Computational approach for modeling and testing NF-κB binding sites

Marcin Pacholczyk[1], Karolina Smolińska[1], Marta Iwanaszko[1], and Marek Kimmel[1,2]

[1]Institute of Automatic Control, Silesian University of Technology, Gliwice, Poland
{marcin.pacholczyk, marta.iwanaszko}@polsl.pl
[2]Department of Statistics, Rice Univeristy, Houston, TX, USA
kimmel@rice.edu

Abstract. Motifs of Transcription Factor Binding Sites (TFBS) in DNA are commonly represented by the Position Weight Matrices (PWM). Recently Alamanova et al. devised a method for creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. We modify Alamanova et al. approach using volume-fraction corrected DFIRE-based energy function (DDNA3) as a model of protein-DNA interaction. The resulting new PWM matrices for NF-κB family show similarity to TRANSFAC matrices and comparable predictive capabilities. Presented approach is general and applicable to any TF for which crystallographic structure of its complex with DNA is available.

Keywords: TFBS, PWM, NF-κB

1 Introduction

NF-κB family is one of the most important Transcription Factor (TF) families in eukaryotic cells. It takes part in regulation of innate immunity, in carcinogenesis, and interacts with other important families such as p53 and HSF. Understanding of transcription regulation of NF-κB is important not only for biology but also for medicine. On the other hand, developing novel bioinformatics, physical modeling and evolutionary analysis tools and techniques applicable to NF-κB and its targets, will significantly aid research on other transcription factor families.

DNA-binding site models exist for about 500 vertebrate TFs and about 900 known Transcription Factor Binding Sites (TFBS) in human and 700 in mouse. Total number of binding sites in the multicellular genomes could be at least an order of magnitude higher than the number of coding genes [1]. Motifs in DNA are commonly represented by the Position Weight Matrices (PWM) and Phylogenetic Motif Models (PMM). PWM scanners score subsequences in the DNA data with respect to their similarity to the TFBS profile, as coded in the PWM. The simple scheme that is commonly used assumes an additive contribution from each position towards the score. Experimentally derived PWM models of TFBS profiles are usually deposited in Jaspar [2] of TRANSFAC [3] databases. Recently Alamanova et al. [4] devised a method for creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. The atomistic detail model of TF-DNA interaction would depend on the knowledge of relative spatial configuration of TF amino acids and DNA bases upon binding and a method to evaluate compatibility and strength of TF-DNA interaction. Increasing although still limited number of high quality crystallographic models of TF-DNA complexes deposited in the PDB database allow for detailed study of binding modes and details of contact interfaces. Some recent works report successful structure based predictions of TF binding sites in DNA. Molecular modeling methods require only the 3D structure of the TF-DNA complex. The binding specificity to given DNA motif can be predicted by molecular dynamics [5, 6], protein-DNA docking [7] or knowledge-based statistical potential [8]. Alamanova et al. successfully applied such methodology to create PWM matrices of NF-κB family namely p50p50, p50RelB and p50p65 dimers. Homology modeling can be used for TF-DNA complexes for which crystallographic data is not yet available. The original Alamanova et. al approach was implemented as 3DTF web-server available at http://cogangs.biobase.de/3dtf/ [4, 9]. Physical models are a natural intermediate stage between purely bioinformatics-based models and experimental techniques such as ChIP-Chip or ChIP-Seq.

We modified Almanova approach using volume-fraction corrected DFIRE-based energy function DDNA3 [8] as a model of protein-DNA interaction. The resulting new PWM matrices for NF-κB family show similarity to TRANSFAC matrices and comparable predictive capabilities.

2 Methods

2.1 Knowledge based protein-DNA potentials

In contrast with physics-based potentials we used before to describe protein-ligand interaction [10], the area of protein-DNA interaction analysis in dominated by knowledge-based or statistical potentials. The thermodynamic statistical potential mimics the free energy of binding taking into account the protein-DNA interface contact distances and the chemical atom types.

