Homework #4: Predicting the Presence of RNA Polymerase II

Anže Pečar (63060257)

December 24, 2012

1 Introduction

In homework #4 we try to predict regions where RNA polymerase II binds to the DNA. We do this by implementing an algorithm that constructs a hidden Markov model and predict the polymerase presence with the Viterbi algorithm.

2 Methods

Viterbi We have tried many versions of the Viterbi algorithm with little success. The problem with this approach is that the probabilities become too small when the sequence is longer.

Viterbi log Instead of multiplying probabilities we sum up their logs. We have tried implementing an algorithm based on the implementation from wikipedia and from Refrence [1], but results weren't as good as Bayes.

Forward We have also implemented the forward algorithm from Refrence [1]. The results were not better than Bayes.

Bayes In order to evaluate other features we used the NaiveBayes learner from Orange. Predicting polymerase II presence was done by comparing probabilities, the hidden state with the highest probability was chosen.

Bayes Average Because "absent presence" is never chosen by the basic Bayes method, we calculate the average probability for each hidden state. If the probability of one state is above it's average it can be chosen.

Bayes Mediana Instead of comparing to average probability we tried comparing the median probability. With this we managed to predict some absent (1) presence but the overall score was lower than Bayes and Bayes Average.

RandomForest Beside NaiveBayes we also tried the RandomForest learner. We used the same method for classification as with the Bayes learner, but the score was not as good.

3 Results

In Table 1 we can see our Kaggle results. We have not submitted results for Viterbi log and forward algorithms as the resulting file would definitely score lower.

Table 1: Kaggle scores

Submission n.	Method	Result	Comment
1	Viterbi	0.48911	Viterbi
2	Bayes	0.68788	Normal Bayes
3	Bayes Average	0.67975	Bayes with averages
4	Bayes Mediana	0.60958	Bayes with mediana
5	Forest	0.65413	Random forest with 100 trees

4 Conclusion

The simplest of all algorithms ended up being the best. When we added extra complexity (either by using RandomForrest or by using average/median probabilities with Bayes) the score ended up being lower. We have made a total of 7 submissions.

Honor Code

My answers to homework are my own work. I did not make solutions or code available to anyone else. I did not engage in any other activities that will dishonestly improve my results or dishonestly improve/hurt the results of others.

References

[1] Nello Cristianini, Matthew W. Hahn (2007) Introduction to computational genomics: A case study approach, Cambridge University Press, Cambridge, UK.