On Discriminating a TATA-box from putative TATA boxes: A Case Study Using Plant Genome

Raja Loganantharaj
The Bioinformatics Research Lab
Center for Advanced Computer Studies
University of Louisiana at Lafayette
logan@cacs.louisiana.edu

Prediction of a promoter is one of the many active areas of bioinformatics. The outcome of a promoter detection algorithm is directly or indirectly influenced by the success of identifying the location of a TATA box in a promoter sequence. A profiling technique is very often used to find putative TATA boxes, but discriminating a TATA box from putative TATA boxes is still a challenging problem. In this work, we formulate the problem and provide solutions using both a linear and a non linear classifiers.

1. Introduction

The outcome of a promoter prediction algorithm is dependent on correctly identifying a TATA box since the location of a TATA box influences the location of a transcription start site (TSS). The core TATA box, which is defined by the consensus sequence 5'-TATAWAW-3', matches with several subsequences of a genomic sequence and these matched subsequences are called putative TATA boxes. The *W* stands for either *A* or *T*. The problem of detection of a TATA box becomes the problem of discriminating a TATA box from putative TATA boxes. Our previous study on TATA and TATA less promoters[1] revealed that the surrounding substrings have the key in discriminating a TATA box from putative TATA boxes. To study this problem, we have selected the annotated plant promoter sequences from PlantProm database that consists of plant TATA and TATA less promoters. The main purpose of this investigation is to study the composition of the subsequences surrounding a TATA box and their abilities to identify a TATA box from a set of putative TATA boxes. Also we have focused on studying the influence of the length of the subsequences in the discrimination.

The problem of discriminating a TATA box from putative TATA boxes becomes a classification problem using the surrounding substrings of a putative TATA box. We have selected a linear classifier naïve Bayes[2] and a non linear classifier an artificial neural network (ANN)[2], for their effectiveness in classifying patterns. The training and testing data sets were created from the positive and negative instances of the surrounding genomic substrings.

The effectiveness the naïve Bayes depends on the correct estimates of the prior probabilities and many systems compute the prior probabilities from the training sets which do not reflect the true prior probabilities. Instead of finding a way to correctly estimate the prior probabilities, we focused on finding a threshold decision point that maximizes the overall prediction accuracy [3]. With the optimal decision threshold, the naïve Bayes classifier had outperformed the results of that of a neural network.

2. Review on Basic Techniques

Let a string, say $e_{-k},e_{-k+1},\ldots e_{-1}$,Core-Tata,$e_1,..,e_k$, represents k nucleotides ($e_{-k},e_{-k+1},\ldots e_{-1}$) upstream and ($e_1,..,e_k$) downstream from a core TATA box. Then $P(tata|e_{-k},e_{-k+1},\ldots e_{-1},e_1,..,e_k)$ and $P(\neg tata|e_{-k},e_{-k+1},\ldots e_{-1},e_1,..,e_k)$ respectively represents the conditional probability of a TATA or a non TATA for the surrounding sequence $e_{-k},e_{-k+1},\ldots e_{-1},e_1,..,e_k$. Each e_r takes one of $\{a,c,g,t\}$.

$P(tata|e_{-k},e_{-k+1},\ldots e_{-1},e_1,..,e_k)$
is rewritten using Bayes theorem [2] as
$=P(tata,e_{-k},e_{-k+1},\ldots e_{-1},e_1,..,e_k)/P(e_{-k},e_{-k+1},\ldots e_1,e_1,..,e_k)$
Using chain rule and applying positional independence
$=C".P(e_k|tata).P(e_{k-1}|tata)..P(e_{-k}|tata)$ where $C"$ is $P(tata) / P(e_{-k},e_{-k+1},\ldots e_1,e_1,..,e_k)$
$Log(P(tata/sub\_sequence)$
$= C + \sum(\log(P(e_r|tata)/ P(e_r|\neg tata))$ for all r from –k to k, where C is a constant representing $\log(P(tata)/ P(e_{-k},e_{-k+1},\ldots e_1,e_1,..,e_k))$.

To determine whether a sequence surrounds a TATA, we compare the posterior probabilities. If $P(tata| e_{-k},e_{-k+1},\ldots e_1,e_1,.., e_k) > P(\neg tata|e_{-k},e_{-k+1},\ldots e_1,e_1,.., e_k)$, the sequence is classified as a TATA as has been defined by maximum posteriori hypothesis or a MAP hypothesis.