Statistical potential developed by Robertson and Varani [11] was previously used by Alamanova et al. to obtain PWM matrices for NF-κB family [4]. The probability of an interatomic contact is expressed in terms of the likelihood of observing a particular distance between a protein and a DNA atom in a native-like complex. The logarithm of this probability of correctness $P(C|D)$ of the interatomic distances describes the Gibbs free energy of the complex [11]:

$$G \approx -\ln P\left(C\middle|D\right) = -\sum_{i}^{N_P}\sum_{j}^{N_D}\ln P\left(C\middle|d_{ij},t_i,t_j\right) \quad (1)$$

where D is the set of atomic distances d_{ij} between the interface atoms, t_i and t_j correspond to the chemical types of the atoms, N_P and N_D represent the number of protein and DNA atoms in the complex. The probability of an individual atomic contact is modeled as the likelihood of observation of a separation d_{ij} between atoms t_i and t_j in a native-like protein-DNA complex:

$$P\left(C\middle|d_{ij},t_i,t_j\right)=P\left(C\right)\frac{P\left(d_{ij},t_i,t_j\middle|C\right)}{P\left(d_{ij},t_i,t_j\right)} \tag{2}$$

where $P(C|d_{ij}, t_i, t_j)$ is the likelihood function, $P(d_{ij}, t_i, t_j)$ the marginal probability, and $P(C)$ the Bayesian prior representing the probability of observing a native-like protein-DNA complex. Finally, the likelihood of observation a native-like interatomic distance d_{ij} can be expressed with the formula:

$$P\left(d_{ij},t_i,t_j\middle|C\right)\approx f\left(d_{ij},t_i,t_j\right)=\frac{N_{obs}\left(d_{ij},t_i,t_j\right)}{\sum_{d_{ij}}N_{obs}\left(d_{ij},t_i,t_j\right)} \tag{3}$$

where $N_{obs}(d_{ij}, t_i, t_j)$ is the number of contacts observed between two atoms of type t_i and t_j separated by distance d_{ij}.

The volume-fraction corrected DFIRE-based energy function DDNA3 used in our modification of original Alamanova et al. approach is based on different theoretical background and trained using different dataset. The first statistical energy function based on a distance-scaled, finite, ideal-gas reference (DFIRE) state was published in 2002 [12]:

$$\overline{u}_{i,j}^{DFIRE}\left(r\right)=-RT\ln\frac{N_{obs}\left(i,j,r\right)}{\left(\dfrac{r}{r_{cut}}\right)^{\alpha}\left(\dfrac{\Delta r}{\Delta r_{cut}}\right)N_{obs}\left(i,j,r_{cut}\right)} \tag{4}$$

where R is the gas constant, $T = 300$ K, $\alpha = 1.61$, $N_{obs}(i,j,r)$ is the number of ij atom pairs within the spherical shell at distance r observed in a given structure database, $\mathrm{r}_{\mathrm{cut}}$ is the cutoff distance, $\Delta\mathrm{r}_{\mathrm{cut}}$ is the bin width at $\mathrm{r}_{\mathrm{cut}}$. In our approach we use third generation of DFIRE-based energy function DDNA3 [8] in the form:

$$\overline{u}_{i,j}^{\mathrm{DDNA3}}\left(r\right)=-\eta\ln\frac{N_{\mathrm{obs}}\left(i,j,r\right)}{\left(\dfrac{f_{\mathrm{i}}^{\mathrm{v}}\left(r\right)f_{\mathrm{j}}^{\mathrm{v}}\left(r\right)}{f_{\mathrm{i}}^{\mathrm{v}}\left(r_{\mathrm{cut}}\right)f_{\mathrm{j}}^{\mathrm{v}}\left(r_{\mathrm{cut}}\right)}\right)^{\beta}\left(\dfrac{r}{r_{\mathrm{cut}}}\right)^{\alpha}\left(\dfrac{\Delta r}{\Delta r_{\mathrm{cut}}}\right)N_{\mathrm{obs}}^{\mathrm{lc}}\left(i,j,r_{\mathrm{cut}}\right)} \tag{5}$$

