PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Predicting nucleic acid binding interfaces from structural models of proteins

Iris Dror,[1] Shula Shazman,[1] Srayanta Mukherjee,[2] Yang Zhang,[2] Fabian Glaser,[3] and Yael Mandel-Gutfreund[1]*

[1] Faculty of Biology, Technion – Israel Institute of Technology, Haifa, Israel 32000

[2] Center for Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, Michigan 48109

[3] Bioinformatics Knowledge Unit, Technion – Israel Institute of Technology, Haifa, Israel 32000

ABSTRACT

The function of DNA- and RNA-binding proteins can be inferred from the characterization and accurate prediction of their binding interfaces. However, the main pitfall of various structure-based methods for predicting nucleic acid binding function is that they are all limited to a relatively small number of proteins for which high-resolution three-dimensional structures are available. In this study, we developed a pipeline for extracting functional electrostatic patches from surfaces of protein structural models, obtained using the I-TASSER protein structure predictor. The largest positive patches are extracted from the protein surface using the patchfinder algorithm. We show that functional electrostatic patches extracted from an ensemble of structural models highly overlap the patches extracted from high-resolution structures. Furthermore, by testing our pipeline on a set of 55 known nucleic acid binding proteins for which I-TASSER produces high-quality models, we show that the method accurately identifies the nucleic acids binding interface on structural models of proteins. Employing a combined patch approach we show that patches extracted from an ensemble of models better predicts the real nucleic acid binding interfaces compared with patches extracted from independent models. Overall, these results suggest that combining information from a collection of low-resolution structural models could be a valuable approach for functional annotation. We suggest that our method will be further applicable for predicting other functional surfaces of proteins with unknown structure.

Proteins 2011; 00:000–000.

Key words: electrostatic patches; structural models; protein surface; function prediction; nucleic acid binding.

INTRODUCTION

Sequencing the genomes of numerous organisms has changed the face of biology. Nevertheless, the knowledge of linear sequences of the genes and proteins themselves can only partially explain the functions of the proteins in the cell. Further insight could come from information about the three-dimensional (3D) structure of proteins and its interactions with other molecules in different cells or developmental stages. Since the protein structure could directly reveal the mechanistic determinants of its function, the availability of structural information about a given protein is generally believed to contribute towards predicting its function. However, the main pitfall of various structure-based function prediction methods is that they are all limited to a relatively small number of proteins for which high-resolution 3D structures are available. Thus, a predictive method that could overcome this limitation would be of great value.

In the past decade, several methods for *de novo* prediction of protein structures from sequence have been developed (reviewed in Ref. [1]). For example, UNRES is a physics-based folding algorithm in which conformation space is searched by global optimization.[2] Rosetta, developed by Baker and coworkers,[3] assembles protein structures using small fragments (3- and 9-mers) from the PDB structures. Using a different approach, Skolnick and coworkers developed the TOUCHSTONE II, which constructs protein structures guided by spatial restraints extracted from nonhomologous templates with the conformational space searched on a lattice-based modeling system.[4] In I-TASSER, developed by Zhang and coworkers[5,6] we first identify nonhomologous templates using multiple threading algorithms.[7] The continuous fragments are then excised from the threading alignment which is

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: Israeli Science Foundation, ISF; Grant number: 1297/09; Grant sponsor: NSF Career Award; Grant number: DBI 0746198; Grant sponsor: National Institute of General Medical Sciences; Grant number: R01GM083107.
Iris Dror and Shula Shazman contributed equally to this work.
*Correspondence to: Yael Mandel-Gutfreund, Faculty of Biology, Technion – Israel Institute of Technology, Haifa, Israel 32000. E-mail: yaelmg@tx.technion.ac.il
Received 12 July 2011; Revised 27 September 2011; Accepted 30 September 2011
Published online 12 October 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.23214

then used to reassemble the full-length models by iterative Monte Carlo simulations. Because of the combination of threading and *ab inito* folding methods, I-TASSER has an advantage in automated modeling of both template-based and free-modeling targets. As tested comprehensively in Helles *et al.*,[1] among 18 algorithms for protein structure predictions, I-TASSER was found to have the best performance in terms of both speed and accuracy. I-TASSER was also ranked as the best performing structure modeling server at several CASP meetings (CASP7, CASP8, and CASP9).[6]