2.2 Detection of TATA Box

The consensus sequence of a TATA core, which is given by 5'-TATAWAW-3' where W is either a or t, alone does not help to detect a TATA box. The TATA box is usually determined by positional weighted matrix constructed from a profile. We have used the profile of a TATA box of plant genome presented in [4]. The background probability of a nucleotide at any position is taken to be 0.25.

2.3 Artificial Neural Networks

A feed forward multi-layer back propagating neural networks have been used for many applications involving pattern recognition and prediction purposes in bioinformatics [2].

The training and the test sequences consist of nucleotides of offset *k* around TATA box or putative TATA boxes and each of these sequences is modeled as a string of length *2k* with alphabets *a,c,g* and *t* representing possible nucleotides.

A neural network tends to trap into a local minimum that prevents the network to achieve the best convergent state during training. By using momentum with training, we can reduce the local minima trap. By trial and error approximation, we have set the learning rate to 0.05 and the momentum coefficient to 0.1.

3. Experiments

We have downloaded promoter sequences of plant genome from PlantProm DB (http://mendel.cs.rhul.ac.uk/mendel.php?topic=plantprom), an annotated non-redundant collection of proximal promoter sequences.

We started the experiment with the collection of all TATA boxes in the annotated promoters. The neighboring strings in the upstream and the downstream of a TATA promoter become the positive training subsequences. Similarly, we have detected all the putative TATA boxes from all the non TATA promoters and collected the substrings from both the upstream and the downstream of the putative TATA boxes. These strings form the negative instances.

We have computed mean values of classification accuracy, true positive and false positive along with their standard deviations for the 30 test sets (note that the training and the test set are disjoint). The results are shown in Figure 1.

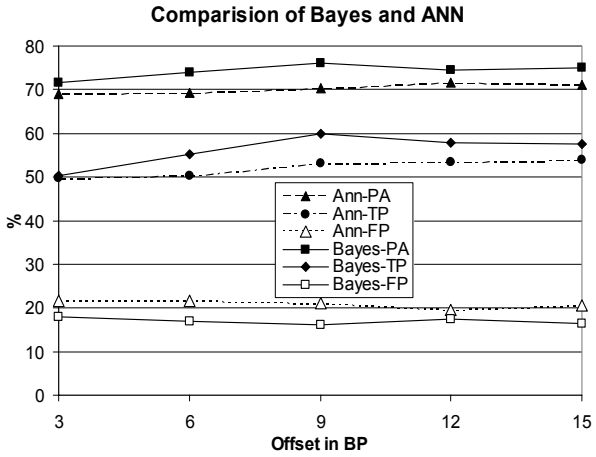

Figure 1: Comparison of Bayes with ANN

4. Summary and conclusion

Both the naïve Bayes and a three layered neural network were trained and tested with the same training and test data sets. From the graph in Figure 1, it is clear that the naïve Bayes with optimal threshold outperformed the neural network. From the empirical tests, we can draw few conclusions: (1) the naïve Bayes method outperformed a neural network classifier, (2) the length of the substrings surrounding a putative TATA box have some influence on the outcome and the best results occurred with the offset of 9 bp. Even though, the results are specific to the plant genome, the approach is general enough to be used for other applications involving discriminating any pattern from putative patterns found in sequences.

5. References

[1] R. Loganantharaj, M. E. Karim, and A. Lakhotia, "Recognizing TATA promoters based on discriminating frequency analysis of neighborhood tuples," presented at Biot-04: First Biotechnology and Bioinformatics Symposium: A Community and Academic Forum, Colorado Springs, Colorado, 2004, September.

[2] P. Baldi and S. Brunak, *Bioinformatics: the machine learning approach*, 2nd ed. Cambridge, Mass.: MIT Press, 2001.

[3] R. Loganantharaj, "An Approach to Estimate the Prediction Accuracy of Data Mining: A Case Study with Human Genome," presented at the 9th World Multiconference on Systemics, Cybernetics and Informatics (WMSCI 2005), Orlando, Florida, 2005.

[4] I. A. Shahmuradov, A. J. Gammerman, J. M. Hancock, P. M. Bramley, and V. V. Solovyev, "PlantProm: a database of plant promoter sequences," *Nucleic Acids Res*, vol. 31, pp. 114-7, 2003.