Journal of Molecular Graphics and Modelling 92 (2019) 86–93

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Prediction of ATP-binding sites in membrane proteins using a two-dimensional convolutional neural network

Trinh-Trung-Duong Nguyen [a], Nguyen-Quoc-Khanh Le [b], Rosdyana Mangir Irawan Kusuma [a], Yu-Yen Ou [a, *]

[a] Department of Computer Science and Engineering, Yuan Ze University, Chung-Li, 32003, Taiwan
[b] School of Humanities, Nanyang Technological University, 48 Nanyang Ave, 6397983, Singapore

ARTICLE INFO

Article history:
Received 30 April 2019
Received in revised form
24 June 2019
Accepted 13 July 2019
Available online 15 July 2019

Keywords:
Membrane protein
Bioinformatics
Deep learning
Convolutional neural network
Position specific scoring matrix
Imbalanced data

ABSTRACT

Membrane proteins, the most important drug targets, account for around 30% of total proteins encoded by the genome of living organisms. An important role of these proteins is to bind adenosine triphosphate (ATP), facilitating crucial biological processes such as metabolism and cell signaling. There are several reports elucidating ATP-binding sites within proteins. However, such studies on membrane proteins are limited. Our prediction tool, DeepATP, combines evolutionary information in the form of Position Specific Scoring Matrix and two-dimensional Convolutional Neural Network to predict ATP-binding sites in membrane proteins with an MCC of 0.89 and an AUC of 99%. Compared to recently published ATP-binding site predictors and classifiers that use traditional machine learning algorithms, our approach performs significantly better. We suggest this method as a reliable tool for biologists for ATP-binding site prediction in membrane proteins.

1. Introduction

Membrane proteins are the proteins that interact with, or are an integral part of biological membranes. They account for about 30% of the total proteome of an organism [1,2]. Membrane proteins play several important roles in living organisms such as transporting ions and molecules across the membrane, binding small molecules in the extracellular matrix, recognizing components of immune system, communicating inter-cellular signals, and transducing energy. A large portion of the membrane proteins bind adenosine triphosphate (ATP), which is then hydrolyzed to adenosine-5′-diphosphate (ADP), releasing energy which fuels the downstream functions.

ATP-binding membrane proteins play several important roles. Human ATP-binding Cassette (ABC) transporters are membrane proteins that serve as drug targets for cancer, metabolic disorders, and gene therapy applications [3]. These transporters can expel chemotherapeutic agents from tumor cells, thus contributing to cancer drug resistance. They also contribute to expulsion of antibiotics from bacterial cells, thus rendering them resistant to antibacterial drugs [4]. Similarly, ABC transporters mediate extrusion of an array of substrates like amino acids, lipids, and xenobiotics. Therefore, they are essential pharmacokinetics, particularly in drug-drug interaction and adverse drug effects [5]. Application of ABC membrane proteins has extended to veterinary medicine. ABC transporters may affect the bioavailability and elimination of various drugs and other xenobiotics in domestic animals [6].

Over the past decade, many research groups have worked on determining the location of ATP-binding sites [7–16]. Chauhan et al. developed a Support Vector Machine (SVM)-based model using primary sequence of proteins and obtained maximum Matthews correlation coefficient (MCC) of 0.33 with an accuracy of 66.25% [8]. This study included a dataset of 168 non-redundant ATP-binding proteins, denoted as ATP168, which became a benchmark dataset used by other research groups. Chen et al. built ATPsite and NsitePred based on much larger benchmark dataset [9,10] denoted as ATP227, including 227 non-redundant ATP-binding proteins, and better prediction results were achieved with 0.46 MCC and 0.86 AUC (Area under the curve). Interestingly, Dong-Jun Yu group combined sequence evolutionary information, bi-profile sampling of multi-view sequential features, and the sequence-derived structural features [16]. In another study by the

* Corresponding author.
E-mail address: yienou@gmail.com (Y.-Y. Ou).

https://doi.org/10.1016/j.jmgm.2019.07.003
1093-3263/

same group [14], the imbalanced learning problem was resolved by using random under-sampling. Both the studies used both the datasets mentioned above and the best AUC achieved with ATP168 and ATP227 were 86% and 88%, respectively. A recent work published by the same group [12] includes a combination of sequence-profiling and structure-based comparisons in ATPbind predictor with AUC 92%. In this study, a dataset containing 429 non-redundant ATP-binding proteins, denoted as ATP429, was created and published.

