Supplement for the manuscript entitled

"Prediction and Analysis of Nucleotide Binding Residues Using Sequence and Sequence-derived Structural Descriptors"

by Ke Chen, Marcin Mizianty and Lukasz Kurgan

FEATURE-BASED SEQUNCE REPRESENTATION

For each residue in the input sequence we compute the following features using the sliding window of size 17:

- *Predicted secondary structure* generated by PSIPRED (1). We use probabilities of the 3 secondary structure states (helix/strand/coil) for each residue in the window (total of 3×17 = 51 features).
- *Predicted relative solvent accessibility* generated by Real-SPINE3 (2). We use the real values, which quantify the fraction of the surface area of a given residue that is accessible to the solvent, for each residue in the window (total of 17 features).
- *Predicted dihedral angles* generated by Real-SPINE3 (2). We utilize two real values, which represent *phi* (involving the backbone atoms C'-N-C^{α}-C') and *psi* (involving the backbone atoms N-C^{α}-C'-N) angles, for each residue in the window (total of 2×17 = 34 features).
- *PSSM profile* generated by PSIBLAST (3) with default parameters using the NCBI non-redundant database. We normalize these inputs with $1/(1+2^{-x})$, where x is the raw value from the PSSM profile; this transformation is commonly used in the secondary structure prediction. For a window centered at R_i residue at i^{th} position, we calculate 17×20 features $f_{i+k,j}$ where $k = -8, -7, \ldots, 7, 8$ is the index of the position in the window and $j = 1, 2, \ldots, 20$ is the index of the PSSM column. We average the values to the left and to the right of the central residue $g_{i+z,j} = (f_{i+z,j} + f_{i-z,j})/2$ where $z = 0, 1, \ldots, 8$. As a result, the original 17×20 values are transformed to 9×20 values (total of 9×20 = 180 features).
- *AA groups* including hydrophobic residues (Ala, Cys, Ile, Leu, Met and Val), negatively charged (Asp and Glu), positively charged (His, Lys, Arg) and carboxamide-containing amino acids (Asn and Gln), are used to aggregate the normalized and averaged 9×20 PSSM values. The PSSM values for the AA types from a given group for a given position $z = 0, 1, \ldots, 8$ are averaged (total of 9×4 = 36 features).
- *Terminus indicator* is set to 1 for the first and the last 3 residues in the sequence, and it equals to 0 for the other positions (total of 17 features).
- *Secondary structure segment indicators* for helix/strand/coil predictions from PSIPRED on both sides of the window are calculated. If at least 4 / 3 consecutive residues on the left / right side of the window are predicted as helix (strand), then we set the helix (strand) indicator as 1 for the left/right side. If helix and strand indicators equal 0, then the coil indicator is set as 1 (total of

3(helix/strand/coil)×2(left/right) = 6 features).

- *Residue conservation scores* are calculated using the PSSM values for each position based on the Shannon entropy, and based on using two formulas proposed in (4, 5) which incorporate the background frequency of the amino acids (total of 3×17 = 51 features).
- *Collocation of AA pairs* is calculated for the residues in the window. This is motivated by results for the membrane proteins where certain amino acid pairs are over-represented (6). Similarly, several sequence motifs occur frequently in the ATP binding sites. To accommodate for mutations in these motifs, we use collocated AA pairs (pairs with gaps) to characterize these motifs. We only consider pairs formed between the central residue in the window and another residue up to 5 positions away. This results in 20×20×10=4000 frequencies (for 20 AA types and 10 positions; 5 on each side). The same as in (6), we calculated *p*-values that indicate the significance of the association between a given amino acid pair and the nucleotide binding annotations. A low *p*-value indicates a low probability that the association between the corresponding amino acid pair and the nucleotide binding annotations is a coincidence. When analyzing 4000 randomly distributed variables, we expect to observe by chance one instance of a difference from expected value with significance $p < 0.00025$ (1/4000). We exclude the amino acid pairs with $p \geq 10^{-6}$, since based on the Engelman's study (6) their association with ATP-binding event would be random.

DEFINITION OF THE SPREAD INDEX

We hypothesize that the difficulty of the sequence-based prediction of the nucleotide binding sites depends on the degree to which the corresponding binding residue are spread over the sequence. The sites that are composed of a single segment of consecutive binding residues should be easier to predict than the sites for which the binding residues are sparsely distributed over a large, relative to the total number of binding residues, stretch of the sequence. We quantify the corresponding degree of "difficulty" using a spread index which follows two rules: 1) a site that consists of a single segment of consecutive binding residues has the spread equal 0; 2) the spread increases with the number of the non-binding residues that are introduced between the binding residues. The spread index is defined as follows

$$spread = \frac{\sum_{i=1}^{n-1} gap(R_i, R_{i+1})}{n-1}$$

where R_i and R_{i+1} are the i^{th} and $(i+1)^{th}$ binding residue, respectively, in a given binding site that consists of n residues. Given that the binding residues are sorted by their residue number in the protein sequence, the $gap(R_i, R_{i+1})$ is defined as

$$gap(R_i, R_{i+1}) = \begin{cases} N_{i+1} - N_i - 1 & (\text{if } N_{i+1} - N_i < 13) \\ 12 & (\text{otherwise}) \end{cases}$$

where N_i and N_{i+1} are the residue numbers in the protein chain of the R_i and R_{i+1} binding residues, respectively The "$N_{i+1} - N_i - 1$" quantifies the number of the non-binding residues between R_i and R_{i+1}. Moreover, we assume that a given pair of consecutive binding residues that are separated by 12 or more non-binding residues is not likely to form local interactions in the fold and thus the corresponding $gap(R_i, R_{i+1})$ value is rounded down to 12. This cut-off threshold is based on the definition of the long-range interactions, which are defined as contacts formed between residues that are at least 12 positions away in sequence (7).

REFERENCES

1. McGuffin, L.J., Bryson, K. and Jones, D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404-5.
2. Faraggi, E., Xue, B. and Zhou, Y. (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**, 847-56.
3. Altschul, S.F., Madden, T.L., Schäffer, A.A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389-402.
4. Wang, K. and Samudrala, R. (2006) Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinformatics*, **7**, 385.
5. Capra, J.A. and Singh, M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875-82.
6. Senes, A., Gerstein, M. and Engelman, D.M. (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**, 921-36.
7. Tegge, A.N., Wang, Z., Eickholt, J. and Cheng, J. (2009) NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Res*., **37**, W515-8.
8. Firoz, A., et al. (2011) Residue propensities, discrimination and binding site prediction of adenine and guanine phosphates. *BMC Biochem*. 2011; **12**: 20.
9. Chauhan, J.S., et al. (2009) Identification of ATP binding residues of a protein from its primary sequence. *BMC Bioinformatics*, 10, 434.
10. 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

Table 1. The optimized parameters of SVM models used by the NsitePred method together with the corresponding values of the best achieved AUCs based on the 5-fold cross-validation on Dataset 1 for each type of nucleotides.

Nucleotide type	Polynomial kernel			RBF kernel		
	degree	complexity constant *C*	best achieved AUC	gamma	complexity constant *C*	best achieved AUC
ATP	2	4	0.845	0.5	4	0.861
ADP	1.5	2	0.880	0.5	4	0.893
AMP	2	2	0.823	0.5	2	0.829
GTP	2	4	0.896	0.5	4	0.910
GDP	1.5	2	0.832	0.5	2	0.844

Table 2. Comparison of the sensitivities (SENS) and specificities (SPEC) achieved by NsitePred and the competing predictors on Dataset 1. The cutoff thresholds, which are used to binarize the predictions, are chosen such that the sensitivities (shown in columns) achieved by a given method are between 0.1 and 0.9 with step of 0.1.

