Enhancing Protein-ATP and Protein-ADP Binding Sites Prediction Using Supervised Instance-Transfer Learning

Jun Hu
Nanjing university of science and technology
Xiaolingwei 200, Nanjing, China, 210094
junh_cs@126.com

Zi Liu
Nanjing university of science and technology
Xiaolingwei 200, Nanjing, China, 210094
liuzi189836@163.com

Dong-Jun Yu
Nanjing University of Science and Technology
Xiaolingwei 200, Nanjing, China, 210094
njyudj@njust.edu.cn

***Abstract*—Protein-ATP and protein-ADP interactions are ubiquitous in a wide variety of biological processes. Accurately identifying ATP-binding and ADP-binding sites or pockets is of significant importance for both protein function analysis and drug design. Although much progress has been made, challenges remain, especially in the post-genome era where large volume of proteins without being functional annotated are quickly accumulated. In this study, we report an instance-transfer-learning-based predictor, ATP&ADPsite, to target both ATP-binding and ADP-binding residues from protein sequence and structural information. ATP&ADPsite first employs evolutionary information, predicted secondary structure, and predicted solvent accessibility to represent each residue sample. In the above feature space, a supervised instance-transfer-learning method is proposed to improve the ATP-binding/ADP-binding residues prediction by combining ATP-binding and ADP-binding proteins. Random under-sampling is lastly employed to solve the imbalanced data learning problem. Experimental results demonstrate that the proposed ATP&ADPsite achieves a better prediction performance and outperforms many existing sequence-based predictors. The ATP&ADPsite web-server is available at http://csbio.njust.edu.cn/bioinf/ATP&ADPsite.**

Keywords-Protein-ATP binding site prediction; Protein-ADP binding site prediction; Instance transfer learning; Random under sampling

I. INTRODUCTION

Both protein-ATP and protein-ADP interactions are indispensable for biological activities and play important roles in a wide variety of biological processes, such as membrane transport, muscle contraction, replication and transcription of DNA, and various metabolic processes [1-3]. Hence, accurately localizing the protein-ATP and protein-ADP binding sites is of significant importance for both basic experimental biology and drug discovery studies [4]. Many efforts have been made to uncover the intrinsic mechanism of protein-ATP and protein-ADP [5] and thousands of protein-ATP and protein-ADP interaction structure complexes have been deposited in the Protein Data Bank (PDB) database [6, 7]. Due to the importance of protein-ATP and protein-ADP interactions and the difficulty of experimentally identifying of ATP-binding and ADP-binding sites and pockets, developing computational methods for protein-ATP and protein-ADP binding residues prediction has become a hot spot in recent bioinformatics field [8-10].

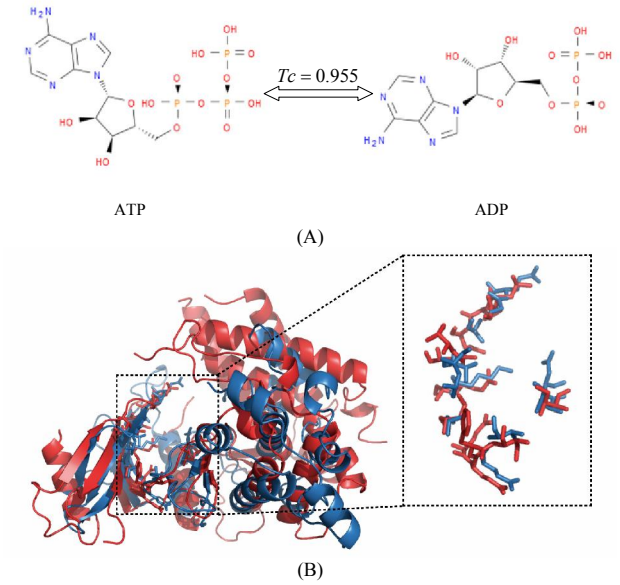

Figure 1. (A) The *Tanimoto coefficient* (*Tc*) between ATP and ADP is 0.955; (B) The PS-score between ATP-binding pocket in 2yaaA protein (Blue color) and ADP-binding pocket in 3mesB protein (Red color) is 0.552 with P-value=2.9×10^{-6}, which is calculated by APoc algorithm.

