BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 3 2012, pages 331–341
doi:10.1093/bioinformatics/btr657

Structural bioinformatics

Advance Access publication November 29, 2011

Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors

Ke Chen[1,2], Marcin J. Mizianty[2] and Lukasz Kurgan[2,*]

[1]School of Computer Science and Software Engineering, Tianjin Polytechnic University, No. 63 Chenglin Road, Hedong District, Tianjin 300160, P. R. of China and [2]Department of Electrical and Computer Engineering, 2nd floor, ECERF (9107 116 Street), University of Alberta, Edmonton, AB, Canada T6G 2V4

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Nucleotides are multifunctional molecules that are essential for numerous biological processes. They serve as sources for chemical energy, participate in the cellular signaling and they are involved in the enzymatic reactions. The knowledge of the nucleotide–protein interactions helps with annotation of protein functions and finds applications in drug design.

Results: We propose a novel ensemble of accurate high-throughput predictors of binding residues from the protein sequence for ATP, ADP, AMP, GTP and GDP. Empirical tests show that our NsitePred method significantly outperforms existing predictors and approaches based on sequence alignment and residue conservation scoring. The NsitePred accurately finds more binding residues and binding sites and it performs particularly well for the sites with residues that are clustered close together in the sequence. The high predictive quality stems from the usage of novel, comprehensive and custom-designed inputs that utilize information extracted from the sequence, evolutionary profiles, several sequence-predicted structural descriptors and sequence alignment. Analysis of the predictive model reveals several sequence-derived hallmarks of nucleotide-binding residues; they are usually conserved and flanked by less conserved residues, and they are associated with certain arrangements of secondary structures and amino acid pairs in the specific neighboring positions in the sequence.

Availability: http://biomine.ece.ualberta.ca/nSITEpred/

Contact: lkurgan@ece.ualberta.ca

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

Received on April 05, 2011; revised on October 24, 2011; accepted on November 15, 2011

1 INTRODUCTION

Nucleotides are multifunctional molecules that are essential for numerous biological processes. The nucleotides are structural units of nucleic acid chains, and they serve as sources for chemical energy, participate in the cellular signaling and they are involved in the enzymatic reactions. As of June 2010, 5293 proteins in the Protein Data Bank (PDB) are annotated as 'nucleotide binding', and nucleotides constitute ~15% of biologically relevant ligands included in this database (Dessailly *et al*, 2008; Goto *et al*., 2002). This demonstrates the ubiquity and the substantial interest in the protein–nucleotide interactions.

Past two decades observed a substantial effort in identification and characterization of the nucleotide-binding sites. Most of these approaches are based on the analysis of known nucleotide-binding sequences and structures, which were used to identify conserved motifs in protein sequences and structures. For instance the Walker A and B sequence motifs were identified for the adenine nucleotide-binding proteins (Walker *et al*., 1982). A fuzzy recognition template was proposed for the characterization of the adenylate–protein interactions (Moodie *et al*., 1996). The Johnson motif was reported to cover one-third of the adenine mononucleotide-binding proteins (Denessiouk and Johnson, 2000). Mao *et al.* proposed a motif that interacts with the adenine and is shared by five different protein folds (Mao *et al*., 2004). Thornton's group applied structural motifs in identification and prediction of adenine-binding sites for functionally uncharacterized proteins (Nobeli *et al*., 2001). Moreover, an empirical scoring function was developed for prediction of the nucleotide-binding sites in protein structures (Saito *et al*., 2006). The above methods characterize the sequence motifs for a relatively narrow range of the nucleotide–protein interactions, usually only for a selected interaction mode for a single nucleotide type, or they require tertiary protein structure as the input, which substantially limits their utility. The large number of protein chains with unknown structure motivates the development of computational tools for high-throughput sequence-based annotation of the nucleotide-binding residues for a wide range of the nucleotides.

Currently, there is only one method for the prediction of binding sites/residues from the protein sequence for a comprehensive set of nucleotides (Firoz *et al*., 2011). Two other methods predict the ATP- and GTP-binding residues, respectively (Chauhan *et al*., 2009, 2010). These two methods input information extracted from the sequence and the corresponding sequence profile using a window centered on the predicted residue into a machine learning classifier that predicts propensity of this residue to interact with the ATP or GDP. We propose an ensemble of predictors of binding residues for five common nucleotides in the PDB including ATP, ADP, AMP, GTP and GDP. Each of these nucleotides binds to at least 50 diverse proteins, i.e. their chains share pairwise sequence similarity at <40%, which means that they cannot be easily annotated using the sequence alignment. At the same time, availability of 50 chains provides sufficient amount of annotated interactions to build a well-performing predictor. Although the interaction modes for some

*To whom correspondence should be addressed.

nucleotides, e.g. ATP and AMP, are relatively similar they bind to different residues. For instance, the ATP molecule contains three phosphates while AMP contains only one phosphate, and even though ATP and AMP may bind in the same pocket more residues will interact with the ATP than with the AMP. This means that a predictor designed for the ATP binding residues cannot be simply re-used to predict the binding residues for the other nucleotides. Therefore, we propose five models that predict the binding residues for the five most common nucleotides in the PDB. In contrast to the existing methods (Chauhan *et al*., 2009, 2010; Firoz *et al*., 2011), the proposed NsitePred is characterized by the following three novel aspects. First, in addition to the sequence and the sequence profile that are used in the existing predictors, our method also uses residue conservation scores, predicted secondary structure, predicted relative solvent accessibility and predicted dihedral angles to build a comprehensive and custom-designed set of input features. These additional inputs allow for significantly more accurate predictions when compared not only with the two existing predictors but also with popular tools including sequence alignment and residue conservation scoring. Second, our analysis shows that the predictions by a machine learning-based classifier that uses the abovementioned inputs are complementary to the predictions based on the sequence alignment. To this end, NsitePred implements a consensus of the machine learning-based and the alignment-based predictors. Third, analysis of our model reveals several sequence-derived hallmarks of the nucleotide-binding residues, which are related to the residue-level conservation and certain arrangements of secondary structures and amino acid pairs in the vicinity of the nucleotide-binding residues.

2 METHODS

2.1 Dataset

The nucleotides that were considered in this study contain at least one of the five nucleobases, a 5-carbon sugar and 1–3 phosphates. We extracted all complexes from PDB that included these nucleotides; we need these structures to obtain annotation of the binding residues to build and evaluate our predictor. The maximal pairwise sequence identity of the resulting protein chains for each of the nucleotides was reduced to 40% with CD-hit (Li and Godzik, 2006). We include the nucleotides with at least 50 chains in the corresponding set. The relatively low identity assures that these nucleotides bind a wide range of protein chains, which makes it challenging to find the binding residues using the sequence alignment. The availability of at least 50 chains provides us with a sufficient amount of annotated binding residues to build and evaluate a well-performing predictor.

Dataset 1 includes 227, 321, 140, 56 and 105 chains that were released in PDB before 10 March 2010 and that bind to ATP, ADP, AMP, GTP and GDP, respectively. Similar to the annotation of the DNA- and small ligand-binding residues (Chen and Kurgan, 2009; Luscombe *et al*., 2001), a given residue is annotated as 'nucleotide binding' if at least one of its non-hydrogen atom is <3.9 Å away from a non-hydrogen atom of the nucleotide. As suggested in (Luscombe *et al*., 2001), atoms within 3.9 Å are considered to interact through the van der Waals contacts. Dataset 1 includes 4688 ADP-binding, 3393 ATP-binding, 1756 AMP-binding, 853 GTP-binding and 1577 GDP-binding residues, and 121158, 80409, 44009, 18888 and 36561 non-binding residues, respectively.

