Improved core promoter prediction using ensembles of support vector machines

Thomas Abeel, Yvan Saeys and Yves Van de Peer

Bioinformatics and Evolutionary Genomics, Department of Plant Systems Biology, Flanders
Interuniversity Institute for Biotechnology (VIB), Ghent University
Technologiepark 927, B-9052 Gent, BELGIUM
E-mail : thomas.abeel@psb.ugent.be

Introduction

Computer-aided gene prediction is one of the hot topics in genome analysis because it allows for computational annotation of genomes. The region before a gene is called the promoter. This promoter, and in particular the core promoter is responsible for initiation of the transcription of a gene. The identification of gene promoters and their regulatory elements is one of the biggest challenges in bioinformatics. The core promoter is the region close around the transcription start site (TSS) (we took a region of -200, +50 around the TSS). Core promoter prediction techniques try to locate the core promoter region, or even more specific: where the transcription of a gene starts. Machine learning techniques are often used to detect putative transcription start sites. However, there are several problems. First the datasets involved are rather large, ranging from several thousand training instances for the positive data to several hundreds of thousands of negative samples. The second problem is the imbalance between positive and negative samples. In the human genome, only 3% of the sequence codes for genes, and an even a smaller amount of sequence data is located within the core promoter. This large class imbalance is very difficult to model in e.g. support vector machines (SVM) as there is so little positive information.

Results

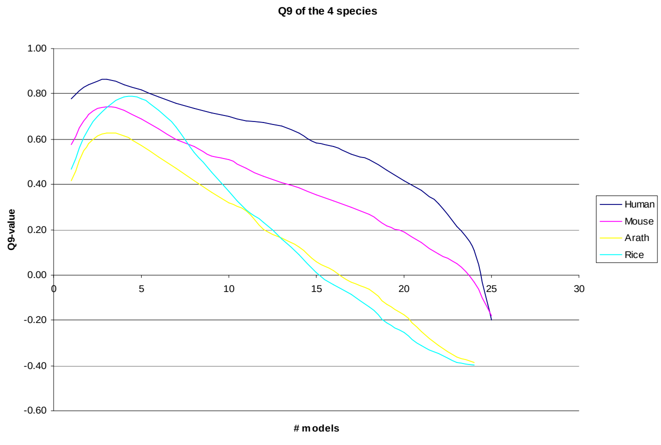
Here, we explore a technique to reduce the training time for support vector machines, while increasing their predicion performance. The used technique is an **ensemble of support vector machines**, each trained on a different training set. Every one of these support vector machines is then validated on a separate validation dataset which is different from the training data. This step is needed to **determine the optimal number of support vector machines that must have a positive output to classify an instance as positive**. While training one radial SVM with increasing large datasets does not increase linearly, training seperate SVM's on parts of the large dataset does increase linearly when we go for single coverage of the original dataset. We have also explored the possibility of selecting a larger number of random subsets of the original data. There is no certainty that all data are used, but due to the larger number of subsets there is a higher chance they are. It is not necessary for all data to be used.

Trainingsize	TP	FP	TN	FN	Q9	Recall	Precision	Time‡
Single (6000)	0.75	0.10	0.90	0.25	0.62	75.00%	88.24%	0d 11:29:11
Single (9000)	0.75	0.11	0.89	0.25	0.61	75.00%	87.21%	1d 09:42:07
Single (12000)	0.75	0.10	0.90	0.25	0.62	75.00%	88.24%	2d 05:20:40
Single (15000)	0.76	0.10	0.90	0.24	0.63	76.00%	88.37%	3d 10:28:47
Single (18000)	0.77	0.10	0.90	0.23	0.65	77.00%	88.51%	4d 09:58:40
Ensembl (83200)	0.95	0.08	0.92	0.06	0.86	94.06%	92.23%	0d 03:27:11

$$Q9 = \begin{cases} \dfrac{TN-FP}{TN+FP}\, if\,(TP+FN=0) \\ \dfrac{TP-FN}{TP+FN}\, if\,(TN+FP=0) \\ 1-\sqrt{2}\sqrt{\left(\dfrac{FN}{TP+FN}\right)^2+\left(\dfrac{FP}{TN+FP}\right)^2}\, if\,(TP+FN\neq 0 \wedge TN+FP\neq 0) \end{cases}$$

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$



We have compared performance on different measures for the ensemble of support vector machines and a single support vector machine. This technique was applied to a core promoter classification task. Here we tried to distinguish core promoters from gene and intergenic sequences. We have performed our analysis on four different species: Rice, Arabidopsis, Mouse and Human. The datasets can be considered as large, with 2500-7000 positive training examples and ten times more negative ones, each sample consisting of 250 features. Training on a dataset of 4500 positive and 13500 negative samples on a single SVM took over 4 days and gave a recall of 77%, precision 88% and a Q9 value of 64%. The validation was done on the trainingdata using a 10-fold cross validation.

Conclusion

Our approach with 26 SVM's each trained on 3200 samples (800 positive and 2400 negative) with a validation of 8000 sequences (80 positive, 7920 negative) performed better. The validation used is much stricter as it is much more difficult for the SVM to predict true positives. The training and validation only took a couple of hours instead of 4 days. We have obtained a recall of 94%, precision 92% and a Q9 value of 86%. This analysis clearly shows that the use of an ensemble of support vector machines is superior to the use of one single SVM, both in time as in classification performance. The optimal number of agreeing models is three for Human, Mouse and Arabidopsis and four for Rice.

References

•Burges, C. J. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2, 2 (1998), 121–167.

•Dietterich, T. G. Ensemble methods in machine learning. In MCS '00: Proceedings of the First International Workshop on Multiple Classifier Systems (2000), Springer-Verlag,pp. 1–15.

•Florquin, K., Saeys, Y., Degroeve, S., Rouze, P., and Van de Peer, Y. Largescale structural analysis of the core promoter in mammalian and plant genomes. Nucl. Acids Res. 33, 13 (2005), 4255–4264.

‡Analysis was done on a single Linux (Linux version 2.4.21-40) machine with 4 Intel(R) Xeon(TM) CPU 2.80GHz processors and 8 Gb memory. The program for the analysis was not written as a multithreaded program, therefore only one CPU was used.