TSC_ATP: A Two-Stage Classifier For Predicting Protein-ATP Binding Sites From Protein Sequence

Bryan J. Andrews, Jing Hu*
Department of Mathematics & Computer Science
Franklin & Marshall College
Lancaster, PA 17604, USA
jing.hu@fandm.edu

Abstract—It is well known that adenine triphosphate (ATP) binds with proteins to play important roles in metabolism, cell signaling, and as cofactor. Therefore it is crucial to identify ATP-binding sites on proteins to understand these mechanisms. In this paper, we present a computational method that can accurately predict ATP-binding sites on proteins using sequence-derived information. The algorithm is organized in a two-layered structure. The method first makes prediction by a K-Nearest Neighbors (K-NN) method using evolutionary profile information of each residue. The output of the first-layer classifier serves as an input feature together with other 11 features to a second-layer classifier. The final method achieved an AUC (area under the ROC curve) of 0.829 and 0.860 respectively on two benchmark datasets.

Keywords— ATP-binding; K-nearest neighbors method; PSSM; Two-stage classifier; Under-Sampling;

I. INTRODUCTION

It is well known that Adenine triphosphate (ATP) plays roles in many important biological processes such as cell signaling, enzymatic cofactor functions, and metabolism [1-5]. The conversion of ATP to ADP (Adenine diphosphate) and AMP (Adenine monophosphate) provides the energy needed for conformation change in many biological interactions. The binding of ATP to proteins is a fundamental step to these biological processes. Thus, it is important to identify ATP-binding sites on proteins for understanding of these processes.

There are several computational methods developed for predicting ATP-binding sites using various sequence and/or structural properties. However, due to high throughput proteome project, huge number of protein sequences has been accumulated without any structural information; therefore it is crucial to predict ATP-binding sites using only sequence-derived information.

Early methods for identifying ATP-binding sites on protein sequences relied on detecting conservative sequence motifs [6]. However, sequence motifs have the limitation of not being able to identify non-conserved or novel binding sites. Many computational methods use machine-learning algorithms to predict ATP-binding sites by exploring various sequence or sequence-based features [7-15]. Among which, most prediction methods are based on Support Vector Machine (SVM). ATPint is the first SVM classifier for predicting ATP interacting residues from its protein sequence and evolutionary information (PSSM) [7]. ATPsite is a SVM method utilizing a comprehensive set of input features that are based on the sequence, evolutionary profiles, and the sequence-predicted structural descriptors including secondary structure, solvent accessibility, and dihedral angels [8]. Firoz et al. explored single amino acid propensities of ATP-binding residues and made predictions using a SVM based on a sliding window of PSSM profiles centering on each residue [9]. Zhang et al. developed a SVM method by combining sequence evolutionary information and bi-profile sampling of multi-view sequential features and the sequence derived structural features [10]. NsitePred is a collection of five predictors that can identify binding residues for ATP, ADP, AMP, GTP and GDP [11]. It makes prediction based on information extracted from the sequence, evolutionary profiles, several sequence-predicted structural descriptors and sequence alignment. TargetATPSite is an ensemble classifier constructed based on SVM from multiple random under-sampling [12][13]. ATPBR predicts ATP-binding residues in proteins from amino acid sequences by using random forests with a novel hybrid feature set, which incorporates a new feature called PSSMPP, the predicted secondary structure and orthogonal binary vectors [14]. CLCLpred is a method which adopts a modified PSSM encoding scheme for ATP-binding prediction as input for SVM [15].

In this research, we have developed TSC_ATP, a two-stage classifier to predict ATP-binding sites on proteins from sequence-based information. At the first layer, a K-nearest neighbors (K-NN) method was used to predict ATP-binding sites using PSSM profile features of the sliding window surrounding each residue. At the second layer, the output of the first layer classifier plus 11 other features served as input to the Threshold Selector classifier. The final method achieved an AUC of 0.829 and 0.860 on two benchmark datasets. Comparison with other published methods indicated that our method outperformed and was comparable to many state-of-the-art ATP prediction methods.

II. METHOD AND MATERIAL

A. Dataset

We used two benchmark datasets (i.e., ATP168 and ATP227) that were used in previous studies.

ATP168: This dataset was originally curated in the study of ATPint [7]. ATP-binding proteins were extracted from...
SuperSite encyclopedia [16] and then removed redundancy by CD-Hit [17] using a sequence identity threshold of 40%. The final dataset has 168 ATP-binding proteins, which contains 3,104 ATP-binding residues and 59,225 non-binding residues.

