Nucleic Acids Research Advance Access published May 19, 2016

Nucleic Acids Research, 2016 **1**
doi: 10.1093/nar/gkw454

BindUP: a web server for non-homology-based prediction of DNA and RNA binding proteins

Inbal Paz, Efrat Kligun, Barak Bengad and Yael Mandel-Gutfreund*

Department of Biology, Technion—Israel Institute of Technology, Technion City, Haifa 32000, Israel

Received March 1, 2016; Revised May 1, 2016; Accepted May 11, 2016

ABSTRACT

Gene expression is a multi-step process involving many layers of regulation. The main regulators of the pathway are DNA and RNA binding proteins. While over the years, a large number of DNA and RNA binding proteins have been identified and extensively studied, it is still expected that many other proteins, some with yet another known function, are awaiting to be discovered. Here we present a new web server, BindUP, freely accessible through the website http://bindup.technion.ac.il/, for predicting DNA and RNA binding proteins using a non-homology-based approach. Our method is based on the electrostatic features of the protein surface and other general properties of the protein. BindUP predicts nucleic acid binding function given the proteins three-dimensional structure or a structural model. Additionally, BindUP provides information on the largest electrostatic surface patches, visualized on the server. The server was tested on several datasets of DNA and RNA binding proteins, including proteins which do not possess DNA or RNA binding domains and have no similarity to known nucleic acid binding proteins, achieving very high accuracy. BindUP is applicable in either single or batch modes and can be applied for testing hundreds of proteins simultaneously in a highly efficient manner.

INTRODUCTION

Nucleic acid binding proteins (NABPs), specifically DNA-binding proteins (DBPs) and RNA-binding proteins (RBPs), play a crucial role in all steps of the gene expression pathway, from RNA transcription via post-transcriptional regulation to protein translation (1,2). In recent years it is becoming apparent that both DBPs and RBPs are also involved in epigenetic regulation (3,4). Understanding the complexity of the gene expression regulation requires the identification of RBPs and DBPs that are involved in these processes and defining their RNA and DNA binding sites. Over the years, high resolution structures of protein–DNA and protein–RNA complexes, solved by x-ray crystallography and nuclear magnetic resonance, have provided crucial information on the properties of DBPs and RBPs and on their modes of interactions with the nucleic acids. In recent years, there has been an enormous advance in the development of high-throughput experimental technologies for detecting NABPs. Recently, several high-throughput proteomic-based methodologies were developed for *in vivo* detection of RBPs in eukaryotes (5–11). These approaches (known as 'RNA interactome capture' experiments) were successfully employed to identify a large fraction of known RBPs, as well as to detect novel RBPs. While DBPs have been extensively studied and characterized (2), *de novo* detection of proteins which bind DNA is generally a hard task. *In vitro* methods for detection of RNA and DNA binding specificities, such as RNAcompete (12) and protein binding microarrays (13), respectively, have been also employed for validating nucleic acid binding preferences.

Despite the advancement of the experimental techniques to discover NABPs, given the high throughput nature of these techniques, these approaches tend to produce a high number of incorrectly detected proteins (including false negatives and false positives). It is thus of great importance to complement the experimental approaches, aiming to discover novel NABPs, with sophisticated computational approaches. Over the years many computational approaches have been developed for the identification and classification of DNA and RNA binding proteins and their binding sites (for recent reviews see (14–17)). The computational methods for classifying DNA and RNA binding proteins can be roughly divided into methods that are based on the protein structure (e.g. (18–21) applied for DNA and (22–25) for RNA) and those that rely on the amino acid sequence alone (such as (26–28)). Moreover, function prediction methods have been employed for predicting the function of NABPs from sequences based on structure, using fold recognition approaches, e.g. (25,29). Several of these methods have been implemented to web servers, such as DBD-hunter (19), Spot-Struct-DNA (30) and Spot-Struct-RNA (25) for identifying DNA and RNA binding proteins from structure. In addition, DBD-threader (29), PSSM-DT (16), RNApred (27) and SPOT-seq-RNA (24) web servers

*To whom correspondence should be addressed. Tel: +972 4 8293958; Fax: +972 4 8225153; Email: yaelmg@tx.technion.ac.il

are available for annotating NA-binding function given the protein sequence using template-based approaches.