Ligand Type	Predictor	SENS	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ATP	NsitePred	SPEC	0.999	0.997	0.993	0.988	0.972	0.941	0.873	0.762	0.443
	SVMPred	SPEC	0.999	0.998	0.993	0.983	0.961	0.922	0.857	0.741	0.478
	Rate4site	SPEC	0.981	0.955	0.926	0.889	0.839	0.765	0.672	0.543	0.350
	BLAST	SPEC	NA	0.993	NA	NA	NA	NA	NA	NA	NA
	ATPint	SPEC	0.951	0.893	0.831	0.765	0.681	0.588	0.48	0.348	0.198
ADP	NsitePred	SPEC	0.999	0.999	0.996	0.994	0.992	0.980	0.939	0.842	0.563
	SVMPred	SPEC	1.000	0.999	0.998	0.996	0.988	0.969	0.919	0.793	0.515
	Rate4site	SPEC	0.980	0.951	0.917	0.875	0.822	0.76	0.674	0.557	0.370
	BLAST	SPEC	NA	NA	0.994	NA	NA	NA	NA	NA	NA
AMP	NsitePred	SPEC	0.997	0.993	0.989	0.972	0.94	0.895	0.785	0.624	0.413
	SVMPred	SPEC	0.999	0.996	0.985	0.961	0.925	0.863	0.790	0.630	0.356
	Rate4site	SPEC	0.981	0.961	0.926	0.874	0.83	0.771	0.674	0.565	0.383
	BLAST	SPEC	0.992	NA	NA	NA	NA	NA	NA	NA	NA
GDP	NsitePred	SPEC	0.999	0.999	0.999	0.998	0.995	0.993	0.976	0.888	0.513
	SVMPred	SPEC	1.000	1.000	0.999	0.998	0.997	0.992	0.968	0.877	0.612
	Rate4site	SPEC	0.964	0.94	0.909	0.873	0.829	0.753	0.664	0.525	0.318
	BLAST	SPEC	NA	NA	NA	0.995	NA	NA	NA	NA	NA
GTP	NsitePred	SPEC	0.999	0.999	0.998	0.994	0.985	0.94	0.839	0.678	0.408
	SVMPred	SPEC	1.000	1.000	0.999	0.993	0.970	0.907	0.802	0.627	0.370
	Rate4site	SPEC	0.963	0.939	0.913	0.871	0.831	0.778	0.693	0.555	0.349
	BLAST	SPEC	NA	NA	0.994	NA	NA	NA	NA	NA	NA
	GTPbinder_seq	SPEC	0.933	0.837	0.762	0.670	0.565	0.455	0.361	0.23	0.115
	GTPbinder_PSSM	SPEC	0.999	0.997	0.988	0.962	0.922	0.852	0.738	0.594	0.383

Table 3. Comparison of the sensitivities (SENS) and specificities (SPEC) achieved by NsitePred and the competing predictors on Dataset 2. The cutoff thresholds, which are used to binarize the predictions, are chosen such that the sensitivities (shown in columns) achieved by a given method are between 0.1 and 0.9 with step of 0.1.

Ligand Type	Predictor	SENS	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ATP	NsitePred	SPEC	1.000	0.999	0.996	0.992	0.978	0.948	0.898	0.816	0.515
	SVMPred	SPEC	1.000	0.999	0.996	0.987	0.965	0.935	0.877	0.763	0.500
	Rate4site	SPEC	0.978	0.956	0.932	0.883	0.836	0.767	0.657	0.576	0.378
	BLAST	SPEC	NA	NA	0.993	NA	NA	NA	NA	NA	NA
	ATPint	SPEC	0.950	0.873	0.814	0.749	0.662	0.565	0.410	0.266	0.133
	Firoz et al.[8]	SPEC	0.943	0.941	0.939	0.922	0.894	0.824	0.824	0.617	0.378
ADP	NsitePred	SPEC	1.000	0.999	0.993	0.992	0.984	0.96	0.924	0.824	0.615
	SVMPred	SPEC	0.999	0.999	0.998	0.991	0.976	0.954	0.907	0.805	0.638
	Rate4site	SPEC	0.982	0.965	0.924	0.893	0.836	0.776	0.679	0.544	0.347
	BLAST	SPEC	NA	0.994	NA	NA	NA	NA	NA	NA	NA
	Firoz et al.[8]	SPEC	0.795	0.793	0.789	0.779	0.765	0.736	0.649	0.488	0.264
AMP	NsitePred	SPEC	0.999	0.998	0.989	0.987	0.963	0.942	0.875	0.796	0.568
	SVMPred	SPEC	0.999	0.999	0.996	0.986	0.962	0.933	0.881	0.803	0.521
	Rate4site	SPEC	0.973	0.934	0.916	0.867	0.836	0.781	0.696	0.547	0.352
	BLAST	SPEC	NA	NA	0.989	NA	NA	NA	NA	NA	NA
	Firoz et al.[8]	SPEC	0.999	0.987	0.960	0.798	0.798	0.798	0.689	0.554	0.428
GDP	NsitePred	SPEC	1.000	0.999	0.998	0.991	0.990	0.984	0.899	0.628	0.324
	SVMPred	SPEC	1.000	0.999	0.998	0.991	0.989	0.955	0.894	0.702	0.325
	Rate4site	SPEC	0.967	0.953	0.916	0.867	0.824	0.762	0.667	0.557	0.374
	BLAST	SPEC	NA	NA	NA	0.99	NA	NA	NA	NA	NA
	Firoz et al.[8]	SPEC	1.000	0.991	0.955	0.943	0.832	0.832	0.748	0.668	0.395
GTP	NsitePred	SPEC	1.000	1.000	0.999	0.994	0.993	0.988	0.969	0.835	0.636
	SVMPred	SPEC	0.999	0.999	0.999	0.997	0.992	0.975	0.93	0.785	0.608
	Rate4site	SPEC	0.968	0.944	0.905	0.873	0.832	0.753	0.654	0.532	0.321
	BLAST	SPEC	NA	NA	NA	0.994	NA	NA	NA	NA	NA
	GTPbinder_seq	SPEC	0.999	0.994	0.978	0.945	0.868	0.758	0.626	0.410	0.255
	GTPbinder_PSSM	SPEC	0.998	0.995	0.990	0.974	0.930	0.876	0.819	0.683	0.481
	Firoz et al.[8]	SPEC	1.000	0.999	0.996	0.994	0.946	0.914	0.830	0.715	0.387

Table 4. Comparison between NsitePred, ATPint, and GTPbinder on the original datasets used to develop 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 [9, 10].

Type	Predictor	AUC	MCC	PREC	REC	SPEC	ACC
ATP	NsitePred	0.86	0.56	0.82	0.72	0.84	0.78
	ATPint	NA	0.51	0.78	0.70	0.80	0.75
GTP	NsitePred	0.93	0.73	0.95	0.75	0.96	0.85
	GTPbinder_PSSM	0.92	0.70	0.93	0.73	0.95	0.84

The tests from Table 4 are run utilizing the protocols from [9, 10]; there protocols were learned based on personal communication with the authors. We first annotate positive samples (binding residues) and negative samples (non-binding residues) in the datasets from [9, 10]. 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. We note that the authors in [9, 10] assumed that users would be interested in evaluation of balanced predictions (equal number of binding and non-binding residues) and thus they designed (parameterized, etc.) their methods for this type of test. This is in contrast to NSitePred which was designed for a prediction on full protein chains. Moreover, the datasets are split per residue, which means that residues from the same chain could be included in the training and test sets.

Table 5. The error rates of NsitePred and the other considered predictors on protein chains from Dataset 3 that do not interact with nucleotides. The error rate is defined as the ratio between the number of false positives and the total number of instances.

Predictor	Error rate					
	ATP	ADP	AMP	GTP	GDP	5 types of nucleotides
NsitePred	0.48%	1.15%	0.76%	0.93%	0.67%	2.67%
SVMPred	0.36%	0.86%	0.58%	0.75%	0.53%	2.13%
BLAST	0.28%	0.59%	0.43%	0.51%	0.34%	1.32%
ATPint	20.1%	NA	NA	NA	NA	20.1%
GTPbinder_seq	NA	NA	NA	3.47%	NA	3.47%
GTPbinder_PSSM	NA	NA	NA	2.94%	NA	2.94%

Table 6. The predictive quality of NsitePred and versions of our method where one type of input is excluded. The inputs are categorized into 5 groups: (1) the BLAST-based prediction; (2) predicted secondary structure (including the secondary structure segment indicator) and dihedral angles; (3) PSSM profile and conservation scores; (4) predicted relative solvent accessibility; and (5) the features calculated from the primary sequence, including the collocation of AA pairs and ter-minus indicator. The evaluation is performed based on the 5-folds cross validation on the Dataset 1 and based on the test on Dataset 2 when the models are build using Dataset 1.