Dataset 2 consists of nucleotide-binding chains that were released after 10 March 2010. The maximal pairwise sequence identity in Dataset 2 was reduced to 40%. Moreover, if a given chain in Dataset 2 shares $>40\%$ identity to a chain in Dataset 1 and both chains interact with the same nucleotide, then we remove the chain from Dataset 2. This assures that the Dataset 2 is independent of the Dataset 1 and can be used to test models developed using Dataset 1. Consequently, Dataset 2 includes 17, 25, 18, 6 and 9 chains that bind to ATP, ADP, AMP, GTP and GDP, respectively.

Dataset 3 consists of chains that do not interact with nucleotides, and is used to evaluate whether the NsitePred would 'overpredict' nucleotide-binding residues. We use the pre-culled list of 1853 PDB chains generated by the PISCES server (Wang and Dunbrack, 2003) at 20% sequence identity, which correspond to high-quality structures with maximal resolution of 1.6 Å and maximal R-factor of 0.25. Next, among this set of representative proteins, we remove all chains that (potentially) interact with nucleotides. Any chain that shares $>40\%$ identity to any chain in Dataset 1 (which is used to build our predictive model), or which is annotated as nucleotide-binding in the Gene Ontology database (Ashburner *et al*., 2000), or which binds to nucleotides among the depositions in the PDB is removed. As a result, we extracted 1372 chains that do not interact with the nucleotides.

The datasets can be found at http://biomine.ece.ualberta.ca/nSITEpred/

2.2 Evaluation criteria and test procedure

We use 5-fold cross validation to assess predictions on Dataset 1. Dataset 2 and 3 are used as independent datasets to assess the prediction models that are built utilizing Dataset 1. The sequences in Dataset 1 are randomly divided into 5-folds, of which four are used for training and the one for testing; each of the 5-folds is used once as the test fold. We evaluate (i) the binary value that defines whether a given residue does or does not bind to a given nucleotide; and (ii) the real value that quantifies the probability of binding to the nucleotide. The binary predictions were assessed using five measures: $Precision(PREC)=TP/(TP+FP)$; $Recall\ (REC)=TP/(TP+FN)$; $Specificity(SPEC)=TN/(FP+TN)$; $Accuracy(ACC)=(TP+TN)/(TP+FP+TN+FN)$; $MCC=(TP*TN--FP*FN)/sqrt[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)]$ where TP (true positives) and TN (true negatives) are the counts of correctly predicted binding and non-binding residues, respectively, FP (false positives) are non-binding residues that were predicted as binding, and FN (false negatives) are binding residues that were predicted as non-binding. The precision, recall and specificity evaluate quality of predictions for the predicted binding residues, native binding residues and native non-binding residues, respectively. The Matthews correlation coefficient (MCC) evaluates the overall predictive quality. MCC values are between -1 and 1 with higher values for better predictions; 0 means that all residues are predicted as binding (or non-binding).

The receiver operating characteristic (ROC) curves were used to examine the predicted probabilities. For each value of probability P achieved by a given method (between 0 and 1), all residues with probability $\geqslant P$ are set as the binding residue and all other residues are set as the non-binding residues. Next, the TP$-$rate$=$TP$/$(TP$+$FN) and the FP-rate$=$FP$/$(FP$+$TN) are calculated to draw the ROC curve and we use the area under the curve (AUC) to quantify the predictive quality. Unlike the measures that assess the binary predictions, which depend on the cutoff threshold to define binding/non-binding residues, the AUC value considers all possible thresholds and thus it provides a more comprehensive evaluation.

We analyzed statistical significance of the differences in the MCC and AUC values between predictions generated by NsitePred and the other considered methods. The MCC values are available for all methods, while the AUC values cannot be calculated for an alignment-based predictor that provides only a binary annotation. The MCC and AUC values are calculated per sequence (using the cross-validated predictions) for each method and we compare them using a paired test. Since these values are not normal, as tested using Shapiro–Wilk test at the 0.05 significance, we use the non-parametric Wilcoxon rank sum test to measure the differences between the paired MCC (AUC) values calculated for two predictors. We annotate the difference as significant when the P-value <0.01.

2.3 Architecture

For a given protein sequence we use PSIPRED (McGuffin *et al*., 2000) to predict the secondary structure, REAL Spine3 (Faraggi *et al*., 2009) to predict

the relative solvent accessibility (RSA) and dihedral angles, and PSIBLAST (Altschul *et al*., 1997) to generate the PSSM profile. These inputs together with the sequence are processed using a sliding window to compute a set of numeric features that describe the residue in the center of the window; the features are inputted into a support vector machine (SVM) classifier, which outputs probability of nucleotide binding for this residue. We use the 'one-against-the rest' strategy to build the SVM models. This machine learning-based approach is named as SVMPred. Moreover, we run the BLAST-based alignment between the predicted sequence and sequences in the training dataset for a given nucleotide type. The residues in the predicted sequence that were aligned with the binding residues in the best aligned training chain are predicted as the binding residues, i.e. they are assigned with probability that equals 1 while the other residues are assigned with probability that equals 0. The proposed NsitePred method implements a consensus of SVMPred and the alignment-based predictor by averaging the probabilities generated by SVMPred and the alignment-based predictor.

2.4 Feature-based sequence representation

The SVMPred utilizes both sequence and predicted structural descriptors, including the secondary structure, dihedral angles and RSA, to generate features. We utilize a sliding window of size 17 centered on the predicted residue to extract the features, which include

- *Predicted secondary structure* generated by the PSIPRED for each residue in the window.
- *Predicted RSA* and *dihedral angles* (*phi* and *psi* angles) generated by the REAL Spine3 for each residue in the window.
- *PSSM profile* generated by PSIBLAST with default parameters using the NCBI non-redundant database. We include the scores for each of the 20 substitution amino acid types and we also compute the average substitution score over all amino acid types in each of the following four groups, hydrophobic (Ala, Cys, Ile, Leu, Met and Val), negatively charged (Asp and Glu), positively charged (His, Lys, Arg) and carboxamide containing (Asn and Gln).
- *Terminus indicator* is set to 1 for the first and the last three residues in the sequence, and it equals 0 for the other positions.
- *Secondary structure segment indicators* for helix/strand/coil on both sides of the predicted residue, which annotate whether a helix or a strand segment (or neither) is predicted to the left or right of the residue in the center of the window.
- *Residue conservation scores* are calculated from the PSSM values for each position based on the Shannon entropy (referred to as conservation A) and based on two formulas that incorporate background frequency of amino acids (Capra and Singh, 2007; Wang and Samudrala, 2006), which are named conservation B and C, respectively.
- *Collocation of significant AA pairs* for the residues in the window. This involves finding the frequency of the amino acids pairs with gaps, as defined in (Chen *et al*., 2007, 2009), formed between the residue in center of the window and another residue up to five positions away. Similarly as in (Senes *et al*., 2000), we use *P*-values to select the collocated pairs that are significantly associated with nucleotide-binding residues.

