Supplementary Text S1

The classifier's performance on the datasets used by Bhardwaj et al.[1] and Stawiski et al.[2]

Stawiski et al., developed a neural net classifier on a dataset of 54 DBPs[3] and 250 nDBPs. Using leave one out cross validation, the classifier reached sensitivity of 81% and specificity of 94%. We examined our descriptors on the same dataset and reached sensitivity of 87% at specificity of 94%, using 10-fold cross validation.

Bhardwaj et al.[1] examined their classifier on a dataset of DBPs collected from previous studied[2,4,5] and the set nDBPs used by Stawiski et al.[2] The dataset was further filtered by allowing maximum of 20% sequence identity between each pair of proteins. Using 5-fold cross validation on the filtered dataset, their support vector machine reached sensitivity of 67.4% and specificity of 94.9%.
Using 5-fold cross validation on this dataset we reached sensitivity of 73.6% at specificity of 94.9%.

Culling the Bhardwaj dataset

Using the PISCES server we culled the PDB chains of the entries given at the web site which accompanied the paper (http://proteomics.bioengr.uic.edu/pro-dna/).[1] From the resulting list, we removed redundancy using PSI-CD-HIT[6] with a sequence identity cutoff of 20%. Finally we had a list of 87 DBPs and 216 nDBPs.

References:

1. Bhardwaj, N., Langlois, R. E., Zhao, G. & Lu, H. (2005). Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nucleic Acids Res.* 33, 6486-93.
2. Stawiski, E. W., Gregoret, L. M. & Mandel-Gutfreund, Y. (2003). Annotating nucleic acid-binding function based on protein structure. *J. Mol. Biol.* 326, 1065-79.
3. Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. (2000). An overview of the structures of protein-DNA complexes. *Genome Biol.* 1, REVIEWS001.1-001.37.
4. Ahmad, S., Gromiha, M. M. & Sarai, A. (2004). Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics* 20, 477-86.
5. Jones, S., Shanahan, H. P., Berman, H. M. & Thornton, J. M. (2003). Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.* 31, 7189-98.
6. Li, W. & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658-9.

Supplementary Table S1:

Score cutoff	Sensitivity	Specificity
0.49	0.906	0.713
0.50	0.906	0.725
0.51	0.891	0.734
0.52	0.884	0.745
0.53	0.884	0.757
0.54	0.877	0.779
0.55	0.870	0.790
0.56	0.862	0.800
0.57	0.855	0.810
0.58	0.848	0.821
0.59	0.833	0.830
0.61	0.819	0.840
0.62	0.812	0.851
0.63	0.804	0.861
0.64	0.775	0.868
0.65	0.732	0.873
0.66	0.717	0.883
0.68	0.717	0.893
0.69	0.710	0.904
0.70	0.667	0.921
0.72	0.645	0.929
0.75	0.623	0.937
0.77	0.587	0.943
0.78	0.580	0.954
0.80	0.536	0.958
0.83	0.464	0.969
0.85	0.428	0.975
0.87	0.377	0.979
0.90	0.333	0.985
0.91	0.283	0.987
0.93	0.217	0.989

The sensitivity and specificity of the classifier at different score cutoffs. The values were measured on a dataset of 138 DBPs and 843 nDBPs.

Supplementary Table S2:

The list of PDB entries in N-Func predicted as DBPs by the classifier, sorted in decreasing order of the confidence measure.

