Use of structural DNA properties for the prediction of regulator binding sites with conditional random fields

Pieter Meysman[1], Thanh Hai Dang[2], Kris Laukens[2], Riet Desmet[1], Yan Wu[1], Kristof Engelen[1] and Kathleen Marchal[1]

[1]Department of Microbial and Molecular systems, K.U.Leuven, Kasteelpark Arenberg 20, B-3001 Leuven Heverlee, Belgium. [2]Intelligent Systems Laboratory, Department of Mathematics and Computer Science, Middelheimlaan 1, B-2020 Antwerpen, Belgium.

Molecular recognition of genomic target sites by regulator proteins is a vital process in the transcription regulation of genes in living cells. The types of physical interactions that contribute to the recognition of binding sites by a protein can roughly be divided into those enabling direct read-out and those that allow for indirect read-out [1]. The former comprises base-specific recognition, such as stabilizing hydrogen bonds between regulator amino acids and a set of conserved bases in the genomic DNA sequence, while in the case of the latter variations within the DNA structure will be used as the basis for recognition. It is the direct form of recognition that is the focus of most current endeavors to model regulator binding sites, usually by modeling a conserved set of nucleotides, e.g. a position weight matrix (PWM). However by considering only a single recognition mechanism, these models overlook any information concerning binding site identity that can be derived from the use of indirect read-out by the regulator. It was therefore our goal to create a binding site model based on this second type of recognition which involves interactions between the regulator protein and the molecular structure of the DNA molecule.

Methodology

We have developed a model which scores sites for the likelihood that they are bound by a certain transcription factor based on specific characteristics in the DNA sequence or structure. These characteristics must first be learned by the model during a training step from known binding sites of the transcription factor.

The structural DNA properties of the binding sites, needed to construct the model, are derived from their nucleotide sequence using a number of higher-order value look-up functions, so-called structural scales, which are based on experimental data (e.g. X-ray crystallography of various DNA molecules). Thirteen profiles representing different structural DNA properties, such as DNA rigidity or stability, were used together with the DNA sequence as input data to train a model representing the common structural features shared by all known binding sites of a specified regulator. This was done using conditional random fields (CRF), a discriminative machine learning method designed to label sequential data [2]. Two novel extension algorithms were included in the training of the models, namely an optimization method which allows the CRF to work with structural DNA properties, and a correction method which can compensate for any bias in the training set towards nucleotide conservation. Once trained, the models could be used to evaluate the likelihood of regulator binding for any given DNA sequence. We have named this general modeling framework CRoSSeD (Conditional Random fields of Smoothed Structural Data) [3] and an overview can be found in figure 1.