Most studies mentioned above used SVM-based approach to build ATP-binding site prediction tools. Although promising, traditional machine learning techniques have limited capability to process raw data [17]. These impediments have been overcome with the advent of deep learning technology. Regarding binding site prediction problems, convolutional neural network, a deep learning architect, has been successfully employed [18–31]. Among these studies, depending on the types of the extracted features, one of these flavors of convolutional neural networks (CNNs) was used: one-dimensional CNN, two-dimensional CNN, three-dimensional CNN, multiple CNN, and hybrid CNN. In our previous study, state-of-the-art performance on identifying GTP binding sites in rab proteins was achieved with the use of two-dimensional CNN (2D-CNN) and the evolutionary information residing in position-specific scoring matrix (PSSM) [31]. Before the arrival of deep learning applications in bioinformatics, the PSSM profile has been used as the most discriminative and effective features in many prediction models. When being input to traditional machine learning algorithms, the PSSM profile is often transformed into a vector. However, the PSSM profile can also be treated as a matrix of pixels of an image and thus can be used as an input for a two-dimensional convolutional neural network. In computer vision research, convolutional neural network is a powerful tool for analyzing visual imagery and has been widely applied with impressive classification performance on a benchmark image dataset [32]. Many technology companies such as Facebook, Google, and Amazon have been utilizing 2D-CNNs as the essence of their services. In both research and industry, 2D-CNNs has been proved to be the de-facto models for many classification tasks.

Many life-essential biology mechanisms can be understood by identifying accurately the ATP-binding sites in these proteins. Though existing ATP-binding site predictors can be used to predict ATP-binding membrane proteins but lack of specificity reduces their potential. Furthermore, there are much less ATP-binding site predictors based on deep learning. Therefore, in the present study, we developed a model named DeepATP to identify ATP interacting residues in membrane proteins using a 2D-CNN. Additionally, in this study, we addressed the specific and greatly imbalanced dataset issue of ATP-binding sites in membrane proteins and achieved a remarkable AUC performance of 0.991 on independent test data.

2. Methods

Our methodology includes three sub processes: data collection, feature set generation, and model evaluation (Fig. 1). The details of the proposed method are described as follows.

2.1. Data collection

We collected 15,543 membrane protein sequences that had clear target annotations and had been deposited into the UniProt database (release 2017_08) before August 30, 2017. From the description files of these proteins, we filtered out the proteins for which ATP-binding sites were experimentally determined (proteins with evidence code ECO:0000250 were not retrieved). After this, 755 remaining ATP-binding membrane proteins were used for further evaluation. Subsequently, BLAST [33] was applied to exclude the sequences with a sequence identity of more than 20%. Finally, 216 membrane proteins having ATP-binding residues were used to form the dataset for the present study, namely ATPMembrane. Further, details of our dataset is presented in Dataset section on web server at http://www.biologydeep.com/deepatp/.

We further carried out experiments on the above-mentioned ATP168, ATP227 and ATP429 datasets to validate the effectiveness of our apporach. Among these datasets, ATP168 and ATP277 are the benchmark datasets that have been used in several ATP-binding site predictors [8,10,11,14,15] whereas ATP429 is the latest benchmark dataset used in a study published in 2018 [12]. The sequence identity of any two proteins in ATP168, ATP227 and ATP429 is less than 40%. By applying deep learning on datasets about general ATP-binding proteins, besides the main dataset about ATP-binding membrane proteins, further insights about these datasets can be gained. Finally, each dataset is divided into 2 parts: the first part (80%) for training and performing model selection via 5-fold cross-validation technique and the second part (20%) for testing. Table 1 shows the statistical composition of each dataset.

2.2. Feature generation

PSSM profile is a matrix for decoding the evolutionary information of a protein sequence. A PSSM profile for a query protein is an Nx20 matrix, where N is the length of the query protein sequence. It assigns a score P_{ij} for the jth amino acid in the ith position of the query sequence with a large value indicating a highly conserved position and a small value indicating a weakly conserved position. Many studies show its outstanding discriminative capability for many prediction problems in bioinformatics. Additionally, almost all the sequence-based methods utilize PSSM profiles as input. Chen et al. reported that the exclusion of PSSM profiles leads to a larger decrease in prediction performance than the exclusion of other input features [9,10]. Nine out of ten recent research on ATP-binding residue prediction used PSSM profiles with or without additional features to build ATP-binding site predictors and achieved satisfactory results [7–16]. This suggests that the position specific scoring matrix plays a key role in identification of the nucleotide binding sites.

