A holistic in silico approach to predict functional sites in protein structures

Segura, Joan; Jones, Pamela F; Fernandez-Fuentes, Narcis

Published in:
Bioinformatics

DOI:
10.1093/bioinformatics/bts269

Publication date:
2012

Citation for published version (APA):
Segura, J., Jones, P. F., & Fernandez-Fuentes, N. (2012). A holistic in silico approach to predict functional sites in protein structures. *Bioinformatics*, *28*(14), 1845-1850. https://doi.org/10.1093/bioinformatics/bts269

BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 14 2012, pages 1845–1850
doi:10.1093/bioinformatics/bts269

Structural bioinformatics

Advance Access publication May 4, 2012

A holistic *in silico* approach to predict functional sites in protein structures

Joan Segura[1], Pamela F. Jones[2] and Narcis Fernandez-Fuentes[1,3,*]

[1]Leeds Institute of Molecular Medicine, Section of Experimental Therapeutics, [2]Leeds Institute of Molecular Medicine, Section of Molecular Gastroenterology, St. James's University Hospital, University of Leeds. Leeds, LS9 7TF and [3]Institute of Biological, Environmental and Rural Science, Aberystwyth University, Gogerddan Campus. Aberystwyth, SY23 3EB, UK

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Proteins execute and coordinate cellular functions by interacting with other biomolecules. Among these interactions, protein–protein (including peptide-mediated), protein–DNA and protein–RNA interactions cover a wide range of critical processes and cellular functions. The functional characterization of proteins requires the description and mapping of functional biomolecular interactions and the identification and characterization of functional sites is an important step towards this end.

Results: We have developed a novel computational method, Multi-VORFFIP (MV), a tool to predicts protein-, peptide-, DNA- and RNA-binding sites in proteins. MV utilizes a wide range of structural, evolutionary, experimental and energy-based information that is integrated into a common probabilistic framework by means of a Random Forest ensemble classifier. While remaining competitive when compared with current methods, MV is a centralized resource for the prediction of functional sites and is interfaced by a powerful web application tailored to facilitate the use of the method and analysis of predictions to non-expert end-users.

Availability: http://www.bioinsilico.org/MVORFFIP

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: naf4@aber.ac.uk; narcis.fernandez@gmail.com

Received on February 7, 2012; revised on April 25, 2012; accepted on April 30, 2012

1 INTRODUCTION

In order to fulfill their cellular functions, proteins interact with other proteins and biomolecules. Protein–protein interactions, including peptide-mediated interactions, are the basis of the formation of macromolecular complexes required to coordinate and perform complex cellular functions, as well as the regulation and coordination of signaling pathways (Petsalaki and Russell, 2008; von Mering *et al*., 2002). Protein–DNA interactions mediate a wide range of cellular functions including gene expression and regulation, DNA replication, DNA repair and DNA recombination (Jones *et al*., 1999; Luscombe *et al*., 2000). Likewise, protein–RNA interactions are central to a number of crucial cellular processes such as post-translational gene regulation, protein synthesis, alternative splicing and RNA processing and metabolism (Draper, 1995; Jones *et al*., 2001). Thus, deciphering and dissecting biomolecular interactions are central to fully understand the function of proteins and their role in cells. Furthermore, the prediction of functional sites can be used to improve the selection of structural models (Chelliah and Taylor, 2008).

Residues located in functional sites present a number of unique structural and physicochemical properties that vary across the different types of interfaces; e.g. protein-binding interfaces are different from DNA-binding ones (Glaser *et al*., 2001; Jones *et al*., 1999). These distinctive features are used by a number of computational tools to predict binding sites. A common denominator for most of recent computational methods is the use of machine-learning algorithms to combine heterogeneous information. This is mainly because simple or composite scoring functions are either not suitable or cannot be fully optimized due to the incomplete understanding of the biophysical events underpinning interactions between biomolecules. Thus, statistical models are better to combine and unify data of diverse nature and a number of methods have been proposed to predict protein- (Segura *et al*., 2011; Sikic *et al*., 2009), peptide- (Petsalaki *et al*., 2009), DNA- (Bhardwaj *et al*., 2005; Tjong and Zhou, 2007; Xiong *et al*., 2011), RNA-binding site (Cheng *et al*., 2008; Liu *et al*., 2010; Maetschke and Yuan, 2009; Terribilini *et al*., 2007) and more generally functional sites (Bray *et al*., 2009; Innis, 2007; Pettit *et al*., 2007).

