W180–W185 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 11 June 2012*
doi:10.1093/nar/gks551

3DTF: a web server for predicting transcription factor PWMs using 3D structure-based energy calculations

R. Gabdoulline[1], D. Eckweiler[2], A. Kel[3,4,5] and P. Stegmaier[4,*]

[1]Heinrich-Heine University of Duesseldorf, Universitaetstr. 1, 40225 Duesseldorf, [2]Helmholtz Center for Infection Research, Inhoffenstrasse 7, 38234 Braunschweig, [3]GeneXplain GmbH, Am Exer 10 b, 38302 Wolfenbüttel, [4]BIOBASE GmbH, Halchtersche Str. 33, 38304 Wolfenbüttel, Germany and [5]Institute of Chemical Biology and Fundamental Medicine, Russian Academy of Science, 10 Lavrentyev Ave, 630090 Novosibirsk, Russia

Received March 8, 2012; Revised May 14, 2012; Accepted May 16, 2012

ABSTRACT

We present the webserver 3D transcription factor (3DTF) to compute position-specific weight matrices (PWMs) of transcription factors using a knowledge-based statistical potential derived from crystallographic data on protein–DNA complexes. Analysis of available structures that can be used to construct PWMs shows that there are hundreds of 3D structures from which PWMs could be derived, as well as thousands of proteins homologous to these. Therefore, we created 3DTF, which delivers binding matrices given the experimental or modeled protein–DNA complex. The webserver can be used by biologists to derive novel PWMs for transcription factors lacking known binding sites and is freely accessible at http://www.gene-regulation.com/pub/programs/3dtf/.

INTRODUCTION

Position-specific weight matrices are an important tool to analyze regulatory DNA sequences with regard to interacting transcription factors. Often, position-specific weight matrices (PWMs) are derived from alignments of known binding sites for certain transcription factors (TFs). Such binding sites may have been determined individually, e.g. for a particular target gene of interest, or on a genome-wide scale using a chromatin-immunoprecipitation (ChIP) assays. Building a PWM from known binding site sequences faces difficulties when only few or no binding sites have been described or when ChIP data are not available. An alternative experimental method, the Protein Binding Microarray, has been developed by Bulyk *et al*. (1) and uses microarray technology followed by statistical analysis to define the binding pattern for a TF of interest.

Previously, we developed a method to compute a PWM using solely information about the structural protein–DNA complex. This further extends the scope of TFs for which a PWM can be obtained, possibly even without requiring any additional experimental efforts, e.g. when the structural complex of interest is already available or by estimating a structure by homology modeling. The method has been described in (2) and the force field for protein–DNA binding energy calculation in (3). The protein–DNA binding affinity is calculated using a statistical potential (3) calibrated on known protein–DNA bound complexes available from the PDB database. Briefly, we generate all possible single point substitutions of the DNA chain in the protein–DNA crystallographic complex using a rigid rotamer library that aims to minimize possible steric conflicts upon nucleotide substitutions. As a result, the protein–DNA binding affinity is computed for each complex of the protein with the altered DNA. Following this a DNA mutation matrix is constructed where all DNA residues that are not mutated are marked with zeros and the mutated residues are marked with one. The matrix multiplied with a vector of the still unknown weights of the DNA residues (this is the PWM vector) gives the binding affinity vector calculated in the previous step. By solving this simple linear equation system, we obtain the PWM of the particular transcription factor. For details please refer to the original publication (2). Here, we report on a webserver, 3D Transcription Factor (3DTF), which enables researchers to easily carry out the necessary computational analysis to build custom 3D structure-based PWMs.

There are several studies and web servers addressing similar issues and complementing our approach. Most

*To whom correspondence should be addressed. Tel: +49 53 3185 8443; Fax: +49 53 3185 8470; Email: philip.stegmaier@biobase-international.com

closely related is the 3dfootprint database (4) which derives the sequence specificity of DNA-binding proteins on the basis of counting molecular contacts contributing to recognition. Here, both the method of estimating binding strength as well as calculating PWM are different from our approach and thus presents an alternative and complementary 3D-based methodology. The same group developed a server to model protein–DNA complexes (5). There are other ways to infer protein–DNA interactions, for example, direct and indirect readout energy (6), Rosetta forcefield (7) and using compressed sensing methods (8). These methodologies can be directly incorporated into the 3DTF server, provided that the energy calculations are sufficiently fast.