PDB id and chain	The proportion of votes as DBPs
2DBBA	0.9822
1PC6A	0.9565
1G2RA	0.9453
2HJ3A	0.9308
2GTVX	0.9305
2NRKA	0.9118
2ESHA	0.9107
1ZG2A	0.9067
2O38A	0.8953
1VKWA	0.8832
1VBKA	0.8775
1NOGA	0.8697
1WVTA	0.8563
1WJ9A	0.8526
1XD7A	0.844
1SDIA	0.8417
2CPXA	0.8278
1VQSA	0.8247
2I3FA	0.8245
2CYYA	0.8237
1TLJA	0.8194
1K3RA	0.8181
1LN4A	0.8152
2I76A	0.811
2GSCA	0.809
2HIYA	0.7965
2I2OA	0.7903
1U3EM	0.7864
1VQYA	0.7832
2GMYA	0.783
1UILA	0.7822
1P9QC	0.7758
1J27A	0.7732
2HH7A	0.7664
1ZEEA	0.7663
2IRUA	0.7604
2D59A	0.7596
2F6SA	0.757
2NWIA	0.7561
1WHXA	0.7551
2AP3A	0.7535
2A8EA	0.7529
2FNAA	0.7525
2FSWA	0.7507
1WOZA	0.7498
1ZN6A	0.7446
1YKUA	0.7443
1RW1A	0.7423
1L9GA	0.7417
2B6CA	0.7377
1V8DA	0.737
2FRNA	0.7339
1MWWA	0.7264
2B78A	0.7228
1VDYA	0.7225
2AEUA	0.72
1WW1A	0.7198
1TO0A	0.7182
2AR1A	0.7174
2IL5A	0.7168
2HXJA	0.7088
1S4KA	0.7036
1Y0KA	0.7018
1XBWA	0.6983
2ATZA	0.697
1YDMA	0.6943
2GUKA	0.6929
2IPQX	0.6925
2I51A	0.691
1Z85A	0.6901
2AEGA	0.6881
1X0MA	0.688
1K7JA	0.6871
1VAVA	0.6828
2GM3A	0.6828
1YX1A	0.6814
2IM9A	0.68
2A1VA	0.6797
1YW1A	0.6784
2O3AA	0.6783
2CPHA	0.6735
2BDTA	0.6715
1XJCA	0.6698
1WHRA	0.6679
2IF6A	0.6668
2GNXA	0.6665
1JRMA	0.6613
2DP9A	0.66
1XNEA	0.6589
1VPYA	0.658
2HWJA	0.6557
1YLNA	0.6556
1XA0A	0.6551
1KU9A	0.6544
2GMQA	0.6534
1RFZA	0.651
2GZ4A	0.6506
2HZTA	0.6501
2IBDA	0.648
1J26A	0.6479
2FEFA	0.647
1UANA	0.6453
2DJ6A	0.6449
2H3RA	0.6429
2OHWA	0.6392
1XPJA	0.6338
2OD0A	0.6331
1V30A	0.6322
2B25A	0.6299
2HH8A	0.6268
1RTYA	0.6206
1UGJA	0.6196
2CU5A	0.6195
1VGJA	0.6162
2G7ZA	0.6138
2ETSA	0.6134
2I9IA	0.6121
1IUKA	0.6049
2ICUA	0.6049
2IKBA	0.6039
1ZD0A	0.6029
2GKPA	0.6027
2HD9A	0.6024
1NG6A	0.6018
1VPVA	0.6004
2BDVA	0.6001
2G7JA	0.5994
1T06A	0.5973
1VMJA	0.5973
2CWQA	0.5971
2I6HA	0.5944
2AMHA	0.5928
1SMBA	0.5923
1WXXA	0.5921
2FUPA	0.5916
1WOLA	0.59
1RTTA	0.5896
1IUJA	0.5888
1UFOA	0.5878
2AV9A	0.5872

1T95A	0.5855
1VPHA	0.5839
2CRRA	0.5836
1UJRA	0.5808
1IHNA	0.5788
2IM8A	0.5788
1SH8A	0.5782
1U5WA	0.5749
1WY7A	0.5748
1O6DA	0.5726
2ALIA	0.5712
1ZMBA	0.5707
2CV9A	0.5707
1TUHA	0.57
2DLXA	0.5685
1T6SA	0.5679
2GBSA	0.5668
1O3UA	0.5659
1VLMA	0.5646
1VKMA	0.5627
1MK4A	0.5626
2FBLA	0.5626
1P90A	0.5625
2FUJA	0.5609
2IJCA	0.5534
2D13A	0.5533
2IAYA	0.5527
1Y81A	0.5525
1T6AA	0.5519
2NRQA	0.5512
1Y6IA	0.5476
2GX8A	0.546
2F46A	0.5453
2IMJA	0.545
1WICA	0.5425
1XM7A	0.5422
1R3DA	0.5401
2NWUA	0.5394
1PZXA	0.5375
2IVYA	0.5374
2GJGA	0.5365
1VPQA	0.5339
1ZXUA	0.5327
2AH6A	0.5312
1LFPA	0.5307
1WEKA	0.5292
1O50A	0.5283
2HQYA	0.528
2NLVA	0.528
1SJ5A	0.5276
1W8IA	0.5275
2F06A	0.5269
1VJLA	0.5266
2ETDA	0.5264