Dataset	Input type	ATP		ADP		AMP		GTP		GDP	
		AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC
Dataset 1	NsitePred	0.861	0.46	0.893	0.572	0.829	0.377	0.844	0.562	0.91	0.675
	Excluding BLAST	0.854	0.433	0.885	0.555	0.82	0.36	0.836	0.551	0.905	0.655
	Excluding predicted secondary structure and dihedral angles	0.852	0.435	0.881	0.546	0.823	0.364	0.835	0.549	0.902	0.651
	Excluding PSSM profile and conservation scores	0.836	0.324	0.853	0.376	0.805	0.272	0.823	0.418	0.886	0.527
	Excluding predicted relative solvent accessibility	0.858	0.449	0.89	0.563	0.826	0.372	0.838	0.557	0.906	0.658
	Excluding collocation of AA pairs and terminus indicator	0.854	0.43	0.883	0.549	0.822	0.356	0.83	0.537	0.902	0.646
Dataset 2	NsitePred	0.875	0.476	0.893	0.512	0.876	0.501	0.909	0.64	0.867	0.576
	Excluding BLAST	0.868	0.451	0.886	0.5	0.87	0.478	0.887	0.602	0.855	0.553
	Excluding predicted secondary structure and dihedral angles	0.87	0.458	0.887	0.495	0.87	0.474	0.892	0.613	0.857	0.556
	Excluding PSSM profile and conservation scores	0.838	0.359	0.856	0.373	0.851	0.405	0.867	0.465	0.833	0.426
	Excluding predicted relative solvent accessibility	0.872	0.467	0.891	0.508	0.872	0.493	0.904	0.626	0.861	0.57
	Excluding collocation of AA pairs and terminus indicator	0.867	0.453	0.887	0.501	0.872	0.487	0.89	0.612	0.857	0.558

Table 7. The predictive quality of the NsitePred versions that use only one feature group as inputs. The features are categorized into 5 groups: (1) the BLAST-based prediction; (2) predicted secondary structure (including the secondary structure segment indicator) and dihedral angles; (3) PSSM profile and conservation scores; (4) predicted relative solvent accessibility; and (5) the features calculated from the primary sequence, including the collocation of AA pairs and ter-minus indicator. The evaluation is performed based on the 5-folds cross validation on the Dataset 1 and based on the test on Dataset 2 when the models are build using Dataset 1.

Dataset	Input type	ATP		ADP		AMP		GTP		GDP	
		AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC
Dataset 1	BLAST	NA	0.359	NA	0.439	NA	0.222	NA	0.461	NA	0.564
	Predicted secondary structure	0.752	0.181	0.768	0.211	0.705	0.128	0.762	0.212	0.838	0.312
	Predicted relative solvent accessibility	0.628	0.073	0.627	0.071	0.649	0.08	0.594	0.063	0.704	0.133
	PSSM profile and conservation	0.833	0.28	0.828	0.276	0.796	0.211	0.816	0.328	0.864	0.41
	Collocation of AA pairs and terminus indicator	0.682	0.264	0.703	0.318	0.587	0.162	0.685	0.353	0.773	0.568
Dataset 2	BLAST	NA	0.422	NA	0.376	NA	0.339	NA	0.539	NA	0.454
	Predicted secondary structure	0.765	0.193	0.749	0.206	0.751	0.188	0.786	0.224	0.778	0.226
	Predicted relative solvent accessibility	0.632	0.076	0.615	0.07	0.664	0.087	0.635	0.081	0.625	0.073
	PSSM profile and conservation	0.838	0.292	0.823	0.266	0.829	0.283	0.856	0.378	0.832	0.281
	Collocation of AA pairs and terminus indicator	0.705	0.277	0.682	0.284	0.674	0.171	0.713	0.352	0.696	0.267

Table 8. The recall of different NsitePred models (e.g., the NsitePred_ATP model predicts the ATP-binding residues) on subsets of chains that bind specific nucleotide types (e.g., 'ATP chains' indicates a subset of chains that interacts with ATP) measured on Dataset 2. Recall is defined as TP/(TP+FN), where TP (FN) is the number of residues that natively bind to a given nucleotide type that were (were not) predicted as nucleotide-binding by the corresponding model.

	NsitePred_ATP	NsitePred_ADP	NsitePred_AMP	NsitePred_GTP	NsitePred_GDP
ATP chains	**0.460**	0.419	0.203	0.131	0.106
ADP chains	0.416	**0.474**	0.234	0.151	0.113
AMP chains	0.232	0.225	**0.423**	0.065	0.073
GTP chains	0.167	0.184	0.067	**0.604**	0.56
GDP chains	0.141	0.181	0.083	0.461	**0.585**

Table 9. List of features that were selected for at least 3 types of nucleotides.

Feature group	Description of individual features	Value
Predicted secondary structure	Probability of prediction as coil at the position 1 residue left from the central residue	Real value
	Probability of prediction as helix at the position 7 residues left from the central residue	Real value
	Probability of prediction as helix at the position 6 residues left from the central residue	Real value
	Probability of prediction as helix at the position 5 residues left from the central residue	Real value
	Probability of prediction as helix at the position 4 residues left from the central residue	Real value
	Probability of prediction as helix at the position 3 residues left from the central residue	Real value
	Probability of prediction as helix at the position 1 residue left from the central residue	Real value
	Probability of prediction as helix at the central position	Real value
	Probability of prediction as strand at the position 8 residues left from the central residue	Real value
	Probability of prediction as strand at the position 7 residues left from the central residue	Real value
	Probability of prediction as strand at the position 4 residues left from the central residue	Real value
	Probability of prediction as strand at the position 3 residues left from the central residue	Real value
Predicted relative solvent accessibility	Predicted RSA value for the position 4 residues left from the central residue	Real value
	Predicted RSA value for the position 3 residues left from the central residue	Real value
	Predicted RSA value for the position 2 residues left from the central residue	Real value
	Predicted RSA value for the position 1 residue left from the central residue	Real value
Predicted Psi angle	Predicted Psi angle for the position 7 residues left from the central residue	Real value
	Predicted Psi angle for the position 6 residues left from the central residue	Real value
	Predicted Psi angle for the position 5 residues left from the central residue	Real value
	Predicted Psi angle for the position 4 residues left from the central residue	Real value
	Predicted Psi angle for the position 3 residues left from the central residue	Real value
	Predicted Psi angle for the position 2 residues left from the central residue	Real value
Conservation score A calculated by Shannon entropy	The conservation score A for the position 5 residues left from the central residue	Real value
	The conservation score A for the position 4 residues left from the central residue	Real value
	The conservation score A for the position 3 residues left from the central residue	Real value
	The conservation score A for the central position	Real value
	The conservation score A for the position 2 residues right from the central residue	Real value
	The conservation score A for the position 3 residues right from the central residue	Real value
	The conservation score A for the position 4 residues right from the central residue	Real value
Conservation score B calculated by formula in ref. 4	The conservation score B for the position 3 residues left from the central residue	Real value
	The conservation score B for the position 1 residue left from the central residue	Real value
	The conservation score B for the position 2 residues right from the central residue	Real value
Conservation score C calculated by formula in ref. 5	The conservation score C for the position 5 residues left from the central residue	Real value
	The conservation score C for the position 4 residues left from the central residue	Real value
	The conservation score C for the position 3 residues left from the central residue	Real value
	The conservation score C for the position 2 residues left from the central residue	Real value
	The conservation score C for the position 1 residue left from the central residue	Real value
	The conservation score C for the central position	Real value
	The conservation score C for the position 1 residue right from the central residue	Real value
	The conservation score C for the position 2 residues right from the central residue	Real value
	The conservation score C for the position 3 residues right from the central residue	Real value
	The conservation score C for the position 4 residues right from the central residue	Real value
Hydrophobic group	Hydrophobic group at the position 2 residues left from the central residue	Real value
	Hydrophobic group at the position 1 residue left from the central residue	Real value
	Hydrophobic group at the central position	Real value
Negatively charged group	Negatively charged group at the position 5 residues left from the central residue	Real value
	Negatively charged group at the position 1 residue left from the central residue	Real value
	Negatively charged group at the central position	Real value
Carboxamide-containing group	Carboxamide-containing group at the position 4 residues left from the central residue	Real value
	Carboxamide-containing group at the position 3 residues left from the central residue	Real value
	Carboxamide-containing group at the position 2 residues left from the central residue	Real value
	Carboxamide-containing group at the position 1 residue left from the central residue	Real value
Secondary structure	Strand on the left side of the central residue	Binary