We accordingly utilized the PSSM profiles to develop our 2D-CNN model. For each n residues long ATP-binding sequence, its PSSM profile was constructed (n rows and 20 columns) using the PSI-Blast version 2.2.26 [33] to search against non-redundant database (nr; version 2015) for three iterations with cutoff e-value of 0.001. If a window_size of 17 is used, then the matrix size is $17 \times 20 = 340$ (because the number of calculated values for each amino acid is 20). Among different window_sizes from 3 to 21, a window_size of 17 generated optimal results, which is in concordance with the previous findings [8,9,11,14,15].

2.3. Imbalanced dataset

A common problem in binding site prediction research is the imbalance of the dataset(s). It means that the number of negative samples is much higher than the number of positive samples. In the current study, the numbers of negative examples are 98.8, 19.3, 23.4 and 25.1 times higher than those of positive examples for ATP-Membrane, ATP168, ATP227 and ATP429, respectively (Table 2). A predictor trained by such a highly skewed dataset would inevitably mispredict many binding sites as non-binding ones [34].

A data processing approach (resampling the training set) was adopted to overcome the imbalanced dataset problem. Resampling the training set comes in two flavors: oversampling and

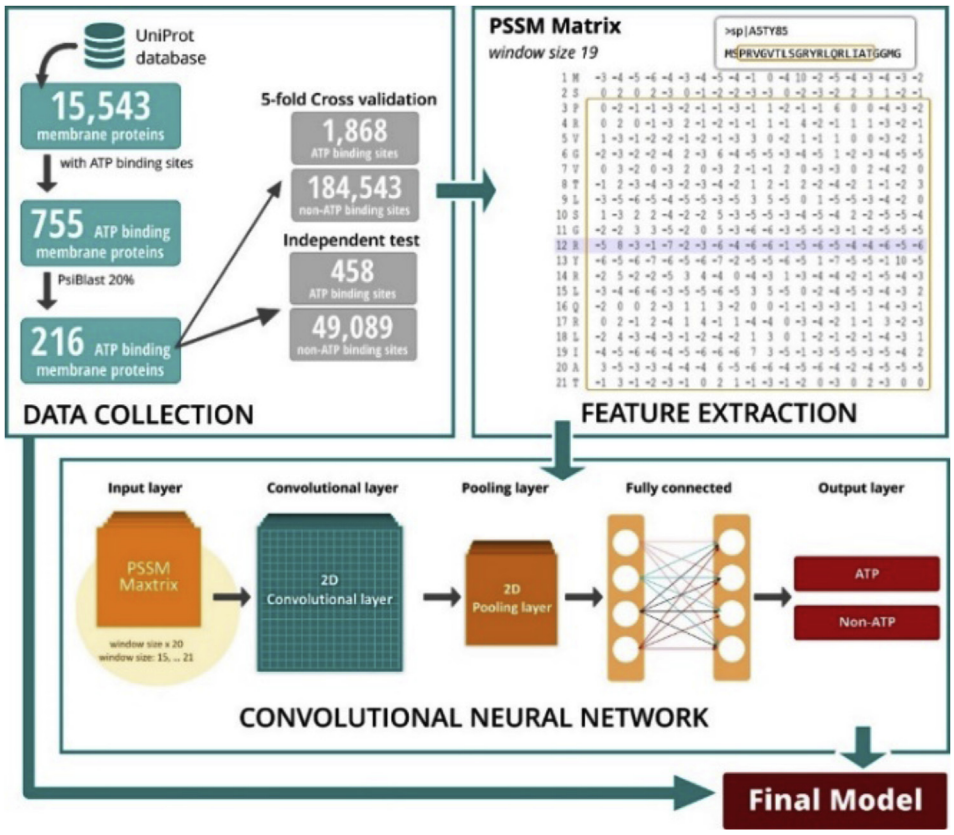

Fig. 1. The whole architecture for identifying ATP-binding sites in membrane proteins using 2D-CNN and PSSM profile.

Table 1
Statistical composition of the training and independent data sets.

Dataset	Original data	Training	Independent
ATPMembrane	216	173	43
ATP168	168	132	36
ATP227	227	171	56
ATP429	429	388	41

Table 2
Statistical Composition of Binding and Non-binding sites in 4 datasets in training data.

Dataset	Number of protein sequences	numP[a]	numN[b]	ratio[c]
ATPMembrane	173	1868	184,543	98.8
ATP168	132	2440	47,058	19.3
ATP227	171	2440	57,032	23.4
ATP429	388	5657	142,086	25.1

[a] numP represents the number of positive samples.
[b] numN represents the number of negative samples.
[c] ratio = numN/numP.

undersampling.