Details concerning the calculation of the features are given in the Supplementary Material. We note that the terminus and the secondary structure segment indicators, collocation of the amino acid pairs and the predicted secondary structure, RSA, and dihedral angles were never before used to predict the nucleotide-binding residues.

2.5 Feature selection and parameterization

The same features, except for the collocated amino acids pairs are considered to predict binding residues for each of the five nucleotides. Some of these features may not be relevant to the prediction of the nucleotide-binding residues and they could be also redundant (correlated) with each other. Therefore, we performed feature selection to remove the irrelevant and redundant features. The selection was performed using the 5-fold cross validation separately for each of the five nucleotide types. First, the biserial correlation (Tate, 1954) between each of the features and the binary annotation of the binding residues was calculated for each of the five training sets. The averaged, over the five training sets, correlation values were used to rank the features. We used a wrapper-based feature selection with the forward best first search. More specifically, for a given list of feature $F = [f_i \text{ where } i = 1, 2, \ldots, n]$ sorted in the descending order by their average biserial correlation and an empty list S that stores the selected features, we add the top-ranked feature from F to S and run a linear SVM (Fan *et al*., 2005, 2008) with default parameters (i.e. linear kernel and complexity constant $C = 1$) using feature set S in the cross validation regime. If the addition of the top-ranked feature improves the average AUC value over the five test folds, then this feature is retained in S; otherwise it is removed. We repeat that until F is empty, i.e. we scan the entire feature set once. Next, the SVM classifier is parameterized on the selected feature set. We considered the polynomial and the Radial Basis Function (RBF) kernels. For the polynomial kernel, the complexity constant C is initially fixed at 1 and the degree of the polynomial is adjusted between 0.5 and 5 with step = 0.5. The degree that results in the highest cross-validated AUC value is selected, and next we adjust C using consecutive powers of 2 between 2^{-3} and 2^5. Similarly for the RBF kernel, the *gamma* parameter is first optimized using the 2^{-7}–2^3 range when C is fixed at 1, and next C is adjusted using the 2^{-3}–2^5 range. We selected the parameters that maximize the cross-validated AUC and we performed a separate parameterization each of the five nucleotide types; the optimized parameters for each nucleotide type are given in the Supplementary Table S1.

2.6 Considered baseline predictors

The NsitePred is compared with the current predictors for the ATP and GDP, ATPint (Chauhan *et al*., 2009) and GTPbinder (Chauhan *et al*., 2010), the method proposed by Firoz *et al*. (2011) and three baseline predictors based on the residue conservation, sequence alignment and a simple classifier similar to the methods in (Chauhan *et al*., 2009, 2010) that uses evolutionary profile:

- *Rate4site* program (Pupko *et al*., 2002) predicts functional sites by finding conserved residues. We first run PSIBLAST with the query sequence against the NCBI non-redundant database. For chains with at least three significant matches, we created alignments of the best 50 sequences, which is the default for the web version of Rate4site (Ashkenazy *et al*., 2010), using ClustalW (Larkin *et al*., 2007) and we inputted them into Rate4site. Rate4site generates conservation score for each residue, and the residues with the lower scores, which indicate higher conservation, have a higher probability to be binding residues. We use these scores to compute ROC curves and the corresponding AUC values. We threshold these scores by maximizing the MCC value on the entire dataset to obtain binary predictions. The AUC and MCC values are computed separately for each nucleotide type.
- *Sequence alignment using BLAST* identifies similar sequences or segments from an annotated (with the nucleotide-binding residues) dataset for a given query sequence. This approach predicts the binding residues by using the nucleotide-binding annotations from the best aligned sequence, i.e. sequence with lowest *E*-value. We execute the BLAST-based alignment between a query sequence and all other sequences (except the query sequence itself) in the dataset for a given nucleotide type. The residues in the query sequence that were aligned with the binding residues on the best aligned chain are predicted as the binding residues.
- *PSSM profile* is widely used in related sequence-based predictors, including ATPint (Chauhan *et al*., 2009) and GTPbinder (Chauhan *et al*., 2010). We build a simple predictor that uses SVM (with the

same parameters as the corresponding SVM in SVMPred) and takes PSSM profile as the input to validate the effectiveness of the sequence representation proposed in this work. This allows us to estimate improvements provided by the new features based on conservation scores and predicted secondary structure, RSA, and dihedral angles.

3 RESULTS

3.1 Comparison with the existing methods

Table 1 compares the NsitePred with the ATPint (using on the web server at www.imtech.res.in/raghava/atpint/), GTPbinder (www.imtech.res.in/raghava/gtpbinder/) and the three baseline predictors based on the alignment, conservation scoring and evolutionary profiles. We use the tripeptide-based GTPbinder, which outperforms the single-residue and dipeptide-based versions (Chauhan *et al*., 2010), in two configurations including the GTPbinder_PSSM that utilizes PSSM profiles and the GTPbinder_seq that is based solely on the protein sequence.

For Dataset 1, across predictions for the five nucleotide types, the NsitePred obtains AUC $\geqslant 0.83$, MCC $\geqslant 0.38$ and accuracy > 0.96. Our method outperforms the other approaches by a statistically significant margin for both AUC and MCC measures. Although some other approaches provide higher precision, recall or specificity, the NsitePred provides favorable balance between these three measures. The sensitivities and specificities, for which predictions are binarized with different cutoff thresholds, achieved by NsitePred and the competing methods are given in the Supplementary Table S2. Based on the MCC, which provides an overall estimate of the quality of the binary predictions, the NsitePred is superior to SVMPred and the BLAST-based predictor, followed by the PSSM profile-based predictor, and the Rate4site. We note that the consensus-based NsitePred achieves higher AUC and MCC values and a better balance between the precision and recall than SVMPred. This suggests that the predictions from SVMPred are complementary to the alignment-based predictions. Since the Rate4site only considers the residue conservation, its relatively low predictive performance could be explained by the fact that the conserved residues could also include binding residues for other types of ligand such as the metal ions, carbohydrates, peptides, etc. This explanation is supported by the relatively high recall (i.e. high fraction of the correctly predicted native binding residues) coupled with the low specificity (which indicates an over-prediction of the binding residues) which are achieved by the Rate4site.

The AUC, MCC, precision, specificity and accuracy of ATPint are lower than the values achieved by NsitePred and the three baseline predictors, see Table 1, and they are also lower than it was reported in (Chauhan *et al*., 2009). The likely reason for that is the fact that the ATPint authors used a balanced number of binding and non-binding residues to design and evaluate their method, which resulted in the lower predictive quality when applied here to the full chains. Our results indicate that the ATPint over-predicts the ATP-binding residues, which is evidenced by the low specificity and precision, i.e. a high number of false positives. We show that, as expected and as shown in (Chauhan *et al*., 2010), GTP_binder that utilizes the evolutionary profile (GTPbinder_PSSM) outperforms the version that does not use this information (GTPbinder_seq). The PSSM-based GTP_binder achieves AUC $= 0.8$ and MCC $= 0.39$ that are lower than the values achieved by the NsitePred (by the statistically significant margin) and the BLAST-based predictor, and higher than the values achieved by the PSSM profile-based predictor and Rate4site.

The ROC curves based on the predictions on Dataset 1 are shown in Figure 1. The figure focuses on the FP rates <0.05 since only $\sim$4% of residues bind to nucleotides; the full ROC is given in the Supplementary Figure S1. The BLAST-based predictor does not provide the probabilities, and thus we include a single point that corresponds to its binary predictions. The ROC curves reveal that NsitePred provides higher TP rates for the FP values between 0.01 and 0.05 when compared with the other methods for each of the five types of the ligands.