	Helix on the left side of the central residue	Binary
Collocation of AA pairs	GXXXS pair	Binary
	GXG pair	Binary
	GXS pair	Binary
PSSM profile	The value for Ala at the position 1 residue away from the central residue	Real value
	The value for Arg at the position 4 residues away from the central residue	Real value
	The value for Arg at the position 2 residues away from the central residue	Real value
	The value for Arg at the position 1 residue away from the central residue	Real value
	The value for Arg at the central position	Real value
	The value for Asn at the position 4 residues away from the central residue	Real value
	The value for Asn at the position 3 residues away from the central residue	Real value
	The value for Asn at the position 2 residues away from the central residue	Real value
	The value for Asn at the position 1 residue away from the central residue	Real value
	The value for Asp at the position 4 residues away from the central residue	Real value
	The value for Asp at the position 3 residues away from the central residue	Real value
	The value for Asp at the position 2 residues away from the central residue	Real value
	The value for Asp at the position 1 residue away from the central residue	Real value
	The value for Gln at the position 5 residues away from the central residue	Real value
	The value for Gln at the position 4 residues away from the central residue	Real value
	The value for Gln at the position 3 residues away from the central residue	Real value
	The value for Gln at the position 1 residue away from the central residue	Real value
	The value for Gln at the central position	Real value
	The value for Glu at the position 3 residues away from the central residue	Real value
	The value for Glu at the position 2 residues away from the central residue	Real value
	The value for Glu at the position 1 residue away from the central residue	Real value
	The value for Glu at the central position	Real value
	The value for His at the position 3 residues away from the central residue	Real value
	The value for Ile at the central position	Real value
	The value for Lys at the position 3 residues away from the central residue	Real value
	The value for Lys at the position 2 residues away from the central residue	Real value
	The value for Lys at the position 1 residue away from the central residue	Real value
	The value for Met at the central position	Real value

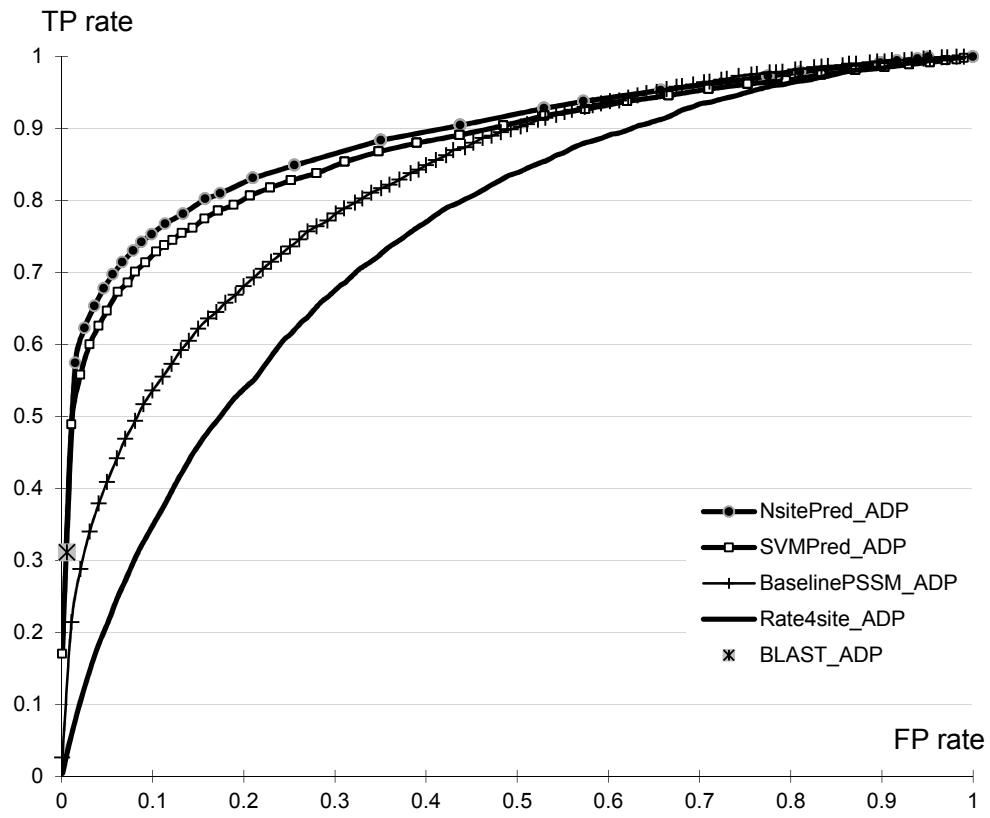


Figure 1A. The ROC curves of the NsitePred, SVMPred, Rate4site, and the predictor based on the PSSM with SVM classifier for prediction of the ADP-binding residues on Dataset 1. The results are based on 5-folds cross validation. The BLAST-based predictor is shown using a single point that corresponds to the binary predictions.

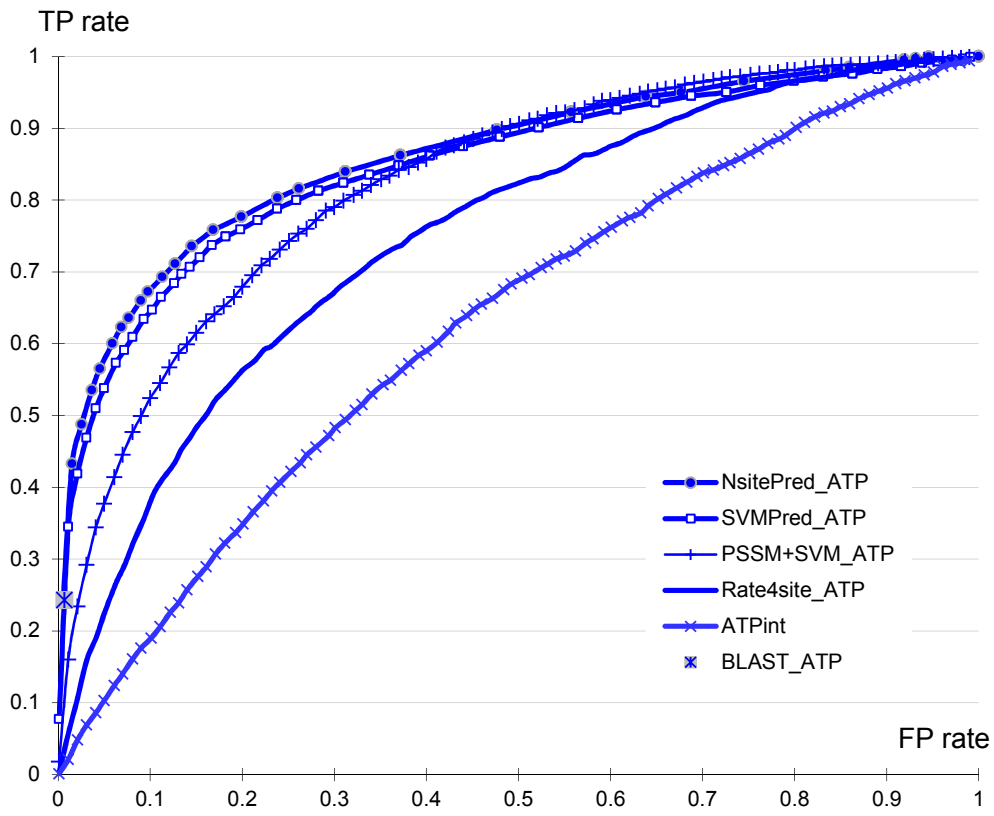


Figure 1B. The ROC curves of the NsitePred, SVMPred, ATPint, Rate4site, and the predictor based on the PSSM with SVM classifier for prediction of the ATP-binding residues on Dataset 1. The results are based on 5-folds cross validation. The BLAST-based predictor is shown using a single point that corresponds to the binary predictions.