which includes atom type dependent volume-fraction correction:

$$f_i^v(r) = \frac{\sum_j N_{ij}^{\text{Protein-DNA}}(r)}{\sum_j N_{ij}^{\text{All}}(r)} \tag{6}$$

and low count correction made to $N_{\text{obs}}(i,j,r)$:

$$N_{\text{obs}}^{\text{lc}}(i,j,r) = N_{\text{obs}}(i,j,r) + \frac{75\sum_{i,j} N_{ij}^{\text{Protein-DNA}}(r)}{\sum_{i,j,r} N_{ij}^{\text{Protein-DNA}}(r)} \tag{7}$$

The parameters in (5) are set to $\beta = 0.5$, $\eta = 0.01$, $\Delta r = 5$ Å and $r_{\text{cut}} = 15$ Å. Please refer to [8] for further details.

2.2 Computational approach to PWM matrices

Starting form crystallographic data available for NF-κB family: the p50 homodimer (PDB entry 1NFK), p50RelB (2V2T), and the p50p65 heterodimer (1VKX) we follow the workflow of structure-based PWM calculation described in [4] modified using volume-fraction corrected DFIRE-based energy function DDNA3 [8] as a model of protein-DNA interaction. A 3D structure of a transcription factor bound to its target DNA sequence is retrieved from the PDB databank. For each DNA sequence of length *N* as found in the corresponding crystal structure we generate $4N + R$ random sequence fragments of the same length (*R* should be limited for computational efficiency to e.g. 10^2). The crystal structures of the DNA chains taken from the corresponding TF-DNA complexes were mutated using the MMTSB (Multiscale Modeling Tools for Structural Biology) [13] script mutateNA.pl by fixing the chain backbone and substituting one base pair at each step. Sterical inaccuracies should be avoided as the script uses a library of torsion parameters for the correct residue rotations. For each of the $4N+R$ random sequences the free energy of binding to the TF was computed using the *DDNA31* software [8]. All weights $w(i, u)$ in the PWM (*i*-position in the sequence ranging from 1 to *N*, *u* - nucleotides A, C, G, T), such that the binding energy predicted by the PWM would maximally correlate with the energy computed with the statistical potential can then be predicted by solving the linear equation:

$$\mathbf{Ax} = \mathbf{b} \tag{8}$$

where **x** is a vector of 4*N* dimensions of the estimated weights and **A** is a binary matrix of dimensions ($4N$, $4N + R$), which contains information on all random DNA sequences whose free binding energy was computed. The free binding energy vector **b** consists of $4N + R$ values obtained with the protein-DNA scoring procedure described above. Linear equation can be conveniently solved e.g. by least squares optimization in the *Matlab* package.

The statistical potential used for scoring TF-DNA interaction has been calibrated for native-like interatomic coordinates. In case of significantly deformed protein and/or DNA structure in the complexes a reasonable protein-DNA configuration should be obtained with the help of docking allowing for chain flexibility as implemented in HADDOCK [14]. Alternatively, a short initial optimization (100-300 molecular mechanics steps) of the protein-DNA complex could be performed using a software like AmberTools [15].