The vast majority of the computational approaches (both structural based and sequence based methods), available to-date for classifying NABPs, rely on homology and thus are less effective for identifying novel DNA and RNA binding proteins. Recently, Zhou *et al*. applied their template-based SPOT-seq-RNA algorithm to the entire human proteome, correctly identifying 42.6% of all annotated RBPs (31). Similar results were achieved when testing their method on the RBPs which were discovered by the interactome capture experiment conducted in Human HeLA cells (6). We have previously developed a machine learning approach named NAbind for classifying DNA and RNA binding proteins given the protein structures (20,23). NAbind is a machine learning approach, based on extracting the largest positive electrostatic patch on the protein surface, implemented in our PFPlus web server (32). Notably, while the majority of NABPs bind the DNA or RNA via a continuous positive interface, some proteins employ different strategies, for example the tRNA binding proteins that possess a significant large negative patch and bind their tRNA substrate via two distinct positive patches (33). Such proteins that do not rely on the large positive surface for NA-binding are likely to be mispredicted by the algorithm. Nevertheless, the great advantage of the NAbind algorithm is that it is based on the overall physicochemical and structural features of the NABPs, learnt from the three-dimensional (3D) structures of known DBPs and RBPs, and does not rely on either sequence or structural homology to the known NABPs. NAbind has been recently trained on a large set of non-redundant DNA and RNA binding proteins (sharing <25% sequence identify) from the protein data bank (PDB) (34) and has been modified to be applicable for both experimentally solved protein structures as well as low resolution structural models, derived from protein sequences.

Here we describe a new web server, BindUP, for predicting NABPs based on the electrostatic patches on the protein surfaces, employing the NAbind algorithm (20,23). The server was tested on a completely independent dataset of DNA and RNA binding protein structures, achieving an Area Under ROC Curve (AUC) value of 0.94, with 0.71 sensitivity and 0.96 specificity. Moreover, BindUP was successfully applied on non-homology-based structural models of novel RBPs. Further, we show that the positive electrostatic patches extracted by BindUP highly overlap with the NA-binding regions, suggesting that BindUP can also be employed for identifying binding interfaces. The information on the largest positive patches, as well as the negative patches, is provided both graphically and in text. BindUP is significantly more efficient than PFPlus (32) in providing the electrostatic patch information, as all patches are pre-calculated and stored in a database. The server currently holds information for all 117 882 protein structures in the PDB (as of 18 April 2016). BindUP is applicable in either single or batch mode and can be applied for testing hundreds of proteins simultaneously in a highly efficient manner. BindUP is freely accessible via the website http://bindup.technion.ac.il/.

BindUP METHODOLOGY

The algorithm for predicting NABPs, given a 3D structure of a protein or a structural model, is based on our NAbind algorithm, originally developed and trained on high resolution structures of DBPs (20) and further on RBPs (23). The NAbind algorithm is based on the unique features of the proteins' electrostatic surface patches, implemented in the PatchFinder web server(32). The PatchFinder algorithm (20) automatically assigns surface (positive and negative) patches by looking for adjacent points on the protein surface that meet a given electrostatic potential cut-off (2 or -2 kT/e, for positive and negative patches, respectively). The algorithm is built of several steps. In the first step, the electrostatic potential of the protein is calculated on a 3D grid, using the Poisson Boltzmann equation. The electrostatic potential is calculated using the APBS software (35) with a grid spacing of 1Å. Hydrogen atoms are added prior to the calculations using PDB2PQR (36). We further define the grid points that fall on the protein surface, using the DMS open source http://www.cgl.ucsf.edu/Overview/software.html#dms to calculate the surface accessibility, based on the Lee and Richards algorithm (37), ignoring all non-surface points. We then extract continuous electrostatic patches on the protein surface by selecting all 3D patches of adjacent grid points which meet the defined cut-off. Finally, we select the largest electrostatic patches for each protein chain and assign the protein residues related to the positive and negative patches.

As aforementioned, the NAbind algorithm employs the information from the largest positive and negative patches, extracted by PatchFinder, as well as other structural features of the protein. Among these features are the molecular weight, the overall surface accessibility and the moment dipole of the protein chain. The patch features include the size of the patch (positive and negative), the overall potential and surface accessibility of the largest positive electrostatic patch as well as the overlap between the largest positive patch and the largest cleft on the protein surface (a detailed description of the features is found in (23)). To distinguish NABPs from non-NABPs, we use the GIST Support Vector Machine (SVM) classifier http://www.chibi.ubc.ca/gist/, trained with a linear kernel function using the default parameters. All properties are fed into the input matrix as numeric features with no manipulations conducted on the matrix. The SVM was trained on a non-redundant set of 450 protein chains, including 90 DNA-binding, 60 RNA-binding and 300 non-NA-binding, extracted from the PDB database. The dataset was generated by selecting from the PDB all DNA and RNA binding protein chains and further removing redundancy by employing the BLASTClust program, which uses the BLAST local alignment algorithm for pairwise comparison and clustering (38). We further selected the representative protein structure from each of the clusters (sharing <25% sequence identity between them) with the best resolution. An equivalent dataset of proteins which do not bind nucleic acids was generated in a similar manner. Finally, the representative datasets of NA and non-NA binding proteins were manually curated, ensuring that the selected proteins chains are the binding or non-binding chains, respectively. Notably, the uniqueness of the NAbind