DNA- and RNA-binding proteins play a central role in all stages of the gene expression pathway from transcription to translation.[8,9] On the basis of the classical assumption that structure infers function, homology modeling and threading approaches have been shown to be very effective in classifying DNA- and RNA-binding proteins,[10,11] as well as in predicting the nucleic acid binding sites (e.g., Refs. [12] and [13]). However, while homology-based approaches are very valuable for function prediction, they are limited to cases for which a structural homologue is available.[14–16] In the past decade, several attempts have been made to predict DNA- and RNA-binding function, exploiting various structural and electrostatic features, as for example.[17–20] In addition, propensity-based approaches have been successfully applied to predict RNA-binding interfaces, for example Ref. 21. We have recently developed a Nucleic Acid (NA)-binding predictor that focuses on the properties of electrostatic patches on the protein surfaces to distinguish NA-binding proteins from non-nucleic acid binding proteins. The concept behind the NA-binding predictor is that the large electrostatic patches on these proteins have unique features that distinguish them from other proteins having similar electrostatic properties. We have shown that large patches of positive charges extracted from high-resolution structures of proteins-NA complexes highly overlap with the actual nucleic acid binding interface.[22–24] Based on the high overlap between the largest positive patch and the real binding interface using a novel differential geometric approach we have recently succeeded, for the first time, to accurately distinguish DNA from RNA-binding interfaces in experimentally solved protein structures.[25]

Here, we examine the competence of the patchfinder algorithm to predict NA-binding interfaces from structural models generated by I-TASSER. As a first step towards our goal, we looked at whether a predicted protein structure (a model) could be used to define correctly the largest positive patch on the protein model surface. We found a high degree of overlap between the largest positive patch extracted from NA-binding proteins' high-resolution 3D structure and their structural models. Furthermore, we show that the patchfinder algorithm can predict the NA-protein binding interface from structural models of proteins with relatively high accuracy. Overall, this study presents a novel approach for correctly identifying NA-binding interfaces given the protein sequence alone. We propose that the method can be further applicable for predicting other functional interfaces in a genomic scale.

MATERIALS AND METHODS

Dataset construction

A nonredundant set of 74 NA-binding protein structures, sharing less than 25% sequence identify, (Table S1) was selected from the Protein Data Bank (PDB) comprised of 35 RNA-binding proteins (RBP) and 39 DNA-binding proteins (DBP). For each protein, we ran the I-TASSER structure predictor, deliberately ignoring the information from the known structure. Consequently, 55 of the 74 proteins (33 DBP and 23 RBP, see Table S1) for which five models were available from I-TASSER (see below) were used for further investigation. Detailed information on each of the models generated by I-TASSER, including the C-scores of the five best models as well as the sequence identity and coverage of the top 10 templates used to build the models are available for each PDB chain given in Table S1 via the link http://zhanglab.ccmb.med.umich.edu/RNA_project/ defining the PDB code and chain number, for example http://zhanglab.ccmb.med.umich.edu/RNA_project/1jidA/.

Creating models using I-TASSER

The I-TASSER prediction pipeline includes four general steps: template identification; structure reassembly; atomic model construction; and final model selection. Initially, the query sequence is threaded through a PDB library to identify appropriate local fragments that are then adopted for further structural reassembly. Subsequently, continuous fragments are used to assemble full-length models with unaligned loop regions built by *ab initio* modeling. The structure assembly simulations (for both sets) are guided by a unified knowledge-based force field.[26] The cluster centroids from I-TASSER are reduced models, with each residue represented by its Cα and side-chain center only. The full-atomic models are built by optimizing the H-bond networks, and the highest scoring models are finally clustered and selected. I-TASSER can produce between one and five different models per sequence, ranked by a confidence score that is defined based on the quality of the threading alignments and the convergence of its structural assembly simulation.[6] In addition a maximum homology between the query and templates can be defined based on sequence identity cutoff, that is, the number of identical residues between template and query divided by the total number of residues in the query sequence. In this work, we ran our database several times through the I-TASSER engine with different levels of homology with six different cutoffs, including in each set all templates that had lower sequence identity than the defined cutoff.

Constructing patches and interfaces

The patchfinder algorithm was employed to extract the largest positive patches on the protein's surfaces. The patchfinder algorithm implemented in a new improved version of the PFPlus web server (version 2.0) http://pfp.technion.ac.il/ uses the Poisson-Boltzmann equation in order to calculate the electrostatic potential of a protein and then to construct the largest continuous positive patch on the protein surface.[22,24] Here, we used a local PFPlus version. Interface residues were calculated using Intervor http://cgal.inria.fr/abs/Intervor/, which detects interface atoms using the Voronoi cells approach.[27]

