Predicting a Transcription Start Site: Case Study with Different Genomes

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

Prediction of a transcription start site (TSS) is one of the many active research areas in bioinformatics. The main purpose of this paper is to study the ability of linear classifiers for predicting a TSS. Also we have focused on the relationship between the length of the subsequences surrounding TSS and their effectiveness.

1. Introduction

Prediction of a transcription start site is one of the many active research areas in bioinformatics. If we can identify a transcription start site (TSS) from a given DNA sequence, we can infer the start of a coding region of a gene. The detection of a promoter, which involves the identification of all the relevant transcription binding sites and the TSS, is a challenging task and the many approaches that were proposed to solve the problem yield high false positive ratio in the order of 29 to 72%[1] .

To understand and to improve the very high false positive ratio of promoter detection, we investigate the composition of the sequences around transcription start sites of the following genomes: human, rat and mouse. The have developed an algorithm using positional weighted matrix (PWM) to estimate the prediction accuracy of a TSS using any linear classifiers and we have verified the results using a popular linear classifier, Naïve Byes.

For this study, we have used the annotated promoter sequences of human genome (33,961 promoters), rat genome (5,705 promoters) and mouse genome (22,549 promoters) that were downloaded from the site http://biowulf.bu.edu/zlab/promoser/download.html, each contains 2000bp upstream and 100bp downstream of a transcription start site.

2. The Algorithm and the Method

The objective is to understand the composition of the neighboring subsequences of a transcription start site of all the promoters in a genome and the effect of the size of the sequences surrounding a TSS in distinguishing themselves from the rest of the sequences. In this section, we briefly outline the basis of PWM [2] and naïve Bayes method [2]. The PWM involves the positional frequency distribution of each nucleotide associated with a pattern and it also involves in comparing the distribution of a pattern with the background (the absence of the pattern).

Let a string, say $e_{-k},e_{-k+1},\ldots e_{-1},e_1,..,e_k$, represents *k* nucleotides upstream and downstream from a transcription start site (TSS). Then $P(tss|e_{-k},e_{-k+1},\ldots e_{-1},e_1,..,e_k)$ and $P(\neg tss|\ e_{-k},e_{-k+1},\ldots e_{-1},e_1,..,e_k)$ respectively represent the conditional probability of the sequence being a TSS or not. Note that each e_r takes one of $\{a,c,g,t\}$. The likelihood ratio

$P(tss|e_{-k},e_{-k+1},\ldots e_{-1},e_1,..,e_k)/P(\neg tss|e_{-k},e_{-k+1},\ldots e_{-1},e_1,..,e_k\ )$

is rewritten using Bayes theorem [2] as

$=P(tss,e_{-k},e_{-k+1},\ldots e_{-1},e_1,..,e_k)/P(\neg tss,e_{-k},e_{-k+1},\ldots e_{-1}, e_1,..,e_k\ )$

Using chain rule and applying positional independence

$= P(e_k|tss)...P(e_{k-1}|tss)..P(e_{-k}|tss).P(tss)\ /(P(e_k|\neg tss).P(\ e_{k-1}\ |\neg tss)..P(\ e_{-k}\ |\neg tss).P(\neg tss))$

The log likelihood ratio becomes

$Log(P(tss/sub\_sequence)/P(\neg tss/sequence))$

$= C + \sum(log(P(e_r|tss)/\ P(e_r|\neg tss))$ for all r from –k to k, where C is a constant representing $log(P(tss)/P((\neg tss)$.

The log likelihood ratio of a nucleotide at a particular location, say nucleotide *g* at position *r* denoted by $log(P(e_r{=}g|tss)/P(e_r{=}g|\neg tss))$, is represented by a weight in a matrix corresponding to the nucleotide and the location (g and r).