There have emerged several computational methods for identifying whether each residue in a query protein is an ATP- or ADP-binding residue or not, such as NsitePred [1], TargetATPsite [11], TargetS [2], and TargetNUCs [12]. However, there still exists room to improve the prediction performance. All of these existing methods only employ the ATP-binding and ADP-binding proteins to predict ATP-binding and ADP-binding residues, respectively. While, similar ligands are very likely to have similar interactions with the pocket, e.g., which might contain a common anchor and variable region [13]. As Fig. 1 shown, the highly significant *Tanimoto coefficient* (*Tc*) [13] between ATP and ADP is 0.955, which means ATP and ADP are extremely similar; the PS-score, which is calculated by APoc[14], between ATP-binding and ADP binding pockets (extracted from 2yaaA and 3mesB protein, respectively) is 0.552 with P-value=2.9×10^{-6}, that is, the two pocket is significantly similar. Therefore, we can believe that the similarity between ATP and ADP can impact the similarity in their binding sites prediction

problems. Based on this phenomenon, transfer learning, which can intelligently apply knowledge learned previously to solve new problems faster or with better solutions [15, 16], is a good choice to fuse the knowledge of ATP-binding and ADP-binding sites for enhancing their prediction performance.

In this study, we want to exploit that the data information of ATP- and ADP-binding proteins whether or not can help each other to enhance the corresponding prediction performance. We employ a supervised instance-transfer learning method to combine the ATP-binding and ADP-binding information to predict ATP-binding and ADP-binding residues. To solve the imbalanced data learning problem, we utilize the random under-sampling method to remove some non-binding residues (negative samples). Then, we develop a new predictor, called ATP&ADPsite, to predict ATP-binding and ADP-binding residues. Experimental results demonstrate that the proposed ATP&ADPsite achieves a better prediction performance and outperforms many existing sequence-based predictors.

II. Materials and Methods

A. Benchmark Datasets

To demonstrate the efficacy of the proposed ATP&ADPsite, we construct a dataset of 5,439 protein sequences (which have binding at least one of ATP and ADP), which had clear target annotations in the Protein Data Bank (PDB) [7] before March 09, 2016, from BioLip [17]. The proteins of BioLip are collected primarily from PDB [17]. Then, maximal pairwise sequence identity of the dataset was culled to 30% with CD-hit software [18] and the non-redundant dataset consists of 623 protein sequences (including 269 ATP-binding and 358 ADP-binding protein sequences). It is noted that there are four protein sequences (i.e., 3LFZA, 1II0B, 1MABB, and 4F93B), which are binding both ATP and ADP, in this dataset. According to the non-redundant dataset, we respectively construct ATP-binding and ADP-binding training datasets, called ATP255 and ADP339, by selecting 255 ATP-binding and 339 ADP-binding protein sequences, which were released into PDB before March 09, 2015. ATP255 and ADP339 both contain the above four special protein sequences. The remaining ATP-binding and ADP-binding protein sequences (which were released into PDB after March 09, 2015) are separately employed to form the independent test datasets, called ATP14 and ADP19. There are 3,408 ATP-binding residues and 99,293 non-ATP-binding residues in ATP255. ADP339 contains 4,589 ADP-binding residues and 134,549 non-ADP-binding residues. ATP14 consists of 178 ATP-binding residues and 6,315 non-ATP-binding residues. ADP19 includes 272 ADP-binding residues and 10,403 non-ADP-binding residues.