APT227: This dataset was originally curated in the study of ATPsite [8]. Complexes including ATP-binding protein chains were extracted from PDB. Using CD-Hit [17] and a sequence identity threshold of 40%, 227 ATP-binding proteins were left, among which 3,393 residues are ATP-binding and the rest 80,409 residues are non-binding.

B. First-layer K-NN Classifier and Features

For the first-layer classifier, we used a similar classification model as proposed by the study of Joo et al. [18].

For each protein, we performed PSI-BLAST [19] against the non-redundant protein database with 3 iterations and 0.001 of E-value cutoff to generate PSSM profiles. For each residue in a protein, a sliding window of 15 residues centered on the target residue was chosen to create a PSSM matrix (15 x 20). For any two residues A and B the distance between them was calculated as

$$d_{AB} = \sqrt{\sum_{j} w_{j}(P_{x}^A - P_{x}^B)^2}$$

where $P_{x}^j$ is the $j$th ($j = 1, 2, ..., 20$) value of the PSSM vector of residue $x$ ($x = A$ or $B$) and $w_{j} = (8 - 8/j^2)$ as proposed by Joo et al. [18], which gives more weight to the center residue and less weight to residues further away from the center.

K-NN is an instance-learning algorithm. A typical K-NN method classifies the query sample by the majority voting strategy. For each query sample, K-NN finds its $K$ nearest neighbors in the training dataset, and then assigns it to the class to which most of its neighbors belong. The K-NN method used in this study differs from standard K-NN in several ways. First, for each query residue we calculated its distances to all residues in the training dataset using the distance defined in formula (1). Second, after $K$ nearest neighbors were selected, we calculated the z-score $z_i$ for each distance $d_i$ as

$$z_i = \frac{(\bar{d} - d_i)}{\sigma}$$

where $i = 1, 2, 3, ..., K$ and $\bar{d}$ is the average distance between the query residue and all residues in the dataset and $\sigma$ is the standard deviation of all these distances. Then we calculated the $u$-score for each class $c$ (i.e., ATP-binding or non-binding) as

$$u_c = \sum_{i \in S_c} z_i^\alpha$$

where $S_c$ is the set of residues in $K$ nearest neighbors of the query residue which belong to the class $c$ (notice that $c$ can be ATP-binding or non-binding and the union of $S_{\text{ATP-binding}}$ and $S_{\text{non-binding}}$ is the set of $K$ nearest neighbors of the query residue) and $\alpha$ is a parameter needs to be tuned. Then we calculated the propensity of the query residue belonging to the class $c$ as

$$P_c = u_c / \sum u_c$$

where $P_c$ is in the range of 0-1. By default the query residue was predicted as the class which has the higher propensity value.

C. Under-Sampling

Both benchmark datasets used in this study were highly imbalanced. For example, the ratio of positive samples (i.e., ATP-binding residues) vs. negative samples (i.e., non-binding residues) was ~1:19 for ATP168 dataset and ~1:23 for ATP227 dataset. Learning from datasets that contain very few instances of the minority class usually produces biased classifiers that have a higher predictive accuracy over the majority class [20]. Under-sampling of the majority class has been proved by many previous studies to be efficient in improving the prediction performance. Due to the property of K-NN method (i.e., the query residue needs to calculate its distances to all training residues to find closest neighbors), under-sampling has an additional advantage of speeding up the prediction process. In this study, under-sampling was only performed on training dataset while all samples in the test set was kept intact. An under-sampling ratio 1:n in our study means that all minority class samples were untouched, however majority class samples were removed at random until the number was $n$ times the number of minority class samples. Various $n$ values ranging from 1-10 were investigated.

D. Second-layer Classifier and Features

The output of K-NN classifier (i.e., propensity of being ATP-binding) further served as one of features of the second-layer classifier. In addition we have investigated 11 more features which have been indicated useful in predicting protein functions and functional sites by previous studies. For each residue of each protein, the entropy was extracted from the HHSSP database [21] and the disorder score was calculated by DISOPRED [22]. The remaining 9 features were AAIndex values of a target residue [23]. Same as [24], the following AAIndex indices were used: BULH740101 (transfer free energy to surface), EISD840101 (consensus normalized hydrophobicity), HOPT810101 (hydrophilicity value), RADA880108 (mean polarity), MCMT640101 (refractivity), BHAR880101 (average flexibility indices), CHOC750101 (average volume of buried residue), COSI940101 (electron-ion interaction potential values).