Statistical analysis

To assess the ability of correctly predicting amino acids (AA) in a patch, we computed the following parameters: sensitivity, representing the proportion of the residues in the actual patch that were predicted correctly; Positive Predicted Value (PPV), representing the proportion of amino acids correctly predicted to be in the patch relative to the real patch; and the Matthew's Correlation Coefficient (MCC), which combines both specificity and sensitivity using the following equations:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{MCC} = \frac{(\text{TP} * \text{TN}) - (\text{FP} * \text{FN})}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FN})(\text{TN} + \text{FP})}}$$

where TP (true positives) is the number of AAs predicted correctly to be included in a positive electrostatic patch (either the patch extracted from the experimentally solved structure "real patch" or the interface as defined from the protein-NA complex), FP (false positives) is the number of AAs falsely predicted to be part of the patch/interface, TN (true negatives), is the number of AAs not in the patch/interface that are predicted correctly, and FN (false negatives) is the number of AAs in the patch/interface that were mispredicted and were not included in the patch. PPV and sensitivity values range from 0 (all wrong predictions) to 1 (perfect prediction), while the values of MCC range from 1 (perfect prediction) to −1 (all wrong predictions).

RESULTS AND DISCUSSION

Extracting positive patches on proteins surfaces based on structural models

To enable us to compare the results produced by patchfinder on structural models to those obtained for the solved structures, we first selected a nonredundant set of NA-binding proteins for which 3D structure has been solved experimentally in the *holo* state and have no more than 25% sequence identity between them (see Materials and Methods). For each protein in the dataset, we generated 3D models using the I-TASSER predictor, deliberately ignoring the information from the known structure and their close homologues (>95% sequence identity). Further, we computed the largest positive patch for each I-TASSER model and compared it to the patch calculated for the experimentally solved protein structure and to the NA-protein binding interface extracted from the NA-protein complex. Consequently, we checked the overlap between the "model patch" (calculated by patchfinder applied to five different models generated by I-TASSER) and the "real patch" (calculated by patchfinder applied to the experimentally solved protein structure available in PDB). Overall, we obtained a relatively high overlap between the "model patch" and the "real patch" with sensitivity and PPV ranging from 0.64/0.60 to 0.68/0.66, respectively. The median MCC value for all models was 0.3 (see Supporting Information Fig. S1). Interestingly, although the I-TASSER method implies that the quality of the models declines from the first to the fifth model, we did not notice a significant decline in the overlap between the "model patch" and the "real patch" when calculating the patch from lower ranked I-TASSER models (Fig. S1). To further explore whether the quality of the models influences patch predictions, we calculated the overlap between the "model patch" versus the "real patch" when considering independently models with low (<0.5) and high (>0.5) TM-scores (a score used to estimate the structural similarity between the models and the native structure[28]). As illustrated in Figure S2, while we did notice a higher overlap between the "model patch" and the "real patch" among the models with higher TM-scores (considered to share a similar fold with the templates), still the differences were not remarkable (P-value=0.002 applying Wilcoxon ranked test on the MCC). Notably, models with TM-score <0.17 were removed from this study as in this range the models are considered to have random folds. Furthermore, we detected a weak significant difference in the MCC (P-value = 0.01 applying Wilcoxon ranked test) when comparing models with relatively high RMSD (>2.5 Å) to the more accurate models with lower RMSD (<2.5 Å) (Fig. S3). Overall, the lack of strong correlation, observed here, between the patch prediction and the model quality reinforces our previous observations that while the electrostatic patch prediction is highly sensitive to details of the structure (resolution, rotamers, etc.),[24] a rough but still valuable patch prediction could be obtained from relatively low-resolution models. Based on these results we were encouraged to test if the patchfinder algorithm could be boosted by using information from multiple models.

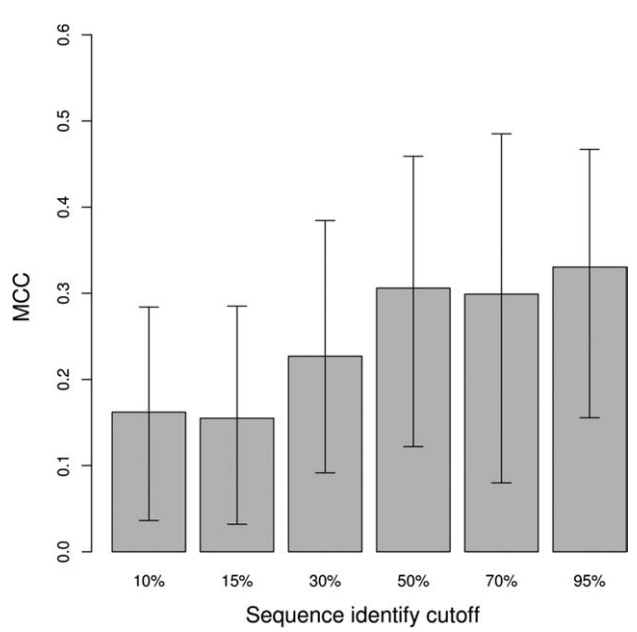

Figure 1

Comparison between the "model patch" and the "real patch". The histogram represents the mean MCC between the largest positive patch calculated using each I-TASSER model and the solved structure. Standard deviations are shown.