THE 3DTF WEBSERVER

The method was implemented as a standalone tool performing automated processing of a protein–DNA complex, invoking third party programs and returning the computed PWM. Before calculating the energies, 3DTF analyzes the protein–DNA interface and can automatically identify the binding site on the DNA in the complex. Automatic detection works when the binding interface can be determined unambiguously and appears to have less than 30 nt positions. The scripts for the server are heavily optimized, allowing calculation of a typical matrix of 10–12 positions in less than a minute. Convergence of the full model (which includes an additional weight for each position) requires longer calculations, which can be invoked by a special option in the submission page.

3DTF has an easy to use interface that enables the user to perform three major task modes related to PWM calculation. For all three task modes, the required input is a structure model in PDB format, containing one protein chain and two complementary DNA chains. Additional parameters can be set in the Task mode 3, described further below. An example file with the proper format is available on the website.

In Task mode 1, the user can check whether the provided structure model is suitable for PWM calculation. Here, 3DTF parses the necessary information, such as chains and types of chains, from the PDB file. It also automatically determines the segment of DNA in close contact with the protein component. The Task mode 1 output is a plain text page that shows the detected protein and DNA chains as well as the parsed-out sequence of the binding site. This output part is followed by a detailed description of the bases contacted by protein with reference to corresponding chains, chain IDs and individual base identifiers. Failing conditions in Task mode 1 include absence or corruption of chains, e.g. if the file does not contain a DNA chain, unpaired bases in the binding site or unconventional base numbering in either strand.

Aside from testing of the PDB file, Task mode 1 is also of interest for users who wish to anticipate how the provided information is going to be processed. It should be noted that the information compiled in the Task mode 1 output is also incorporated in the output of Task mode 2, but the latter requires more complex calculations.

Task mode 2 as well as Task mode 3 calculate a PWM for a given protein–DNA complex. In Task mode 2, 3DTF computes the binding profile on the basis of the automatically defined DNA segment (see Task mode 1). For Task mode 3, the user can specify chains, start base numbers and desired length of the binding site in order to enforce a particular binding site to be modeled. This provides greater flexibility to obtain custom PWMs based on prior knowledge.

Task mode 3 is important when the binding interface cannot be defined unambiguously from the structure (for example, when there is more than one binding site), thus disqualifying the structure from being used via Task 2. Another possible application of this user-defined mode is to calculate matrices for long binding sites. A long site can be divided into shorter segments to be handled by 3DTF. PWMs from each segment can then be concatenated into a larger model. This yields the same result as calculating the whole matrix, since within the applied model of protein/DNA interactions energy contributions of positions are independent from one another. An example output of 3DTF is shown in Figure 1.

ANALYSIS OF 3D STRUCTURE-BASED PWMS

A few changes to the published algorithm were implemented in 3DTF. To ensure better convergence of results, generation of random sequences was modified to guarantee that at each position any of four bases appears at least once. With this modification, the calculated matrices converge already with $4N$ (N is the number of positions) sequences, as the applied model is additive with respect to positions (9). To avoid possible problems with perturbations in modeled DNA structures, we set the required number of sequences to $4N+5$.

The method is able to derive more information than only positional nucleotide preferences, because the calculated energies reveal the relative importance of each position of the matrix compared to other positions (10). Notably, specificity of a motif position for certain nucleotides does not necessarily imply that the position contributes more to the binding energy than less specific ones. Calculations carried out by 3DTF yield information beyond the classical PWM. The average energy contribution of each position is estimated by the logarithm of the sum of Boltzmann factors of each base (LSBF) as

$$\mathrm{LSBF} = \log\left(\sum_{b=A,G,T,C} \exp(-E_b/beta)\right) \quad (1)$$

where E_b is the energy contribution of base b in respective position and $beta = k_B T$. This quantity shows the expected correlation with the information content of a motif position (11,12), although the contributions from positions of maximal specificity (1 0 0 0, for example) may be very different (Figure 2A). In all our examples, the more informative positions located at the core of the binding site.