In order to have sufficient data for the deep learning method and avoid losing valuable information, we utilized oversampling approach which was successfully applied in imbalanced data problem in our previous work [35]. Accordingly, multiple datasets comprising increasing balanced amounts of positive examples and negative examples were evaluated and the results were recorded. More specifically, the dataset with ratio n would contain n replications of minority class samples and the same amount of randomly chosen majority class samples. We tried this method on each dataset and chose the best ratio in respect of AUC for each dataset. Table S2 in Supplementary materials presents the best ratio found for each dataset.

2.4. Convolutional neural network construction

The lower portion of Fig. 1 illustrates the fundamentals of layer construction for a simplified CNN. In general, CNN is composed of multiple layers with each layer performing a specific function of transforming its input into a useful representation. All layers were combined using a specific ordering to form the architecture of our CNN model. Three major types of layers are commonly observed in CNN architectures: the convolutional layer, activation layer, and pooling layer. Different layers used in CNN for the current study include:

(1) Input layer: The input of the CNN is a PSSM profile, which is a matrix of numbers of size: window_size x 20
(2) Zero padding layer: Zero-padding is the process of symmetrically adding zeros to the input matrix that allows the size of the input to be adjusted to certain requirements. In the model presented in the current study, zero values were added at the beginning and ending of the window_size x 20 matrices. This allowed us to apply the filter to the border positions of the matrices.
(3) Convolutional layer: A convolutional layer is used to extract features encoded in the 2D input matrix via convolution operations. The convolutional layer takes a sliding window that moves in stride across the input transforming the values into representative values. During this process, convolution operation preserves the spatial relationship between

numeric values in the PSSM profiles by learning useful features using small squares of input data. When constructing convolutional layer, two values need to be determined: the number of kernels (filters) and the kernel size. The common choices for kernel size are 3×3, 5×5 or 7×7. As the input of our CNN model is a matrix of small size (window_size x 20), compared to the image size used in computer vision, we used the kernel size of 3×3 to deduce more information.

(4) Activation layer: An additional non-linear operation, called ReLU (Rectified Linear Unit) was performed after every convolutional operation. Its output is defined by the formula:

$$f(x) = \max(0,x) \quad (1)$$

where, x is input value.

(5) The purpose of ReLU is to introduce non-linearity in our CNN and help our model learn better from the data.
(6) Pooling layer: The pooling layer is usually inserted among the convolutional layers with the aim at reducing the size of matrix calculation for the next convolutional layer. The operation performed by this layer is also called "down-sampling" as it removes certain values leading to less computational operations and over-fitting control while still preserving the most relevant representative features. The "down-sampling" is either performed by taking the maximum value in the window (max pooling), or by taking the average of the values (average pooling). In our study, we performed max pooling and down sampled the data by a factor of 2 through selection of the maximum value over a window of 2×2.
(7) Batch normalization layer was included to allow a deeper network, which means that more hidden layers could be added.
(8) Dropout layer was added to enhance the predictive performance of the present model and preventing overfitting.
(9) Flatten layer: The flatten layer is always included before the fully connected layers to convert the input matrix into a vector.
(10) Fully connected layer: In a fully connected layer, each node is fully connected with all the nodes of the previous layers. Fully connected layers are typically used in the last stages of the CNN. In the present model, two fully connected layers were included. The first layer connected all the nodes into flatten layer to allow our model gain more knowledge and perform better. The second layer connected the first fully connected layer to the output layer. As predicting ATP-binding sites can be considered as a binary classification problem, the number of nodes in the output layer is equal to 2.
(11) Softmax: The output of the model was evaluated through a softmax function by which the probability for each possible output was determined. Softmax function is a logistic function defined by the formula:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}} \quad for\ j = 1, \ldots K \quad (2)$$

z in the above formula indicates the K-dimensional vector that is input to the last layer, $\sigma(z)_j$ is real values in the range (0, 1) and jth class is the predicted probability from sample vector x. K is considered equal to 2 in the present study.

2.5. Hyperparameter optimization

Hyperparameters are architecture-level parameters and are different from parameters of a model trained via backpropagation. When building a deep learning model, the choice of these hyperparameters is governed by a number of factors: How many layers should be stacked? How many filters each layer should have? How to choose the dropout rate? Should we apply batch normalization layers? Certainly, the choice of hyperparameters significantly affects the performance of the model.

Accordingly, in this study, the following approach was used for tuning hyperparameters:

(1) Choosing the set of hyperparameters for speeding up the training process
(2) Choosing the set of hyperparameters for preventing overfitting

As suggested by Chollet [36], each step of the above hyperparameter-tuning approach was integrated into the hyperparameter-tuning process as follows:

(1) Choose a set of hyperparameters.
(2) Build the corresponding model.
(3) Fit the training data to the model, and measure the final performance on the validation dataset.
(4) Try the next set of hyperparameters.
(5) Repeat.
(6) Eventually, measure performance on independent dataset.