In this work, we present Multi-VORFFIP (MV), a structure-based, machine learning, computational method developed to predict four different types of interactions or functional sites in proteins: protein–protein, protein–peptide, protein–DNA and protein–RNA binding sites. MV integrates a wide range of structural, evolutionary, energy-based and experimental data (i.e. crystallographic B factors) into a common probabilistic framework. The different functional sites (e.g. peptide-binding sites) are predicted using statistical models developed and tailored to that end. The method compares favorably with recently described methods. Moreover, the mapping of functional sites (e.g. protein- and DNA-binding sites) within the same protein is highly accurate and selective. MV is accessible through a user-friendly web application available at http://www.bioinsilico.org/MVORFFIP. The web application features a powerful and convenient graphic interface that allows the visualization and analysis of the different predictions simultaneously.

*To whom correspondence should be addressed.

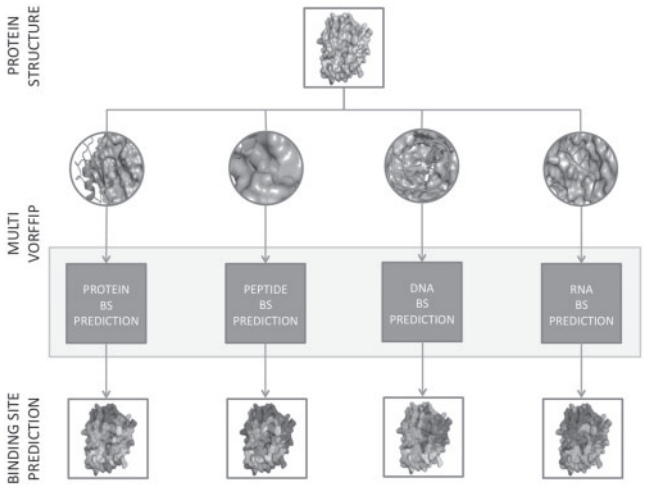

Fig. 1. Overall flowchart of the prediction process. The algorithm has been trained using four different types of interactions: protein–protein, peptide–protein, DNA–protein, RNA–protein interactions. BS: binding site

2 METHODS

2.1 Prediction algorithm: MV

The novel algorithm, MV, builds on our previous method, VORFFIP, developed to predict protein-binding sites in protein structures (Segura *et al*., 2011). Briefly, the method is composed of a two-step Random Forest (RF) ensemble classifier that integrates structural features, energy terms, evolutionary information, normalized crystallographic B factors, environmental-based metrics derived from Voronoi diagrams (VDs), and scores and score-derived metrics as described in the original publication (Segura *et al*., 2011); (see Supplementary Data for the complete list of input variables). Besides predicting protein-binding sites, the newest implementation of the method, MV, also includes three novel RFs, each of them trained to predict peptide-, DNA- and RNA-binding sites in protein structures (Fig. 1). Thus, all statistical models use the same input variables but were trained to distinguish different functional sites.