B. Feature Representation

Previous studies have demonstrated that lots of sequence-based or sequence-predicted feature views, such as protein specific scoring matrix (PSSM) [19, 20], predicted secondary structure (PSS) [1], predicted solvent accessibility (PSA) [1], and amino acid physiochemical characteristics [21], can be effectively used to predict the ATP-binding or ADP-binding residues. In this study, we utilize a sliding window of size 17 centered on the predicted residue to extract the above features as follows:

- PSSM profile generated by PSI-BLAST with default settings using the Swiss-Prot database [22]. We include the scores for each of the 20 substitution amino acid types and we also employ logistic function $f(x)=1/(1+e^{-x})$ to normalize each element in the profile.
- Predicted secondary structure (PSS) generated by PSIPRED [23] for each residue in the window.
- Predicted solvent accessibility (PSA) generated by SANN program [24], which can be downloaded at http://lee.kias.re.kr/~newton/sann/, for each residue in the window.

Finally, the feature representation of a residue is formed by serially combining its three corresponding feature vectors, i.e., PSSM, PSS, and PSA. We denoted the final feature vector as PSSM+PSS+PSA.

C. Supervised Instance-Transfer Learning

To use the data information of ATP-binding and ADP-binding datasets to help each other to improve the prediction performance, the instance-transfer learning is a good choice to do that. In this study, we design a supervised instance-transfer learning method, SITI, as follows:

Let S_{tar} be the target domain dataset, S_{src} be the source domain dataset. In order to improve the prediction performance of target domain, we hope to select lots of available data in S_{src} to join in S_{tar}. We first pre-train an evaluation model (*EvaModel*) using the original target domain dataset, as follows

$$EvaModel = trainModel(S_{tar}) \quad (1)$$

Then, we utilize the evaluation model *EvaModel* to predict the probability scores (P_{src}) of all the samples in S_{src} to evaluate their contribution for the target domain prediction task.

$$P_{src} = predicted(EvaModel, S_{src}) \quad (2)$$

Lastly, we sort the dataset S_{src} based on P_{src}. We select the top N ($|S_{src}|/2$ in this study) samples from S_{src} to join in S_{tar} to compose the final target domain dataset FS_{tar}.

$$FS_{src} = SortSelect(S_{src}, P_{src}, N) \quad (3)$$

The final target domain prediction model will be trained on the final target domain dataset FS_{tar}.

D. Imbalanced Data Learning

From the perspective of machine learning, both protein-ATP and protein-ADP binding sites predictions are

two exemplary class imbalance learning problems [25]. For example, the number of non-ATP-binding residues (i.e., negative class samples) is more than 29 times of that of ATP-binding residues (i.e., positive class samples) in the ATP-binding training dataset ATP255. Directly using the traditional machine learning algorithms (e.g., support vector machine (SVM) [26]), which assume that samples in different classes are balanced, often leads to a poor performance.

In order to solve the class imbalance learning problems embedded in ATP-binding and ADP-binding sites predictions, we utilized random under-sampling method (termed as RUS) to change the distribution of samples in different classes. Concretely, RUS randomly selects a set of negative samples, denoted as $\hat{S}_{nega}$, from the original dataset $S = S_{nega} \cup S_{posi}$ to relieve the degree of imbalance, where S_{nega} and S_{posi} means the negative and positive subsets of the original dataset, respectively. The new dataset can be denoted as $\hat{S} = \hat{S}_{nega} \cup S_{posi}$. For convenience, the imbalance coefficient of a dataset is defined as $\beta = |\hat{S}_{nega}| / |S_{posi}|$, where $|\cdot|$ represents the cardinality of the corresponding set. Obviously, the imbalance coefficient β reflects the severity of imbalance of a dataset, i.e., $\hat{S}$. In this study, we set $\beta = 1$, which can make us to obtain a completely balanced dataset from the original imbalanced dataset. After obtaining the balanced dataset, we choose SVM algorithm to train each prediction model. LIBSVM [27] is employed to implement SVM function. Here, radial basis function is chosen as the kernel function. The important two parameters, i.e., the kernel width parameter σ and the regularization parameter γ, are optimized based on five-fold cross-validation using a grid search strategy in the LIBSVM tool. In this study, the final γ and are 2.5 and 0.044, respectively.