Different from other standard structure-prediction comparative modeling algorithms, I-TASSER creates a 3D model based on structural fragments from different templates.[6] Generally, I-TASSER will choose the best available fragment to build a model based on the sequence identity between query and templates. Nevertheless, I-TASSER allows limiting the quality of templates chosen from the template library. To examine the overlap between the patches of the models and solved structures as a function of model quality, we tested six different groups of models differing in the cutoff defined for the best sequence identity of the model to the template (10, 15, 30, 50, 70, and 95% sequence identity). For each group, we compared the overlap between the "model patch" and the "real patch". As shown in Figure 1, the MCC dropped significantly below the 30% sequence identify cutoff (0.16 in 10%), suggesting that prediction quality decreases when the templates available are not evolutionary related. However, MCC values did not vary significantly between the 95, 75, and 50% cutoffs. Taken together, these results suggest that when a close homologue (>30% sequence identity) is available, the patchfinder is not highly sensitive to the model's accuracy. However, for low-quality models patchfinder cannot be used as an accurate predictor of the electrostatic patch. On the basis of these results we decided to select the default option of 95% cutoff for further investigation. It is important to emphasize that the models further used in the study were of variable qualities, ranging from 21% to 95% sequence identify to the template with variable coverage lengths (Table S1). Detailed information on each of the individual models is available at http://zhanglab.ccmb.med. umich.edu/RNA_project/ (see example in Materials and Methods).

Combined information from several models contributes to patch sensitivity

As demonstrated in Figure S4, the I-TASSER models calculated for a given protein can differ considerably from one another resulting in different electrostatic patches. We speculated that if an AA is predicted to be included in the largest positive patch in several different models, it is more likely to be part of the "real patch" (i.e., the patch predicted on the solved structure). To investigate this, we defined a new term, we named: "combined patch," which includes only AAs that were predicted to be included in the "model patch" in at least x models (x ranging from 1 to 5). Further, we compared the "combined patch" with the "real patch" and evaluated the overall sensitivity and specificity (PPV) of the method to detect the real patch (Fig. 2). Notably, the analysis was performed only for the 55 proteins for which five I-TASSER models were available. As illustrated in Figure 2(A), the positive predicted value between the "real patch" and the "model patch" increased as the "combined patch" was constructed from a larger number of models. On the other hand, as expected, the sensitivity of the "combined patch" improved when we were less stringent (e.g., the patch was constructed from AAs found in the patch in at least one model) [see Fig. 2(B)]. Overall, as demonstrated in Figure 2(C), the MCC, which takes into account both sensitivity and specificity, reached its peak (median = 0.38) when including at least three models in the "combined patch" and declined gradually when requiring that the AAs will be present in all four or five "model patches".

On the basis of the latter results, two possible approaches for constructing the largest positive patch from multiple structural models could be considered. The more conservative approach would be to construct a patch using AAs that were found in all five "model patches." In this case, we expect that most of the AAs selected will also be found in the "real patch." However, one would miss a large fraction of the AAs which should be in the patch (i.e., higher specificity but lower sensitivity). The second, more permissive approach would be to construct the largest positive patch using AAs that were found in at least one patch. In this case, we would improve sensitivity at the expense of specificity (PPV). In order to account for both sensitivity and specificity, we suggest defining an AA in the "model patch" if it was found in the patch of at least three of the five best models. The example shown in Figure 3 emphasizes the strength of the latter approach. In this example we show the electrostatic patch calculated on the solved structure of the Ebola virus matrix protein vp40 n-terminal do-