We used Keras framework library (version 2.1.1) with a Tensorflow backend [37] and Python (version 2.7.13) to build the DeepATP model. Grid search was performed and AUC was used as a metric to choose the next set of hyperparameters. Earlier, our initial study showed that among the 6 optimization algorithms Adam [38], Adadelta [39], Adagrad [40], Stochastic Gradient Descent (SGD) [41], RMSprop [42], and Adamax supported by Keras API, Adadelta gave the best performance (See Table S1 in the Supplementary materials). Accordingly, Adadelta was selected as the optimization algorithm and used for tuning other hyperparameters. Table 3 below shows the hyperparameters used for tuning the model. We also presented the best set of hyperparameters found for each datasets in Table S2 in Supplementary materials.

2.6. Assessment of predictive ability

The most important purpose of the present study was to predict whether or not an amino acid at a specific position is an ATP-binding site; therefore, we used "Positive" to define the location of an ATP-binding site, and "Negative" to define the location of a non-ATP-binding site. For each dataset, we first trained the model by applying 5-fold cross validation technique on the oversampled training dataset. Based on the 5-fold cross validation results, hyperparameter optimization process was employed to find the best model for each dataset. Finally, the independent dataset was used to assess the predictive ability of the current model.

For evaluating the performance of the methods, some standard metrics were used, such as accuracy (Acc), sensitivity (Sen), specificity (Spec), Matthews's correlation coefficient (MCC), precision (Pre), and F_1 Score or F-measure (F_1) using below given formulae (TP, FP, TN, FN are true positive, false positive, true negative, and false negative values, respectively):

Table 3
Hyperparameters used for tuning the model.

Hyperparameters for speeding up the training process	
Number of epoch	10 to 500
Batch size	32, 64, 128
Learning rate	1, 1.2, 0.8
Hyperparameters for preventing dropout	
Dropout rate	0.1, 0.2, 0.3, 0.4, 0.5
Weight regularization	1×10^{-5}, 5×10^{-5}, 1×10^{-4}, 5×10^{-4}, 1×10^{-3}, 5×10^{-3}, 1×10^{-2}

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$F1 - Score = \frac{2TP}{2TP + FP + FN} \quad (8)$$

3. Results and discussions

In this study, sequence logos and *n*-gram model were used to analyze sequence motifs in the generated dataset. We further compared the proposed method with previous methods using SVM, Random forest, kNN classifiers. Finally, the performance of our approach was compared with existing ATP-binding site predictors.

3.1. Interacting and non-interacting residue frequency analysis

We analyzed the frequency of interacting and non-interacting ATP-binding sites in training data by extracting ATP-binding fragments and non-ATP-binding fragments of the same size using a window_size of 17. Seq2Logo web server [43] was used to generate the sequence frequency logo as shown in Fig. S1 for ATPMembrane, Fig. S2 for ATP168, Fig. S3 for ATP227, and Fig. S4 for ATP429 datasets.

In each sequence logo, we have a stack of letters at each position. The relative sizes of the letters indicate their frequency in the sequence while the total height of the letters depicts the information content of the specific position. In Figs. S1a, S2a, S3a and S4a, some differences can be observed among the ATP-binding proteins in ATPMembrane, ATP168, ATP227 and ATP429 datasets. For example, the position having the most information content in ATP-binding proteins in ATPMembrane is 9th while those of ATP168, ATP227 and ATP429 are 7th, 7th and 8th, respectively. Additionally, for non-ATP-binding proteins in ATPMembrane, ATP168, ATP227 and ATP429 (Figs. S1b, S2b, S3b and S4b), amino acid E contributes more significantly in the non-ATP-binding proteins in ATP168, ATP227 and ATP429 in comparison to ATPMembrane. We also observed that amino acid S contributes more significantly in non-ATP-binding proteins in ATPMembrane than in ATP168, ATP227 and ATP429 datasets.

3.2. Analysis of the important sequence motifs

In this analysis, we tried to observe the motifs that more frequently appear in the surveyed protein sequences. Only motifs of length from 1 to 5 residues were analyzed and were called unigram, bigram, trigram, four-gram, and five-gram, respectively. The number of times (Freq) these *n*-grams appeared in the protein sequences were also counted and showed in Table S3. As observed from Table S3, it is interesting that ATP168 and ATP227 have highly similar n-gram distribution. For example, both the datasets have HHH, LEE, ELL, EEL, LLD as the most frequently occurring 3-g. In addition, ATP168, ATP227, and ATP429 share 8 out of 10 most regular 2-g (LL, AL, LA, LE, VL, EL, AA, LK). Among 4 datasets, ATP-Membrane shares the least commonly appearing n-grams. This indicates that there are some certain amino acid patterns or motifs distinguishing ATP-binding membrane proteins and general ATP-binding proteins.