E. Assessment of Predictive Ability

We employ five routinely used evaluation indexes (i.e., Sensitivity (*Sen*), Specificity (*Spe*), Accuracy (*Acc*), and the Mathew's Correlation Coefficient (*MCC*)) to measure the effectiveness of the proposed method as follows:

$$Sen = \frac{TP}{TP + FN} \tag{4}$$

$$Spe = \frac{TN}{TN + FP} \tag{5}$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{6}$$

$$MCC = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP + FN) \cdot (TP + FP) \cdot (TN + FN) \cdot (TN + FP)}} \tag{7}$$

where *TP*, *FN*, *FP*, and *TN* represent the counts of true positive, false negative, false positive, and true negative prediction, respectively. However, the above five evaluation indexes are all threshold-dependent [12]. How to objectively use these evaluation indexes to impartially evaluate the performance of different methods is an important problem, especially in the imbalance data learning. In this paper, the four above-mentioned indexes are reported with two special thresholds, i.e., $T_{Balance}$ and T_{MaxMCC}. The $T_{Balance}$ threshold can make *Sen* is approximately equal to *Spe* of the predictions on the cross-validation tests [28]. The T_{MaxMCC} threshold can maximize the *MCC* value of the predictions on the cross-validation tests [28]. Here, for convenience, the evaluation results obtained under $T_{Balance}$ and T_{MaxMCC} will separately be termed as *Balanced Evaluation* and *MaxMCC Evaluation* in the subsequent descriptions. Furthermore, we also plot the ROC curves (which plot *TPR* (true positive rate, i.e., *Sen*) against *FPR* (false positive rate, i.e., 1-*Spe*)) for the predictors and calculated the area under ROC curve (AUC) to evaluate the overall performance of the predictors. Finally, the area under precision-recall curve (AUPRC) is used to verify the performance of predictors when restricting low false positive ratesss. The precision-recall curve, which plots *Pre* against *Sen*, is more sensitive to false positives than ROC curve.

III. Experimental Results and Analysis

A. SITL Helps to Improve Prediction Performance

In this section, we will investigate the efficacy of the proposed SITL for improving the performance of protein-ATP and protein-ADP binding sites predictions. In protein-ATP binding sites prediction task, the ADP-binding sites in ADP339 are employed to be the source domain dataset S_{src} in SITL for improving the prediction performance. In protein-ADP binding sites prediction task, the ATP-binding sites in ATP255 are used to be the source domain dataset S_{src} in SITL for enhancing the prediction performance. Tables I and II separately demonstrate the results of without- and with-SITL on two prediction tasks, i.e., protein-ATP binding sites prediction and protein-ADP binding sites prediction, over five-fold cross-validation under *Balanced Evaluation* and *MaxMCC Evaluation*. It is noted that the negative sample number is roughly equal to the positive sample number in each training phase of cross-validation tests.

TABLE I. Performance Comparisons Between Without- and With-SITL Over Five-Fold Cross-Validation under *Balanced Evaluation*

Task	Without-/ With-SITL	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*
T [a]	Without-SITL	79.37	79.51	79.51	0.253
	With-SITL	82.54	82.75	82.74	0.296
D [b]	Without-SITL	80.85	81.26	81.25	0.273
	With-SITL	82.92	83.17	83.16	0.301

[a] T represents the protein-ATP binding sites prediction task on ATP255.
[b] D represents the protein-ADP binding sites prediction task on ADP339.