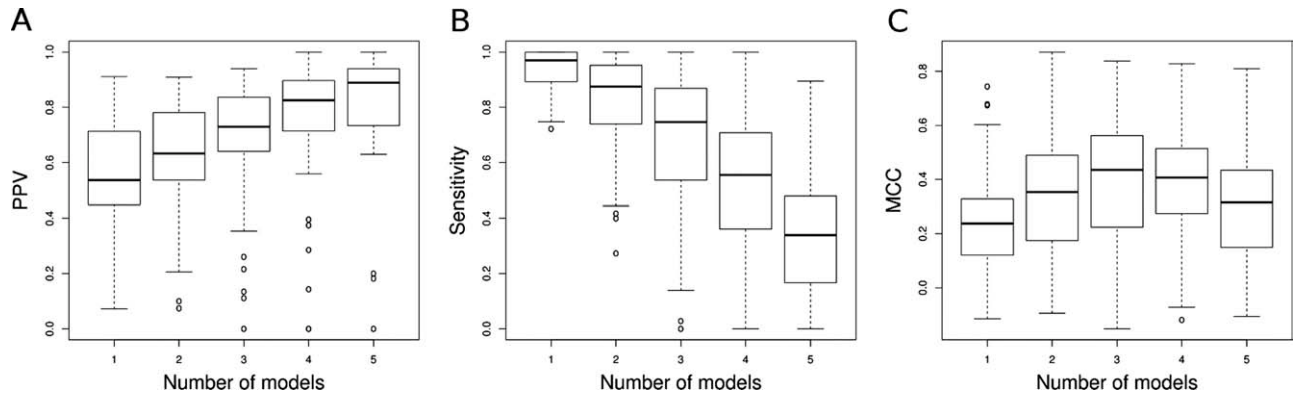

Figure 2

Comparing the different "combined patch" approaches. The box plots illustrate the (A) PPV (B) sensitivity and (C) MCC (see Material and Methods) between five different "combined patches" and the "real patch". Each bar represents results obtained when considering a different number of models required to define the AAs included in the patch (see text for details).

main (PDB 1h2c) compared with different predicted patches constructed using different stringencies. As shown in Figure 3(B), when including AAs in the "model patch" that were found in all five models, we obtained a small defined patch with very high specificity but low sensitivity. Very interesting, in this example the small patch based on all five models [Fig. 3(B)] highly overlaps with the real RNA-binding interface [shown in orange in Fig. 3(C)].

Nucleic acid binding interfaces can be predicted based on the "model patch"

Previous studies have shown that in general, the largest positive patch calculated on the protein structure highly overlaps with the nucleic acid binding interface.[22–24] Clearly, the ultimate goal would be to use the patch extracted from the model as a predictor of the nucleic acid binding interfaces. As shown in the example given in Figure 3, when including in the patch AAs found in multiple "model patches," we increased the positive prediction value of the "combined patch". To test whether multiple models could be useful for defining nucleic acid binding interfaces, we calculated the overlap between the patch calculated from I-TASSER models and the nucleic acid binding interface.

As shown in Figure 4(B), when including an increasing number of models to construct the patch, we noticed a decrease in sensitivity values compared to the real binding interface. However, the PPV values, which capture the specificity of the prediction, clearly improved. As shown, the PPV obtained from the "combined patch" (including all five models) was notably higher than the PPV obtained when comparing the "real patch" to the "real interface." Additionally, the MCC values calculated for patches constructed from the structural models

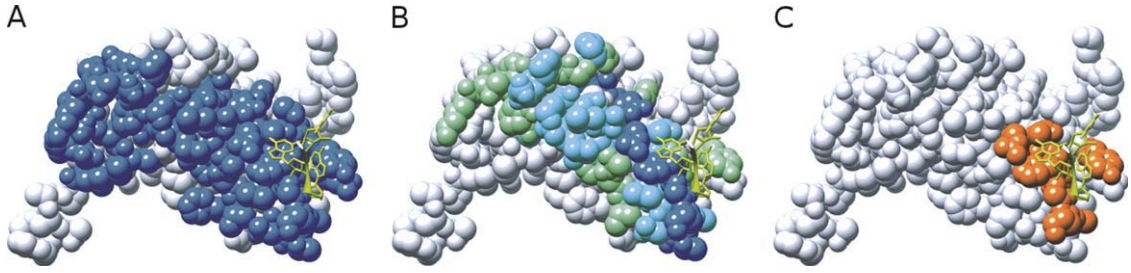

Figure 3

Illustrating the combined patch approach on the Ebola virus matrix protein vp40 n-terminal domain in complex with RNA (PDB 1h2c). The protein is shown in CPK representation and the RNA in yellow sticks. **A**: "Real patch" produced by patchfinder on the PDB structure (AAs belonging to the patch are colored blue). **B**: Three different "combined model" patches calculated based on a different number of I-TASSER models (dark blue, AAs that were found in five patches; cyan, AAs that were found in four patches; and light green, AAs that were found in three patches). **C**: The calculated protein-nucleic acid interface extracted from the PDB using the intervor algorithm (colored orange).