2.2 Datasets and benchmarking

Three different datasets, PEP-set, DNA-set and RNA-set, extracted from recent publications, were used to benchmark MV. Benchmark 4.0 dataset (Hwang *et al*., 2010), named PROT-set, was also used to assess the selectivity of the predictions. The PROT-set is a dataset of 176 protein–protein complexes specifically compiled for docking evaluation. No two single pairs of complexes belong to the same SCOP family. The PEP-set is a dataset of protein–peptides complexes compiled by Petsalaki *et al*. (2009) and it is composed of a non-redundant set [i.e. does not include protein–peptide complexes that belong to the same SCOP family (Murzin *et al*., 1995)] of 405 protein–peptides structure complexes solved both in bound and unbound conformation. The DNA-set is a dataset of protein–DNA complexes (Xiong *et al*., 2011) that consists of 206 protein–DNA complexes sharing <25% sequence identity and featuring both in unbound and bound conformations. The RNA-set is a dataset of protein–RNA complexes (Liu *et al*., 2010), comprising 205 protein–RNA complexes where RNA and protein sequences among the set share <60% and 25% sequence identity, respectively. Finally, a combined set, COMB-set, containing 17 proteins that have more than one functional site, e.g. a DNA- and a protein-binding site, was used to assess the selectivity of predictions. The list of PDB codes included in the COMB-set is given in the Supplementary Data.

The datasets have almost empty intersections, PROT-set shares one structure with DNA-set and two with PEP-set. Thus, the potential bias introduced by complexes that might score well in different type of binding site predictions is negligible. The benchmarking of MV, including the definition of interaction interfaces (i.e. binding sites), was performed following the same procedure described in each publication where the datasets were described. Thus, protein–peptide interfaces were defined as protein residues within a distance of 6 Å from the peptide (PEP-set); protein–DNA interfaces were defined by the protein residues with a relative surface accessibility area >10% and within 4.5 Å of the DNA (DNA-set); and RNA binding sites were defined using ENTANGLE (Allers and Shamoo, 2001) as in the original work (Liu *et al*., 2010) (RNA-set). In the case of the PROT-set, interfaces were defined using DIMPLOT (Wallace *et al*., 1995) as described in de Vries *et al*. (2006).

2.3 Assessing the predictive power of MV

A number of statistical measures were used to assess and compare the performance of MV to original sources. These included the F_1 score (1), the Mathew correlation coefficient (MCC) (2) and area under the receiver operating characteristic (ROC) curve (AUC).

$$F_1 = \frac{2TP}{2TP+FN+FP} \quad (1)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \quad (2)$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. Statistical measures were calculated using the prediction values of single residues. *P*-values to assess the significance of the differences observed in performance and distribution of scores were calculated using the Wilcoxon rank-sum test.

2.4 The server

A dedicated web server to interface MV was developed. The web application is composed of a collection of Perl packages, JavaScript and java applets [Jmol (http://www.jmol.org/)], a number of applications that include PSAIA (Mihel *et al*., 2008), FoldX (Guerois *et al*., 2002), al2co (Pei and Grishin, 2001), QHULL (Barber *et al*., 1996), Psi-Blast (Altschul *et al*., 1997), 3DCA (Landgraf *et al*., 2001) and the R package *randomForest* (Breiman, 1984; Liaw and Wiener, 2002) to compute RFs. Currently, the training sets used for predictions are the PROT-set, PEP-set, DNA-set and RNA-set described here. All the databases required during prediction, such as the NR database (Pruitt *et al*., 2007) used to compute sequence profiles, are updated on a weekly basis.

3 RESULTS

3.1 Predictive performance and competitiveness

MV has been compared with current state-of-the-art methods to assess its competitiveness in the prediction of peptide-, DNA- and RNA-binding sites. In all tests, MV was trained and tested under the same conditions as the compared method and interfaces were defined using the same criteria (see Section 2). This ensured that: (i) benchmarking and assessment of MV was performed against validated sets; and (ii) benchmarking results were comparable with those described in the original publications, thus providing a measure of competitiveness of MV with respect to current methodologies.

In general, the accuracy of the predictions increases as more information is included (Supplementary Tables S1 and S2), which is in agreement with observations in our previous work (Segura *et al*., 2011). Indeed, the performance of peptide-, DNA- and RNA-binding

site predictions in term of AUC, MCC, F_1-score, Precision (P) and recall (R) values (Supplementary Tables S1 and S2) improved as structure, energy, conservation and crystallographic B factors were added to the predictions. The only exception was *P*-values, which in the case of the PEP-set and RNA-set dropped slightly when all the features were combined although MCC values were higher, i.e. better R (Supplementary Table S2).