TABLE II. PERFORMANCE COMPARISONS BETWEEN WITHOUT- AND WITH-SITL OVER FIVE-FOLD CROSS-VALIDATION UNDER MaxMCC EVALUATION

Task	Without-/With-SITL	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*
T [a]	Without-SITL	45.69	97.41	95.70	0.393
	With-SITL	52.08	98.11	96.58	0.486
D [b]	Without-SITL	43.87	98.69	96.88	0.467
	With-SITL	50.38	98.71	97.12	0.522

[a] T represents the protein-ATP binding sites prediction task on ATP255.
[b] D represents the protein-ADP binding sites prediction task on ADP339.

By observing Table I, we can find that the values of *Sen*, *Spe*, *Acc*, and *MCC* for with-SITL are consistently superior to those for without-SITL throughout the both prediction tasks. Taking MCC as example, which is the index measuring the overall prediction performance of a prediction model, the average improvement of 3.55% is observed, after using SITL on the both prediction tasks. Under *MaxMCC Evaluation* (refer to Table II), the similar results can also be observed.

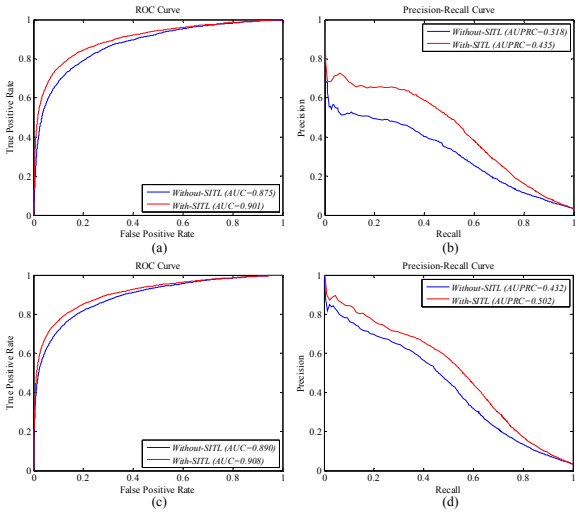

Figure 2. The ROC and precision-recall curves of *With-SITL* and *Without-SITL*, respectively, over five-fold cross validation. (a) ROC curves for protein-ATP binding sites prediction task; (b) Precision-recall curves for protein-ATP binding sites prediction task; (c) ROC curves for protein-ADP binding sites prediction task; (d) Precision-recall curves for protein-ADP binding sites prediction task.

Furthermore, Fig. 2 directly demonstrates the ROC curves and precision-recall curves of the predictions without- and with-SITL, respectively, on both protein-ATP and protein-ADP binding sites prediction tasks over five-fold cross-validation. By observing Fig. 2, we can also find that the overall performances (i.e., AUC and AUPRC) of the two tasks are further improved by using SITL.

B. Comparisons with existing predictors

In this section, we will experimentally demonstrate the efficacy of the proposed ATP&ADPsite by comparing with other popular ATP-binding and ADP-binding sites predictors via the following two different tests:

1) Independent Test on Validation Datasets: Table III shows the performance comparison of the proposed ATP&ADPsite with other protein-ATP and protein-ADP binding sites predictors, including TargetS[2], TargetATPsite[11], and NsitePred[1], on the corresponding independent test datasets, i.e., ATP14 and ADP19. It is easily found that ATP&ADPsite consistently outperforms other predictors on both datasets concerning MCC and AUC evaluation indexes. Taking ADP19 as an example, the MCC and AUC of ATP&ADPsite are 0.605 and 0.931, which are about 1.5% and 2.7% better than that of the second-best method, TargetS, respectively. The other two evaluation indexes of ATP&ADPsite are all slightly higher than other methods.

TABLE III. Performance Comparison of The Proposed ATP&ADPsite with Other Protein-ATP and Protein-ADP Binding Sites Predictors on The Corresponding Independent Test Datasets.