Table 2 compares NsitePred with the existing methods, including the recent method proposed by Firoz *et al*. (2011), and baseline predictors on Dataset 2, which consists of chains that were released after the NsitePred was designed and which are dissimilar to chains in the Dataset 1 that was used to build the predictive models. We did not evaluate the method by Firoz *et al.* on Dataset 1 because the training set used by these authors overlaps with our Dataset 1. Similar to the results on Dataset 1, NsitePred achieves significantly higher AUC and MCC values when compared with the other methods, including ATPint, GTPbinder and the method by Firoz *et al*. (2011), for all five nucleotide types. NsitePred improves by 0.01–0.02 in AUC and 0.01–0.04 in MCC, depending on a nucleotide type, over the predictions from SVMPred by implementing the consensus of SVMPred and the BLAST-based alignment. These improvements are shown to be statistically significant. A detailed summary of sensitivities and specificities achieved by NsitePred and the competing methods are shown in the Supplementary Table S3. The ROC curves of NsitePred and the other methods on Dataset 2 are given in the Supplementary Figures S2 (for the FP rates <0.05) and S3 (entire range of FP rates). The ROC curves reveal that NsitePred provides higher TP rates for the FP values between 0.012 and 0.05 when compared with the other methods for each of the five types of the nucleotides. We also evaluated the predictive quality of the considered methods for prediction of all nucleotide-binding residues. In this case, a residue is defined as a 'nucleotide-binding' if it interacts with any of the five nucleotides, and a residue is predicted as a 'nucleotide-binding' when a given method predicts that this residue interacts with any of the five nucleotides; see the 'All' row in Table 2. Similarly as for the prediction of individual nucleotide types, NsitePred achieves higher AUC, MCC and recall than the remaining methods, while the BLAST-based predictor achieves higher precision, specificity and accuracy, see Table 2.

Based on a request from a reviewer, we test the NsitePred on the original datasets that were used to develop and test ATPint and GTPbinder. We use the same inputs and parameterization for the NsitePred (as for the Datasets 1, 2 and 3) and perform 5-folds cross validation that duplicates the tests done in (Chauhan *et al*., 2009, 2010). Specifically, we first annotate positive samples (binding residues) and negative samples (non-binding residues). Next, we randomly select a subset of the non-binding residues that equals to the number of binding residues. Finally, the binding and non-binding residues are combined and divided (per residue) into 5-folds to perform cross validation. The results of NsitePred, ATPint and GTPbinder are given in the Supplementary Table S4. We note that NsitePred generates higher AUC, MCC, precision, sensitivity and specificity than ATPint and GTPbinder.

Table 1. Comparison of the quality of the sequence-based prediction of the ATP, ADP, AMP, GDT and GTP-binding residues between the NsitePred and the related predictors of nucleotide-binding residues, including ATPint and GTPbinder that predict ATP- and GTP-binding residues, respectively, and predictors based on the alignment (utilizing BLAST), conservation scoring (utilizing Rate4site), evolutionary profiles (utilizing PSSM and SVM classifier) and SVMPred (utilizing the same feature representation as NsitePred and SVM classifier) on Dataset 1

Type	Predictor	AUC		MCC		PREC	REC	SPEC	ACC
		value	sig.	value	sig.				
ATP	NsitePred	**0.861**		**0.46**		0.519	0.444	0.982	0.96
	SVMPred	0.854	+	0.433	+	0.564	0.361	0.988	**0.962**
	Rate4site	0.749	+	0.182	+	0.132	0.446	0.87	0.852
	PSSM + SVM	0.824	+	0.27	+	0.262	0.354	0.957	0.933
	BLAST	NA	NA	0.359	+	**0.578**	0.243	**0.993**	**0.962**
	ATPint	0.627	+	0.078	+	0.061	**0.539**	0.651	0.648
ADP	NsitePred	**0.893**		**0.572**		0.633	**0.544**	0.988	0.971
	SVMPred	0.885	+	0.555	+	**0.704**	0.458	0.993	**0.973**
	Rate4site	0.749	+	0.161	+	0.106	0.472	0.844	0.83
	PSSM + SVM	0.826	+	0.296	+	0.344	0.298	0.978	0.953
	BLAST	NA	NA	0.439	+	0.658	0.311	**0.994**	0.969
AMP	NsitePred	**0.829**		**0.377**		0.511	0.304	0.988	0.962
	SVMPred	0.82	+	0.36	+	**0.667**	0.208	**0.996**	**0.966**
	Rate4site	0.755	+	0.174	+	0.107	**0.562**	0.799	0.79
	PSSM + SVM	0.788	+	0.203	+	0.142	0.46	0.889	0.873
	BLAST	NA	NA	0.222	+	0.395	0.145	0.992	0.959
GDP	NsitePred	**0.91**		**0.675**		0.734	**0.646**	0.991	0.976
	SVMPred	0.905	+	0.655	+	**0.716**	0.623	0.989	**0.977**
	Rate4site	0.733	+	0.17	+	0.11	0.516	0.823	0.811
	PSSM + SVM	0.879	+	0.442	+	0.433	0.502	0.972	0.952
	BLAST	NA	NA	0.564	+	0.780	0.426	**0.995**	0.972
GTP	NsitePred	**0.844**		**0.562**		0.706	0.473	0.991	0.968
	SVMPred	0.836	+	0.551	+	**0.848**	0.373	**0.997**	**0.97**
	Rate4site	0.748	+	0.18	+	0.108	**0.569**	0.806	0.796
	PSSM+SVM	0.801	+	0.308	+	0.331	0.346	0.968	0.941
	BLAST	NA	NA	0.461	+	0.689	0.327	0.994	0.968
	GTPbinder_seq	0.548	+	0.03	+	0.055	0.177	0.876	0.849
	GTPbinder_PSSM	0.802	+	0.388	+	0.655	0.246	0.995	0.965

We report the average values over the 5-folds cross validation. The highest values for each ligand type and each quality index, including AUC, precision (PREC), recall (REC), specificity (SPEC), accuracy (ACC) and MCC, are set in bold. The significance of the differences between NsitePred and the other methods are measured for the AUC and MCC and they are given in the 'sig.' columns. The significance tests compare paired per-sequence prediction quality over a given benchmark dataset. The + and − mean that the NsitePred is statistically significantly better/worse with $P\text{-}value < 0.01$, and = means that results are not significantly different. The 'NA' means that the corresponding value could not be computed, i.e. BLAST generates only the binary predictions.