algorithm is that it does not rely on either sequence or structural homology and thus can be applied for predicting novel NABPs. Nevertheless, it is important to note that NAbind was trained and tested on single NA-binding chains and thus it may fail in predicting NA-binding in cases where proteins bind the nucleic acid as large multimeric complexes and rely on the large electrostatic patch, induced by the complex formation, as for example in the case of the DnaQ-like 3'-5' exonuclease (39).

BindUP DESCRIPTION

Input

BindUP server has two modes of usage, a single protein mode and a batch mode. In the single protein mode, BindUP predicts the NA-binding propensity for a given protein structure. The user can choose whether to calculate all the protein chains of the structure (each chain is calculated separately) or to select a specific chain identifier. The structure can be provided as either a PDB ID or as a user-defined coordinate file (of a known structure or a structural model) in PDB format. In case the input is provided as PDB ID, BindUP retrieves the results from a database of pre-calculated predictions. Other calculation options enable the user to control the type and number of electrostatic patches that will be displayed in the results. By default, BindUP displays the largest positive electrostatic patch. However, it is possible to choose whether to display only positive patches or negative patches (up to three patches together) or the combination of both (one positive and one negative patch).

In the batch mode, BindUP gets a list of protein structures and calculates the requested electrostatic patches and the NA-binding prediction for each structure. The structures should be provided as PDB IDs only, pasted into the browser or uploaded as a text file. The number of entries is unlimited. The four-letter PDB ID may be followed by a chain identifier, to indicate that the calculation should be performed on the specific chain exclusively. Otherwise, the calculation will be performed on all the protein chains of the structure. The list of PDB entries may be combined of four-letter PDB ID entries (e.g. 1d66) and PDB IDs including a chain identifier (e.g. 1d66A) mixed together.

In both modes it is optional to add an e-mail address to which the results will be automatically sent when the analysis is completed.

Output

BindUP calculation is performed on each protein chain separately. The results, for each requested protein chain, include the NA-binding prediction and the requested electrostatic patches on the protein surface. In the single protein mode, BindUP results are provided both in a web-based presentation and in downloadable text files. In case the user has requested to calculate a specific chain or if the PDB file contains only one protein chain, BindUP presents the results for this chain exclusively (Figure 1A). In case BindUP calculates more than one protein chain, it initially presents the results for the first protein chain (Figure 1B). Using a drop-down menu, it allows the user to interactively switch between all the protein chains and display the results per each chain. The last option in the drop-down menu is 'all', which displays the results for all the protein chains together (Figure 1C and D). Both the graphic presentation and the downloadable text files change according to the chain selection. Notably, when selecting the 'all' option, the NA-prediction is not displayed, as results may differ between chains. In this case, the prediction appears in the results text file.

The web-based presentation includes the NA-binding prediction for the selected chain and a visualization of the electrostatic patches, requested by the user, using Jmol: an open-source Java viewer for chemical structures in 3D (http://www.jmol.org/). In addition to the graphic presentation, BindUP provides two text files for download. The first file is a summary of the results. It contains the NA-binding prediction and the residues composing the electrostatic patches requested by the user. The second file is the coordinate PDB file, with the patches annotation inserted to the B-factor (temperature) column (the color-coding is described in the manual section of the website). These two downloadable files interactively change according to the user's choice. In the batch mode, the results for each protein structure are provided as downloadable text files only. The two text files are the same as described above and include the results for one chain or for all the protein chains, according to the input provided by the user. The first link in the results page refers to a summary text file, including the results for all the protein structures that have been submitted in the current job together.

RESULTS AND DISCUSSION

In the last few years there has been a great advancement in experimental (*in vivo* and *in vitro*) technologies for the detection of RNA and DNA binding proteins (5,7,9–13,40). However, given the many new roles expected for these proteins, it is estimated that many other proteins are yet to be discovered. In previous studies we have developed the NAbind algorithm for predicting novel DNA and RNA binding proteins from the structure of the protein, without relying on either sequence or structural homology (20,23). Given the enormous expansion in the number of DBPs and RBPs in the PDB (including proteins bound in complex with the nucleic acid as well as proteins solved in the unbound state) and the advancement in methods for non-homology-based protein structure predictions (41), we have added to the algorithm an additional feature, enabling the prediction of NABPs given a structural model of the protein predicted from sequence. Furthermore, while our previous studies have considered DNA and RNA binding proteins separately (training the algorithms on each group of proteins independently), the current version of NAbind, implemented in our new web server BindUP, was trained on a mixed set of DBPs and RBPs and does not attempt to distinguish between the two types of NABPs. This is consistent with the growing knowledge of proteins that bind both DNA and RNA (42), as well as our previous work showing that DBPs and RBPs can be weakly distinguished when considering only double stranded DBPs versus single stranded RBPs (43) and recent studies showing that algorithms for predicting DNA and RNA binding sites are un-