3.3. The receiver operating characteristic (ROC) curves

The receiver operating characteristic (ROC) curves are often used to examine the predictive ability. For a cutoff threshold (between 0 and 1), all the residues with output probability P greater than the cutoff threshold are designated as binding residues and all other residues are designated as non-binding residues. Next, the sensitivity (TP/(TP + FN)) and the false positive rate (FP/(FP + TN)) are calculated to draw the ROC curve.

To compare different classifiers, ROC performance should be reduced to a single scalar value representing expected performance. Commonly, area under the curve (AUC) is used to quantify the predictive quality. Unlike the measures that assess the binary predictions, the AUC value considers all possible thresholds and thus provides a more comprehensive evaluation of predictive ability.

Fig. 2a and Fig. 2b illustrate the ROC curve for predicting ATP and non-ATP-binding sites in each dataset on the Cross-validation data and the independent data, respectively. We easily observe that with our 2D-CNN model, the ROC curve was absolutely reaching the ideal point at AUC = 0.962 and AUC = 0.99, respectively for the cross-validation data and independent data of ATPMembrane dataset. These performance values obviously surpassed those of other datasets. Therefore, this method is extremely appropriate for predicting the ATP-binding sites in membrane proteins.

3.4. Performance of proposed method on 4 datasets

In order to get a more specific evaluation, the threshold (0.019) was chosen to determine the predictive performance of our 2D-CNN model on 4 datasets. Table 4 shows the results of the model with best values highlighted in bold.

As shown in Table 4, it is interestingly confirmed that our 2D-CNN models work well on membrane proteins but is not very significant on general proteins. This can be explained by the use of ATP in only membrane as our main training set, and the model has good

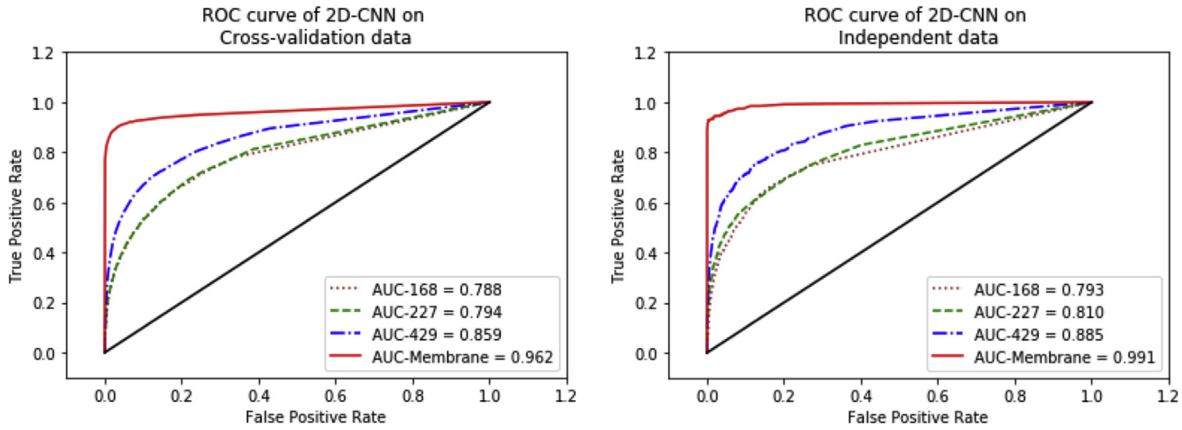

Fig. 2. a. The receiver operating characteristic (ROC) curves demonstrating the predictive performance of DeepATP on 4 datasets on the Cross-validation data. **b**. The receiver operating characteristic (ROC) curves demonstrating the predictive performance of DeepATP on 4 datasets on the Independent data.

Table 4
Performance of 2D-CNN models on 4 datasets.

