Statistical potential based method for modeling and testing of Nf-κB binding sites

Karolina Smolinska[1], **Marcin Pacholczyk**[1], **and Marek Kimmel**[1,2]
[1] Institute of Automatic Control, Silesian University of Technology, Gliwice, Poland
[2] Department of Statistics, Rice Univeristy, Houston, TX, USA

Introduction: Binding proteins is a huge group of proteins which are able to bind with specific, short DNA sequences. One representative of this class are transcription factors (TFs). They create the complexes with transcription factor binding site (TFBS). Alamanova et al. devised a computational approach for creating models of TFBS using statistical potential developed by Robertson and Varani to estimate TF-DNA free binding energy. These models take form of Position Weight Matrices (PWMs).

Method: We propose a modification of Alamanova et al. approach and use a volume-fraction corrected DFIRE-based energy function to calculate interaction energy between protein and DNA. The algorithm uses the crystal structure of TF-DNA complexes obtained from Protein Data Bank. Our data set contains dimers from NF-κB family: p50p50, p50p65 and p50RelB. We also present a method, based on the receiver-operator characteristic (ROC), of improving the matrices quality by manipulating values of parameters in statistical potential. For positive set we selected 41 human promoter sequences with experimentally verified NF-κB binding sites. We also prepared four negative sets. Sequences were scanned with *NuqleoSeq 2.0* algorithm, which detects TFBSs. We counted recognized experimentally confirmed binding sites in positive set (True positive rate) by iterating over all PWM scores (1-100). We use the area under the curve (AUC) as quality measure of PWM.

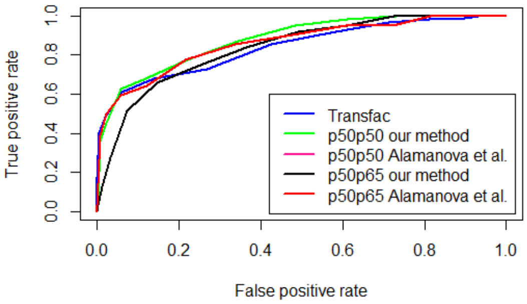

Figure 1: The PWM accuracy in distinguishing true positive from false positive TFBSs.

Results and discussion: We constructed PWMs based on the largest value of AUC for each factor. To test presented technique we compared matrices constructed for our method, Alamanova et al. approach and from TRANSFAC database. We also evaluated the quality of all matrices by calculating ROC and AUC values. The p50p50 computed in this work had the largest AUC of 0.8798. The AUC values of rest of PWMs performed slightly worse. We also used *NuqleoSeqc 2.0* to examine an ability to detect verified TFBSs by different PWMs for minimum PWM score of 80. Matrices detected similar number of experimentally confirmed binding sites.

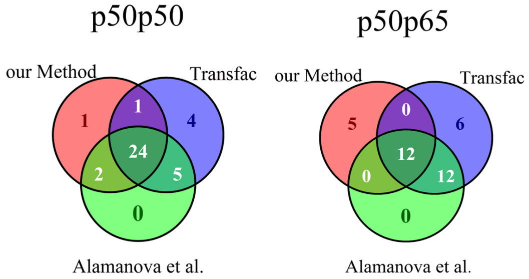

Figure 2: Results of the PWM scan of 58 experimentally confirmed NF-κB binding sites

Conclusion: The comparison shows significant similarity and comparable performance between calculated and experimental matrices. The proposed approach can be a promising alternative to experimental techniques of detecting TFBSs.

Keywords: NF-κB, TFBS, PWM

References:

Alamanova, D., P. Stegmaier, A. Kel (2010). Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies *Gene 11*, 225.

Jaksik, R., J. Rzeszowska-Wolny (2012). The distribution of GC nucleotides and regulatory sequence motifs in genes and their adjacent sequences *BMC Bioinformatics 492(2)*, 375–381.

Robertson, T. G. Varani (2007). An all-atom, distance-dependent scoring function for the prediction of protein-DNA interactions from structure *Proteins 66(2)*, 359–374.

Zhao, H., Y. Yang, Y. Zhou (2010). Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function *Bioinformatics 26(15)*, 1857–1863.