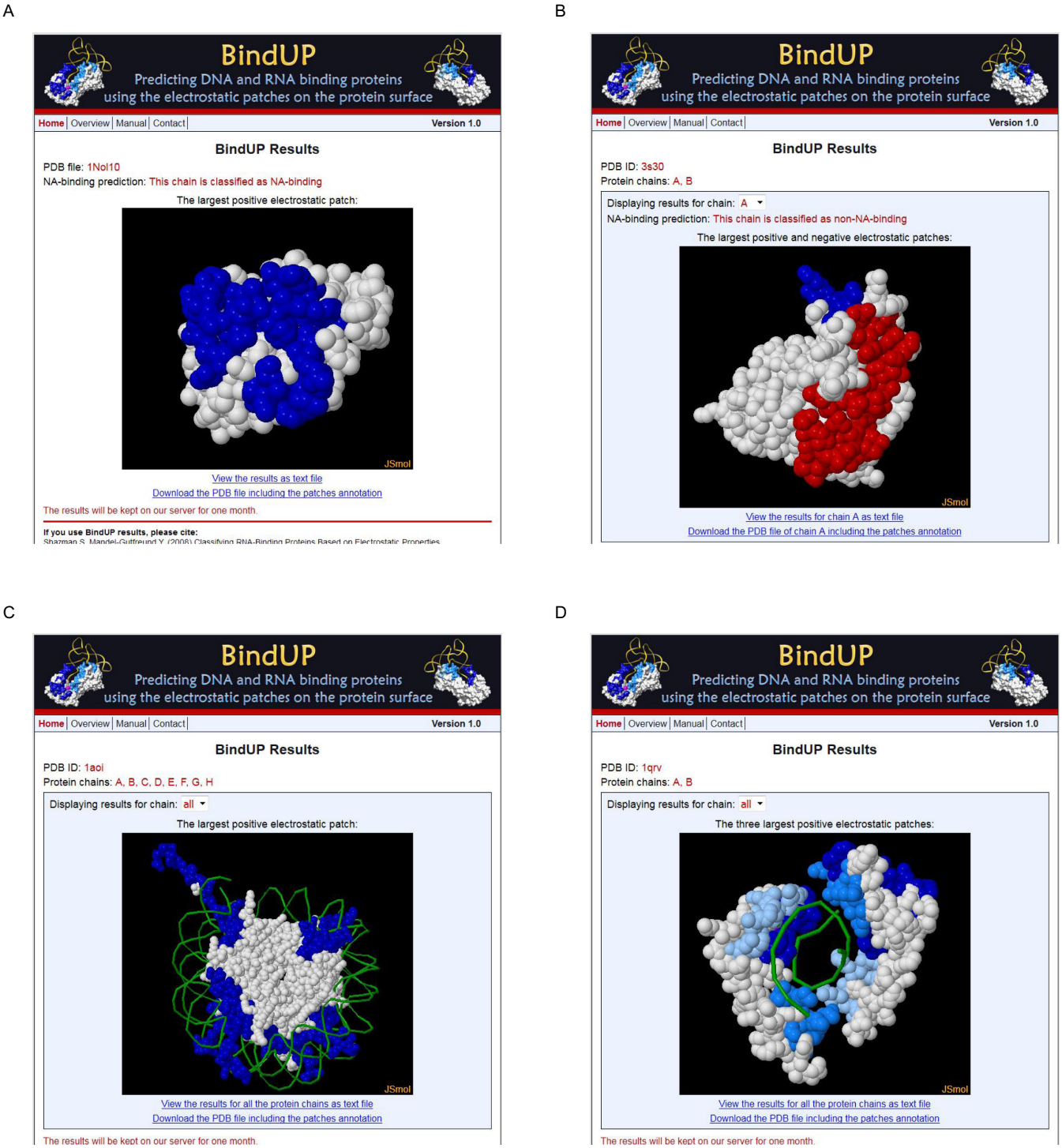

Figure 1. Examples of BindUP results pages. (**A**) A presentation of the largest positive patch, calculated on a structural model of NOL10, constructed by I-TASSER. The model is predicted to be NA-binding. (**B**) A presentation of the largest negative and positive patches, calculated on chain A of PDB ID: 3S30, which is predicted to be non-NA-binding. (**C**) A presentation of the largest positive patches, calculated on eight protein chains of PDB ID: 1AOI, displayed together with the DNA chains. (**D**) A presentation of the three largest positive patches, calculated on the two protein chains of PDB ID: 1QRV, displayed together with the DNA chains.

able to distinguish between the binding sites of the different NABPs (15,17).

