Improving the prediction of transcription factor binding sites by flexible inclusion of biophysical properties

http://bioit.irc.ugent.be/physbinder

Pieter De Bleser PIETERDB@IRC.VIB-UGENT.BE
Arne Soete ARNE.SOETE@IRC.VIB-UGENT.BE
Inflammation Research Center, VIB, Ghent, Belgium

Department of Biomedical Molecular Biology, Ghent University, Ghent, Belgium

Yvan Saeys YVAN.SAEYS@IRC.VIB-UGENT.BE
Inflammation Research Center, VIB, Ghent, Belgium

Department of Respiratory Medicine, Ghent University Hospital, Ghent, Belgium

Keywords: transcription factor binding site prediction, random forest, biophysical properties

Abstract

The most important mechanism in the regulation of transcription is the binding of a transcription factor (TF) to a DNA sequence called the transcription factor binding site (TFBS). Most binding sites are short and degenerate, which makes predictions based on their primary sequence alone somewhat unreliable.

We present a web tool that implements a flexible and extensible algorithm for predicting TFBS. The algorithm makes use of both direct (the sequence) and several indirect readout features of protein-DNA complexes (biophysical properties such as bendability or the solvent-excluded surface of the DNA). This algorithm significantly outperforms state-of-the-art approaches for in silico identification of transcription factor binding sites.

Users can submit FASTA sequences for analysis and choose from more than 77 different TF binding models. Based on previously published work (Hooghe et al., 2012; Broos et al., 2013).

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).

1. Introduction

DNA-binding specificity of TFs consists of a direct and an indirect readout component. The direct readout mechanism is well-defined and involves recognition of specific DNA bases by amino acids. Indirect readout is much less well-defined but takes into consideration protein-DNA interactions that depend on base pairs that are not directly contacted by the protein. These protein-DNA interactions essentially reflect the influence of the structure and thermodynamic properties of the DNA before or upon binding by the TF. We present a sequence-based method that uses the Random Forest (RF) algorithm with features that cover either nucleotide positional dependencies (the NPD model) or nucleotide-sequencedependent structural characteristics of the TFBS and its flanking sequences (the structural model). Our method optimally combines both models and tries to integrate the PWM score in the combined model. We compare our models with each other and compare their accuracy with the accuracy of high-quality PWMs and with an alternative method, CRoSSeD (Meysman et al., 2011). 77 vertebrate TF models are available for binding site predictions. Results can be visualized together with the ENCODE TFBS data track in the UCSC Genome Browser (http://genome.ucsc.edu/).

2. Method

Figure 1 gives an overview of the approach. A, the input from which models are built consists of the two

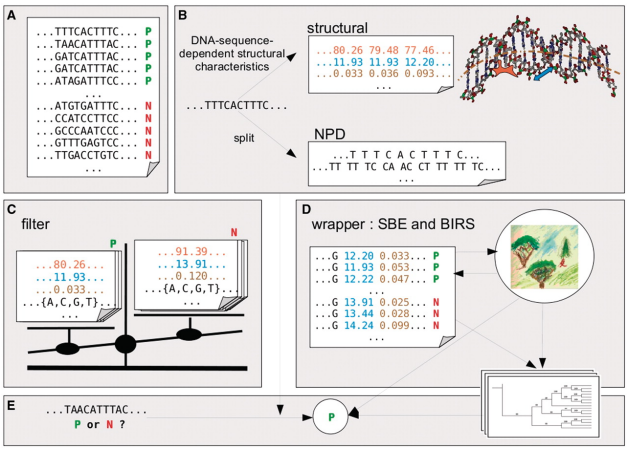


Figure 1. Overview of the approach.

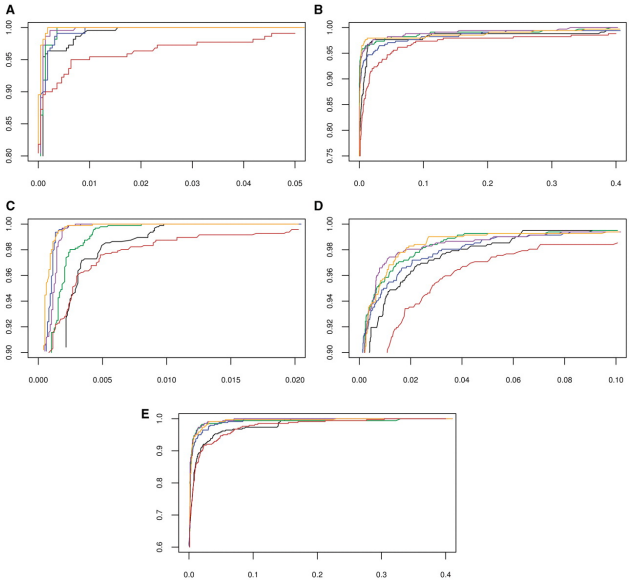


Figure 2. ROC curves for various TFs: (A). HIF1 (B) P53; (C) SP1; (D) STAT1; (E) TBP. Models: PWM (black), NPD (green), struct (blue), NPD_struct (purple), NPD_struct_PWM (orange), CRoSSeD (brown).

classes of nucleotide sequences. One class contains positive sequences (P, green); the other contains negative sequences (N, red). B, Each sequence is converted into multiple series of values providing values for a specific DNA structural characteristic at all positions of the TFBS and its context (structural model), or consists of one-base or two-base parts of the sequence (NPD). C, Basic selection of positions is made by statistical comparison of distributions of values for positive and negative sequences. D, Further selection is performed through wrapper-based feature selection. Several models with one characteristic might be merged through a best incremental ranked subset scheme (BIRS). The final NPD model and final structural model can be merged into one integrative model. E, The resulting model can be used by Random Forest to predict the likelihood that a nucleotide sequence is a TFBS.

3. Results

With respect to the classification accuracies, for the eukaryotic TFs (Figure 2), both structural and NPD models perform better than the PWM for four out of five TFs (HIF1, SP1, STAT1, TBP). Overall, the NPD model performs better than the structural model (four out of five cases). This is logical because the structural method almost exclusively captures the shape readout mechanisms of DNA-binding specificity. The NPD model, in contrast, is expected to capture base readout, as well as some portions of the shape readout that can be derived from nucleotide positional dependencies. Nevertheless, the structural models alone perform better than PWM in four out of five cases.

4. Conclusions

A web tool is presented that implements a flexible and extensible algorithm for predicting TFBS. We show that the integrative approach used by the algorithm combining two or three different approaches establishes the best possible prediction of TFBSs.

References

Broos, S., Soete, A., Hooghe, B., Moran, R., van Roy, F., & De Bleser, P. (2013). Physbinder: Improving the prediction of transcription factor binding sites by flexible inclusion of biophysical properties. *Nucleic Acids Research*, *41*, 531–534.

Hooghe, B., Broos, S., van Roy, F., & De Bleser, P. (2012). A flexible integrative approach based on random forest improves prediction of transcription factor binding sites. *Nucleic Acids Research*, *40*, e106.

Meysman, P., Thanh, H. D., Laukens, K., Smet, R. D., Yan, W., Marchal, K., & Engelen, K. (2011). Use of structural dna properties for the prediction of transcription-factor binding sites in escherichia coli. *Nucleic Acids Research*, *39*, e6.