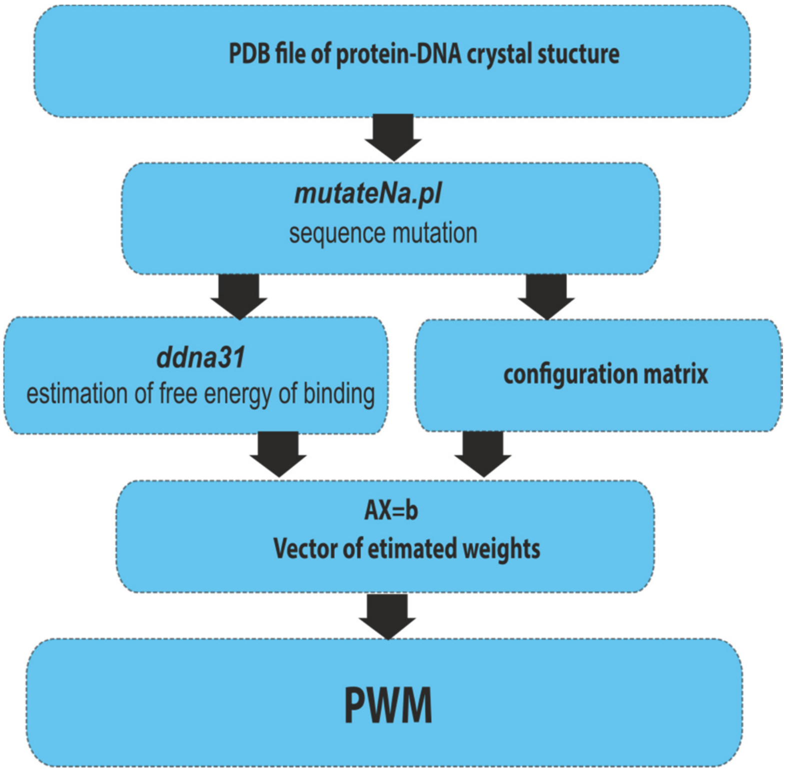
PDB file of protein-DNA crystal stucture
mutateNa.pl
sequence mutation
ddna31
estimation of free energy of binding
configuration matrix
AX=b
Vector of etimated weights
PWM

Fig. 1. Workflow of structure-based PWM computation

3 Results and Discussion

We applied the described approach to data available for NF-κB family: the p50 homodimer (PDB entry 1NFK), p50RelB (2V2T), and the p50p65 heterodimer (1VKX). See Fig. 2 - 4 for the corresponding sequence logos calculated for each PWM (created with *enoLOGOS* [16]).

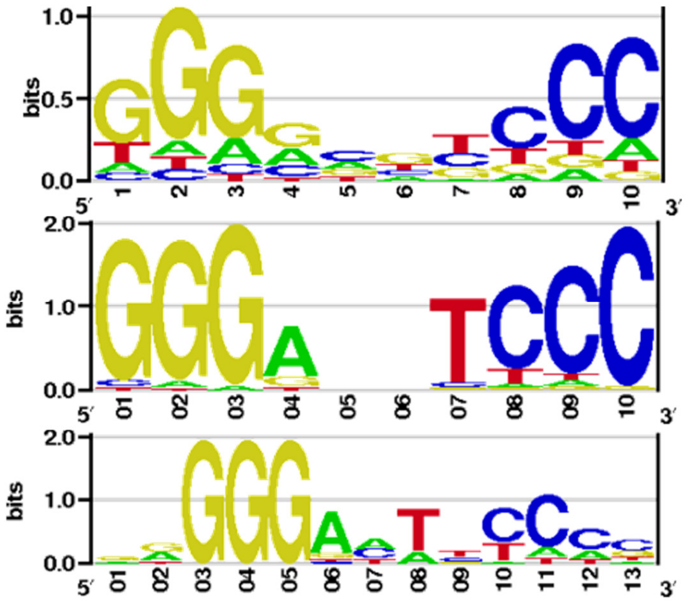

Fig. 2. The p50p50 logos for our method (top), original Alamanova et al. (middle), TRANSFAC (bottom)