Cross-validation						
Dataset	Acc	Sen	Spec	MCC	Pre	F1-score
ATPMembrane	**99.6**	**79.3**	**99.8**	**0.8**	**80.5**	**79.9**
ATP168	95.2	20.9	99.1	0.32	54.4	30.2
ATP227	95.4	16.0	99.6	0.303	66.2	25.7
ATP429	96.7	27.3	99.4	0.406	63.4	38.2
Independent Dataset	Acc	Sen	Spec	MCC	Pre	F1-score
ATPMembrane	**99.8**	**90.4**	**99.9**	**0.887**	**87.2**	**88.7**
ATP168	95.1	20.8	99.2	0.329	58.5	30.7
ATP227	96	25.6	99.3	0.384	62.7	36.4
ATP429	96.3	28.6	99.6	0.451	75.4	41.5

Values in bold are the best performance

performance on this specific dataset. The second, ATP-binding sites are ubiquitous in membrane proteins, so the development of this dataset will help us generate special features for ATP-binding sites that can not be generated with general proteins. It will help to narrow down the problem and provide a significant model for biologists who would like to identify ATP-binding sites in membrane only.

3.5. Comparison between the proposed method and traditional machine learning methods

We performed comparative evaluation of the predictive ability of our 2D-CNN model and other classifiers such as kNN, RandomForest and SVM on the ATPMembrane dataset. We also determined the optimal hyperparameters for each of these classifiers using grid search technique. Both the validation and independent datasets were included for comparison and the results were evaluated assuming the standard threshold. Table 5 shows the performance comparisons of the proposed method and the other classifiers with best results highlighted in bold. The results of comparative analysis showed that the proposed method performs better compared to the other classifiers on both the validation and independent data.

3.6. Comparison between the proposed method and other predictors

Previous studies have provided web-servers for predicting ATP-binding sites and ATP-binding residues. We used the independent test dataset of ATPMembrane for ATP-binding site prediction using ATPint, NsitePred and ATPbind web server [8,10,12] for comparing the predictive performance of our 2D-CNN model. The comparison is showed in Table 6 with the best results highlighted in bold. From the comparative analysis, we found that ATP-binding residues in membrane proteins can be much better identified using a specific model like DeepATP. As DeepATP was trained on membrane proteins and used for predicting ATP-binding residues in membrane proteins, it can benefit biologists interested in these important proteins more than the existing general predictors.

In case of the studies that did not implement a web server or for which the provided web server was not working, we used the reported AUC when comparing predictive performance. The model proposed in the present study achieved the highest AUC at 0.99.

3.7. Web server for predicting ATP-binding sites in membrane proteins

A web server 'DeepATP' was built based on the method proposed in this study. The user may submit the amino acid sequence(s) in 'FASTA' format. The server then generates the evolutionary profile of all the submitted sequences and predicts the interacting residues. In the output, ATP-interacting residues are displayed in red. The web server is freely available at http://www.biologydeep.com/deepatp/. The specific two-dimensional neural network architecture used for building DeepATP and the prediction time comparison among different classifiers are mentioned in Table S4 and Table S5, respectively in the Supplementary materials.

Table 5
Comparison of the DeepATP's ability to identify ATP-binding site in membrane proteins with other classifiers *(SVM: c = 8, g = 0.0004882, RandomForest: estimator = 277, max_feature = "auto", kNN: k = 500)*.

Cross-validation							
Classifer	Acc	Sen	Spec	MCC	Pre	F1-score	AUC
RF	99.2	73.5	99.5	0.661	59.8	65.9	0.865
kNN	93.1	68.8	93.4	0.241	9.5	16.7	0.852
SVM	99.5	**83.3**	99.7	0.772	72.1	77.3	0.915
2D-CNN	**99.6**	79.3	**99.8**	**0.8**	**80.5**	**79.9**	**0.962**
Independent							
Classifer	Acc	Sen	Spec	MCC	Pre	F1-score	AUC
RF	99.5	82.8	99.6	0.742	67.1	74.1	0.912
kNN	91.8	91.5	91.8	0.278	9.4	17.1	0.916
SVM	99.7	**93.4**	99.7	0.847	77.1	84.5	0.966
2D-CNN	**99.8**	90.4	**99.9**	**0.887**	**87.2**	**88.7**	**0.99**

Table 6
Comparative performance of DeepATP with other existing tools for identifying ATP-binding site in membrane proteins.

Predictor	Acc	Sen	Spec	MCC	Pre	F1-score
ATPint	84.7	86.7	84.6	0.186	5.0	9.5
NsitePred	98.0	27.0	99.7	0.431	71.2	39.1
ATPbind	98.8	73.4	99.0	0.548	41.8	53.3
2D-CNN	**99.8**	**90.4**	**99.9**	**0.887**	**87.2**	**88.7**