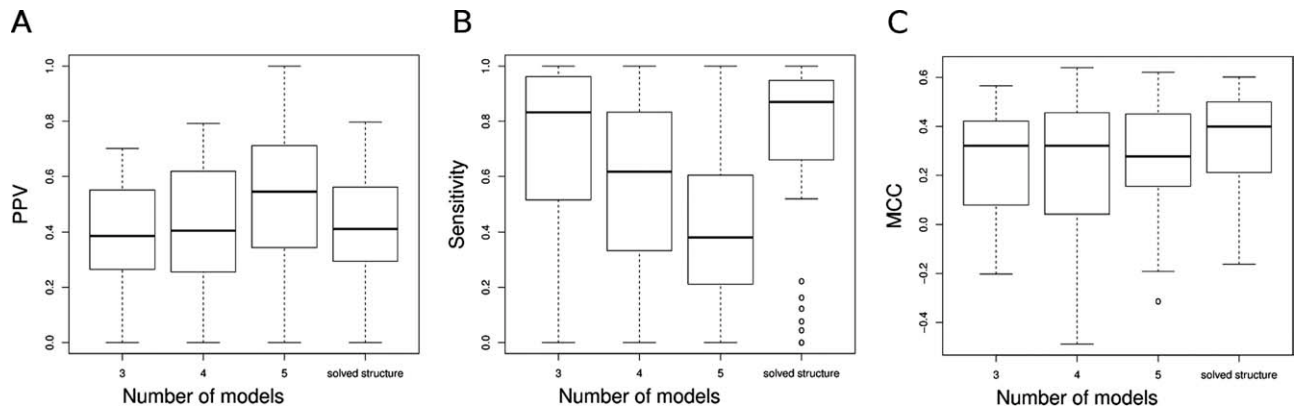

Figure 4

The overlap between the patch and the NA-binding interface. (**A**) PPV (**B**) sensitivity, and (**C**) MCC evaluating the overlap between the largest positive patch, calculated from the models or the solved structure, and the protein-nucleic acid interface. Each bar represents a different number of models required to define the AAs included in the patch (see text for details).

(median MCC of 0.32, 0.32, and 0.28, when combining three, four, and five models, respectively) were only slightly lower than the MCC calculated for the "real patch" and the NA-binding interface (MCC = 0.39). Overall, these results show that the "combined patch" is a good indicator of NA-binding interface. Very interestingly, the sensitivity of the "combined patches," including AAs that were found in at least three models, in identifying the real binding interface was similar to the overall sensitivity of the "real patch". These results suggest that including multiple models could be useful in defining the nucleic acid binding region even in cases in which the crystal structure is known.

To further investigate the contribution of the multimodel approach to identify NA-binding interfaces from structural models, we examined the overlap between the model's patch and real interface in each of the proteins for which I-TASSER produced five models. The heat maps in Figure 5 show the PPV [Fig. 5(A)] and sensitivity [Fig. 5(B)] calculated for 43 out of the 55 proteins included in our dataset. The 12 proteins not shown in the heatmap are proteins for which we could not extract a "combined patch" due to the lack of overlap between the patches in the different models. From the results presented in Figure 5, it is evident that PPV generally improves when including in the "combined patch" AAs that appear in a higher number of models, while (as expected) sensitivity has the opposite trend. Overall when employing the "combined approach", including AAs that were found in the positive patch of at least three models, the average overlap with the real binding interface was 70% (Table S2) which is slightly lower than the average overlap between the patch of the solved structure and the real interface (74%) and comparable to our earlier findings on RNA and DNA-binding unique datasets (68% and 80% for RNA and DNA, respectively).[22–24] As shown in Figure 5, in 14/43 cases (excluding the 12 proteins for which the combined patch approach was not available) the results of the "combined 3 patch" were improved relative to the "real patch". Interestingly, in 13/43 cases the sensitivity improved as more models were used to define the patch and in five of these cases both the PPV and sensitivity improved. An intriguing example is the human Pumilio 1 (PUM1 1m8wA) homology domain (P53r in Table S1). As shown in Figure S5, the largest electrostatic patch calculated by patchfinder (Fig. S5(A), cyan) completely misses the real binding interface (Fig. S5(B), red), while the patch that was calculated when including information from 1 model only (Fig. S5(C), blue) or combining 3, 4, and 5 models (Fig. S5(D) S5(E) S5(F) S5(B-D), respectively) overlapped with the real binding interface (see Fig. 5S(B-D)). The latter case is a unique case where patchfinder fails to predict the real binding interface in the bound crystal structure of the PUM1 domain due to a segmentation of the large positive patch on the protein surface in the bound state of the protein, while in each of the best 5 models the patch was predicted correctly. Nevertheless, when considering both the sensitivity and specificity the combined patch approach usually improves the prediction of the real interface compared to the one model approach. Clearly, the most accurate interface prediction would be from the experimentally solved structure; however, currently for the majority of the proteins this information is not available.