Predictor	Dataset	*Sen* (%)	*Acc* (%)	*MCC*	*AUC*
ATP&ADPsite	ATP14	45.5	98.0	0.567	0.869
TargetS		56.7	97.4	0.533	0.865
TargetATPsite		42.1	97.8	0.516	0.837
NsitePred		42.7	96.2	0.361	0.860
ATP&ADPsite	ADP19	54.4	98.2	0.605	0.931
TargetS		54.0	98.1	0.590	0.904
NsitePred		46.7	97.4	0.466	0.883

2) Performing Comparison on Dataset That Have Been Used By Other Predictors: Except for the independent validation test performed above, we will try to further demonstrate the efficacy of the proposed ATP&ADPsite by comparing it with other predictors based on the same benchmark dataset that has been used by the compared predictors. The dataset 1 [1] includes 227 and 321 sequences that bind to ATP and ADP, respectively, and the maximal pairwise sequence identity of the sequences among each type of ATP and ADP is less than 40 percent. We compare ATP&ADPsite with TargetS, TargetATPsite, and NsitePred on dataset 1 over five-fold cross-validation as done in [1] and the comparison results are listed in Table IV. By observing Table IV, we find that the ATP&ADPsite obtains AUC >0.92 and MCC >0.62 for both ATP and ADP. The AUC and MCC values of ATP&ADPsite are consistently superior to that of other predictors and the averaged improvement of 5.7 and 2.5 percent are achieved, respectively, if compare with the second performer TargetS.

TABLE IV. PERFORMANCE COMPARISON OF THE PROPOSED ATP&ADPSITE WITH OTHER PROTEIN-ATP AND PROTEIN-ADP BINDING SITES PREDICTORS ON ATP227 AND ADP321 OVER FIVE-FOLD CROSS-VALIDATION.

Predictor	Dataset	*Sen* (%)	*Acc* (%)	*MCC*	*AUC*
ATP&ADPsite	ATP227	59.06	97.25	0.622	0.927
TargetS		44.60	96.70	0.531	0.896
TargetATPsite		44.50	96.60	0.520	0.881
NsitePred		44.40	96.00	0.460	0.861
ATP&ADPsite	ADP321	62.80	97.64	0.654	0.938
TargetS		58.70	97.50	0.631	0.918
NsitePred		54.40	97.10	0.572	0.893

IV. CONCLUSIONS

In this study, we have designed and implemented a new sequence-based ATP-binding and ADP-binding residues predictor, called ATP&ADPsite. ATP&ADPsite employs the supervised instance-transfer learning method to combine the ATP-binding and ADP-binding dataset information to predict ATP-binding and ADP-binding residues, and utilizes random under-sampling method to solve the imbalanced data learning protein. Experimental results demonstrate that the proposed ATP&ADPsite achieves a better prediction performance and outperforms many existing sequence-based predictors. The good performances of the ATP&ADPsite come from several reasons include good benchmark dataset, more discriminative feature design, and careful construction of the prediction model. The ATP&ADPsite web-server is available at http://csbio.njust.edu.cn/bioinf/ATP&ADPsite.

Besides ATP and ADP studied in this study, there are lots of other ligands (e.g., RNA and Ca^{2+}) that can bind protein targets, it is also necessary to investigate how to effectively discriminate the different ligands to reveal the various molecular binding mechanisms.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (No. 61373062 and 61772273), the Natural Science Foundation of Jiangsu (No. BK20141403) China Scholarship Council (No. 201606840087), and Jiangsu University Graduate Research and Innovation Project (KYLX16_0457). Dong-Jun Yu is the corresponding author for this paper.

REFERENCES

[1] K. Chen, *et al.*, "Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors," *Bioinformatics,* vol. 28, pp. 331-41, Feb 1 2012.

[2] D. Yu, *et al.*, "Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 10, pp. 994-1008, 2013.

[3] A. Maxwell and D. M. Lawson, "The ATP-binding site of type II topoisomerases as a target for antibacterial drugs," *Current topics in medicinal chemistry,* vol. 3, pp. 283-303, 2003.

[4] P. Schmidtke and X. Barril, "Understanding and predicting druggability. A high-throughput method for detection of drug binding sites," *Journal of medicinal chemistry,* vol. 53, pp. 5858-5867, 2010.

