

Modeling the bacterial transcription regulatory jigsaw

Scott David Mann¹
sdman@deakin.edu.au

Yi-Ping Phoebe Chen^{1,2}
phoebe@deakin.edu.au

¹ Faculty of Science and Technology, Deakin University, Victoria, Australia

² ARC Centre of Excellence in Bioinformatics

Abstract

The computational approach for identifying promoters on increasingly large genomic sequences has led to many false positives. Toward developing a novel method in this field, a computational analogue of the biological process was sought. In comparison to RNA polymerase, the model architecture attempts to replicate the recognition of promoter elements. The significant contribution involved the hybrid formed via aggregation of the profile hidden markov model (recognition elements) with the artificial neural network (aggregator), via viterbi scoring optimisations. The benefit obtained using this architecture include the modeling ability of the profile hidden markov model with the ability of the artificial neural network to associate elements composing the promoter.

Keywords: prokaryote promoter identification, profile hidden Markov model, artificial neural network

1 Introduction

Techniques to achieve promoter identification typically rely on the detection of consensus motifs. This scheme works well when identifying highly conserved motifs in short biological background sequences, however, motifs quickly get lost and produce many false positives in larger background sequences. The most effective use of consensus identification is toward seeking relatively large motifs, e.g. CpG islands, the classical introduction to hidden Markov models (HMMs).

Key to overcoming these issues is the maximisation of the information provided in the promoter, typically containing low informational content. The bacterial promoter consists of several sequences with varying degrees of conservation that enable recognition by the RNA polymerase complex. To address the situation of incorporating additional neighbouring biological information into computational models, an effective motif extraction technique, such as the HMM, needs its results placed in a biological context, as can be achieved by the artificial neural network (ANN). By filtering HMM results via an ANN, the false positives obtained by an HMM-only scheme can be reduced. This approach effectively builds a profile for the regulatory region using two of the most powerful pattern searching techniques available. The stimuli for this hybrid approach is founded in the properties of each model as outlined in prior publication [2].

2 Method and Results

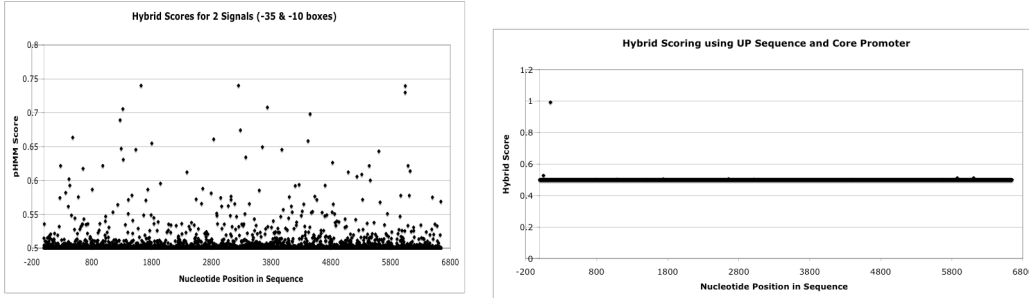
Methodology supporting the hybrid architecture centers on two distinct phases. The recognition elements (pHMMs) are trained on multiple sequence alignments of motifs. In this study the motifs included UP sequences, -35 and -10 box sequences. Multiple sequence alignment via ClustalW allowed for greater information content per position in the motif. The pHMM scores for both positive and negative promoter sequences formed the stimuli for training the artificial neural network.

The HMM-ANN hybrid transfer function for the hidden layer that receives the pHMM values is:

$$f(v) = \frac{1}{1 + e^{-v_k(N) \in \pi}}$$

Results generated by the hybrid served to eliminate false positives in the promoter region as shown in Figure 1.

Figure 1: Hybrid scoring distribution over sequence AB102735 (Left) classical Pribow box, (Right) incorporation of UP element and Pribnow box.



Comparative results of the pHMM-ANN hybrid against other methods of promoter recognition are described in [2] and presented in Table 1.

Table 1: Comparative measures of promoter identification performance (bacterial UP sequence rRNA annotated promoter dataset)

Method	Sensitivity (S_n)*	Specificity(S_p)*	Precision (P)*
NNPP [3]	0.4	0.0625	0.210526316
SAK [1]	0.4	0.1875	0.235294118
pHMM-ANN hybrid	0.700	0.833	0.777

3 Discussions

The model has been successfully applied to various biological motifs containing multiple sub-components of conservation. The generality and the ‘pluggable’ nature of the architecture will enable the broader application of the technique to feature detection beyond promoter sequences.

References

- [1] Gordon L, Chervonenkis AY, Gammerman AJ, Shahmuradov IA, Solovyev VV: Sequence alignment kernel for recognition of promoter regions. *Bioinformatics* 2003, 19(15):1964-1971.
- [2] Mann S, Li J, Chen YP: A pHMM-ANN based discriminative approach to promoter identification in prokaryote genomic contexts. *Nucleic Acids Res* 2007, 35(2):e12.
- [3] Reese M: Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Computers & Chemistry* 2001, 26(1):51-56.