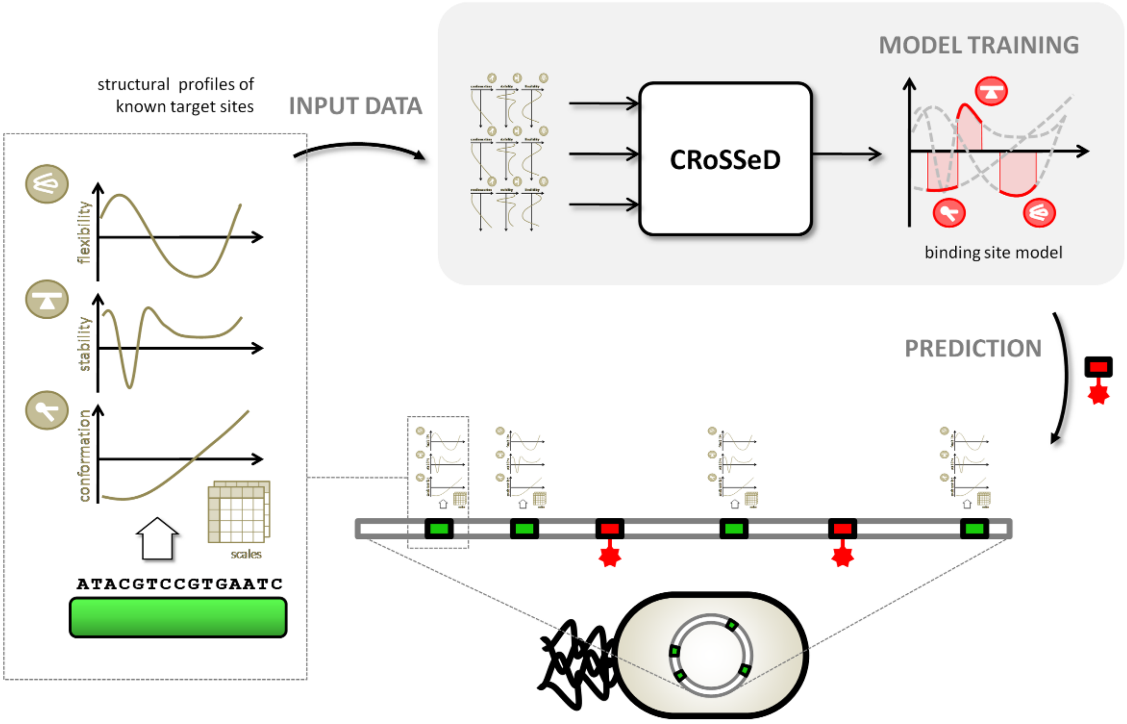

Figure 1: Overview of the CRoSSeD methodology. The sequence of known TF binding sites (green) are collected and used to create different structural profiles by applying structural scales. These structural scales are then used as input for the CRoSSeD model which will create a binding site model featuring strongly conserved structural profile characteristics in specific regions at the binding sites. These binding site models can then be used to predict other binding sites (red) for the given TF in the genome.

Cross-validation analysis

The classification performance of the CRoSSeD methodology was demonstrated and compared to several existing binding site models on a number of synthetic and biological data sets through cross-validation analysis. The synthetic data set was constructed so that the samples fit a predefined structural profile as shown in figure 2a. The ROC curve generated after the cross-validation on the synthetic data set is illustrated in figure 2b. The biological data sets consist of the known binding sites for 27 different *Escherichia coli* transcription factors. For the majority of the biological data sets the CRoSSeD models had a performance that was equal to or greater than that of all the other tested models. From the results of these cross-validations, we were able to summarize that the CRoSSeD methodology had an overall better predictive power than a standard PWM model and a previously proposed structural property model [4] for both the synthetic and the biological data sets. The improved classification performance of the CRoSSeD model could not be replicated used a higher-order sequence-only CRF method (named CRFseq), thus demonstrating that the structural profiles contain additional information about regulator binding which cannot be derived from any comparable sequence representation.

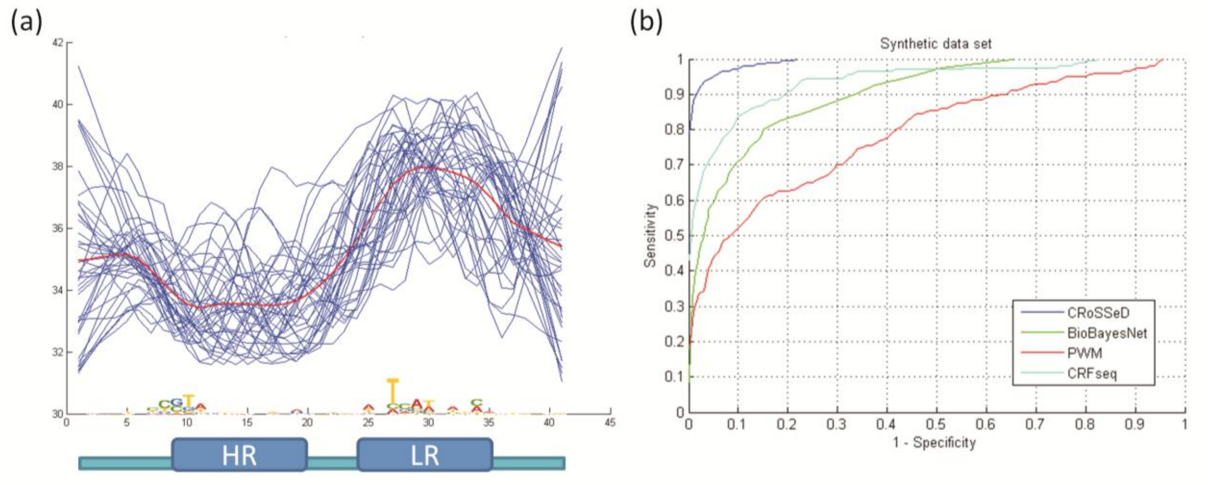

Figure 2: (a) Flexibility profiles of all 40 positive synthetic samples (blue lines) as measured by the B-DNA twist scale, (lower values correspond to more flexible regions). The red line is the average profile. For comparison, the sequence conservation logo is also given for each position. At the bottom of the figure is the structural characteristic that was simulated (HR: high rigidity, LR: low rigidity). (b) ROC curve displaying the average result of five ten-fold cross validations for the CRoSSeD (blue line), BioBayesNet (green line), PWM (red line) and CRFseq (cyan line) model when applied to the synthetic data set.