Although all predictors use the same input features, the relative importance of each variable varied across the different predictions. These differences are to be expected as physicochemical and structural properties vary across different binding sites (Glaser *et al*., 2001; Jones *et al*., 1999; London *et al*., 2010). For instance, sequence conservation was a powerful feature to discriminate DNA-binding sites as total accessible surface area (ASA) was in the case of RNA-binding site predictions. Backbone hydrogen bonding contributed highly in peptide-binding predictions and the average protrusion index (CX) was a valuable feature in RNA-binding prediction. Finally, other features were more equally distributed among predictors (Supplementary Fig. S1).

3.1.1 Protein–peptide binding site prediction PEP-set (Petsalaki *et al*., 2009) was used to assess the performance in the prediction of peptide-binding sites. According to original work, the optimal *P*-value cut-off in a one-leave-out cross-validation experiment was 0.04, representing an MCC value of 0.24 according to the reported false positive and true positive rates (Petsalaki *et al*., 2009). MV achieved a MCC value of 0.55 on a 5-fold cross-validation experiment that is more disadvantageous than the leave-one-out validation (as in the original publication) because the latter implies a larger training set and thus a better statistical model. However, MV predicts peptide-binding interfaces whereas the method of Petsalaki *et al*. takes into account the sequence of the peptide, i.e. predict the interface based on the sequence of the peptides, which is a more difficult prediction. Hence, the MCC values between MV and Petsalaki's method are not directly comparable and so MV was compared with a random predictor. Under this scenario, MV performed substantially better in both MCC (0.55 versus 0.00 – expected value in a random prediction) and AUC (0.86 vs. 0.50 – expected value in a random prediction).

3.1.2 Protein–DNA binding site prediction The performance of MV in protein–DNA binding site prediction was assessed using the DNA-set and performing benchmark tests as previously described (Xiong *et al*., 2011). The first test consisted of a 5-fold cross-validation using the entire DNA-set. In terms of F_1 scores and AUC values, MV (F_1: 0.49; AUC: 0.86) and the method described by Xiong *et al*. (F_1: 0.51; AUC: 0.82) performed at comparable levels and the differences in performance were not significant ($P > 0.01$). In a second test, the DNA-set was derived into two subsets as described in the original work (Xiong *et al*., 2011). One of the subsets was used as the training set while the other subset, including the bound and unbound conformations, was used as the test set. Again the performance of MV in terms of F_1 (bound: 0.50; unbound: 0.44) and AUC (bound: 0.85; unbound: 0.80) values were comparable with those reported in the original publication (F_1; bound: 0.51; unbound: 0.44; AUC: bound: 0.84; unbound: 0.78).

3.1.3 Protein–RNA binding site prediction The RNA-set was used to assess the performance of MV in protein–RNA binding site prediction. This set was recently derived to benchmark a RNA-binding site prediction method (Liu *et al*., 2010). The first test consisted of a 5-fold cross-validation on the entire RNA-set. In terms of F_1-scores and AUC values, MV (F_1-score: 0.80; AUC: 0.88) slightly underperformed in comparison with the method of Zhi-Ping *et al*. (F_1-score: 0.85; AUC: 0.92) although the differences were marginal and not significant ($P > 0.01$). The second test consisted on the prediction of RNA-binding sites on a randomly chosen independent set of 100 complexes. In order to compare the performance of MV under the same conditions, the same 100 complexes were selected. In this comparison, the original method of Zhi-Ping (F_1-score: 0.79; MCC: 0.49) performed marginally better than MV (F_1-score: 0.79; MCC: 0.43) although again differences were minimal and not significant ($P > 0.01$).

