論文内容の要旨

論文題目 Modeling Architectural Patterns in Promoter Sequences for Tissue-specific Expression Prediction
(組織特異的発現のためのプロモーター構造パターンのモデル化)

氏名 バンデンボン　アレクシス

Introduction

In eukaryotes, the activity of genes and their products is regulated on many levels. Regulation of transcription is the first step in the cascade of regulation, and as such it 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 factor (TF) binding motifs in regulatory sequences, we can make the assumption that regulatory regions containing similar sets of motifs are bound by similar sets of TFs, and thus drive similar expression profiles.

Here I present three studies on the analysis and modeling of the architecture of regions regulating the initiation of transcription. In these three studies, models are presented that aim at capturing some common structural features from promoter sequences that drive similar expression profiles. Subsequently, the trained models are used for finding other promoter sequences that share similar structural features and result in similar expression profiles.

Results and Discussion

1. A Markov chain-based promoter structure model

In a first study, a Markov chain-based promoter structure model was introduced. The promoter sequences are divided into two regions, reflecting the positional preferences of regulatory sites with regard to the translation or transcription start site (TSS). Next, a first order Markov chain is constructed for each region, capturing order and orientation of the sites in each region. After training, the model is used to score a genomic set of promoter sequences. High-scoring promoters are assumed to have a structure similar to the input sequences, and are thus expected to drive similar expression patterns. Figure 1 shows a visual representation of the scoring process.

First, we trained our model on a set of promoters driving expression in pharyngeal muscle cells in *Caenorhabditis elegans*. Using available annotation data we confirmed that

high-scoring non-input promoters were enriched for promoters driving expression in pharyngeal muscle cells (P-value = 0.0025) and in muscle tissue in general (P-value = 0.0072), illustrating the validity of the model. Second, we trained the model on a set of muscle-specific promoters in the sea squirt *Ciona intestinalis*. For four high-scoring non-input genes *in situ* hybridization experiments were conducted, confirming expression in muscle tissue for three of them.

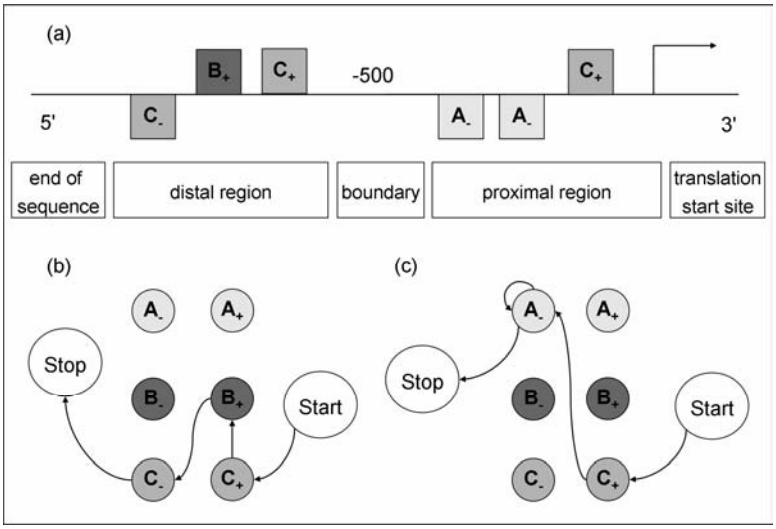

Figure 1. The scoring process of the Markov chain-based promoter structure model. (**a**): A promoter sequence to score. The squares represent predicted sites for motifs A, B and C, with '+' and '-' indicating their orientation. Here, the boundary between the proximal and distal regions is set at -500 bp. (**b**) and (**c**): The promoter model during the scoring process of the distal region and the proximal region, respectively. The states of the model are shown as circles. Each of the two regions has a 'start' and a 'stop' state, in addition to states for each motif type in both orientations. Arrows indicate transitions used during the scoring of the sequence shown in (a). Values of these transitions are captured in the Markov chains during training of the model.

2. A model based on simple rules on presence and positioning of motifs

In a second study, structural rules were constructed concerning the presence of regulatory sites, their positioning relative to the TSS, and the relative positioning between pairs of sites. During training, a large number of such patterns are extracted from a set of training promoters. Subsequently, a Genetic Algorithm (GA) is used to select from this pool of patterns a small subset of highly meaningful patterns optimizing performance on a second set of training samples. This final subset of structural patterns represents our promoter structure model.