The LSBF is included in the output matrix as recommended weighting for positions, where it can be found in

Protein-DNA interface analysis program output:

```
number of chains= 3, chain IDs:  ABC
 number of bases/residues= 63 11 11
          number of atoms= 513 220 225

chain ID, number of bases, ifdna(0/1): A       0     0
KHYRGVRQRPWGKFAAEIRDPAKNGARVWLGTFETAEDAALAYDRAAFRMRGSRALLNFPLRV
chain ID, number of bases, ifdna(0/1): B      11     1
TAGCCGCCAGC
chain ID, number of bases, ifdna(0/1): C      11     1
GCTGGCGGCTA

chain, chain ID and number of paired bases:  1  A     0
chain, chain ID and number of paired bases:  2  B    11
chain, chain ID and number of paired bases:  3  C    11


chain, chain ID and number of bases contacting protein  2  B     7
B0003 B0004 B0005 B0006 B0007 B0008 B0009
C0024 C0024 C0022 C0021 C0020 C0019 C0018

chain, chain ID and number of bases contacting protein  3  C     8
C0015 C0016 C0017 C0018 C0019 C0020 C0021 C0022
B0012 B0011 B0010 B0009 B0008 B0007 B0006 B0005

 Detected binding site
 chain, range of basepairs contacting protein, length
 2:   3 ->   12  length: 10
 B0003 C0024 ->  B0012 C0015
```

Binding sequence:

```
TAGCCGCCAG
```

Binding sequence:

```
TAGCCGCCAG
```

Sequence vs. energy

Full energy model

Calculated PWM (download in TransFac format)

```
AC  M3D000
P0      A      C      G      T         195.3
01     33    960      1      5      C     71
02      0    556     79    363      Y     70
03      0      0    999      0      G    105
04      0    999      0      0      C    144
05      0   1000      0      0      C    207
06      0      0    999      0      G    106
07    854    126     14      4      A     67
08      0    787      0    212      C     82
09      1    990      8      0      C     78
10    459    113    411     15      R     65
XX
```

Sequence logo:

Figure 1. Outputs of Task modes 2 and 3 are plain text pages that feature the sequences, for which binding energies have been computed, a tabular description of the derived PWM as well as a PWM logo. The PWM can further be downloaded in the TRANSFAC-like format. In 3DTF, the consensus column is complemented with calculated binding energy contributions of each position (see below). The Task mode 2 output encompasses in addition the results of evaluating the input PDB file as described for the Task mode 1.

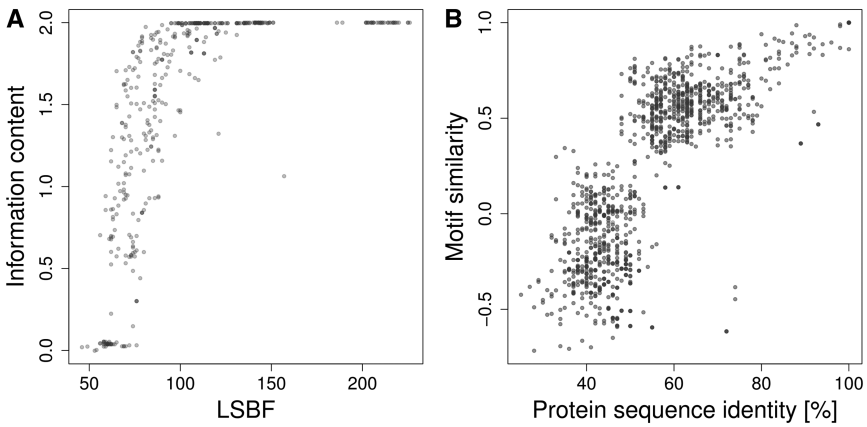

Figure 2. (**A**) LSBF versus information content of the position (from data in Application example 1). (**B**) PFM pairwise similarity of matrices derived from models versus pairwise sequence identity of modeled proteins.

the rightmost column appended to the consensus. These weights of the positions can be used in the PWM-based site search procedures, similar to the information vector used in the MATCH program (12). An example with LSBF values is shown in Figure 1 for the matrix of *Arabidopsis thaliana* ERF2 modeled on 1GCC, ATERF1. All frequencies and weights are normalized to 1000. The last value of the first matrix row (195.3) is the sum of unnormalized LSBFs.

