
| 14 | 11675 | 203895 | 1815 | 5.42% | 86.55% | 10.19% |
| 15 | 9936 | 205916 | 3021 | 4.60% | 76.68% | 8.68% |
| 16 | 7803 | 2964 | 7307 | 72.47% | 51.64% | 60.31% |
| 17 | 14660 | 11301 | 3867 | 56.47% | 79.13% | 65.91% |
| 18 | 3939 | 3797 | 4573 | 50.92% | 46.28% | 48.49% |
| 19 | 14851 | 8710 | 2074 | 63.03% | 87.75% | 73.36% |
| 20 | 7414 | 34908 | 3884 | 17.52% | 65.62% | 27.65% |
| 21 | 3957 | 7294 | 891 | 35.17% | 81.62% | 49.16% |
| 22 | 6736 | 6125 | 1488 | 52.38% | 81.91% | 63.89% |
| X | 11352 | 41441 | 2484 | 21.50% | 82.05% | 34.08% |
| Y | 381 | 5484 | 213 | 6.50% | 64.14% | 11.80% |
| **Genome** | **265563** | **862841** | **113029** | **23.53%** | **70.14%** | **35.24%** |

### Conclusion

Ensembles of classifiers provide a fast and accurate way to identify promoters in the human genome. While there is certainly room for improvement, in particular for some chromosomes, these first results provide a good insight in the possibilities that machine learning has to offer to automated annotation of promoter regions.

### Further research

Further research can improve the presented results by applying more advanced ensemble learning techniques like bagging or boosting (Polikar, 2006). Another way to improve the performance of the proposed techniques, are more complex kernels, specifically designed for promoter detection (Sonnenburg, et al., 2006).

### References

1.Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A. M., Taylor, M. S., Engström, P. G., Frith, M. C., *et al.* (2006) *Nat Genet* **38,** 626--635.
2.Polikar, R. (2006) in *IEEE Circuits and Systems Magazine*, pp. 21-45.
3.Sonnenburg, S. o., Zien, A., & Rätsch, G. (2006) *Bioinformatics* **22,** e472--e480.