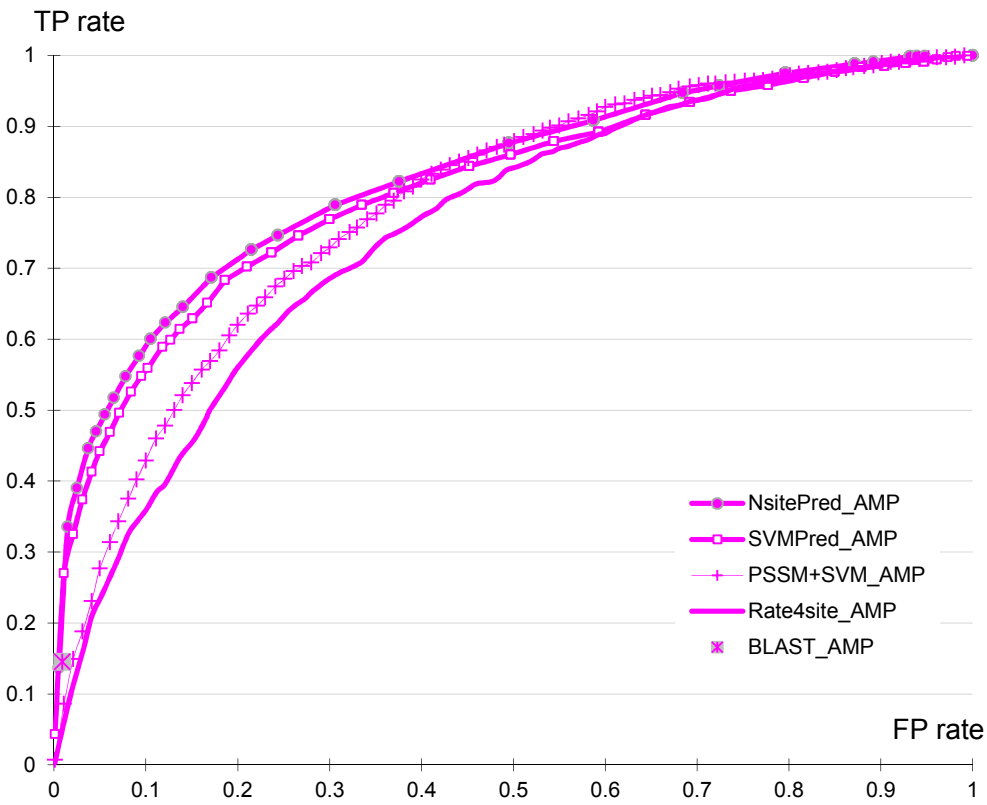


Figure 1C. The ROC curves of the NsitePred, SVMPred, Rate4site, and the predictor based on the PSSM with SVM classifier for prediction of the AMP-binding residues on Dataset 1. The results are based on 5-folds cross validation. The BLAST-based predictor is shown using a single point that corresponds to the binary predictions.

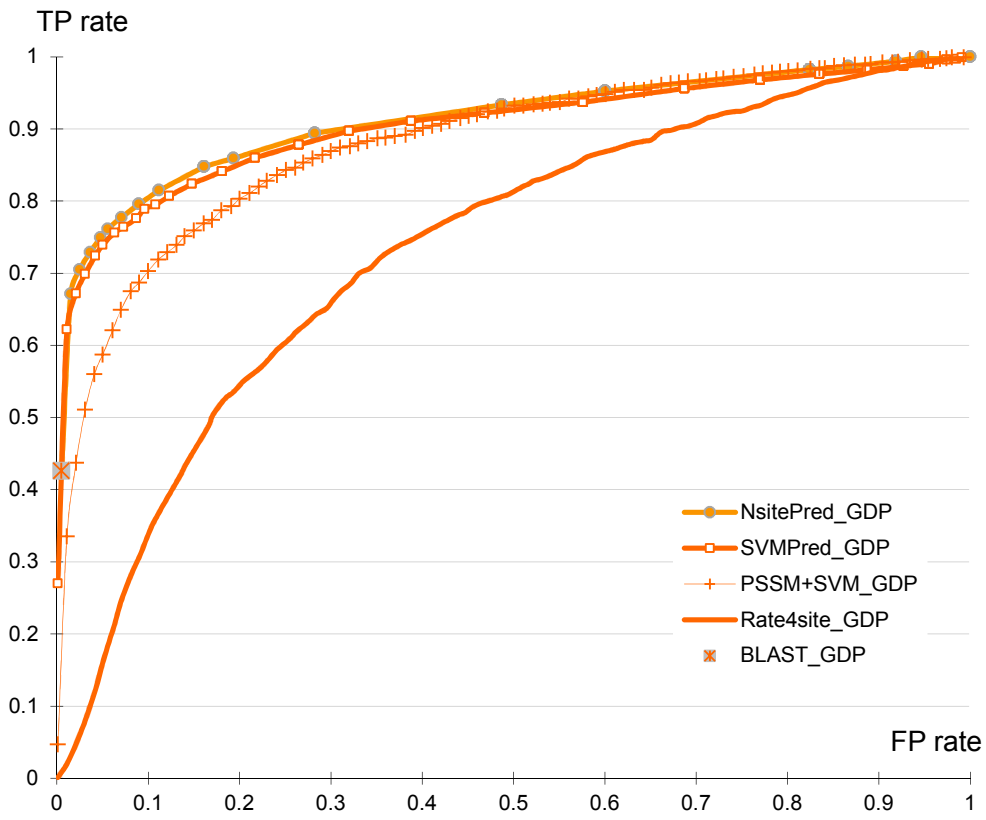


Figure 1D. The ROC curves of the NsitePred, SVMPred, Rate4site, and the predictor based on the PSSM with SVM classifier for prediction of the GDP-binding residues on Dataset 1. The results are based on 5-folds cross validation. The BLAST-based predictor is shown using a single point that corresponds to the binary predictions.

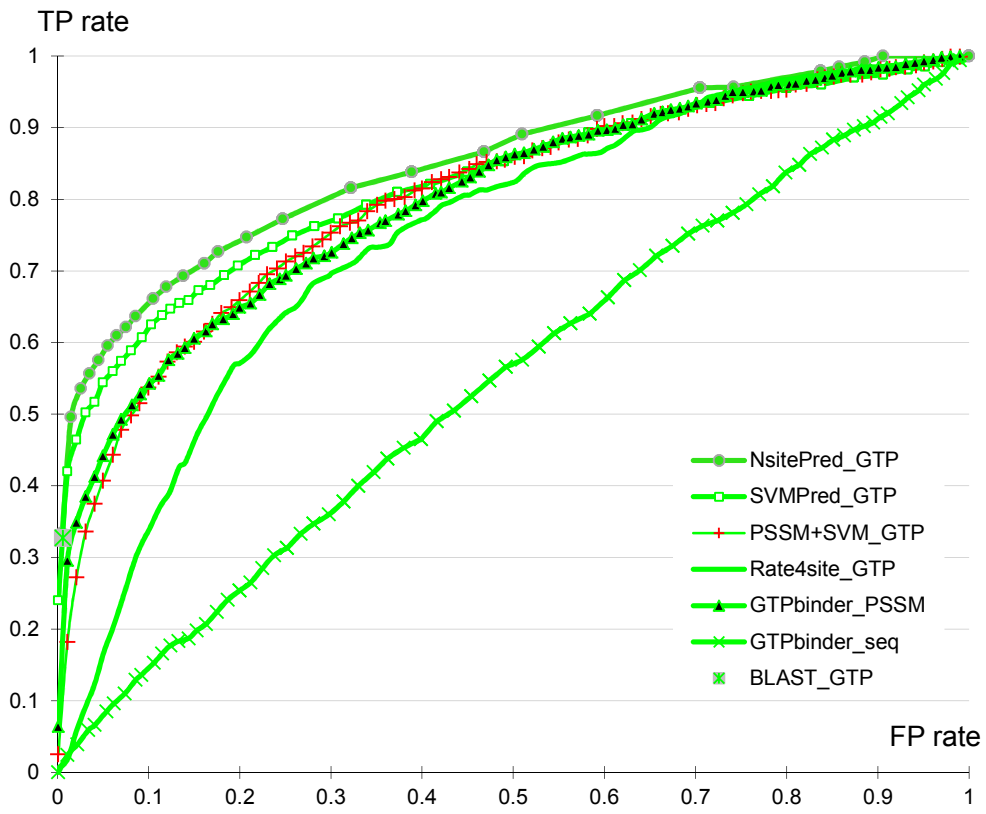


Figure 1E. The ROC curves of the NsitePred, SVMPred, GTPbinder_PSSM (which utilizes PSSM profiles), GTPbinder_seq (which is based solely on the protein sequence), Rate4site, and the predictor based on the PSSM with SVM classifier for prediction of the GTP-binding residues on Dataset 1. The results are based on 5-folds cross validation. The BLAST-based predictor is shown using a single point that corresponds to the binary predictions.