2IA0A	0.5241
2F4NA	0.5239
2O8QA	0.523
2D4GA	0.5214
1Y7PA	0.5212
1UFBA	0.5209
1PULA	0.5191
2GUUA	0.5182
1XXLA	0.517
2O8IA	0.5167
1S4CA	0.5166
2DCEA	0.5164
1XMTA	0.5155
1T3UA	0.5138
1VE0A	0.5125
1RCUA	0.5115
2F20A	0.5095
1SK7A	0.5082
1PVMA	0.5068
1ZTCA	0.5062
1V96A	0.5053
1YYVA	0.5014
2G40A	0.5007
2B0AA	0.5003

Supplementary Table S3:

PDB entry	Prediction score	ProFunc[1]	Dali [2]	Additional support	Predicted as RNA binding	Other
2DBB	0.9822	probable (HTH)	+	InterPro:[3] Winged helix repressor DNA-binding; Transcription regulator, AsnC/Lrp		
1PC6	0.9565			InterPro: Recombinase NinB		
1G2R	0.9453	long shot (DBP Template)			J.Osipiuk et al. (2001)[4]	
2HJ3	0.9308					ERV/ALR Sulfhydryl Oxidase- E.Vitu et al. (2006)
2GTV	0.9305	Possible (RBP Template) long shot (DBP Template)				Chorismate mutase
2NRK	0.9118	long shot (DBP Template)	+			
2ESH	0.9107	long shot (DBP Template) long shot (RBP Template)	+	InterPro: Winged helix repressor DNA-binding; Transcriptional regulator PadR-like		
1ZG2	0.9067		+	InterPro: Excinuclease ABC, C subunit, N-terminal; Pfam: GIY-YIG catalytic domain		
2O38	0.8953	probable (HTH)	+			
1VKW	0.8832	long shot (RBP Template)				InterPro: FMN-dependent nitroreductase-like superfamily
1VBK	0.8775		+		Pfam:[5] THUMP domain - a predicted RNA-binding domain	
1NOG	0.8697		+			InterPro:ATP:cob(I)alamin adenosyltransferase
1WVT	0.8563		+			InterPro:ATP:cob(I)alamin adenosyltransferase
1WJ9	0.8526				A.Ebihara et al. (2006)[6]	
1XD7	0.844	probable (HTH)	+	Prosite: rrf2-type HTH domain. A putative DNA-binding domain.		

1SDI	0.8417	long shot (DBP Template) long shot (RBP Template)				
2CPX	0.8278	probable (RBP Template)			InterPro: RNA recognition motif, RNP-1	
1VQS	0.8247					
2I3F	0.8245					j.g.mccoy et al. - Crystal Structure of a Glycolipid transfer-like protein from Galdieria sulphuraria; InterPro: Glycolipid transfer protein, GLTP superfamily
2CYY	0.8237	probable (HTH)	+	InterPro: Winged helix repressor DNA-binding; Transcription regulator, AsnC/Lrp; Bacterial regulatory protein, ArsR		
1TLJ	0.8194				InterPro: methyltransferase TYW3 (tRNA-yW- synthesising protein 3)	
1K3R	0.8181				T.I.Zarembinski et al. (2003)	
1LN4	0.8152		+		predicted novel class of RNA binding proteins. G.J.Ostheimer et al. (2002)[7]	
2I76	0.811					(Crystallized with NDP)
2GSC	0.809	long shot (DBP Template)			Pfam: S23 ribosomal protein	
2HIY	0.7965					
2I2O	0.7903				e.bitto et al. eif4g-like protein	
1U3E	0.7864	probable (HTH)	+	(Crystallized with dsDNA)		
1VQY	0.7832		+			
2GMY	0.783					
1UIL	0.7822	probable (RBP Template)	+		t.nagata et al. - double-stranded RNA-binding motif	

Table S3:

Analysis of the top ranked PDB entries from N-Func predicted as DBPs. Prediction score, the score assigned by the classifier; ProFunc, indicates helix-turn-helix[8] motifs or similarity with DNA/RNA binding templates[9] (DBP/RBP respectively) found in the query protein; Dali, a '+' sign indicates a significant fold similarity[2] between the query protein and a known DBP; Additional support, specify additional support for the prediction. The last two columns specify whether there is evidence that the protein binds RNA or other molecules.