In summary, MV is competitive in peptide-, DNA- and RNA-binding site predictions when compared with recent individual methods developed for the prediction of these specific functional sites. Thus, having a single method able to predict different interface types and still be competitive makes MV a useful resource. Moreover, the low computational cost of the RF approach makes the prediction process sufficiently fast to be implemented as a web application with almost no waiting time (see below in Section 3.4).

3.2 Selectivity of the predictions

A central consideration during the development of MV was to explore the selectivity or discriminative nature of the predictions. For example, did DNA-binding sites show consistently higher scores when using MV to predict DNA-binding sites that when predictions were made using the specific RNA-binding statistical model? To answer this question, a number of cross-prediction experiments were performed. MV was used with each dataset (PROT-set, PEP-set, DNA-set and RNA-set) to predict protein-, peptide-, DNA- and RNA-binding sites. When the training and testing set were the same, the scores were calculated using a 5-fold cross-validation. The distributions of raw scores of the interfaces residues were plotted against each of the predicted interface types.

As shown in Figure 2, the distribution of scores and median values of interface residues were significantly different ($P < 0.01$) when predicted interfaces and dataset types coincided. For example, the prediction scores for protein-binding interfaces in the PROT-set were higher and median values were significantly different than peptide-, DNA- and RNA-binding site prediction scores. This was also true for the PEP-set, DNA-set and RNA-set. In the case of peptide-binding predictions on the PROT-set, there were a number of outlier protein complexes, i.e. protein–protein interfaces that scored very high (Fig. 2; shown in red).

The analysis of these outliers revealed that protein–protein interactions were mediated by linear stretches of the polypeptide chain, i.e. one of the protein partners binds to the cognate partner via a stretch of residues in an extended conformation. An example is illustrated in Supplementary Fig. S2. The interaction in protein complex formed by PPIase A and protein Gag-Pol is mediated by the recognition of a long and flexible loop of Gag-Pol. Thus, the actual interface mediating the interaction is structurally more similar to a peptide-binding site than a protein-binding site and consequently MV assigned high scores to this region of the protein.

The selectivity of the predictions was further assessed by analyzing the COMB-set. The COMB-set included three different

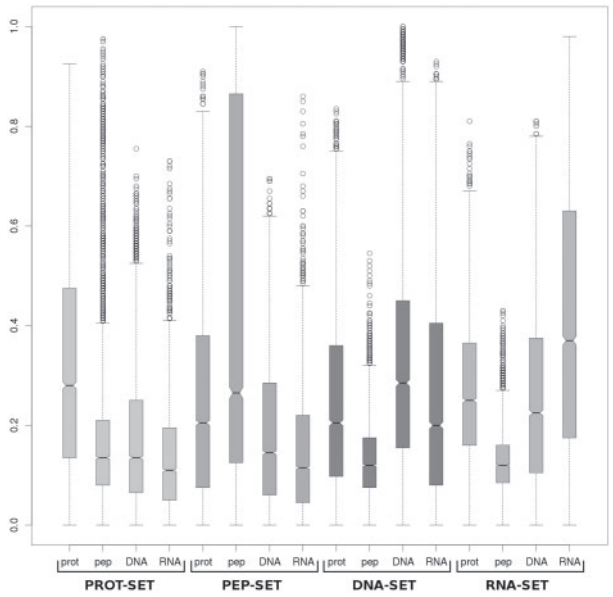

Fig. 2. Residue binding site score box plots. The different colors represent the different datasets: light blue PROT-set, light green PEP-set, red DNA-set and orange RNA-set. In each dataset, four binding site types were predicted as shown in the *X*-axis: prot, pep, DNA and RNA for protein-, peptide-, DNA- and RNA-binding site prediction, respectively. The central horizontal line in the box marks the median and the box edges the first and third quartile; errors bars show minimum and maximum values and outliers are represented by empty circles