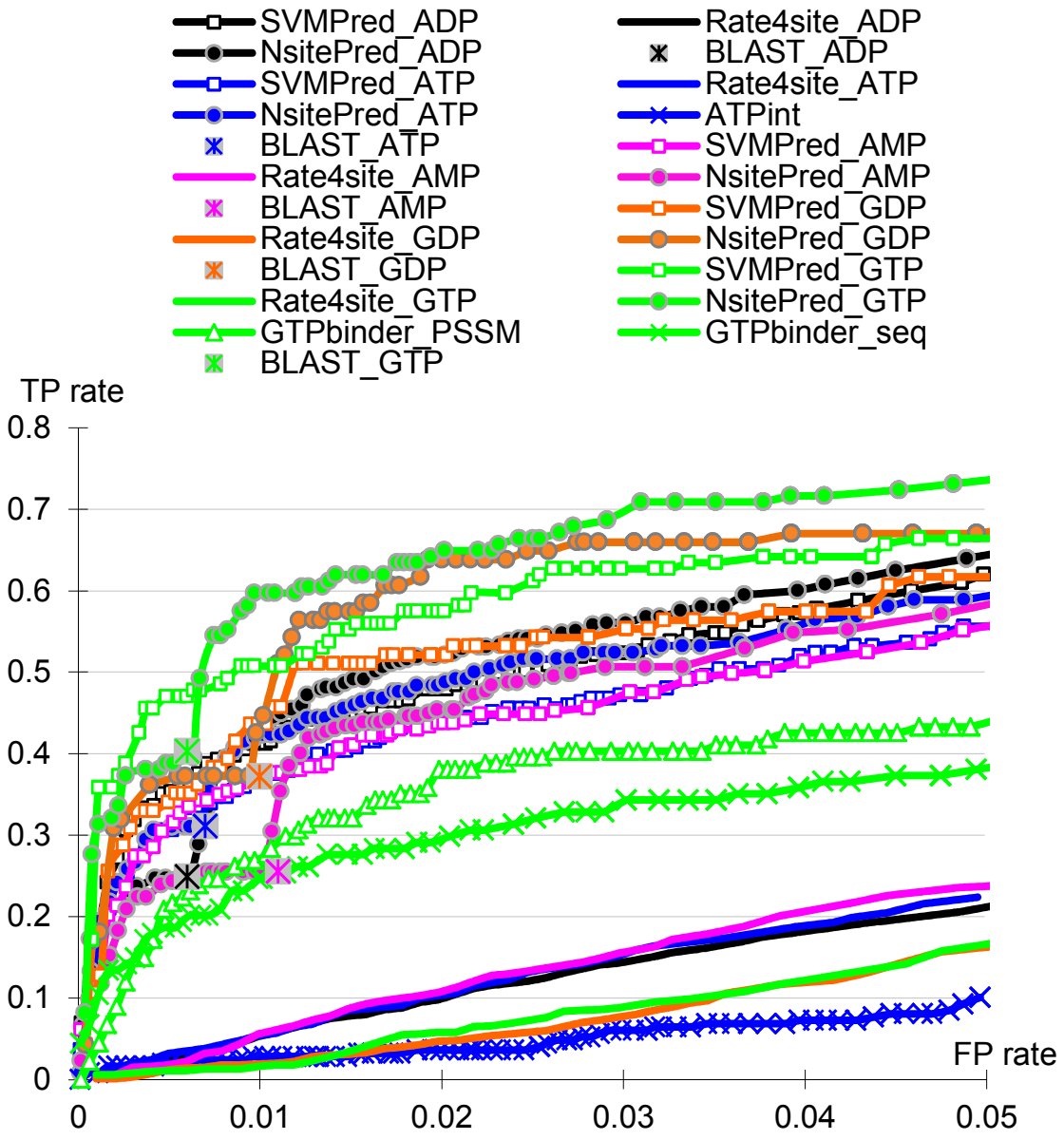


Figure 2. The ROC curves for the NsitePred (denoted using thick solid lines with filled circle markers), SVMPred (denoted using thick solid lines with hollow square markers), ATPint (thick solid line with x markers), GTPbinder (thick solid lines using cross and hollow triangle markers), Rate4site (thick 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 2. Dataset 2 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. The FP-rate is constrained to [0, 0.05] range and the BLAST-based solution is shown using a single point (star marker on grey background) that corresponds to the binary predictions.

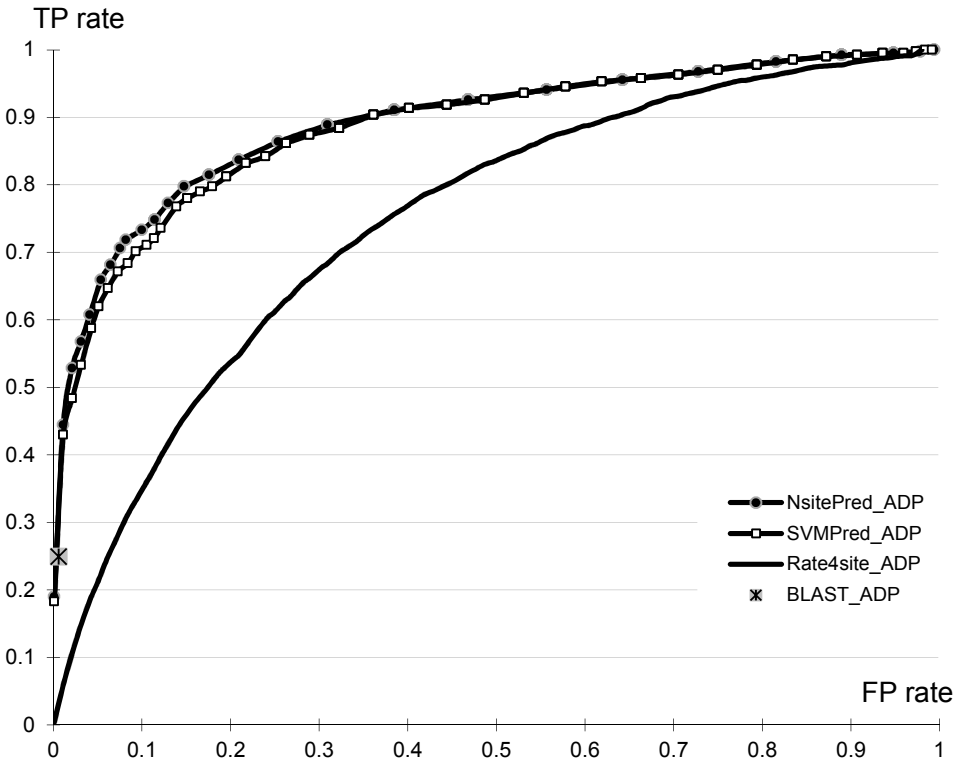


Figure 3A. The ROC curves of the NsitePred, SVMPred, and Rate4site for prediction of the ADP-binding residues on Dataset 2. Dataset 2 consists of chains that were released after the NsitePred was designed and which are dissimilar to chains in the Dataset 1 that was used to build the predictive models. The BLAST-based predictor is shown using a single point that corresponds to the binary predictions.