For the second-layer classifier, we used the Threshold Selector classifier implemented in WEKA [25]. The Threshold Selector classifier chooses a threshold so that F-measure is optimized and then does the classification process based on that threshold and the scores from the logistic classifier.

The detail workflow of the proposed method (TSC_ATP) is shown in Fig. 1.
Performance is measured using sensitivity ($S_a$), specificity ($S_p$), accuracy (Acc), MCC (Matthews Correlation Coefficient), and AUC (area under the ROC curve). Sensitivity is the fraction of ATP-binding residues that are correctly predicted as ATP-binding. Specificity is the fraction of non-binding residues that are correctly predicted as non-binding. False negatives (FN) are ATP-binding residues predicted as non-binding. False positives (FP) are non-binding residues predicted as ATP-binding.

Performance is measured using sensitivity ($S_a$), specificity ($S_p$), accuracy (Acc), MCC (Matthews Correlation Coefficient), and AUC (area under the ROC curve). Sensitivity is the fraction of ATP-binding residues that are correctly predicted as ATP-binding. Specificity is the fraction of non-binding residues that are correctly predicted as non-binding. False negatives (FN) are ATP-binding residues predicted as non-binding. False positives (FP) are non-binding residues predicted as ATP-binding.

\begin{align}
S_a &= \frac{TP}{TP + FN} \\
S_p &= \frac{TN}{TN + FP} \\
\text{Acc} &= \frac{(TP + TN)}{(TP + TN + FP + FN)} \\
\text{MCC} &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)}}
\end{align}

### III. Results and Discussion

#### A. Identification of ATP-binding Residues by K-NN

We investigated the prediction ability of the first-layer K-NN classifier which uses PSSM profiles of a sliding window centered on each residue as features.

For the K-NN method in this study, there are two parameters (i.e., $K$ and $\alpha$) to be tuned. The optimal values of $K$ and $\alpha$ depend on the data and should be found by experiment. To find the best $K$ and $\alpha$, various $K$ values from 1 to 100 with step size of 1 and $\alpha$ values from 0 to 15 with step size of 0.1 were tried to optimize the performance (i.e., AUC). Five-fold cross-validation was used to evaluate the performance. On ATP227 dataset, the best performance was achieved when $K = 49$ and $\alpha = 12$. As can be seen from Table I (row 2), the K-NN method achieved 95.8% accuracy, 29.8% sensitivity, 99.3% specificity, 0.432 MCC, and 0.786 AUC. On ATP227 dataset, K-NN achieved the best performance of 96.8% accuracy, 36.9% sensitivity, 99.4% specificity, 0.498 MCC, and 0.824 AUC when $K = 18$ and $\alpha = 10.4$, as can be seen from Table II (row 2).
Because both datasets are highly imbalanced (i.e., the ratio of positive samples vs. negative samples is ~1:19 for ATP168 dataset and ~1:23 for ATP227 dataset), we then seek to improve the prediction performance and to speed up the prediction process by undersampling majority class samples (i.e., negative samples, or non-binding residues). An undersampling ratio 1:\(n\) means that all minority class samples (i.e., ATP-binding) were untouched, however majority class samples (non-binding) were randomly chosen until the number was \(n\) times the number of minority class samples. Various \(n\) values ranging from 1-10 were investigated. Notice that undersampling process was only performed on the training dataset, while all samples in the test set were kept intact to reflect the real-life prediction task. For both ATP168 and ATP227, the best under-sampling ratio was 1:5. After sampling, K-NN achieved the optimal performance on both datasets when \(K = 99\) and \(\alpha = 7.8\). On ATP168 dataset, K-NN achieved 95.3\% accuracy, 35.3\% sensitivity, 98.5\% specificity, 0.419 MCC, and 0.810 AUC, as can be seen from Table I (row 3). It also achieved 96.1\% accuracy, 43.5\% sensitivity, 98.3\% specificity, 0.455 MCC, and 0.845 AUC on ATP227 dataset, as indicated by Table II (row 3).

B. Prediction by the Second-layer Classifier

The raw output (i.e., propensity) of the first-layer K-NN classifier together with 11 other features (i.e., 9 AAindex features, entropy, and disorder score of each residue) served as input to the second-layer classifier (i.e., Threshold Selector with the Logistic classifier). Using the same sampling and chosen parameters (i.e., \(K = 99\) and \(\alpha = 7.8\)), five-fold cross-validation was performed to evaluate the prediction performance of the two-stage classifier. Our method achieved 94.4\% accuracy, 40.2\% sensitivity, 97.8\% specificity, 0.416 MCC, and 0.829 AUC on ATP168 dataset (see Table I, row 4), and 96.2\% accuracy, 41.9\% sensitivity, 98.5\% specificity, 0.460 MCC, and 0.860 AUC on ATP227 dataset (see Table II, row 4). Figure 1 and 2 show ROC curves of our method on both ATP168 dataset and ATP227 dataset, which indicate that the prediction performance was gradually improved from the first-layer K-NN method, K-NN method after under-sampling, to the second-layer Threshold Selector classifier.