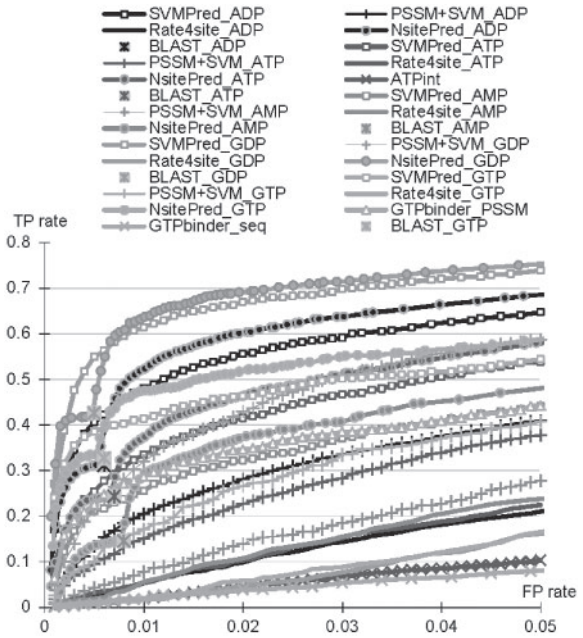

Fig. 1. The ROC curves for the NsitePred (denoted using thick solid lines with filled circle markers), SVMPred (denoted using thick solid lines with hollow square markers), ATPint (thick solid line with x markers), GTPbinder (thick solid lines using cross and hollow triangle markers), Rate4site (thick solid line without markers) and the predictor based on the PSSM with the SVM classifier (thin solid line with cross makers) for predictions on Dataset 1. The FP-rate is constrained to $[0, 0.05]$ range and the BLAST-based solution is shown using a single point (star marker on gray background) that corresponds to the binary predictions.

3.2 Performance on non-binding chains

We also assess the predictive quality of NsitePred and the other methods on protein sequences that do not interact with nucleotides (Dataset 3). We measure the error rate, which is defined as ratio between the number of false positives (FPs) and the total number of residues, for all considered methods; we note that there are no positive (nucleotide-binding) residues in this dataset. The error rates of NsitePred are 0.48, 1.15, 0.76, 0.93 and 0.67% for ATP, ADP, AMP, GTP and GDP, respectively, see Supplementary Table S5. The error rates of NsitePred are slightly higher than the error rates of BLAST-based method and SVMPred, but lower than the error rates of ATPint and GTPbinder. We note that NsitePred predicts 3.6 and 3.1%, 3.2 and 3.1%, 2.4 and 2.2%, 2.6 and 4.0%, and 3.3 and 4.3% of the residues in Dataset 1 and Dataset 2 as ATP-, ADP-, AMP-, GTP- and GDP-binding residues, respectively. These results demonstrate that NsitePred predicts fewer nucleotide-binding residues for the non-binding chains than for the nucleotide-binding chains.

3.3 Contribution of specific input types

We assess contributions of specific input types for the prediction of the nucleotide-binding residues with NsitePred. The inputs are categorized into five groups: (i) the BLAST-based prediction; (ii) predicted secondary structure (including the secondary structure segment indicator) and dihedral angles; (iii) PSSM profile and conservation scores; (iv) predicted relative solvent accessibility; and (v) the features calculated from the primary sequence, including the collocation of AA pairs and terminus indicator. The contributions of these feature groups are assessed in two ways. First, we compare NSitePred with versions of our method where one type of input is excluded; see Supplementary Table S6. Second, we calculate the predictive quality of the models that take only one feature group as inputs, see Supplementary Table S7. The exclusion of PSSM profile and conservation scores leads to a larger decrease in the AUC and MCC than the exclusion of other input types, which suggests that the evolutionary information plays a key role in determination of the nucleotide-binding residues. On the other hand, the removal of the predicted relative solvent accessibility has the smallest impact to the predictive quality, i.e. we observe a decrease of 0.4% in AUC and 0.8% in MCC on average for five nucleotides on two datasets. Similar observations are made when using one feature group as inputs. The group that includes PSSM profile and conservation scores provides higher AUC values than the remaining groups. The BLAST-based features provide the highest MCC values. Moreover, we note that all feature groups, including the lowest scoring predicted relative solvent accessibility, are useful for the prediction of the nucleotide-binding residues, i.e. the AUC and MCC values are above 0.5 and 0 respectively, for each of the five feature groups.

3.4 Similarity between the prediction models for different types of nucleotides

Some of the considered nucleotides (e.g. ATP and ADP) have similar structures, which means that they may bind to the same pocket, while some other nucleotides (e.g. AMP and GTP) have less similar structures. Consequently, we assessed whether the prediction model for one nucleotide identifies the binding residues of other nucleotides. We use the prediction model for a given nucleotide, e.g. ATP, to predict the nucleotide-binding residues for sequences that interact with the other nucleotides, i.e. ADP, AMP, GTP and GDP. Dataset 2 is divided into five subsets with the ATP-, ADP-, AMP-, GTP- and GDP-binding chains, respectively. We calculate recall for each of the five prediction models and each subset of the nucleotide-binding chains; see Supplementary Table S8. Recall quantifies the fraction of the natively binding residues for a given nucleotide type that are predicted by a given model. The results show that the highest recall is obtained when a given model predicts chains that bind the corresponding nucleotide type, e.g. when NsitePred_ATP predicts the ATP-binding chains. This shows that the models are, as expected, specialized to predict binding for their 'own' nucleotide. We note that relatively high recall values are achieved by the ATP predictor when it predicts the ADP-binding residues and by the ADP predictor when it predicts the ATP-binding residues. The same is also observed when the GTP and GDP models predict the GDP- and GTP-binding residues, respectively. This suggests that structural similarity between the ligands, which impacts the similarity in their binding, is also observed among the predictions generated by NsitePred. However, the recall is substantially smaller when a given model is used to predict binding residues for nucleotides that are less similar, e.g. when NsitePred_ATP predicts the AMP-, GTP- and GDP-binding chains. This indicates that when the structures of nucleotides are different, the corresponding prediction models are also different, which motivates the development of consensus-based predictors.

Table 2. Comparison of the quality of the sequence-based prediction of the ATP, ADP, AMP, GDT, GTP and nucleotide-binding (indicated by all) residues between the NsitePred and the related predictors of nucleotide-binding residues, including ATPint and GTPbinder that predict ATP- and GTP-binding residues, respectively, method proposed by Firoz *et al.* (2011), and predictors based on the alignment (utilizing BLAST), conservation scoring (utilizing Rate4site), evolutionary profiles (utilizing PSSM and SVM classifier) and SVMPred (utilizing the same feature representation as NsitePred and SVM classifier) on Dataset 2