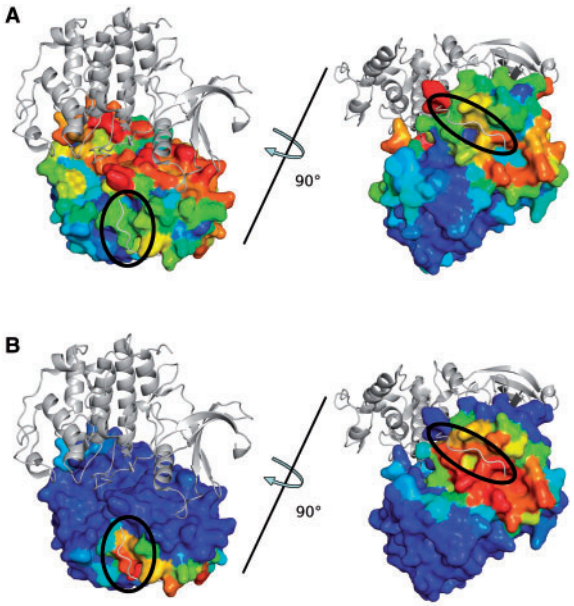

Fig. 3. Structural mapping of protein- and peptide-binding site predictions onto the crystal structure of cyclin-A2 complexed with CDK2 and the CDK2 substrate peptide: Nt-PKTPKKAKKL-Ct (PDB code: 3qhr). Cyclin-A2 is shown in surface representation, while CDK2 and substrate peptide are depicted in ribbon. Cyclin-A2 colored according to prediction scores (s): red $s > 0.8$; orange $0.6 < s < 0.8$; yellow $0.4 < s < 0.6$; green $0.2 < s < 0.3$; light blue $0.3 < s < 0.2$; blue $s < 0.2$. (**A**) protein-binding site prediction; (**B**) peptide-binding site prediction. Peptide-binding site highlighted using a solid ellipse. Figs. 3 and 4 were generated using PyMOL (http://pymol.sourceforge.net)

types of complexes: protein–protein–peptide, protein–protein–DNA and protein–protein–RNA (full list of PDB codes is available in the Supplementary Data). As shown in the Supplementary Data (Supplementary Fig. S3), the prediction scores were consistently higher when interface and prediction type was the same and lower and distributed in a narrower interval when different, e.g. scores assigned to an actual DNA-binding site when predicting a protein-binding site. Two examples of combined predictions are depicted in Sections 3.3 and 3.4.

3.3 Combined predictions

3.3.1 A protein–protein–peptide complex An example of a combined prediction of protein- and peptide-binding sites is depicted in Figure 3. Cyclin-A2 recognizes both a globular protein and a peptide and so can bind to both the cell division kinase 2 (CDK2) and the CDK2 substrate peptide. As shown in Figure 3, when predicting protein-binding sites, MV assigned high scores to the actual interface to CDK2 (red) and low scores to the rest of the exposed surface and the peptide-binding site (blue). On the contrary, when predicting peptide-binding sites, only the region that recognizes the substrate peptide scored high. Therefore, and in accordance to the data shown in Figure 2 and Supplementary Fig. S3, MV was able to discriminate between two different types of interfaces and correctly locate the interaction patches on the surface of the protein.

3.3.2 A protein–protein–DNA complex The crystal structure of an engineered heterodimeric I-CreI endonuclease composed of two subunits V2 and V3 is an example of a protein that interacts both with a protein and DNA. The prediction of both protein- and DNA-binding sites on subunit V2 is depicted in Figure 4. MV predicted with a high accuracy the actual DNA-binding site (red) of the V2 endonuclease (chain A), while scoring low (blue) the interface with V3 endonuclease. Likewise, MV assigned high scores to the actual protein interface between V2 and V3 endonucleases (red), while scoring low the DNA interface (blue). Again, this example shows the discriminative power of the predictions in agreement with the data shown in Figure 2 and Supplementary Fig. S3.