We have tested BindUP on an independent set of 323 structures of DNA and RNA binding proteins (BindUP_NA323) and on a control set of an equal number of non NA-binding proteins extracted from the PDB. Overall, we achieve an AUC value of 0.94, with 0.71 sensitivity and 0.96 specificity (see Table 1 and detailed results in Supplementary Table S1). As expected, among the proteins we mispredicted are proteins that do not rely on a large continuous electrostatic patch to bind the nucleic acid, such as yeast aspartyl-tRNA synthetase (PDB ID: 1ASY), or proteins that bind nucleic acids as large multimeric complexes, such as Ebola virus matrix protein VP40 (PDB ID: 1H2C) that binds the RNA as an octamer. Notably, the proteins in our test set were completely independent from the proteins in the training set, sharing <25% sequence identity among them and when compared to each of the proteins in the training set. To further ensure that BindUP does not rely on structural homology we used the CATH (44) structural classification to generate an additional training set of 230 structures of DNA and RNA binding proteins (BindUP_NA230_struct) and a control set of an equal number of non-NABPs, extracted from the PDB, which do not share structural homology (CATH 'H-level') with any protein in the training set. The results of the structural non-redundant set were very similar to those achieved for BindUP_NA323, with an AUC value of 0.91, 0.7 sensitivity and 0.91 specificity (see Table 1 and detailed results in Supplementary Table S2). To further test BindUP on another independent set of NABP structures, we extracted from the literature a recently compiled dataset of 627 NABPs solved in complex with the nucleic acid (DNA or RNA), generated by Miao and Westhof (45). We have further removed from this set proteins which overlapped with our training set and structures that are defined as obsolete in the RCSB PDB database, ending up with 535 NABP structures. As a control set we used an equal number of non-NABPs, extracted from the PDB. In this case, BindUP achieved less accurate results with an AUC of 0.83 compared to 0.94 in our independent testing set, 0.6 sensitivity and 0.9 specificity (see Table 1 and detailed results in Supplementary Table S3). Given the fact that previous prediction algorithms considered DBPs and RBPs prediction separately, to compare our results with other NA-binding prediction servers, we tested BindUP on independent sets of DBPs (BindUP_D190) and RBPs (BindUP_R127). The two latter sets are subsets of BindUP_NA323, excluding six protein chains that are annotated as both DNA and RNA binding proteins. The results for the independent lists were AUC ROC values of 0.96 and 0.90 for DBPs and RBPs, respectively (Table 1, detailed results in Supplementary Tables S4 and S5). We further compared BindUP predictions to the only two available active servers on the www for predicting NA-binding function from the protein structure: SPOT-Struct-DNA (30) and SPOT-Struct-RNA (25), for DBPs and RBPs, respectively. As shown in Table 1, both programs achieved very similar results to BindUP, with slightly higher sensitivity values attained by BindUP. However, in comparison to SPOT-Struct-DNA and SPOT-Struct-RNA, BindUP runs on batch mode and thus can predict the NA-binding function for an unlimited number of structures in one session. Moreover, as all BindUP calculations are stored in a database, holding information for the entire set of protein structures in the PDB, it runs extremely fast, independent of the size of the proteins and the number of protein chains in the PDB structure. To our knowledge, BindUP is the only web server for prediction of NA-binding function from structure which runs efficiently in a batch mode.