Type	Predictor	AUC		MCC		PREC	REC	SPEC	ACC
		value	sig.	value	sig.				
ATP	NsitePred	**0.875**		**0.476**		0.528	0.46	0.985	0.967
	SVMPred	0.868	+	0.451	+	0.587	0.367	0.991	0.969
	Rate4site	0.741	+	0.167	+	0.107	0.464	0.862	0.849
	BLAST	NA	+	0.422	+	**0.611**	0.31	**0.993**	**0.97**
	ATPint	0.606	+	0.066	+	0.051	0.512	0.66	0.655
	Firoz *et al.*	0.79	+	0.247	+	0.126	**0.714**	0.823	0.82
ADP	NsitePred	**0.893**		**0.512**		0.589	0.474	0.987	0.968
	SVMPred	0.886	+	0.5	+	0.68	0.388	0.993	**0.971**
	Rate4site	0.735	+	0.166	+	0.102	0.521	0.823	0.812
	BLAST	NA	+	0.376	+	**0.608**	0.249	**0.994**	0.966
	Firoz *et al.*	0.664	+	0.155	+	0.081	**0.669**	0.71	0.708
AMP	NsitePred	**0.876**		**0.501**		0.606	0.423	0.987	**0.969**
	SVMPred	0.87	+	0.478	+	**0.721**	0.335	0.994	0.967
	Rate4site	0.752	+	0.175	+	0.114	**0.52**	0.824	0.811
	BLAST	NA	+	0.339	+	0.504	0.255	0.989	0.959
	Firoz *et al.*	0.781	+	0.311	+	0.656	0.16	**0.996**	0.962
GDP	NsitePred	**0.867**		**0.576**		0.598	**0.585**	0.985	0.97
	SVMPred	0.855	+	0.553	+	**0.632**	0.511	0.988	**0.971**
	Rate4site	0.748	+	0.173	+	0.116	0.545	0.793	0.781
	BLAST	NA	+	0.454	+	0.593	0.372	0.99	0.967
	Firoz *et al.*	0.801	+	0.304	+	0.625	0.16	**0.996**	0.965
GTP	NsitePred	**0.909**		**0.64**		0.711	**0.604**	0.988	**0.969**
	SVMPred	0.887	+	0.602	+	0.783	0.485	0.993	**0.969**
	Rate4site	0.745	+	0.168	+	0.103	0.531	0.817	0.806
	BLAST	NA	+	0.539	+	**0.761**	0.403	**0.994**	0.966
	GTPbinder_seq	0.742	+	0.276	+	0.544	0.276	0.988	0.954
	GTPbinder_PSSM	0.822	+	0.418	+	0.597	0.321	0.989	0.957
	Firoz *et al.*	0.844	+	0.552	+	0.744	0.433	0.993	0.966
All	NsitePred	**0.905**		**0.48**		0.386	0.663	0.957	0.946
	SVMPred	0.899	+	0.455	+	0.374	0.62	0.958	0.945
	Rate4site	0.741	+	0.17	+	0.107	0.512	0.83	0.818
	BLAST	NA	+	0.423	+	**0.426**	0.468	**0.974**	**0.955**
	Firoz *et al.*	0.709	+	0.169	+	0.088	**0.7**	0.704	0.704

In the evaluation of nucleotide-binding residue, a residue is defined as a 'nucleotide-binding' if it interacts with any of the 5 nucleotides, and a residue is predicted as a 'nucleotide-binding' when a given method predicts that this residue interacts with any of the 5 nucleotides. The highest values for each ligand type and each quality index, including AUC, precision (PREC), recall (REC), specificity (SPEC), accuracy (ACC) and MCC, are set in bold. The significance of the differences between NsitePred and the other methods are measured for the AUC and MCC and they are given in the 'sig.' columns. The significance tests compare paired per-sequence prediction quality over a given benchmark dataset. The + and − mean that the NsitePred is statistically significantly better/worse with P-value < 0.01, and = means that results are not significantly different. The 'NA' means that the corresponding value could not be computed, i.e. BLAST generates only the binary predictions.

3.5 Evaluation per binding site

Besides the evaluation at the residue level, we investigate the quality of the predictions at the binding site level. A given binding site, which is made of residues that interact with the same molecule, is assumed to be correctly predicted if at least 50% of its residues are correctly predicted. We vary the per-residue precision between 0.05 and 0.8 (the number of correctly predicted binding sites is approximately 0 when the precision >0.8) with 0.05 step to control the number of FPs. This is performed by thresholding the predicted probabilities (we vary the threshold to obtain the binary predictions) for all methods except for the BLAST-based predictor, which is represented with one point that corresponds to its binary prediction. Supplementary Figure S4 shows that NsitePred correctly predicts $\sim$62% of the ADP-binding sites, 38% of the ATP-binding sites, 19% of the AMP-binding sites, 76% of the GDP-binding sites and 37% of the GTP-binding sites at the precision that equals 0.5, i.e. when half of the predicted binding residues are correct. To compare, the PSSM-profile based predictor correctly predicts only 6, 0, 0, 34 and 2% of the binding sites for the ADP, ATP, AMP, GDP and GTP, respectively, when considering the same precision. The Rate4site predictor cannot achieve such high precision for any of the five types of nucleotides, and thus we assume that its success rate equals 0. The ATPint and GTPbinder_seq also cannot correctly predict any sites at precision of 0.5, while the GTPbinder_PSSM correctly predicts $\sim$13% of the GTP-binding sites. When compared with the BLAST-based predictor at the same precision, the NsitePred correctly finds 5–15% more binding sites. Overall, the results indicate that the NsitePred captures more binding sites than the other predictors, especially at the higher precision rates.

3.6 Impact of the degree of spread of the binding residues in the protein chain

Some nucleotide binding sites consist of a single segment in the protein chain, e.g. the p-loop motif GXXXXGKS(T)T, while other sites are composed of binding residues that are sparsely distributed over the sequence. We study the relation between this degree of the spread of the binding residues in the chain and the predictive quality. We quantify this spread/clustering of the binding residues using a spread index that reflects the average number of non-binding residues between the consecutive binding residues in the chain, and that equals zero when a given site consists of a single segment of the consecutive binding residues. In other words, larger spread values correspond to sites that are composed of the binding residues that cover a longer fragment in the input sequence, relative to the total number of the binding residues in a given site. A detailed definition of this index is provided in the Supplementary Material.

We sorted all binding sites for a given nucleotide type in the ascending order according to their spread index values, and we divided them into five equally sized subsets where the first subset contains 20% of sites with the lowest spread. The average spread values for each subset and the corresponding predictive quality for NsitePred calculated based on the 5-folds cross validation on Dataset 1 are shown in Figure 2. Figure 2A shows the average precision (fraction of correct prediction among the predicted binding residues) at the recall that equals 0.5, while Figure 2B gives the average recall (fraction of correctly predicted native binding residues) at the precision that equals 0.5. The results show that both precision and recall decline with the increasing spread value, and that this

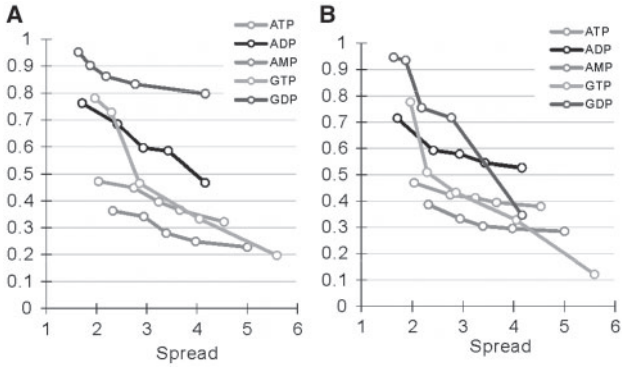

Fig. 2. Relation between the predictive quality (y-axis) and the spread index values (x-axis). The binding sites for a given nucleotide type, which are sorted in the ascending order based on their spread index values, are divided into five equally sized subsets where the first subset (the left-most point) contains 20% of sites with the of the lowest spread, and the fifth subset (the right-most point) with the 20% of sites with the highest values. (**A**) shows the average precision (over the sites in a given subset) at the recall $=0.5$. (**B**) shows the average recall at the precision $=0.5$.

trend is independent of the nucleotide type. The NsitePred performs very well for compact sites, i.e. sites that include residues that are clustered close in the sequence, and its quality declines when the binding residues are spread over a longer fragment of the protein chain. Moreover, this relation also explains the differences in the predictive quality for different nucleotide types. The average spread values for the binding sites for the GDP, ADP, ATP, GTP and AMP are 2.53, 2.93, 3.25, 3.35 and 3.53, respectively. The Pearson correlation coefficients between these spread values and the corresponding AUC and MCC values achieved by NsitePred, see Table 1, equal -0.96 and -0.98, respectively.