C. Comparison With Other Methods

We also compared our final TSC_ATP method with other state-of-the-art methods on both ATP168 and ATP227 datasets. The prediction performances of ATPint [7], ATPsite [8], Zhang et al.’s method [10], and NsitePred [11] were directly obtained from their reports. ATPsite and NsitePred only have results reported on ATP227, while ATPint and Zhang et al.’s method have results on both ATP168 and ATP227. Table III shows that our method (i.e., TSC_ATP) was comparable to ATPint in terms of AUC on ATP168. On ATP227, as can be seen from Table IV, our method outperformed ATPint and ATPsite and was comparable to NsitePred in AUC. We noticed that on both datasets our method had slightly lower AUC than that of Zhang et al.’s method. However, our method had much higher specificities. In a typical ATP-binding protein, there are a lot more non-binding residues than ATP-binding residues. Therefore, high specificity can significantly reduce the number of false positives. In comparison, Zhang et al.’s method achieved higher sensitivities with lower specificity and therefore was capable of identifying more ATP-binding residues by tolerating many more false positives. Both methods complement each other, indicating that a possible ensemble method with higher performance could be developed. Among all methods, our method had very similar performance as that of NsitePred on ATP227 dataset on all criteria. However, no method outperformed all other methods in all criteria.
IV. CONCLUSION

In this study we presented TSC_ATP, a two-stage classifier for identifying ATP-binding residues on proteins using only sequence information. The output (i.e., propensity score) of the first-layer classifier (i.e., K-NN method based on PSSM profiles) plus 11 other features served as input to the second-layer Threshold Selective Classifier to make predictions. TSC_ATP achieved 0.829 AUC and 0.860 AUC on two benchmark datasets (i.e., ATP168 and ATP227). Comparison with other methods indicated that TSC_ATP outperformed some previously published methods and was comparable to some of the highest performing methods. Our future work includes investigating more sequence-based features such as predicted structural information of each residue to improve the performance, performing feature selection and making our method publicly accessible by providing an online web server.

**TABLE III. COMPARISON WITH OTHER METHODS ON ATP168**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Sn</th>
<th>Sp</th>
<th>Acc</th>
<th>MCC</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATPint [7]</td>
<td>74.4%</td>
<td>75.8%</td>
<td>75.1%</td>
<td>0.50</td>
<td>0.823</td>
</tr>
<tr>
<td>Zhang et al.'s method [10]</td>
<td>77.4%</td>
<td>77.6%</td>
<td>77.6%</td>
<td>NA</td>
<td>0.858</td>
</tr>
<tr>
<td>TSC_ATP</td>
<td>40.2%</td>
<td>97.8%</td>
<td>94.4%</td>
<td>0.42</td>
<td>0.829</td>
</tr>
</tbody>
</table>

**TABLE IV. COMPARISON WITH OTHER METHODS ON ATP227**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Sn</th>
<th>Sp</th>
<th>Acc</th>
<th>MCC</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATPint [7]</td>
<td>53.9%</td>
<td>65.1%</td>
<td>64.8%</td>
<td>0.08</td>
<td>0.627</td>
</tr>
<tr>
<td>ATPsite [8]</td>
<td>36.1%</td>
<td>98.8%</td>
<td>96.2%</td>
<td>0.43</td>
<td>0.854</td>
</tr>
<tr>
<td>Zhang et al.'s method [10]</td>
<td>80.1%</td>
<td>80.3%</td>
<td>80.3%</td>
<td>NA</td>
<td>0.881</td>
</tr>
<tr>
<td>NsitePred [11]</td>
<td>44.4%</td>
<td>98.2%</td>
<td>96.0%</td>
<td>0.46</td>
<td>0.861</td>
</tr>
<tr>
<td>TSC_ATP</td>
<td>41.9%</td>
<td>98.5%</td>
<td>96.2%</td>
<td>0.46</td>
<td>0.860</td>
</tr>
</tbody>
</table>

CONTRIBUTION

JH conceived of and designed the study, performed the analysis and drafted the manuscript. BJA contributed to most of the computation. Both authors read and approved the final manuscript.

REFERENCES