Clearly, the most important advantage of NAbind, implemented in BindUP, compared to other NA-binding prediction algorithms, is that it does not rely on homology and thus can contribute to predict novel NABPs. While there is a relatively large number of 'hypothetical proteins' in the PDB resulting from structure genomic initiatives, still the majority of protein structures in PDB are of known function. To examine the ability of BindUP to predict NA-binding function from sequence for novel proteins, we chose to test it on RBPs which were recently identified by the RNA interactome capture experiments (5,7,8), deliberately selecting the subset of proteins which do not possess known RNA or DNA binding domains and have no homologous structure in the PDB. To this end, we extracted a list of 131 protein orthologs, reported by Kwon *et al.* (8) to be common to the three RNA interactome capture experiments, conducted in mammalian cells (5,7,8) and were not previously annotated as RBPs. We further manually curated the list, removing proteins possessing domains annotated in PFAM (46) as related to NA-binding function as well as proteins that shared over 35% identity (>30 amino acid coverage) to any other protein in the PDB, ending up with a final list of 58 novel RBPs. We further extracted all the domains from the proteins and used the I-TASSER non-homology-based protein structure modeler software (41), which we ran locally on our servers, to generate the structural model of the protein domains. The models of each of the domains were tested independently on BindUP. Overall, 86% of the RBPs in our test set had at least one domain predicted as NA-binding by BindUP (Supplementary Table S6). An example of a novel RBP, predicted correctly as NA-binding by BindUP, is Nol10. Nol10 is a WD-repeat nuclear protein with an unclear function, previously known to bind protein complexes in the nucleoli (47) and recently shown to bind RNA in the high throughput interactome capture experiments (5,7,8). We predicted the structure of its two domains using I-TASSER (41) and further submitted them to BindUP. Interestingly, while the Nol10 domains have no significant homology to other structures in the PDB and more so do not share similarity with any RNA binding domain, both domains were predicted by BindUP as NA-binding (Figure 1A). Overall, consistent with the fact that BindUP does not rely on homology, these results strongly support that BindUP can accurately identify novel NABPs. Moreover, while NAbind was originally developed for predicting NA-binding function from structure, we show that BindUP achieves highly accurate results given structural models of proteins that are predicted from sequence, using non-homology-based approaches.

As aforementioned, the NAbind algorithm strongly depends on the features of the electrostatic patches on the protein surface, extracted by the PatchFinder algorithm, implemented in the PFPlus web server (32). In addition to pre-

Table 1. A summary of BindUP results tested on different datasets

Dataset	Algorithm	Sensitivity	Specificity	AUC
BindUP_NA323	BindUP	0.71	0.96	0.94
BindUP_NA230_struct	BindUP	0.70	0.91	0.91
BindUP_R127	BindUP	0.65	0.97	0.90
BindUP_R127	SPOT-Struct-RNA	0.63	0.99	
BindUP_D190	BindUP	0.74	0.95	0.96
BindUP_D190	SPOT-Struct-DNA	0.57	1.00	
RBscore_P627	BindUP	0.60	0.90	0.83

Sensitivity was calculated using the formula TP/(TP + FN). Specificity was calculated using the formula TN/(TN + FP). AUC was calculated using the Gist Support Vector Machine (SVM) classifier (http://www.chibi.ubc.ca/gist/). Results for SPOT-Struct-DNA and SPOT-Struct-RNA were obtaining by running the independent datasets D190 and R127 of the respective web servers. RBscore_P627 was extracted from (15,45). The dataset was processed, removing protein structures that are defined as obsolete in the RCSB PDB database, as well as structures that overlap with BindUP training set.

dicting whether a protein binds nucleic acids, BindUP provides the information on the largest positive patches on the protein surface. We have previously shown that the largest positive patches on proteins, extracted by the PatchFinder algorithm, highly overlap with DNA and RNA binding interfaces (23,32). Moreover, we have shown that PatchFinder can also be applied to structural models, derived from a non-homology-based modeling method, showing high overlap with the known binding interfaces (48). Given the phenomenal growth in the number of protein-NA complexes in the PDB, we repeated the previous tests on the most updated set in the literature of non-redundant protein chains which were solved in complex with nucleic acids (including protein–DNA and protein–RNA complexes) (45). We compared the largest positive patches, calculated by BindUP, to the known nucleic acid binding interface, extracted by the program Intervor, which employs the Voroni interface model (49). This was done by calculating the overlap between the residues composing the one, two or three largest positive patches and the residues that are part of the known binding interface. The results are detailed in Supplementary Table S7. Overall, we show that the largest electrostatic patches highly overlap with the real NA-binding interface (extracted by Intervor). When considering the residues composing the largest positive patch only, the median sensitivity and specificity of the overlap is 0.65 and 0.86, respectively (Supplementary Table S7). When considering the residues in the three largest positive patches on the protein surface, as suggested in (23), the median sensitivity increases to 0.72 and accordingly the specificity reduces to 0.75.

Taken together, BindUP is currently the most accurate and efficient web service for computational prediction of NA-binding function. BindUP is a non-homology-based predictor and can thus be applied for predicting novel NABPs given the protein structure or a structural model derived from sequence. In addition to providing the prediction of the protein function, i.e. whether it is an NA-binding protein or not, BindUP offers information on the largest continuous electrostatic patches on the protein surface. The latter has been shown to correspond with functional regions on the protein surface, specifically to the NA-binding interface, in case of the largest continuous positive patch.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank BindUP users for their useful comments and suggestions for improving the web server. Specifically we would like to thank Fabian Glaser for helpful suggestions in designing the website.