AVAILABLE TEMPLATES

The PDB (13) harbors a large number of high-quality structures that can be used to derive PWMs. At the time of writing this manuscript, we found 375 structure files (PDB files) with a transcription-related protein bound to DNA, 190 of them for TFs from human, mouse, rat. Data in PDB can be highly redundant—many entries in fact are related to similar transcription factors. Therefore, to understand, which part of all transcription factors can be described with our approach, we estimated the possibility of finding a homologous 3D structure for transcription factors belonging to distinct classes. The TF classification was taken from (14), whereas the assignment of homology was obtained from the protein model portal (PMP) (15) using UniProt IDs of TFs. Results of our analysis are compiled in Supplementary Table S1. Altogether, 47% of TF class representatives can be modeled and assigned a PWM based on homologs in PDB having 50% sequence identity. This fraction rises to 70% if the sequence identity threshold is set to 30%.

One can attempt to model the TF–DNA complex starting from an unbound TF structure (16). This would increase the fraction of TF representatives suitable for modeling to 67 and 86%, respectively, at 50 and 30% sequence identity thresholds. At least one representative of 21 out of 33 classes can be processed (50% sequence identity). Hence, the already existing data have great potential to increase the coverage of TFs with known binding profile.

APPLICATION EXAMPLE 1

We used 3DTF to derive PWMs for plant transcription factors. Using the template 1GCC, we modeled 48 *A. thaliana* proteins that have sequence identity to ATERF-1 ranging from 30% to 100%. The list of proteins is given in Supplementary Table S2 and Blast sequence alignment to 1GCC sequence in Supplementary Table S3. Visualization of homology models (Supplementary Figure S1) shows that as a rule the amino acid substitutions appear on the loop regions, i.e. one can expect that the fold and the binding mode are conserved while the binding specificity is altered.

Figure 2B shows that binding matrices produced with 3DTF highly depend on the sequences of the proteins. The similarity between PWMs calculated for ATERF-1 homologs and the PWM of the template protein increases with the protein sequence identity. The motif similarity index was calculated as in (17), without any shifts along positions, since the matrices in this example have the same length and equivalent positions in 3D. This example was designed to show the ability of the approach to deliver divergent PWMs using the same protein–DNA complex as a modeling template.

APPLICATION EXAMPLE 2

We performed calculation of PWMs for all transcription factors in TRANSFAC that have a link to a PDB entry with bound DNA. There were 18 factors with both the PDB link and documented binding sites as well as with defined TRANSFAC PWM. We performed the 3DTF calculations for these factor–DNA complexes and compared energies calculated for random sequences with the energies for the known binding sequences collected in TRANSFAC. As summarized in Table 1, calculated energies for reported binding sequences tend to be on the lower end of the spectrum of binding energies, validating the appropriateness of the forcefield used in our method. The average ranks of energies for known sites show that the applied calculations assign to them low energy values as compared to random sequences. An exceptional case is 2EZD, where the structure contained only a truncated form of HMGIY consisting of the second and third DNA binding domains. A more detailed example is shown in the supplement (Supplementary Figure S2) with energies calculated for binding sequences of C/EBPbeta using PDB entry 1GTW.

Furthermore, we compared 3DTF PWMs obtained for PDB entries listed in Table 1 to PWMs defined for corresponding TRANSFAC TFs. Similarities between two PWMs were expressed as the Pearson correlation coefficient (PCC) of their binding site scores similar to the approach described in (18). The two PWMs were arranged with an overlap of at least five positions. Scores were calculated for a random sequence over the entire segment covered by both PWMs. A sample of 2000 random score pairs was drawn for every possible overlapping arrangement and PWM orientation. The random score pairs were used to calculate the PCC by the following formula, where X and Y are the vectors of $n = 2000$ random scores for each motif and $\bar{x}$ and $\bar{y}$are respective sample mean scores.

$$\mathrm{PCC(X,Y)} = \frac{\sum_{i=1,n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1,n} (x_i - \bar{x})^2 \sum_{i=1,n} (y_i - \bar{y})^2}$$