Novel binding site predictions

To demonstrate its potential as a tool for novel binding site discovery, the CRoSSeD models were used to screen the entire *E. coli* genome for binding sites of the 27 transcription factors. For a comparison the same procedure was followed using a traditional PWM model as binding site representation. Genes where the promoters contained predicted binding sites, were considered as novel targets of the modeled transcription factors and were further validated using gene expression data and an extensive literature and database survey. The predictions made by the CRoSSeD methodology for fourteen of the transcription factors showed significant overlap with the results from a gene expression analysis, while the PWM predictions only displayed overlap in nine cases. Except for a single case, all validated PWM predictions were also made by the CRoSSeD models. This indicates that most novel predictions made using a PWM based method can also be made with a structure-based model, while the reverse is not always true. To verify whether the targets predicted only by the structure-based method might indeed correspond to true targets, we carefully checked the literature for additional validation. Experimental evidence was found to back up several of the CRoSSeD predictions for a number of different transcription factors. For some validated binding sites the applied models indicated that the structural homology is much more pronounced than the sequence homology. This might suggest that poor sequence conservation in a binding site could be compensated by a strong structural profile.

Biological relevance

To further demonstrate that the CRoSSeD methodology was able to uncover specific structural properties that most likely play a biological role in the recognition of the binding sites by the regulator protein, the constructed models were compared with the current knowledge of protein-DNA interactions for two well studied TFs, namely CRP and PurR. The highest-weighted structural profile derived from the CRP model clearly shows the 'primary kink', a bend towards the major groove, and 'secondary kink', a bend towards the minor groove, that the CRP protein commonly induces in the DNA molecule [5]. This profile is given figure 3. The modeled structural characteristics for the PurR regulator confirm the necessity of a very stable binding site in close proximity to the leucine intercalation [6].

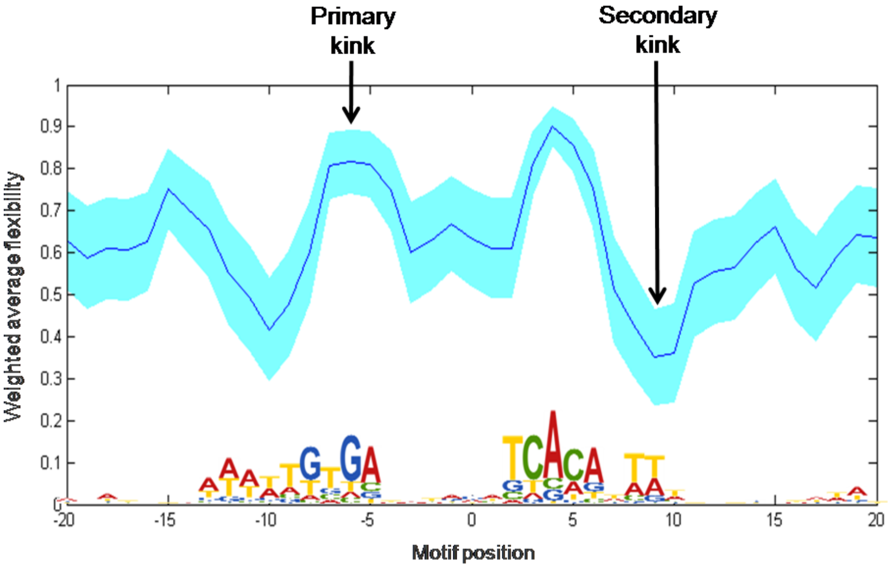

Figure 3: Structural profile corresponding to the DNase-I cutting frequency (flexibility toward major groove) values based on the weights assigned to the CRP model. Plotted in the dark blue line is the weighted average of the property at each position in the motif and surrounding it in the light blue area is the standard deviation on this average for each position. Locations of the primary and secondary kink in CRP-DNA complex are marked.

Conclusions

Our work has shown that the inclusion of DNA structural properties into a binding site modeling framework not only improves classification performance but also identifies additional biological relevant properties that are missed by sequence-only methodologies. While we demonstrated its performance for prokaryotic organisms, the method is generic and can also be applied on eukaryotic transcription factor binding sites or other functional genomic elements.

References

[1] Gromiha,M.M., Siebers,J.G., Selvaraj,S., Kono,H. and Sarai,A. (2005) Role of inter and intramolecular interactions in protein-DNA recognition. Gene, 364, 108-113.

[2] Lafferty,J., McCallum,A. and Pereira,F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc.18th International Conf.on Machine Learning , 282-289. 2001. San Francisco, CA.

[3] Meysman P, Dang TH, Laukens K, De Smet R, Wu Y, 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 (2).

[4] Nikolajewa S, Pudimat R, Hiller M, Platzer M & Backofen R. (2007) BioBayesNet: a web server for feature extraction and Bayesian network modeling of biological sequence data. Nucleic Acid Research 35: W688-W693.

[5] Lawson,C.L., Swigon,D., Murakami,K.S., Darst,S.A., Berman,H.M. and Ebright,R.H. (2004) Catabolite activator protein: DNA binding and transcription activation. Curr. Opin. Struct. Biol, 14, 10-20.

[6] Arvidson,D.N., Lu,F., Faber,C., Zalkin,H. and Brennan,R.G. (1998) The structure of PurR mutant L54M shows an alternative route to DNA kinking. *Nat. Struct. Biol*, 5, 436-441.