3.7 Sequence-derived hallmarks of nucleotide-binding residues

The significant improvements in the quality of the prediction of the binding residues for the five considered nucleotides between the NsitePred and the PSSM profile-based predictions (denoted as PSSM+SVM), see Tables 1, suggest that the increased quality stems from the use of the novel inputs proposed in this work. This means that the nucleotide-binding residues could be characterized using the information concerning their conservation and the predicted secondary structure, RSA, and dihedral angles. We analyze features used by the NsitePred to find the corresponding sequence-derived markers of the nucleotide-binding residues.

We focus on the features that were selected for at least three nucleotide types; they are listed in Supplementary Table S9. We observe that the selected features that are based on the predicted secondary structure, RSA, and psi angles are biased to positions in the sequence that are toward the N-terminus from the predicted residue. We investigate this asymmetry (the lack of use of the positions toward the C-terminus at the same position (relative to the predicted residues) computed for all native nucleotide-binding residues and the non-binding residues, respectively, i.e. ratio=average value of a given feature for the nucleotide-binding residues divided by the average for the non-binding residues. The value close to 1 indicates that similar

average values are observed for the binding and the non-binding residues, and thus the corresponding feature at this position does not differentiate between these two types of residues. The ratios along the 17 positions of the sliding window used by the NsitePred for the probabilities of secondary structures, RSA values, dihedral angles and the three conservation scores are shown in Figure 3. As our feature selection suggests, the plots for the secondary structure, dihedral angles and RSA are asymmetric, which is in contrast to the conservation scores that are symmetric. We note the particularly high ratios for the predicted probabilities of strands at the positions that are 4–5 residues toward the N-terminus. These ratios show that the nucleotide-binding residues are characterized by over twice higher probabilities of the predicted strand residues for these positions when compared with the non-binding residues. Moreover, positions toward the C-terminus show ratios relatively close to 1, i.e. between 0.826 and 1.001. We also observe that the helix is less likely to occur toward the N-terminus when compared with positions toward the C-terminus, which coincides with the above preference toward the strands. Similarly (as expected), the ratios for the phi and psi angles follow the pattern of the secondary structures, although they vary in a smaller range, e.g. ratios for the psi angles vary between 0.926 and 1.196, with the larger values only toward the N-terminus. The RSA values are smaller toward the N-terminus and close to 1 toward the C-terminus, which explains the bias toward the former positions among the selected features. The plot indicates that residues located near by and toward the N-terminus from the nucleotide-binding residues are less likely to be solvent exposed. The ratios for the three conservation scores are symmetrically distributed around the central residue. Their largest values are at the central position, which indicates that the nucleotide-binding residues are more conserved than the non-binding residues. These three plots also reveal that the residues at the adjacent positions have smaller ratios, which means that the nucleotide-binding residues are flanked by residues with a smaller degree of conservation.

We also note that several features extracted from the PSSM profile (including the aggregations using certain amino acid groups) and certain collocated amino acid pairs are included among the selected features, and thus they can be used to formulate sequence-derived hallmarks of the nucleotide-binding residues. The PSSM profile-based features are likely correlated with the formation of certain secondary structure types, e.g. the scores aggregated for the hydrophobic residues is associated with the formation of strands and coils. Three amino acid pairs, GXXXS, GXG and GXS, where 'X' indicates a wild card residue (any amino acid type) and where the right-most residue is located at the center of a sliding window, are found to be strong markers for the nucleotide-binding residues. These collocated pairs are related to the p-loop motif, GXXXXGKS(T), which is characteristic to the interactions with ATP (Saraste *et al*., 1990).

3.8 Case study

We demonstrate the predictions generated by the NsitePred for the chain A of cell division control protein 6 (PDB code: 1FNN) that interacts with the ADP. The predictive quality of the considered methods for this target is similar to their average quality on the entire dataset. The native ADP-binding residues, the binary predictions, and the probabilities predicted by NsitePRED, Rate4site, BLAST and PSSM+SVM methods are shown in Figure 4. The binding

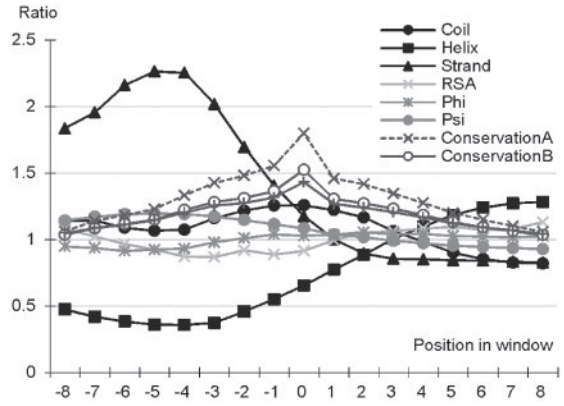

Fig. 3. The ratios, which are calculated as the average of values of a given feature for the nucleotide-binding residues divided by the average for the non-binding residues, at the 17 positions in the sliding window used by NsitePred. The ratios are calculated for the predicted secondary structures (helix, strand and coil), RSA, dihedral angles (phi and psi), and the three conservation scores based on the Shannon entropy (conservation A) and formulas proposed in (Wang and Samudrala, 2006) (conservation B), and in (Capra and Singh, 2007) (conservation C). The *x*-axis shows the positions in the sequence relative to the predicted residues, which is at 0.

residues are clustered into four segments, which include positions 14–22, 54–59, 193–205 and 226–230. The second binding segment consists of six consecutive binding residues while the other three segments include interspersed binding and non-binding residues. The NsitePred correctly predicts 15 out of the 17 ADP-binding residues and produces one FP. The SVMPred generates the same binary predictions as the NsitePred method. The PSSM+SVM method and the BLAST-based predictor predict 5 binding residues and they also produce 2 and 0 FPs, respectively. The Rate4site correctly finds 15 binding residues, but it also generates 49 FPs. The predicted binding residues on the protein surface and the structure of the nucleotide are given in the Supplementary Figure S5. We also observe several of the abovementioned hallmarks of the nucleotide-binding residues. The second binding segment, GTGKTV, includes the collocated amino acid pair GXG. The GXXR and LXXR pairs, which are significantly associated with the ADP-binding residues, are found in the first and the fourth binding segments, respectively. Moreover, the conservation scores for the residues in these pairs are below -0.85, which is the threshold to binarize the predictions from the Rate4site. Consequently, the above collocated amino acid pairs and their conservation scores explain why the first and the fourth binding segment are captured by the NsitePred and the Rate4site predictors; the other two predictors fail to find them since they do not consider this information. We note that strands are predicted 3–7 residues left (toward the N-terminus) from the binding residues that make up the second and third segments; these positions are also characterized by relatively low probability of prediction of the helical conformation.