The highest PCC out of all possible configurations was reported. We calculated similarities between (i) the 3DTF PWM and the TRANSFAC matrices homologous to the PDB protein, (ii) the 3DTF PWM and TRANSFAC matrices linked to other transcription factors listed in Table 1 and (iii) the TRANSFAC matrices associated with the same TF protein. The correlation values show that 3DTF PWMs are in most cases more similar to motifs of the homologous TF than to other motifs (Table 1, columns A and B). In the case of 1IF1 and 2DGC, 3DTF was not able to derive a PWM that would be similar to the PWMs reported in TRANSFAC for the corresponding TF. Furthermore, in several cases, we observe that PWMs produced by 3DTF achieve a level of similarity that is comparable to the similarity among multiple PWMs of the same TF (Table 1, columns A and C), e.g. for 1IGN, 1HCP, 1UBD or 1GTW.

For comparison, we carried out the same calculations for matrices computed by the 3DPWM method (4). The results were similar to those obtained with 3DTF. The similarity of the 3DPWM motif to matrices of the same TF was in most cases higher than to other motifs but often lower than the similarity between TRANSFAC motifs of that TF. This suggests that such results are connected to the structure-based methods. Further, 3DPWM matrices were often more similar to homologous TRANSFAC matrices than the corresponding 3DTF matrix. However, the PCC values in comparisons to TRANSFAC matrices were also higher. To highlight this, we divided PCCs of 3DPWM or 3DTF matrices and homologous TRANSFAC matrices by the PCCs calculated with other TRANSFAC matrices. The results are reported in Table 2 (A/B columns). Here, 3DTF more often achieved a higher ratio than 3DPWM, which suggests that 3DTF motifs tend to be more specific for the particular TF. The reason could be a general bias in either approach to infer more or less informative matrix positions. In the 3DTF approach, this can be controlled, e.g. by adjusting the temperature in the Boltzmann equation.

We also performed calculations with experimental data available from Uniprobe database (19) holding PWMs derived from protein binding microarrays (1). We checked all 418 Uniprobe database entries (April 2012)

Table 1. Ranking of binding sequence energies for a set of TFs with assigned PDB entry and correlation between 3DTF and TRANSFAC PWMs

PDB ID	Transcription factor	Average rank of site energies versus random (%)[a]	A. 3DTF PWM versus homolog TRANSFAC PWMs[b]	B. 3DTF PWM versus other TRANSFAC PWMs	C. Homolog TRANSFAC PWMs
1XBR	Brachyury	0	**0.61**[c]	0.38 ± 0.09	
1CF7	E2F4	3.5	0.44	0.30 ± 0.10	
1SRS	SRF	1.5	0.58 ± 0.03	0.29 ± 0.11	0.80 ± 0.06
1IF1	IRF-1	9.5	0.30 ± 0.04	0.31 ± 0.07	0.76 ± 0.08
1IGN	RAP1p	4.3	**0.67 ± 0.06**	0.25 ± 0.07	0.69 ± 0.10
1HDD	En	3.6	0.61	0.32 ± 0.12	
1HCP	ER	0.1	0.66 ± 0.08	0.35 ± 0.11	0.70 ± 0.08
2DGC	GCN4	8.4	0.36 ± 0.10	0.32 ± 0.12	0.73 ± 0.07
1MDY	E2A	1.3	0.53 ± 0.06	0.28 ± 0.08	0.82 ± 0.05
1FOS	AP-1	4.3	0.54 ± 0.07	0.36 ± 0.13	0.89 ± 0.06
1TUP	P53	0	0.49 ± 0.05	0.37 ± 0.10	0.69 ± 0.12
2EZD	HMGIY	22.6	0.46 ± 0.06	0.24 ± 0.16	0.45 ± 0.12
1UBD	YY1	1.6	0.67 ± 0.12	0.34 ± 0.10	0.74 ± 0.09
1YTB	TBP	14.0	0.55 ± 0.07	0.28 ± 0.11	0.70
1APL	MATalpha2	9.7	0.58 ± 0.16	0.35 ± 0.08	0.43
2BOP	E2	1.8	0.57 ± 0.04	0.23 ± 0.08	0.91 ± 0.04
1BY4	RXRalpha	2.9	0.47 ± 0.10	0.37 ± 0.13	0.48 ± 0.06
1GTW	C/EBPbeta	1.3	**0.70 ± 0.04**	0.30 ± 0.10	0.81 ± 0.06

In column B, higher PCCs than those achieved by 3DPWM (Table 2) are highlighted bold.
[a]Average rank of known sites is calculated from the ranks of energies to known sites in the list of ordered binding energies to 1000 random DNA sequences.
[b]Correlation coefficients were calculated as described in the main text.
[c]Correlation values and/or standard errors may be missing due to lack of data.