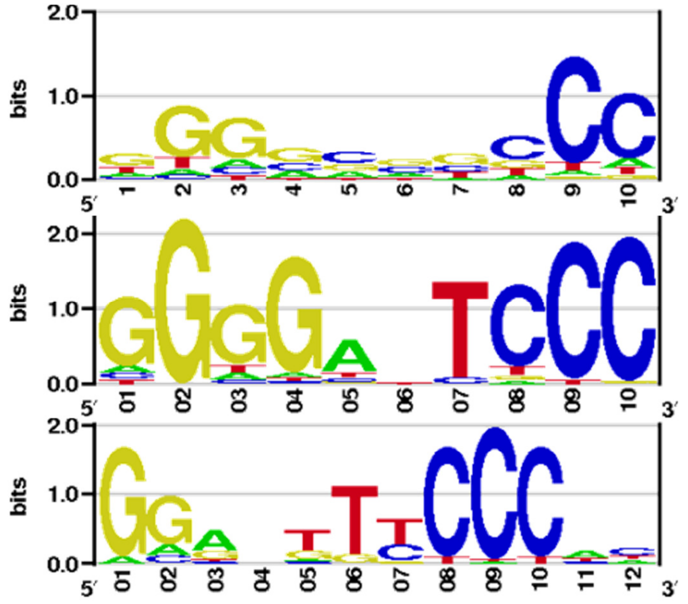

Fig. 3. The p50p65 logos for our method (top), original Alamanova et al. (middle), TRANSFAC (bottom)

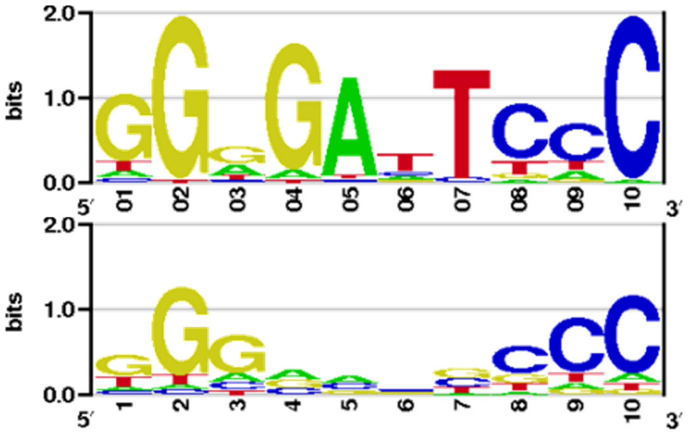

Fig. 4. The p50RelB logos for our method (top), original Alamanova et al. (bottom)

To test the proposed approach we used PWM based TF binding site searching algorithm implemented in *NucleoSeq* 2.0 software [17] to scan human promoters containing known, experimentally verified NF-κB binding sites. The only adjustable parameter in *NucleoSeq* - minimum PWM score was set to 80. We used dataset complied by Alamanova et al. [4]. The dataset consists of 124 human promoter sequences belonging to 69 genes known to be regulated by NF-κB. Experimentally confirmed 58 binding sites can be found only in 31 out of 124 promoter sequences belonging to 25 genes. We compared scan results of p50p50 and p50p65 matrices calculated using our method, original Alamanova et al. implementation and V$P50P50_Q3 and V$P50RELAP65_Q5_01 matrices downloaded from TRANSFAC database. Our p50p50 and p50p65 recovered 30 and 22 sites respectively, Alamanova et. al p50p50 matrix found 31 and corresponding TRANSFAC PWM reported 32 sites. For p50p65 – Alamanova et al. matrix recovered 24 sites while TRANSFAC found 29. Our p50p50 matrix recovered additional one and p50p65 five sites not reported by any other PWM. Results are shown in Fig. 5. We found that while the correlation between protein-DNA scores calculated with Robertson and Varani (used in original Alamanova et al. approach) and DDNA3 potentials is strong (Pearson's r = 0.96) surprisingly sensitivity (understood as change in score magnitude) to DNA mutations is much lower for DDNA3 (data not shown). This may result in less specific PWMs and explain the fact our matrices detect additional sites missed by Alamanova et al. and TRANSFAC PWMs.