4 CONCLUSION

The NsitePred is a collection of five accurate sequence-based predictors that identify binding residues for the five most populated nucleotides in the PDB, including ATP, ADP, AMP, GTP and GDP. Empirical results demonstrate that NsitePred outperforms

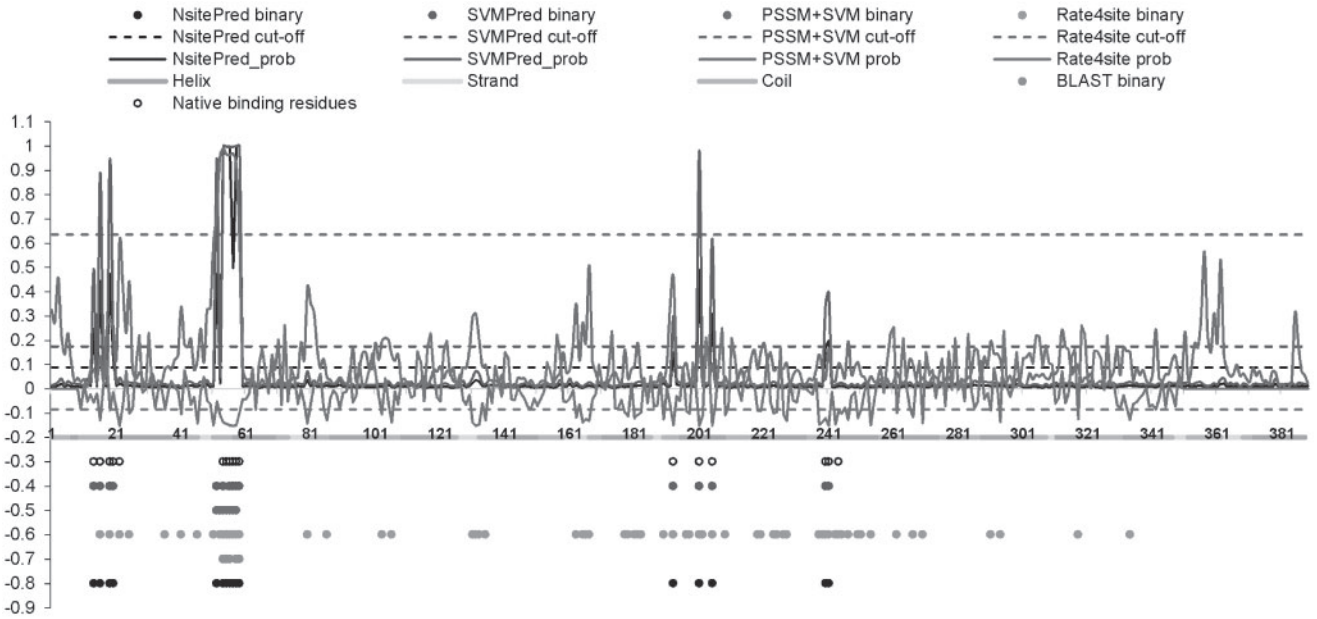

Fig. 4. Comparison of predictions for chain A of cell division control protein 6 (PDB id: 1FNN). Plots at the top show the predicted probabilities for PSSM+SVM (in red), Rate4site (green) and NsitePred (blue). Conservation scores generated by the Rate4site were divided by 10 to fit the figure. The dotted horizontal lines denote the cut-offs used to binarize the probabilities. The corresponding binary predictions are shown using dots (one dot per residue) at the bottom. Black dots denote native ADP-binding residues, and blue, red, green and gray denote predictions from NsitePred, PSSM+SVM, Rate4site and BLAST, respectively. The secondary structure is shown in horizontal line below the *x*-axis with strands in yellow, helices in light green and coil in orange.

the existing ADPint and GTPbinder methods, as well as solutions based on the sequence alignment and residue conservation scoring. The favorable predictive quality stems from the usage of novel custom-designed input features that are based on the sequence, the sequence-derived evolutionary profiles, the sequence-predicted structural descriptors and the BLAST-based alignment. Our study shows that NsitePred performs particularly well for the binding sites in which the binding residues are clustered close together in the sequence. Analysis of the features used in the predictive model reveals several interesting hallmarks of the nucleotide-binding residues, which are related to the arrangement of secondary structures, dihedral angles and certain amino acid pairs in the specific neighboring positions in the sequence. The NsitePred is implemented as a web server that is available at http://biomine.ece.ualberta.ca/nSITEpred/.

ACKNOWLEDGEMENTS

We thank reviewers for their constructive comments that helped in improving both the methodology and presentation of the results.

Funding: NSERC Canada (to L.K., in part); the iCORE and Alberta Ingenuity scholarship (to K.C., in part); the Izaak Walton Killam Memorial scholarship (to M.J.M., in part).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Ashkenazy,H. *et al*. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res*., **38**, W529–W533.

Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.

Chauhan,J.S. *et al*. (2009) Identification of ATP binding residues of a protein from its primary sequence. *BMC Bioinformatics*, **10**, 434.

Chauhan,J.S., *et al*. (2010) Prediction of GTP interacting residues, dipeptides and tripeptides in a protein from its evolutionary information. *BMC Bioinformatics*, **11**, 301.

Chen,K., *et al*. (2007) Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. *BMC Struct Biol*., **7**, 25.

Chen,K., *et al*. (2009) Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. *J. Comput. Chem*., **30**, 163–172.

Chen,K. and Kurgan,L. (2009) Investigation of atomic level patterns in protein-small ligand interactions. *PLoS ONE*, **4**, e4473.

Denessiouk,K.A. and Johnson,M.S. (2000) When fold is not important: a common structural framework for adenine and AMP binding in 12 unrelated protein families. *Proteins*., **38**, 310–326.

Dessailly,B.H. *et al*. (2008) LigASite—a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res*., **36**, D667–D673.

Fan,R.E. *et al*. (2005) Working set selection using second order information for training SVM. *J. Mach. Learn Res.*, **6**, 1889–1918.

Fan,R.E. *et al*. (2008) LIBLINEAR: a library for large linear classification. *J. Mach. Learn. Res.*, **9**, 1871–1874.

Faraggi,E. *et al*. (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins*., **74**, 847–856.

Firoz,A. *et al*. (2011) Residue propensities, discrimination and binding site prediction of adenine and guanine phosphates. *BMC Biochem*., **12**, 20.

Tate,R.F. (1954) Correlation between a discrete and a continuous variable. Point-biserial correlation. *Annals of Mathematical Statistics*, **25**, 603–607

Goto,S. *et al*. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res*., **30**, 402–404.

Larkin,M.A. *et al*. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Luscombe,N.M. *et al*. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res*., **29**, 2860–2874.

Mao,L., *et al*. (2004) Molecular determinants for ATP-binding in proteins: a data mining and quantum chemical analysis. *J. Mol. Biol*., **336**, 787–807.

McGuffin,L.J. *et al*. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.

Moodie,S.L. *et al*. (1996) Protein recognition of adenylate: an example of a fuzzy recognition template. *J. Mol. Biol*., **263**, 486–500.

Nobeli,I. *et al*. (2001) On the molecular discrimination between adenine and guanine by proteins. *Nucleic Acids Res*., **29**, 4294–4309.

Pupko,T. *et al*. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18** (Suppl. 1), S71–77.

Saito,M. *et al*. (2006) An empirical approach for detecting nucleotide-binding sites on proteins. *Protein Eng. Des. Sel*., **19**, 67–75.

Saraste,M. *et al*. (1990) The P-loop - A common motif in ATP-binding and GTP-binding proteins. *Trends Biochem Sci*., **15**, 430–434.

Senes,A. *et al*. (2000) Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J. Mol. Biol.,* **296**, 921–936.

Walker,J.E. *et al*. (1982) Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide-binding fold. *EMBO J.,* **1**, 945–951.

Wang,G. and Dunbrack,R.L. Jr. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Wang,K. and Samudrala,R. (2006) Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinformatics*, **7**, 385.