CONCLUSIONS

In this paper, we combined an electrostatics-based approach with structure-based modeling to extract the largest positive patches on proteins. As shown, the largest positive patches on protein surfaces predicted from structural

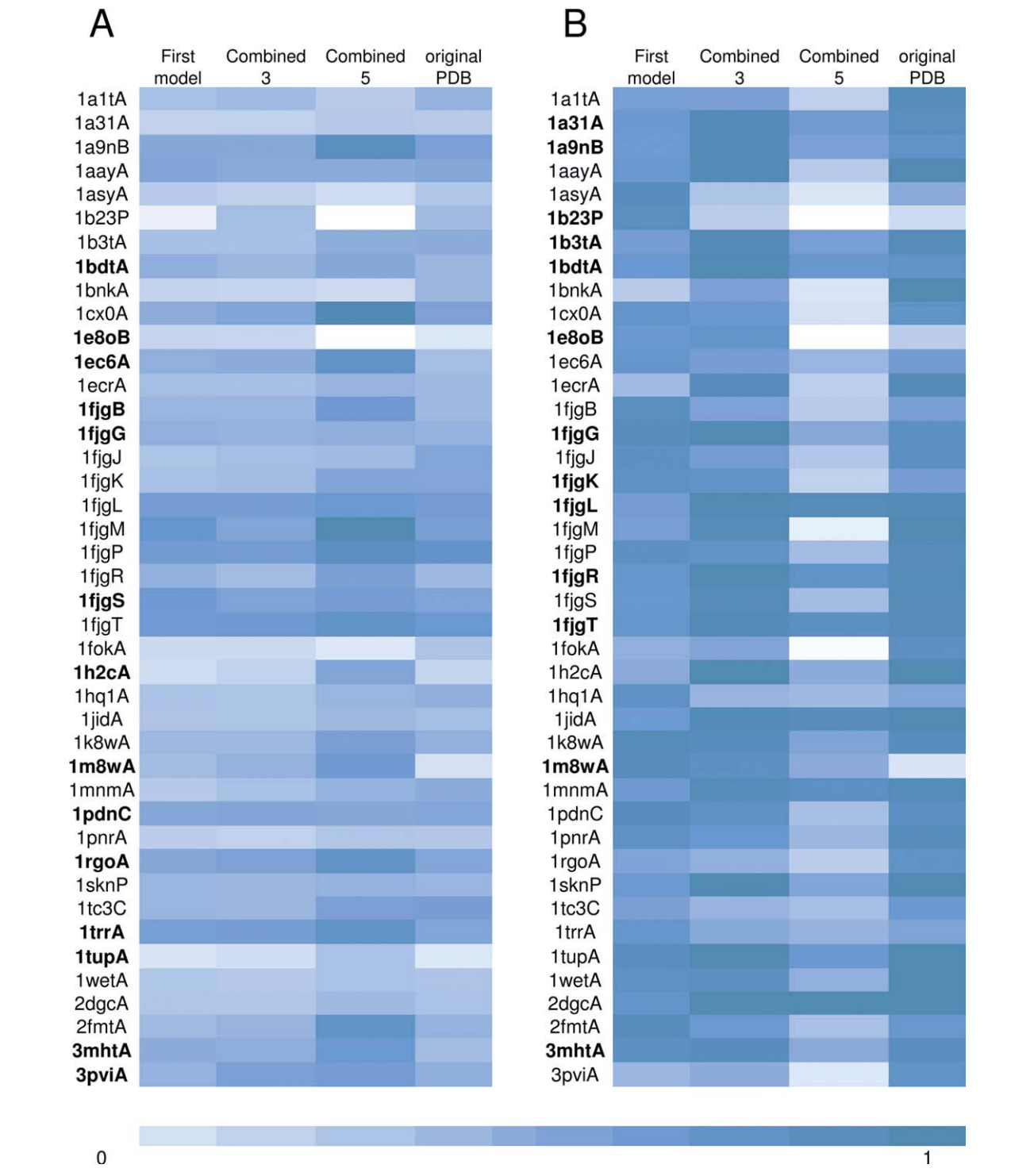

Figure 5

The overlap between the "combined patch" and the NA-binding interface for individual proteins. Heat maps demonstrate the change in PPV (**A**) and sensitivity (**B**) calculated for the overlap between the patch and the NA-binding interface calculated from one I-TASSER model, "combined 3 patch", "combined 5 patch" and solved structure (PDB codes are given in Table S1). Color intensity correlates with PPV and sensitivity values, ranging from 0 to 1. In bold are PDBs which their "combined 3 patch" was in better agreement with the NA-binding interface when compared with the patch calculated from the solved structure.

