

Using simple rules on presence and positioning of motifs for promoter structure modeling and tissue-specific expression prediction

Alexis Vandebon¹
alexvdb@hgc.jp

Kenta Nakai^{1,2,3}
knakai@ims.u-tokyo.ac.jp

¹ Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

² Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

³ Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency, 5-3 Yonbancho, Chiyoda-ku, Tokyo 102-0081, Japan

Abstract

Regulation of transcription is controlled by sets of transcription factors binding specific sites in the regulatory regions of genes. It is therefore believed that regulatory regions driving similar expression profiles share some common structural features. We here introduce a computational approach for finding a small set of rules describing the presence and positioning of motifs in a set of promoter sequences. This rule set is subsequently used for finding promoters that drive similar expression profiles from a genomic set of sequences. We applied our approach on muscle-expressed genes in *Caenorhabditis elegans* and obtained a high predictive performance. We believe that our model can help us to increase our understanding about transcription factor cooperation and transcription initiation.

Keywords: regulation of transcription, promoter modeling, *C. elegans*, muscle tissue

1 Introduction

As a first step in the regulation of gene expression, regulation of transcription is of major importance in determining when, where (e.g. in which tissues), and under what conditions a gene is expressed. Since this process is controlled by transcription factors (TFs) binding motifs in the regulatory sequences, we can make the assumption that regulatory regions containing similar sets of motifs are bound by similar sets of these TFs, and will drive similar expression profiles. Here we present an approach for modeling promoter architectures from a set of input promoter sequences, and for subsequent prediction of regulatory regions with a similar structure and tissue expression using the trained model. Our model takes into account presence or absence of motifs, positional bias with regard to the translation (or transcription) start sites, and positional preferences between pairs of sites. Not only the presence or absence of each of these patterns is considered, but also their number of presences.

2 Methods

We applied our model on a dataset of *Caenorhabditis elegans* muscle-specific promoter sequences [2]. This true positive dataset was divided into two training sets and a test set. In the first training set we performed *de novo* motif prediction using four popular motif prediction algorithms and selected over-represented motifs. Subsequently, we generated a set of patterns concerning the presence and positioning of motifs in the training sequences that might be useful for distinguishing muscle-expressed genes from genes not expressed in muscle. Finally, we used a genetic algorithm approach to find within this set of patterns a small subset of biologically meaningful rules, using the second training set. This small subset of rules constitutes our promoter structure model. This analysis was repeated on 10 bootstrap runs.

3 Results and Discussion

After training the model on muscle-specific promoter sequences, we scored the genomic set of promoters. Using the set of true positive (TP) and true negative (TN) test sequences we constructed an ROC curve, and obtained AUC values of up to 0.76. In the case of the best performing model almost 50% of the TP test sequences scored higher than 90% of the TN test sequences, which outperforms a previous study on the same dataset [2]. We showed that the high scoring non-input promoters are enriched for promoters driving expression in muscle tissues such as body wall muscle (p-value: $7.68e-5$) and vulval muscle (p-value: $4.44e-4$). Finally, we found that motifs fitting the rules in our model show a significant tendency to be present in regions that have previously been shown to be of importance in driving expression in muscle tissue (p-value: 0.0017). Figure 1 shows the 6 rules in the best performing model.

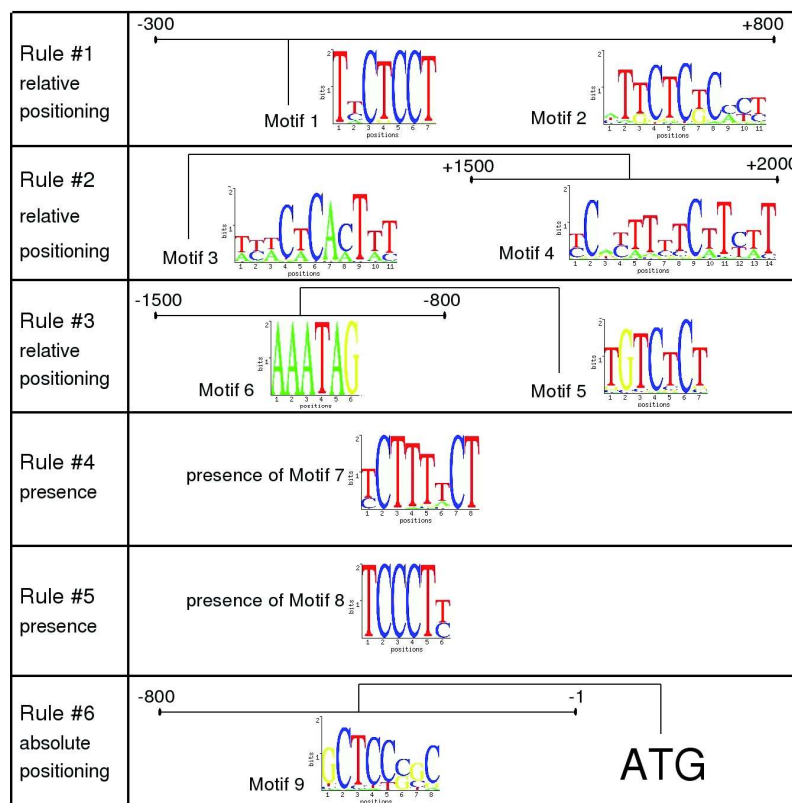


Figure 1: A visual representation of the 6 selected rules in the best performing model. For each rule the motif logo of the motif(s) and the nature of the rule (“presence”, “absolute positioning”, or “relative positioning”) is shown.

4 Concluding Remarks

We have introduced a method for modeling promoter architectures using a simple set of rules on motif presence and positioning. Despite its simplicity our model achieves a high performance. Refer to our full paper [1] and our oral presentation on Wednesday 3 December for a more detailed description.

References

- [1] Vandebon, A., Nakai, K., Using simple rules on presence and positioning of motifs for promoter structure modeling and tissue-specific expression prediction, *Genome Informatics*, 2008.
- [2] Zhao, G., Schriefer, L. A. and Stormo, G. D., Identification of muscle-specific regulatory modules in *Caenorhabditis elegans*, *Genome Res*, 17, 348-57, 2007.