Table 2. Correlation between 3DPWM and TRANSFAC PWMs for a set of TFs with assigned PDB entry

PDB ID	A. 3DPWM versus homolog TRANSFAC PWMs	B. 3DPWM versus other TRANSFAC PWMs	C. Homolog TRANSFAC PWMs	A/B 3DPWM	A/B 3DTF
1XBR	0.55	0.42 ± 0.11		1.31	1.61
1CF7	**0.73**	0.21 ± 0.08		3.52	1.47
1SRS	**0.65 ± 0.04**	0.28 ± 0.08	0.80 ± 0.06	2.31	2.00
1IF1	**0.58 ± 0.06**	0.41 ± 0.12	0.76 ± 0.07	1.40	0.97
1IGN	0.57 ± 0.04	0.29 ± 0.09	0.68 ± 0.10	1.94	2.68
1HDD	**0.69**	0.38 ± 0.15		1.82	1.91
1HCP	0.70 ± 0.10	0.40 ± 0.12	0.71 ± 0.06	1.76	1.89
2DGC	0.35 ± 0.09	0.30 ± 0.10	0.73 ± 0.06	1.16	1.13
1MDY	**0.69 ± 0.06**	0.29 ± 0.08	0.82 ± 0.05	2.33	1.89
1FOS	0.50 ± 0.02	0.35 ± 0.10	0.89 ± 0.06	1.40	1.50
1TUP	0.47 ± 0.07	0.33 ± 0.12	0.69 ± 0.11	1.43	1.32
2EZD	**0.59 ± 0.06**	0.32 ± 0.11	0.44 ± 0.11	1.83	1.92
1UBD	**0.74 ± 0.09**	0.41 ± 0.10	0.74 ± 0.09	1.82	1.97
1YTB	**0.65 ± 0.02**	0.31 ± 0.12	0.69	2.11	1.96
1APL	0.62 ± 0.04	0.38 ± 0.09	0.43	1.66	1.66
2BOP	**0.63 ± 0.02**	0.29 ± 0.10	0.90 ± 0.03	2.19	2.48
1BY4	**0.52 ± 0.08**	0.42 ± 0.15	0.48 ± 0.05	1.22	1.27
1GTW	0.57 ± 0.03	0.32 ± 0.09	0.81 ± 0.06	1.78	2.33

Columns A–C correspond to the same columns in Table 1 with PCC values for 3DPWM. The two rightmost columns contain the PCC of column A divided the PCC of column B of respective table row (Table 1 for 3DTF values). In column A, higher PCCs than those achieved by 3DTF (Table 1) are highlighted bold.

for links to relevant PDB entries. There were 35 that have associated PDB file for reported TF bound to DNA. Five among these had repeated experiments for the same transcription factor, so we can assess the reproducibility of experiment-derived PWMs. This last comparison gave high consistency between experiment-derived PWMs, with similarities, calculated as described above, from 0.830 (UP00321/UP00398 for Rap1, *Saccharomyces cerevisiae*, P11938) and 0.993 (UP00013/UP00408 for Gabpa, *Mus musculus*, Q00422). On average, 3DTF-derived PWMs were less similar to respective Uniprobe PWMs, as in comparisons with TRANSFAC data, possibly indicating importance of several factors such as 3D modeling quality or differences in experimental

conditions. One should also keep in mind possible inconsistencies in mapping binding factors identities from one experiment to another. In some cases, 3DTF-derived PWMs were extremely similar to Uniprobe matrices, for example, Egr1 with PCC of 0.875 and Gabpa with PCC of 0.819. The results on Tables 1–2 also show that structure-derived PWMs show a clear preference to be similar to appropriate experiment-derived PWMs.