FUNDING

Israel Science Foundation (ISF) [1623/12 to Y.M.G.]. Funding for open access charge: Israel Science Foundation [1623/12].

Conflict of interest statement. None declared.

REFERENCES

1. Gerstberger,S., Hafner,M. and Tuschl,T. (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 829–845.
2. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
3. Minarovits,J.M.H.C., Banati,F.H., Szenthe,K.H. and Niller,H.H. (2016) Epigenetic regulation. *Adv. Exp. Med. Biol.*, **879**, 1–25.
4. Cirillo,D., Livi,C.M., Agostini,F. and Tartaglia,G.G. (2014) Discovery of protein-RNA networks. *Mol. Biosyst.*, **10**, 1632–1642.
5. Baltz,A.G., Munschauer,M., Schwanhausser,B., Vasile,A., Murakawa,Y., Schueler,M., Youngs,N., Penfold-Brown,D., Drew,K., Milek,M. *et al.* (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell*, **46**, 674–690.
6. Castello,A., Fischer,B., Eichelbaum,K., Horos,R., Beckmann,B.M., Strein,C., Davey,N.E., Humphreys,D.T., Preiss,T., Steinmetz,L.M. *et al.* (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, **149**, 1393–1406.
7. Castello,A., Horos,R., Strein,C., Fischer,B., Eichelbaum,K., Steinmetz,L.M., Krijgsveld,J. and Hentze,M.W. (2013) System-wide identification of RNA-binding proteins by interactome capture. *Nat. Protoc.*, **8**, 491–500.
8. Kwon,S.C., Yi,H., Eichelbaum,K., Fohr,S., Fischer,B., You,K.T., Castello,A., Krijgsveld,J., Hentze,M.W. and Kim,V.N. (2013) The RNA-binding protein repertoire of embryonic stem cells. *Nat. Struct. Mol. Biol.*, **20**, 1122–1130.
9. Mitchell,S.F., Jain,S., She,M. and Parker,R. (2013) Global analysis of yeast mRNPs. *Nat. Struct. Mol. Biol.*, **20**, 127–133.
10. Beckmann,B.M., Horos,R., Fischer,B., Castello,A., Eichelbaum,K., Alleaume,A.M., Schwarzl,T., Curk,T., Foehr,S., Huber,W. *et al.* (2015) The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nat. Commun.*, **6**, 10127–10135.
11. Matia-Gonzalez,A.M., Laing,E.E. and Gerber,A.P. (2015) Conserved mRNA-binding proteomes in eukaryotic organisms. *Nat. Struct. Mol. Biol.*, **22**, 1027–1033.
12. Ray,D., Kazan,H., Chan,E.T., Pena Castillo,L., Chaudhry,S., Talukder,S., Blencowe,B.J., Morris,Q. and Hughes,T.R. (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.*, **27**, 667–670.

13. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. 3rd and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
14. Si,J., Cui,J., Cheng,J. and Wu,R. (2015) Computational prediction of RNA-Binding Proteins and Binding Sites. *Int J Mol Sci*, **16**, 26303–26317.
15. Miao,Z. and Westhof,E. (2015) A Large-Scale Assessment of Nucleic Acids binding site prediction programs. *PLoS Comput. Biol.*, **11**, e1004639.
16. Xu,R., Zhou,J., Wang,H., He,Y., Wang,X. and Liu,B. (2015) Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. *BMC Syst. Biol.*, **9**(Suppl. 1), S10–S21.
17. Yan,J., Friedrich,S. and Kurgan,L. (2016) A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief. Bioinform.*, **17**, 88–105.
18. Ahmad,S., Gromiha,M.M. and Sarai,A. (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.
19. Gao,M. and Skolnick,J. (2008) DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. *Nucleic Acids Res.*, **36**, 3978–3992.
20. Stawiski,E.W., Gregoret,L.M. and Mandel-Gutfreund,Y. (2003) Annotating nucleic acid-binding function based on protein structure. *J. Mol. Biol.*, **326**, 1065–1079.
21. Nimrod,G., Szilagyi,A., Leslie,C. and Ben-Tal,N. (2009) *J. Mol. Biol.*, **387**, 1040–1053.
22. Ahmad,S. and Sarai,A. (2011) Analysis of electric moments of RNA-binding proteins: implications for mechanism and prediction. *BMC Struct. Biol.*, **11**, 8–20.
23. Shazman,S. and Mandel-Gutfreund,Y. (2008) Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput. Biol.*, **4**, e1000146.
24. Yang,Y.U., Zhao,H., Wang,J. and Zhou,Y. (2014) SPOT-Seq-RNA: predicting protein-RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction. *Methods Mol. Biol.*, **1137**, 119–130.
25. Zhao,H., Yang,Y. and Zhou,Y. (2011) Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic Acids Res.*, **39**, 3017–3025.
26. Kumar,M., Gromiha,M.M. and Raghava,G.P. (2007) Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics*, **8**, 463–472.
27. Kumar,M., Gromiha,M.M. and Raghava,G.P. (2011) SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J. Mol. Recognit.*, **24**, 303–313.
28. Shao,X., Tian,Y., Wu,L., Wang,Y., Jing,L. and Deng,N. (2009) Predicting DNA- and RNA-binding proteins from sequences with kernel methods. *J. Theor. Biol.*, **258**, 289–293.
29. Gao,M. and Skolnick,J. (2009) A threading-based method for the prediction of DNA-binding proteins with application to the human genome. *PLoS Comput. Biol.*, **5**, e1000567.
30. Zhao,H., Yang,Y. and Zhou,Y. (2010) Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function. *Bioinformatics*, **26**, 1857–1863.
31. Zhao,H., Yang,Y., Janga,S.C., Kao,C.C. and Zhou,Y. (2014) Prediction and validation of the unexplored RNA-binding protein atlas of the human proteome. *Proteins*, **82**, 640–647.
32. Shazman,S., Celniker,G., Haber,O., Glaser,F. and Mandel-Gutfreund,Y. (2007) Patch Finder Plus (PFplus): a web server for extracting and displaying positive electrostatic patches on protein surfaces. *Nucleic Acids Res.*, **35**, W526–W530.
33. Tworowski,D.I., Feldman,A.V. and Safro,M.G. (2005) Electrostatic potential of aminoacyl-tRNA synthetase navigates tRNA on its pathway to the binding site. *J. Mol. Biol.*, **350**, 866–882.
34. 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.
35. Baker,N.A., Sept,D., Joseph,S., Holst,M.J. and McCammon,J.A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 10037–10041.
36. Dolinsky,T.J., Czodrowski,P., Li,H., Nielsen,J.E., Jensen,J.H., Klebe,G. and Baker,N.A. (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.*, **35**, W522–W525.
37. Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.
38. Sikic,K. and Carugo,O. (2010) Protein sequence redundancy reduction: comparison of various method. *Bioinformation*, **5**, 234–239.
39. Smith,D.W., Han,M.R., Park,J.S., Kim,K.R., Yeom,T., Lee,J.Y., Kim do,J., Bingman,C.A., Kim,H.J., Jo,K. *et al.* (2013) Crystal structure of the protein from Arabidopsis thaliana gene At5g06450, a putative DnaQ-like exonuclease domain-containing protein with homohexameric assembly. *Proteins*, **81**, 1669–1675.
40. Kustatscher,G., Wills,K.L., Furlan,C. and Rappsilber,J. (2014) Chromatin enrichment for proteomics. *Nat. Protoc.*, **9**, 2090–2099.
41. Yang,J.y.n.e.c. and Zhang,Y. (2015) I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.*, **43**, W174–W181.
42. Hudson,W.H. and Ortlund,E.A. (2014) The structure, function and evolution of proteins that bind DNA and RNA. *Nat. Rev. Mol. Cell Biol.*, **15**, 749–760.
43. Shazman,S., Elber,G. and Mandel-Gutfreund,Y. (2011) From face to interface recognition: a differential geometric approach to distinguish DNA from RNA binding surfaces. *Nucleic Acids Res.*, **39**, 7390–7399.
44. Cuff,A.L., Sillitoe,I., Lewis,T., Redfern,O.C., Garratt,R., Thornton,J. and Orengo,C.A. (2009) The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.*, **37**, D310–D314.
45. Miao,Z.F. and Westhof,E. (2015) Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. *Nucleic Acids Res.*, **43**, 5340–5351.
46. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
47. Wada,K., Sato,M., Araki,N., Kumeta,M., Hirai,Y., Takeyasu,K., Furukawa,K. and Horigome,T. (2014) Dynamics of WD-repeat containing proteins in SSU processome components. *Biochem. Cell Biol.*, **92**, 191–199.
48. Dror,I., Shazman,S., Mukherjee,S., Zhang,Y., Glaser,F. and Mandel-Gutfreund,Y. (2012) Predicting nucleic acid binding interfaces from structural models of proteins. *Proteins*, **80**, 482–489.
49. Cazals,F., Proust,F., Bahadur,R.P. and Janin,J. (2006) Revisiting the Voronoi description of protein-protein interfaces. *Protein Sci.*, **15**, 2082–2092.