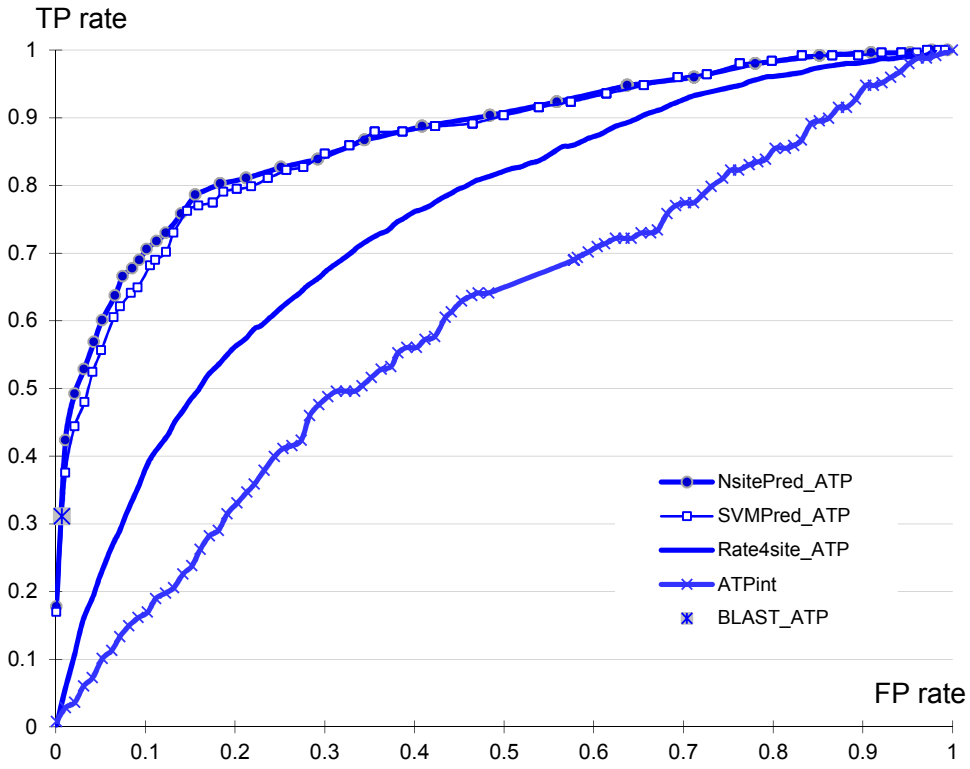


Figure 3B. The ROC curves of the NsitePred, SVMPred, ATPint, and Rate4site for prediction of the ATP-binding residues on Dataset 2. Dataset 2 consists of chains that were released after the NsitePred was designed and which are dissimilar to chains in the Dataset 1 that was used to build the predictive models. The BLAST-based predictor is shown using a single point that corresponds to the binary predictions.

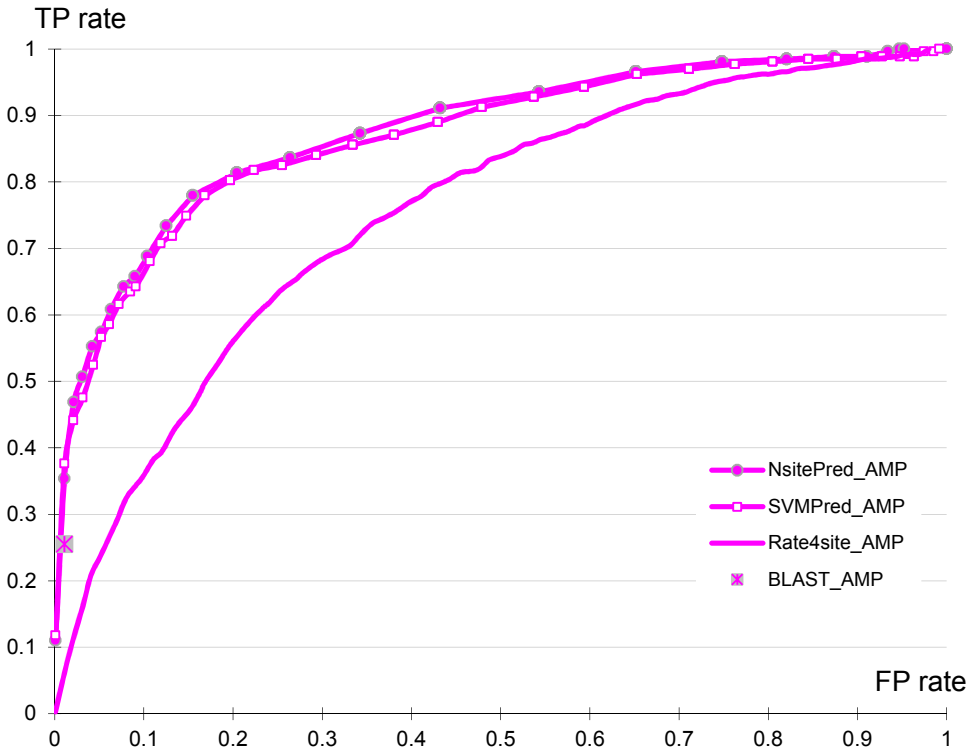


Figure 3C. The ROC curves of the NsitePred, SVMPred, and Rate4site for prediction of the AMP-binding residues on Dataset 2. Dataset 2 consists of chains that were released after the NsitePred was designed and which are dissimilar to chains in the Dataset 1 that was used to build the predictive models. The BLAST-based predictor is shown using a single point that corresponds to the binary predictions.

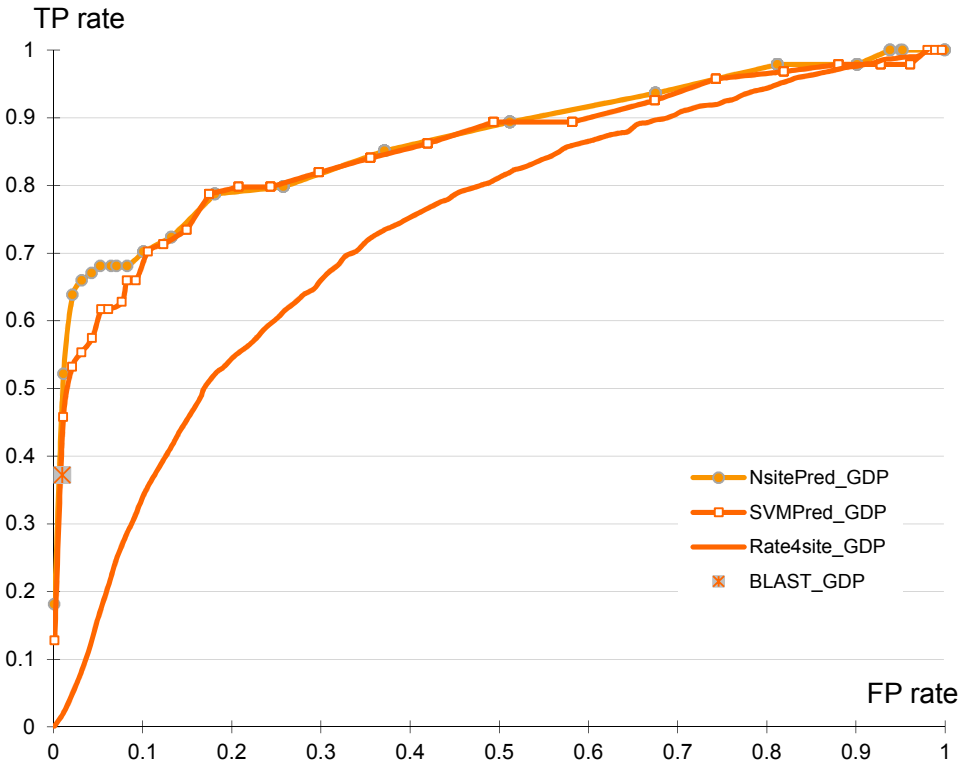


Figure 3D. The ROC curves of the NsitePred, SVMPred, and Rate4site for prediction of the GDP-binding residues on Dataset 2. Dataset 2 consists of chains that were released after the NsitePred was designed and which are dissimilar to chains in the Dataset 1 that was used to build the predictive models. The BLAST-based predictor is shown using a single point that corresponds to the binary predictions.