CONCLUSIONS

The 3DTF webserver provides a possibility to derive DNA binding profiles based on the 3D structure of the protein/DNA complex. There are many structures in PDB that can be used for this purpose. More importantly, PDB structures can be used as templates for homology modeling of thousands of transcription factors having similar fold, but with very different binding specificities. Therefore, we expect that the webserver will be used to generate PWMs for families of transcription factors as well as for specific TFs provided that 3D structures are modeled accurately.

We showed that the method used here successfully reproduces strong binding of TFs to sequences of known binding sites. Therefore, we expect that this and similar structure-based PWM models (4) can further be used to develop novel matching algorithms outperforming existing methods. PWMs produced by 3DTF are correlated with known TF motifs, so that the webserver provides for a valid, alternative path for researchers to obtain PWMs of interest. The 3DTF server will be further developed to use more adequate or updated forcefields that should make its predictions more accurate.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–3 and Supplementary Figures 1–2.

FUNDING

BMBF PLANT-KBBE TRANSNET. The work of A.Kel was funded by the Russian federal program 'Living systems', State Contract #11.519.11.2031 and by FP7 project 'SysCol' and BMBF project 'GerontoShield'. Funding for open access charge: BMBF PLANT-KBBE TRANSNET and SysCol [FP7 no. 258236].

Conflict of interest statement. None declared.

REFERENCES

1. Mukherjee,S., Berger,M.F., Jona,G., Wang,X.S., Muzzey,D., Snyder,M., Young,R.A. and Bulyk,M.L. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.*, **12**, 1331–1339.
2. Alamanova,D., Stegmaier,P. and Kel,A. (2010) Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. *BMC Bioinformatics*, **11**, 225.
3. Robertson,T.A. and Varani,G. (2007) An all-atom, distance-dependent scoring function for the prediction of protein-DNA interactions from structure. *Proteins Struc. Funct.Bioinform.*, **66**, 359–374.
4. Contreras-Moreira,B. (2012) 3D-footprint: a database for the structural analysis of protein-DNA complexes. *Nucleic Acids Res.*, **38**, D91–D97.
5. Contreras-Moreira,B., Branger,P.A. and Collado-Vides,J. (2007) TFmodeller: comparative modelling of protein-DNA complexes. *Bioinformatics*, **23**, 1694–1696.
6. Gromiha,M.M., Siebers,J.G., Selvaraj,S., Kono,H. and Sarai,A. (2004) Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. *J. Mol. Biol.*, **337**, 285–294.
7. Ashworth,J. and Baker,D. (2009) Assessment of the optimization of affinity and specificity at protein-DNA interfaces. *Nucleic Acids Res.*, **37**, e73.
8. AlQuraishia,M. and McAdamsa,H.H. (2010) Direct inference of protein–DNA interactions using compressed sensing methods. *Proc. Natl Acad. Sci. USA*, **108**, 14819–14824.
9. Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.
10. Zhou,Q. (2010) On weight matrix and free energy models for sequence motif detection. *J. Comput. Biol.*, **17**, 1621–1638.
11. Kel,A., Kel-Margoulis,O., Babenko,V. and Wingender,E. (1999) Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *J. Mol. Biol.*, **288**, 353–376.
12. Kel,A.E., Gössling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
13. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
14. Stegmaier,P., Kel,A.E. and Wingender,E. (2004) Systematic DNA-binding domain classification of transcription factors. *Genome Inform.*, **15**, 276–286.
15. Arnold,K., Kiefer,F., Kopp,J., Battey,J.N., Podvinec,M., Westbrook,J.D., Berman,H.M., Bordoli,L. and Schwede,T. (2009) The protein model portal. *J. Struct. Funct. Genomics*, **10**, 1–8.
16. Chen,C.Y., Chien,T.Y., Lin,C.K., Lin,C.W., Weng,Y.Z. and Chang,D.T. (2012) Predicting target DNA sequences of DNA-binding proteins based on unbound structures. *PLoS One*, **7**, e30446.
17. Pickert,L., Reuter,I., Klawonn,F. and Wingender,E. (1998) Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics*, **14**, 244–251.
18. Kielbasa,S.M., Gonze,D. and Herzel,H. (2005) Measuring similarities between transcription factor binding sites. *BMC Bioinformatics*, **6**, 237.
19. Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**, D77–D82.