References:

1. Laskowski, R. A., Watson, J. D. & Thornton, J. M. (2005). ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.* 33, W89-93.
2. Holm, L. & Park, J. (2000). DaliLite workbench for protein structure comparison. *Bioinformatics* 16, 566-7.
3. Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., Courcelle, E., Das, U., Daugherty, L., Dibley, M., Finn, R., Fleischmann, W., Gough, J., Haft, D., Hulo, N., Hunter, S., Kahn, D., Kanapin, A., Kejariwal, A., Labarga, A., Langendijk-Genevaux, P. S., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Nikolskaya, A. N., Orchard, S., Orengo, C., Petryszak, R., Selengut, J. D., Sigrist, C. J., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H. & Yeats, C. (2007). New developments in the InterPro database. *Nucleic Acids Res.* 35, D224-8.
4. Osipiuk, J., Gornicki, P., Maj, L., Dementieva, I., Laskowski, R. & Joachimiak, A. (2001). Streptococcus pneumonia YlxR at 1.35 A shows a putative new fold. *Acta Crystallogr. D Biol. Crystallogr.* 57, 1747-51.
5. Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. & Bateman, A. (2008). The Pfam protein families database. *Nucleic Acids Res.* 36, D281-8.
6. Ebihara, A., Yao, M., Masui, R., Tanaka, I., Yokoyama, S. & Kuramitsu, S. (2006). Crystal structure of hypothetical protein TTHB192 from Thermus thermophilus HB8 reveals a new protein family with an RNA recognition motif-like domain. *Protein Sci.* 15, 1494-9.
7. Ostheimer, G. J., Barkan, A. & Matthews, B. W. (2002). Crystal structure of E. coli YhbY: a representative of a novel class of RNA binding proteins. *Structure* 10, 1593-601.
8. Ferrer-Costa, C., Shanahan, H. P., Jones, S. & Thornton, J. M. (2005). HTHquery: a method for detecting DNA-binding proteins with a helix-turn-helix structural motif. *Bioinformatics* 21, 3679-80.
9. Laskowski, R. A., Watson, J. D. & Thornton, J. M. (2005). Protein function prediction using local 3D templates. *J. Mol. Biol.* 351, 614-26.

Supplementary Table S4:

Method / PDB id	DBS-PRED[1] (sequence based)	DB-MOM[2]	PreDs[3]	Szilágyi and Skolnick[4]	Nimrod et al., 2008
2IVHA	+	+	+	+	+
2JG3A	-	-	+	+	+
2BGWA	-	-	-	+	+
2NQJA	-	-	-	-	+
2UZKA	+	+	+	+	+
2VLAA	+	-	+	+	+
3CLZA	+	-	+	-	+
2V6EA	+	-	+	+	+
2BSQE	-	+	-	+	+
2V1UA	-	-	-	+	+
2IVKA	+	-	+	+	+
Correct predictions	6	3	7	9	11

The success in the identification of 11 new structures of DBPs by 5 different methods. Each column represents a method and each row is an examined PDB entry. The '+' sign (grey shading), represents successful prediction, and a '-' stands for a false prediction. On this small set, our methods identified all the new structure correctly.

References

1. Ahmad, S., Gromiha, M. M. & Sarai, A. (2004). Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics* 20, 477-86.
2. Ahmad, S. & Sarai, A. (2004). Moment-based prediction of DNA-binding proteins. *J. Mol. Biol.* 341, 65-71.
3. Tsuchiya, Y., Kinoshita, K. & Nakamura, H. (2004). Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins* 55, 885-94.
4. Szilagyi, A. & Skolnick, J. (2006). Efficient prediction of nucleic acid binding function from low-resolution protein structures. *J. Mol. Biol.* 358, 922-33.