[5] K. Chen, *et al.*, "ATPsite: sequence-based prediction of ATP-binding residues," *Proteome Science,* vol. 9 Suppl 1, p. S4, 2011.

[6] H. M. Berman, *et al.*, "The protein data bank," *Biological Crystallography,* vol. 58, pp. 899-907, 2002.

[7] P. W. Rose, *et al.*, "The RCSB Protein Data Bank: redesigned web site and web services," *Nucleic Acids Research,* vol. 39, pp. D392-401, Jan 2011.

[8] S. Ahmad, *et al.*, "Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information," *Bioinformatics,* vol. 20, pp. 477-486, 2004.

[9] J. N. Si, *et al.*, "An Overview of the Prediction of Protein DNA-Binding Sites," *International Journal of Molecular Sciences,* vol. 16, pp. 5194-5215, 2015.

[10] L. Wang and S. J. Brown, "BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences," *Nucleic Acids Research,* vol. 34, pp. W243-W248, 2006.

[11] D. J. Yu, *et al.*, "TargetATPsite: A template-free method for ATP-binding sites prediction with residue evolution image sparse representation and classifier ensemble," *Journal of computational chemistry,* vol. 34, pp. 974-985, Apr 2013.

[12] S. Ahmad, *et al.*, "Protein–DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins," *Nucleic Acids Research,* vol. 36, pp. 5922-5932, 2008.

[13] M. Gao and J. Skolnick, "A comprehensive survey of small-molecule binding pockets in proteins," *PLoS Comput Biol,* vol. 9, p. e1003302, 2013.

[14] M. Gao and J. Skolnick, "APoc: large-scale identification of similar protein pockets," *Bioinformatics,* vol. 29, pp. 597-604, 2013.

[15] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering,* vol. 22, pp. 1345-1359, 2010.

[16] H.-C. Shin, *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging,* vol. 35, pp. 1285-1298, 2016.

[17] J. Yang, *et al.*, "BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions," *Nucleic Acids Research,* vol. 41, pp. D1096-103, Jan 1 2013.

[18] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics,* vol. 22, pp. 1658-1659, 2006.

[19] J. S. Chauhan, *et al.*, "Identification of ATP binding residues of a protein from its primary sequence," *BMC Bioinformatics,* vol. 10, p. 434, 2009.

[20] J. Hu, *et al.*, "A New Supervised Over-Sampling Algorithm with Application to Protein-Nucleotide Binding Residue Prediction," *PloS one,* vol. 9, p. e107676, 2014.

[21] D. J. Yu, *et al.*, "Designing Template-Free Predictor for Targeting Protein-Ligand Binding Sites with Classifier Ensemble and Spatial Clustering," *Ieee-Acm Transactions on Computational Biology and Bioinformatics,* vol. 10, pp. 994-1008, Jul-Aug 2013.

[22] R. Anirudh and P. Turaga, "Geometry-Based Symbolic Approximation for Fast Sequence Matching on Manifolds," *International Journal of Computer Vision,* pp. 1-13, 2015.

[23] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of Molecular Biology,* vol. 292, pp. 195-202, 1999.

[24] K. Joo, *et al.*, "Sann: solvent accessibility prediction of proteins by nearest neighbor method," *Proteins-structure Function & Bioinformatics,* vol. 80, pp. 1791–1797, 2012.

[25] D. J. Yu, *et al.*, "Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling," *Neurocomputing,* vol. 104, pp. 180-190, Mar 2013.

[26] V. N. Vapnik, Ed., *Statistical Learning Theory* New York: Wiley-Interscience, 1998, p.^pp. Pages.

[27] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 2, p. 27, 2011.

[28] D. J. Yu, *et al.*, "Enhancing protein-vitamin binding residues prediction by multiple heterogeneous subspace SVMs ensemble," *BMC Bioinformatics,* vol. 15, Sep 5 2014.