A ten-fold cross-validation approach on muscle-specific promoters from *C. elegans* indicated that this model is capable of finding architectural patterns characteristic of muscle-expressed

promoters. On average 27.1% and 36.5% of muscle-expressed promoters scored higher than 95% and 90% of the control promoters, respectively. In the best validation run, we found that almost 50% of muscle-expressed promoters scored higher than 90% of the controls promoters. A visual representation of the six structural patterns selected in this validation run is shown in Figure 2. The selected patterns contain a wide variety of structural information, and some motifs present in the patterns show similarity to known motifs of importance in muscle-specific regulation of transcription. High-scoring non-training sequences were enriched for muscle-expressed genes, and predicted regulatory sites fitting the patterns showed a tendency to be present in experimentally verified regulatory regions (P-value = 0.0017).

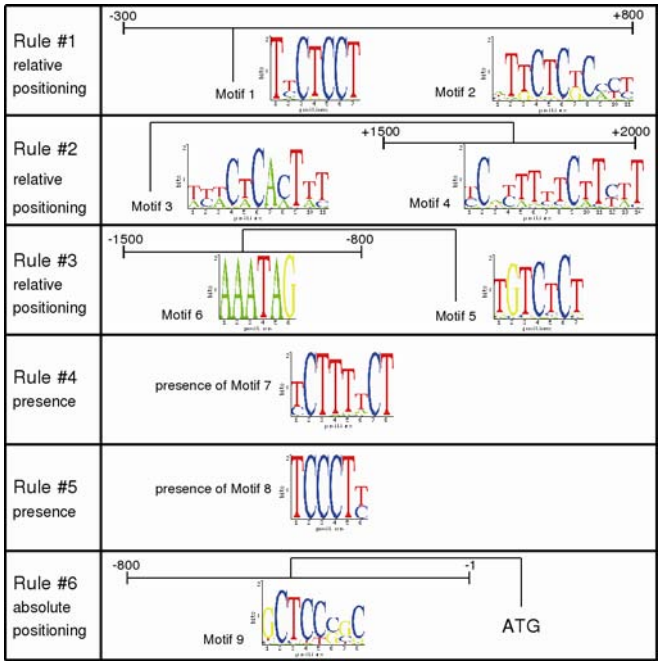

Figure 2. A visual representation of the 6 selected rules in the best cross-validation run for the *C. elegans* muscle promoter model. For each pattern, the sequence logo of the motif(s) and the nature of the pattern are shown.)

3. A large-scale analysis of tissue-specific promoter structures

Finally, we applied an updated version of the rule-based model described above on 26 human and 34 mouse tissues. In this updated version, the GA does not only select a subset of structural patterns, but it also assigns a weight to each pattern, reflecting its importance. Promoter structure models were constructed for each tissue, and ten-fold cross-validation was used to evaluate the ability of each model to distinguish positive test samples from control promoter sequences. As measures for performance, the Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) curves, and the sensitivity at 90% and 95% specificity were used.

We found that the models had statistically significant performance in 35 out of 60 tissues. Models with high performance include those for tongue, (fetal) liver, kidney, and skeletal muscle. Table 1 shows an overview of the five best performing human promoter models.

Further analysis of the important structural patterns in these models revealed that many involve TFs known to be of importance in the tissues in question, such as HNF1 and HNF4 in liver promoter models, and MEF2 in skeletal muscle promoter models. In addition, we found that promoter models of a tissue in one species tend to have high performance when applied on promoter sequences of the same tissue in the other species. For example, the human kidney promoter models are able to recognize mouse kidney-specific promoter sequences, and vice versa. Finally, promoter models of related tissues, such as liver and kidney, tend to have high inter-tissue performance.

Description	Size (No. of seqs)	AUC value (P-value)	Sensitivity at 95% specificity	Sensitivity at 90% specificity
tongue	76	0.81 ($< 6.0 \times 10^{-5}$)	0.46	0.58
fetal liver	89	0.79 ($< 6.0 \times 10^{-5}$)	0.28	0.39
kidney	95	0.71 ($< 6.0 \times 10^{-5}$)	0.20	0.30
skeletal muscle	67	0.70 ($< 6.0 \times 10^{-5}$)	0.21	0.35
liver	276	0.68 ($< 6.0 \times 10^{-5}$)	0.24	0.34

Table 1. Overview of the five best performing human models. A description of each dataset, the number of promoter sequences it contains, the average AUC value of the ROC curves obtained form the 10 cross-validation runs, and a corrected P-value for this value is shown. Finally, the sensitivity at 90% specificity and 95% specificity is shown.

Conclusions

We introduced two approaches for modeling tissue-specific promoter architectures. Predictions of the Markov chain-based model were validated using available annotation data and experimental results. In a second approach, a GA was used to select a small set of simple rules on the presence and positioning of regulatory sites in tissue-specific promoter sequences. Both models are capable of capturing a wider variety of structural features, compared to *cis*-regulatory module-based models. The third study showed that such features can be used to model tissue-specific structural features on a large scale in higher eukaryotes. We believe that our approaches can be useful for finding promising candidate genes for wet-lab experiments, and for increasing our understanding of the regulation of transcription.