models obtained with I-TASSER [6] usually overlap the largest positive patch predicted from the experimentally solved structure. Since the largest positive patch on DNA- and RNA-binding proteins has been shown to correlate strongly with the DNA- and RNA-binding interfaces of these proteins, respectively,[22–24] this implies that this approach could be useful for predicting nucleic acid binding interfaces from sequences. Here, we show that the electrostatic patches extracted from an ensemble of structural models were compatible with the real binding interfaces of the NA-binding proteins tested in this study. Interestingly, we observed that for several proteins, the combined positive patch was in better agreement with the real binding interface compared with the patch extracted from the solved structure. Overall, this study proposes a new approach for predicting nucleic acid binding interfaces from protein sequence alone. Given the recent advancements in successfully distinguishing between DNA- and RNA-binding proteins based on the differential geometric properties of the largest electrostatic patches,[25] we postulate that in the future it will be possible to uniquely predict DNA- and RNA-binding proteins from sequence.

REFERENCES

1. Helles G. A comparative study of the reported performance of ab initio protein structure prediction algorithms. J R Soc Interface 2008;5:387–396.
2. Liwo A, Lee J, Ripoll DR, Pillardy J, Scheraga HA. Protein structure prediction by global optimization of a potential energy function. Proc Natl Acad Sci USA 1999;96:5482–5485.
3. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
4. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J 2003;85:1145–1164.
5. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 2007;5:17.
6. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 2010;5:725–738.
7. Wu ST, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. Nucl Acids Res 2007;35:3375–3382.
8. Kanitz A, Gerber AP. Circuitry of mRNA regulation. Wiley Interdiscip Rev Syst Biol Med 2010;2:245–251.
9. Charoensawan V, Wilson D, Teichmann SA. Genomic repertoires of DNA-binding transcription factors across the tree of life. Nucleic Acids Res 2010;38:7364–7377.
10. Gao M, Skolnick J. A threading-based method for the prediction of DNA-binding proteins with application to the human genome. PLoS Comput Biol 2009;5:e1000567.
11. 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;26:1857–1863.
12. Gao M, Skolnick J. DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. Nucleic Acids Res 2008;36:3978–3992.
13. Zhao H, Yang Y, Zhou Y. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. Nucleic Acids Res 2011;39:3017–3025.
14. Kinoshita K, Nakamura H. Protein informatics towards function identification. Curr Opin Struct Biol 2003;13:396–400.
15. Rigden DJ. Understanding the cell in terms of structure and function: insights from structural genomics. Curr Opin Biotechnol 2006;17:457–464.
16. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. From structure to function: approaches and limitations. Nat Struct Biol 2000;7 Suppl:991–994.
17. Ahmad S, Sarai A. Moment-based prediction of DNA-binding proteins. J Mol Biol 2004;341:65–71.
18. Nimrod G, Schushan M, Szilagyi A, Leslie C, Ben-Tal N. iDBPs: a web server for the identification of DNA binding proteins. Bioinformatics 2010;26:692–693.
19. Nimrod G, Szilagyi A, Leslie C, Ben-Tal N. Identification of DNA-binding proteins using structural, electrostatic and evolutionary features. J Mol Biol 2009;387:1040–1053.
20. Szilagyi A, Skolnick J. Efficient prediction of nucleic acid binding function from low-resolution protein structures. J Mol Biol 2006;358:922–933.
21. Perez-Cano L, Fernandez-Recio J. Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. Proteins 2010;78:25–35.
22. Stawiski EW, Gregoret LM, Mandel-Gutfreund Y. Annotating nucleic acid-binding function based on protein structure. J Mol Biol 2003;326:1065–1079.
23. Shazman S, Mandel-Gutfreund Y. Classifying RNA-binding proteins based on electrostatic properties. PLoS Comput Biol 2008;4:e1000146.
24. Shazman S, Celniker G, Haber O, Glaser F, Mandel-Gutfreund Y. Patch Finder Plus (PFplus): a web server for extracting and displaying positive electrostatic patches on protein surfaces. Nucleic Acids Res 2007;35(Web Server issue):W526–W530.
25. Shazman S, Elber G, Mandel-Gutfreund Y. From face to interface recognition: a differential geometric approach to distinguish DNA from RNA binding surfaces. Nucleic Acids Res 2011;39:7390–7399.
26. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 2008;9:40.
27. Cazals F, Proust F, Bahadur RP, Janin J. Revisiting the Voronoi description of protein-protein interfaces. Protein Sci 2006;15:2082–2092.
28. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.