Naïve Bayes method classifies a subsequence as a TSS, if the conditional probability $P(tss|\ e_{-k},e_{-k+1},\ldots e_{-1}, e_1,..,e_k) > P(\neg tss|e_{-k},e_{-k+1},\ldots e_{-1},e_1,..,e_k)$. To avoid underflow in the computation of the conditional probability, *log* of the conditional probabilities are computed and compared.

When using either one of these methods, the prior probabilities of P(tss) and p(¬tss) must be computed. The objective of any detection algorithm is to improve the prediction accuracy while minimizing the false positive and false negative ratios. From all the positive and negative instances of the training sets, obtain the log likelihood ratios and their positive and negative frequency distributions. If the distributions overlap, the best prediction accuracy occurs at the intersection of these two distributions [3]. The threshold point that maximizes the prediction accuracy of the training set is used for testing.

4. Experiment and Empirical Results

We have used human, rat and mouse genome to understand the composition of the subsequences around a transcription start site of each genome. The background probability distributions of the nucleotides are obtained by computing the occurrences of each nucleotide in the sequences that do not overlap with the TSS sequences.

We have started experimenting with human genome with subsequences of length 20bp around a transcription start site, 10bp in the downstream and 10bp in the upstream of the TSS. From all the 33,961 promoter subsequences around a TSS of human genome, we have computed positional probability distribution of each nucleotide (note that a TSS is at position 1).

We have repeated the experiment for subsequences of length 20bp for the mouse and rat genome and computed the weight distributions for TSS and non TSS for each genome. We computed the prediction accuracy by calculating the areas corresponding to the false positive and false negative as have been illustrated in [3]. The estimated prediction accuracy for sequence of length 20 (-10 to +10) around TSS of different genome is shown in Table 1.

Genome	Naïve Bayes Method		Estimated Prediction accuracy
	Mean Value	Standard Deviation	
Human	61.36%	2.208	62.61%
Rat	65.95%	1.803	65.87%
Mouse	68%	1.934	68.13%

Table 1: Estimated prediction accuracy is compared with that of naïve Bayes.

4.1. Influence of the Length of Subsequences on Prediction Accuracy

Now we can study the influence of the length around TSS in the prediction accuracy. We have conducted experiments for different length of subsequences around the TSS from 6bp to 20bp in the downstream and in the upstream and have estimated the prediction accuracy using the method that we have outlined. The relationship between the prediction accuracy and the length of subsequences are shown for each genome in Figure 1.

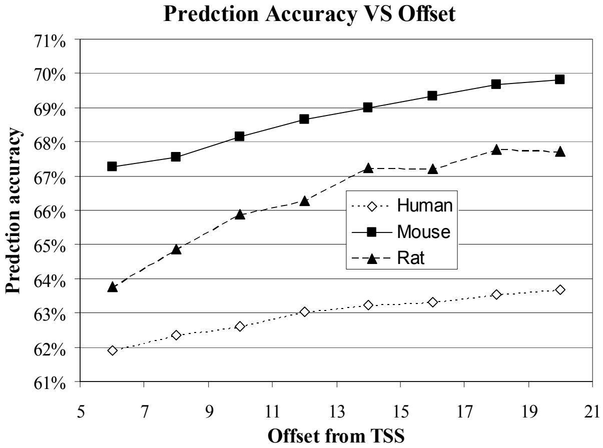

Figure1: Influence of sequence length on the prediction accuracy for different genome

5. Summary and Conclusion

We have presented an algorithm based on PWM to estimate the prediction accuracy of any linear classifiers. Our estimated prediction accuracy agrees with that of obtained by a naïve Bayes classifier as has been shown in Table 1. We also have investigated the influence on the length of the subsequences surrounding a TSS on the prediction accuracy. The subsequences varied from 12 through 40 in steps of 4 (starting from -6 , +6 to -20 , +20 in step of 2). The length has positive influence on the prediction accuracy.

6. References

[1] J. W. Fickett and A. G. Hatzigeorgiou, "Eukaryotic promoter recognition," *Genome Res*, vol. 7, pp. 861-78, 1997.

[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.