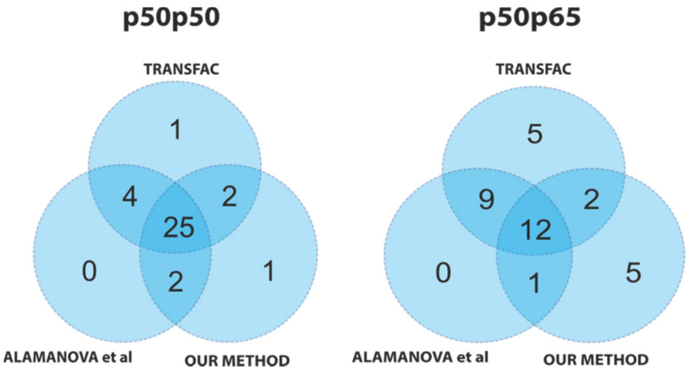

Fig. 5. Results of the PWM scan of 58 experimentally confirmed NF-κB binding sites

4 Conclusions

The resulting new PWM matrices for NF-κB family show similarity to TRANSFAC matrices and comparable predictive capabilities. Presented approach is general and applicable to any TF for which crystallographic structure of its complex with DNA is available. Our results concerning model of HSF1 transcription factor binding motif are detailed in [18].

Acknowledgements. This work has been partially supported by Polish National Science Centre funds under grants Opus DEC-2012/05/B/NZ2/01618 and N-N519-647840 based at the Institute of Automatic Control, Silesian University of Technology, Gliwice, Poland.

References

1. GuhaThakurta D. Computational identification of transcriptional regulatory elements in DNA sequence Nucl Acids Res (2006) 34 (12): 3585-3598
2. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucl Acids Res (2004), 32 Database:D91-94
3. Knüppel R, Dietze P, Lehnberg W, Frech K, Wingender E: TRANSFAC® retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. J Comput Biol (1994), 1:191-198
4. Alamanova D., Stegmaier P., Kel A. Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. BMC Bioinformatics, (2010) May 3;11:225
5. Bauer A. L., Hlavacek W. S., Unkefer P. J., Mu F. Using sequence-specific chemical and structural properties of DNA to predict transcription factor binding sites. PLoS Comput Biol, (2010) Nov 18;6(11)
6. Liu L. A., Bader J. S., Structure-based ab initio prediction of transcription factor-binding sites. Methods Mol Biol, (2009), 541, 23-41.

7. Liu Z., Guo J. T., Li T., Xu Y. Structure-based prediction of transcription factor binding sites using a protein-DNA docking approach. Proteins, (2008) , 72(4):1114-24.
8. Zhao H., Yang Y., Zhou Y. Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function, Bioinformatics, 2010, 26(15), 1857-1863.
9. Gabdoulline R., Eckweiler D., Kel A., and Stegmaier P.: 3DTF: a web server for predicting transcription factor PWMs using 3D structure-based energy calculations. Nucl. Acids Res. (2012) 40 (W1): W180-W185
10. Pacholczyk M, Kimmel M: Exploring the landscape of protein-ligand interaction energy using probabilistic approach. J Comp Biol 2011, 18(6):843-850
11. Robertson, T. and Varani, G. An all-atom, distance-dependent scoring function for the prediction of protein-DNA interactions from structure. Proteins (2007), 66(2) 359-374
12. Zhou H., Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci (2002), 11:2714–2726
13. Feig M, Karanicolas J, Brooks CL. MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. J Mol Graph Model. 2004 22:337-395
14. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: a protein-protein docking approach based on biochemical and/or biophysical information. J Am Chem Soc. 2003 125, 1731-1737
15. Case D.A., Cheatham T.E., Darden T., Gohlke H., Luo R., Merz K.M., Onufriev A., Simmerling C.,Wang B. and Woods R.. The Amber biomolecular simulation programs. J. Computat. Chem. 26, 1668-1688 (2005)
16. Workman CT, Yin Y, Corcoran DL, Ideker T, Stormo GD, Benos PV: enoLOGOS: a versatile web tool for energy normalized sequence logos. Nucleic Acids Res 2005, 33:W389-W392.
17. Jaksik, R., Rzeszowska-Wolny, J. The distribution of GC nucleotides and regulatory sequence motifs in genes and their adjacent sequences (2012) Gene, 492 (2), 375-381
18. M. Iwanaszko, P. Janus, T. Stokowy, M. Kimmel: Changes in heat shock duration influence regulatory schemes of HSF1 activity. Proceedings IWBBIO 2014: International Work-Conference On Bioinformatics And Biomedical Engineering.