3.4 Web-server interface

For end-users two of the most important aspects of a computer-based application are accessibility and ease of use. To achieve that, an *ad hoc* web-server and visualization tool was developed to both allow access to the method and to facilitate the analysis and visualization of the predictions. The interface allows the visualization of the multiple predicted interfaces simultaneously and in the context of the protein structure by using a composite viewer, and thus facilitating the analysis and assessment of the predictions (Supplementary Fig. S4). Each independent viewer is synchronized, such that any structural manipulation (e.g. a rotation) occurs simultaneously in the other views. The web application has other in-built functionalities including choice of surface representation, the selection of raw or normalized scores and clickable residue lists sorted by prediction scores. Prediction scores are mapped onto the structure of the protein and are represented by a color gradient

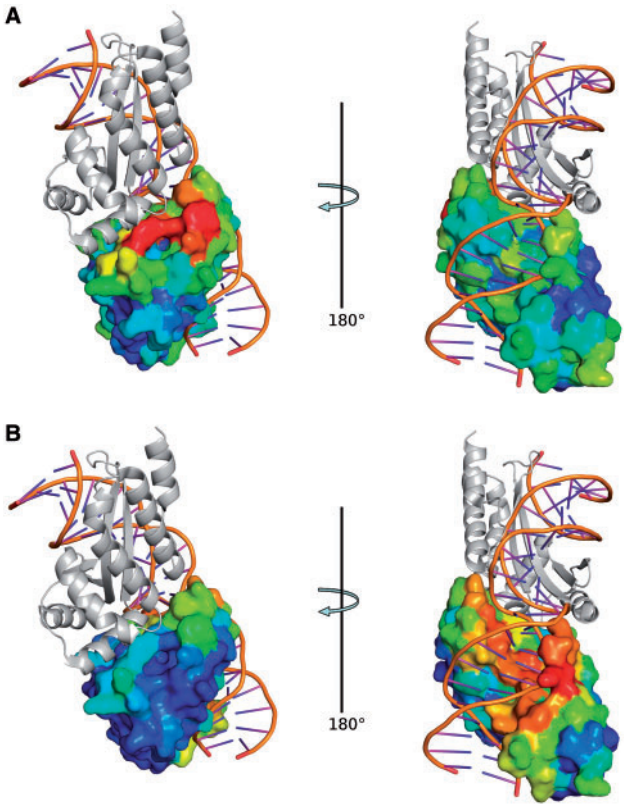

Fig. 4. Structural mapping of protein- and DNA-binding site predictions onto the crystal structure of an engineered heterodimeric I-CreI endonuclease complexed with a 24-bp oligonucleotide of the human RAG1 gene sequence (PDB code: 3mxb). V2 endonuclease is shown in surface representation, while V3 and DNA are shown in ribbon. V2 colored according to prediction scores as described in Fig. 3. (**A**) Protein-binding site prediction; (**B**) DNA-binding site prediction

between 0 (blue) and 1 (red). Furthermore, the atomic coordinates of the protein in PDB format with modified B factors to represent the predicted residue scores and tables containing the predicted scores are available for download.

4 CONCLUSION

A novel computational method, MV, has been developed to predict protein-, peptide-, DNA- and RNA-binding sites. MV has been compared with recently published methods that predict individual types of interactions with a positive outcome. The structural mapping of functional sites is highly selective, allowing multiple sites to be predicted with high accuracy and reliability. A user-friendly web application has been developed to easily access to the method. Prediction results are readily available for download or can be analyzed within the same web-browser by using a web-application and a special graphic visualization viewer.

ACKNOWLEDGEMENTS

N.F.-F. thanks Dr Gendra for critical reading and insightful comments to the manuscript, and Ms Martina and Ms Daniela G. Fernandez for continuing inspiration and motivation.

Funding: This work was supported by the Research Councils United Kingdom (RCUK) Academic Fellowship scheme (to N.F.-F.) and an internal scholarship awarded by the Leeds Institute of Molecular Medicine (to J.S.). Publication costs were provided by the Institute of Biological, Environmental and Rural Science (IBERS).

Conflict of Interest: none declared.

REFERENCES

Allers,J. and Shamoo,Y. (2001) Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J. Mol. Biol.*, **311**, 75–86.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389.