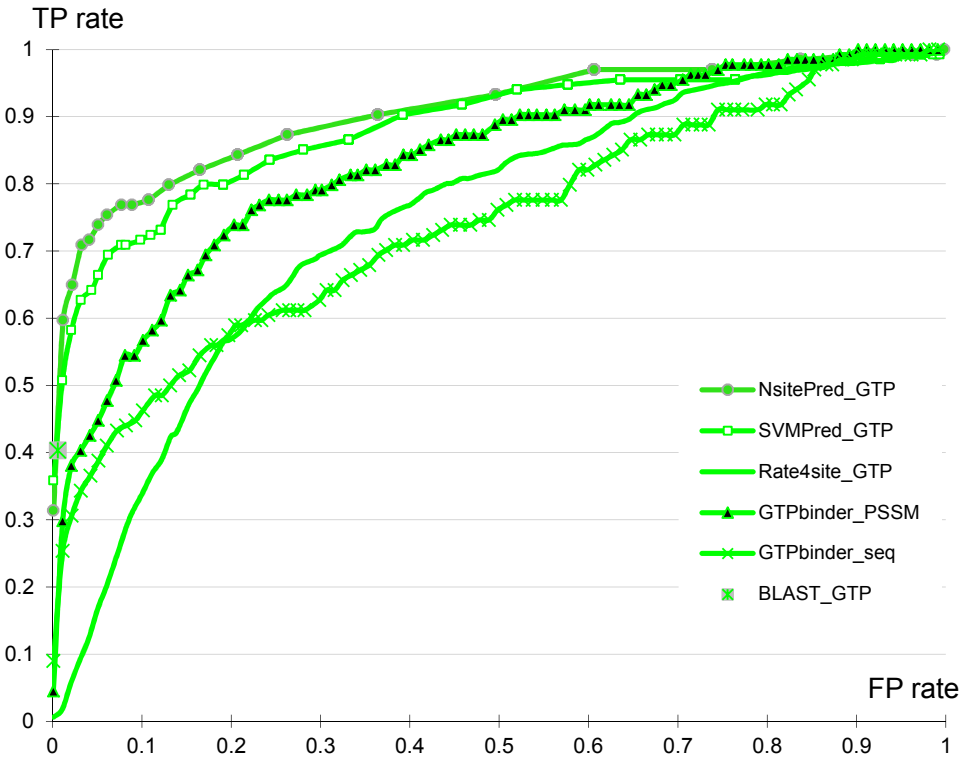


Figure 3E. The ROC curves of the NsitePred, SVMPred, Rate4site, GTPbinder_PSSM (which utilizes PSSM profiles), and GTPbinder_seq (which is based solely on the protein sequence) for prediction of the GTP-binding residues on Dataset 2. Dataset 2 consists of chains that were released after the NsitePred was designed and which are dissimilar to chains in the Dataset 1 that was used to build the predictive models. The BLAST-based predictor is shown using a single point that corresponds to the binary predictions.

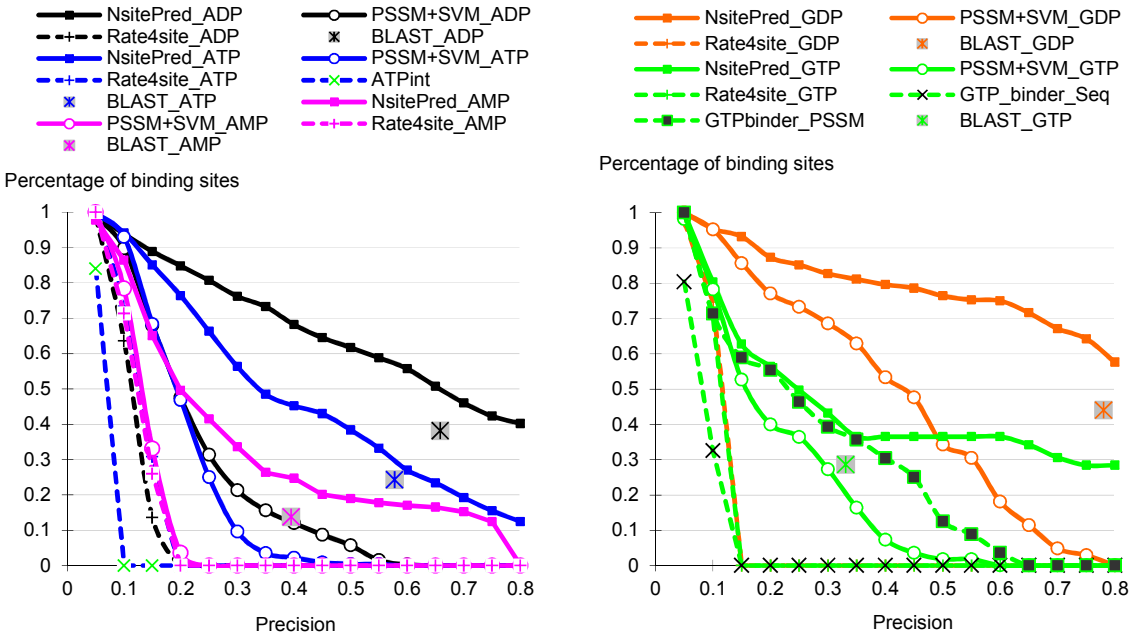
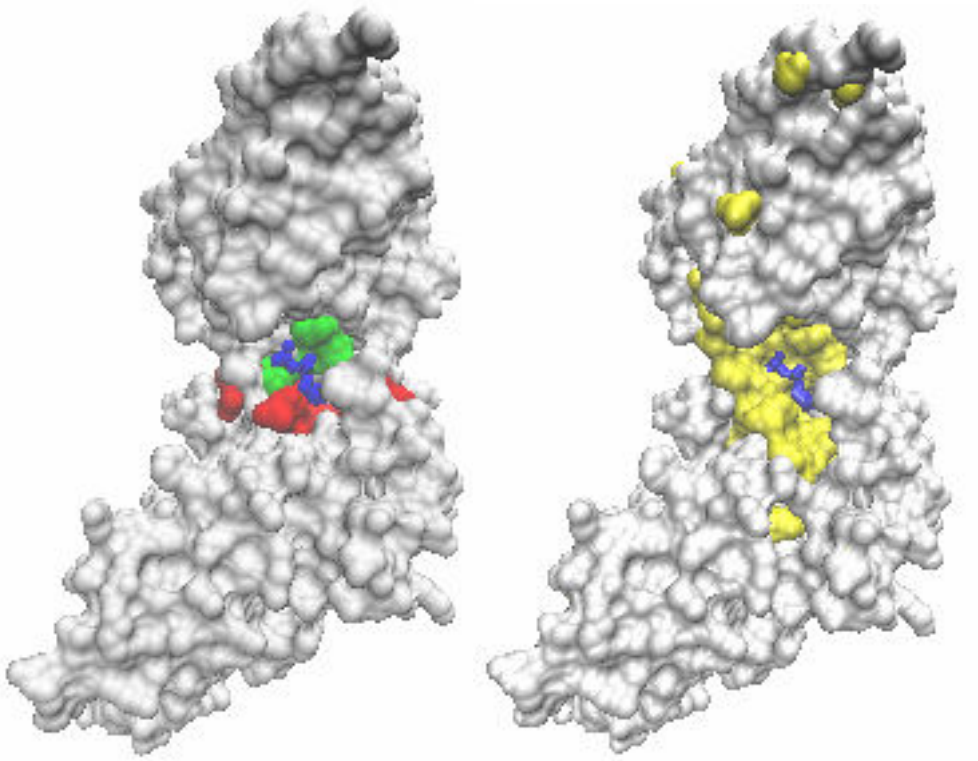


Figure 4. Evaluation of the predictions per binding site on Dataset 1 (based on 5-folds cross validation) for the NsitePred (denoted using solid lines with square markers), ATPint (dashed line with x markers), GTPbinder (dashed lines using x and square markers), Rate4site (solid line with cross markers), and the predictor based on the PSSM with the SVM classifier (solid line with hollow circle makers); the BLAST-based solution is shown using a single point (star marker on grey background) that corresponds to the binary predictions. A given binding site is assumed to be correctly predicted if at least 50% of its residues are correctly predicted. The y-axis shows the percentage of the correctly predicted binding sites. We vary the per-residue precision (x-axis) between 0.05 and 0.8 with 0.05 step to control the number of false positives.



A **B**

Figure 5. A three dimensional view of the predicted binding residues for chain A of cell division control protein 6 (PDB id: 1FNN) by NsitePred, SVMPred, BLAST-based method, and Rate4site. The surface of the non-binding residues is coloured gray and the nucleotide is coloured blue. Panel A shows the predictions of NsitePred and BLAST-based method, where the binding residues predicted by both NsitePred and BLAST-based method are coloured green and the predictions by NsitePred are coloured red. The SVMPred generates the same binary prediction as NsitePred. Panel B shows the predictions by Rate4site (coloured yellow).