Barber,C.B. *et al.* (1996) The Quickhull algorithm for convex hulls. *ACM Trans. Mathematical Softw.*, **22**, 469–483.

Bhardwaj, N. et al. (2005). Structure Based Prediction of Binding Residues on DNA-binding Proteins. *Conf Proc IEEE Eng Med Biol Soc*, **3**, 2611–2614.

Bray,T. *et al.* (2009) SitesIdentify: a protein functional site prediction tool. *BMC Bioinformatics*, **10**, 379.

Breiman,L. (1984) *Classification and Regression Trees*. Chapman & Hall/CRC, London, UK.

Chelliah,V. and Taylor,W.R. (2008) Functional site prediction selects correct protein models, *BMC Bioinformatics*, **9** (Suppl. 1), S13.

Cheng,C.W. *et al.* (2008) Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. *BMC Bioinformatics*, **9** (Suppl. 12), S6.

de Vries,S.J. *et al.* (2006) WHISCY: what information does surface conservation yield? Application to data-driven docking. *Proteins*, **63**, 479–489.

Draper,D.E. (1995) Protein-RNA recognition. *Ann. Rev. Biochem.*, **64**, 593–620.

Glaser,F. *et al.* (2001) Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*, **43**, 89.

Guerois,R. *et al.* (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.

Hwang,H. *et al.* (2010) Protein-protein docking benchmark version 4.0. *Proteins*, **78**, 3111–3114.

Innis,C.A. (2007) siteFiNDER|3D: a web-based tool for predicting the location of functional sites in proteins. *Nucleic Acids Res.*, **35**, W489–W494.

Jones,S. *et al.* (1999) Protein-DNA interactions: a structural analysis. *J. Mol. Biol.*, **287**, 877–896.

Jones,S. *et al.* (2001) Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.*, **29**, 943–954.

Landgraf,R. *et al.* (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.*, **307**, 1487.

Liaw,A. and Wiener,M. (2002) Classification and Regression by random Forest. *R News*, **2**, 18–22.

Liu,Z.P. *et al.* (2010) Prediction of protein-RNA binding sites by a random forest method with combined features. *Bioinformatics*, **26**, 1616–1622.

London,N. *et al.* (2010) The structural basis of peptide-protein binding strategies. *Structure*, **18**, 188–199.

Luscombe,N.M. *et al.* (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.*, **1**, reviews001.001–reviews001.037.

Maetschke,S.R. and Yuan,Z. (2009) Exploiting structural and topological information to improve prediction of RNA-protein binding sites. *BMC Bioinformatics*, **10**, 341.

Mihel,J. *et al.* (2008) PSAIA - protein structure and interaction analyzer. *BMC Struct. Biol.*, **8**, 21.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536.

Pei,J. and Grishin,N.V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.

Petsalaki,E. and Russell,R.B. (2008) Peptide-mediated interactions in biological systems: new discoveries and applications. *Curr. Opin. Biotechnol.*, **19**, 344–350.

Petsalaki,E. *et al.* (2009) Accurate prediction of peptide binding sites on protein surfaces. *PLoS Comput. Biol.*, **5**, e1000335.

Pettit,F.K. *et al.* (2007) HotPatch: a statistical approach to finding biologically relevant features on protein surfaces. *J. Mol. Biol.*, **369**, 863–879.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

Segura,J. *et al.* (2011) Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams. *BMC Bioinformatics*, **12**, 352.

Sikic,M. *et al.* (2009) Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. *PLoS Comput. Biol.*, **5**, e1000278.

Terribilini,M. *et al.* (2007) RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res.*, **35**, W578–W584.

Tjong,H. and Zhou,H.X. (2007) DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res.*, **35**, 1465–1477.

von Mering,C. *et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.

Wallace,A.C. *et al.* (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.*, **8**, 127.

Xiong,Y. *et al.* (2011) An accurate feature-based method for identifying DNA-binding residues on protein surfaces. *Proteins*, **79**, 509–517.