4. Conclusion

Drug transporters are membrane proteins that mediate the uptake and efflux of a wide range of therapeutic agents. These proteins play important roles in adsorption, distribution, metabolism and excretion (ADME) of various clinical drugs [44]. The ATP-binding cassette (ABC) family is a crucial family of drug transporters. ABC family members including P-glycoprotein (P-gp/ABCB1), multidrug resistance associated protein 1 (MRP1/ABCC1) and breast cancer resistance protein (BCRP/ABCG2) have been identified as key determinants of the pharmacokinetics and pharmacodynamics of various drugs [45]. Over-expression of P-gp, MRP1 and/or BCRP in the drug selected model cell lines has been reported to be one of the major mechanisms responsible for multidrug resistance (MDR) [46]. ABC transporters require ATP hydrolysis for substrate transport and they are mainly efflux transporters that mediate the transfer of drugs out of the cells. Hence, accurately identifying the ATP-binding sites in these proteins is highly desirable for facilitating the development of membrane proteins based therapeutic interventions.

In the current state of bioinformatics, deep learning models can be built that can reach near-human-level in image classification, speech recognition, handwriting transcription, and improved machine translation. In this study, we aimed to identify ATP-binding sites in membrane proteins by employing the power of deep learning model via two-dimensional convolutional neural network. As the structures of many proteins are currently not available, the protein sequences are more frequently used in ligand binding sites prediction. When building the predictor, we employed the highly discriminative ability of PSSM profiles and treated these matrices in the vision perspective. This enabled the use of a two-dimensional convolutional neural network.

On the independent test dataset, the method proposed in the present study obtained accuracy, sensitivity, specificity, and MCC values of 99.8%, 90.4%, 99.9%, 0.887, respectively. These results show that our model outperformed other classifiers based on a traditional machine learning algorithm in identifying ATP-binding sites in the membrane proteins. Moreover, ATP-binding residues in membrane proteins can be much better identified using a specific model like DeepATP, which was specifically trained on membrane proteins, compared to other existing general predictors. These results are even more significant given the fact that our dataset is greatly imbalanced. From our results, we can conclude that our model can be used as a reliable tool for predicting ATP-interacting sites in the membrane proteins.

Acknowledgments

This research partially supported by Ministry of Science and Technology, Taiwan, R.O.C. under Grant no. MOST 104-2221-E-155-037 and 105-2221-E-155-065.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmgm.2019.07.003.

References

[1] M.M. Gromiha, Y.Y. Ou, Bioinformatics approaches for functional annotation of membrane proteins, Briefings Bioinf. 15 (2) (2014) 155–168.
[2] S. Tan, H.T. Tan, M.C. Chung, Membrane proteins and membrane proteomics, Proteomics 8 (19) (2008) 3924–3932.
[3] N. Khunweeraphong, T. Stockner, K. Kuchler, The structure of the human ABC transporter ABCG2 reveals a novel mechanism for drug extrusion, Sci. Rep. 7 (1) (2017) 13767.
[4] L.M. Browning, et al., Single gold nanoparticle plasmonic spectroscopy for study of chemical-dependent efflux function of single ABC transporters of single live Bacillus subtilis cells, Analyst 143 (7) (2018).
[5] R.R. Crawford, et al., Beyond competitive inhibition: regulation of ABC transporters by kinases and protein-protein interactions as potential mechanisms of drug-drug interactions, Drug Metab. Dispos. 46 (5) (2018).
[6] G. Virkel, et al., Role of ABC transporters in veterinary medicine: pharmacotoxicological implications, Curr. Med. Chem. 26 (7) (2018).
[7] B.J. Andrews, J. Hu, *TSC_ATP*: a two-stage classifier for predicting protein-ATP binding sites from protein sequence, in: Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on, IEEE, 2015.
[8] J.S. Chauhan, N.K. Mishra, G.P. Raghava, Identification of ATP binding residues of a protein from its primary sequence, BMC Bioinf. 10 (2009) 434.
[9] K. Chen, M.J. Mizianty, L. Kurgan, ATPsite: sequence-based prediction of ATP-binding residues, in: Proteome Science, BioMed Central, 2011.
[10] K. Chen, M.J. Mizianty, L. Kurgan, Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors, Bioinformatics 28 (3) (2012) 331–341.
[11] C. Fang, T. Noguchi, H. Yamana, Simplified sequence-based method for ATP-binding prediction using contextual local evolutionary conservation, Algorithm Mol. Biol. 9 (1) (2014) 7.
[12] J. Hu, et al., ATPbind: accurate protein–ATP binding site prediction by combining sequence-profiling and structure-based comparisons, J. Chem. Inf. Model. 58 (2) (2018) 501–510.
[13] A.N. Mbah, Application of hybrid functional groups to predict ATP binding proteins, ISRN Comput Biol 2014 (2014) 581245.
[14] D.-J. Yu, et al., Improving protein-ATP binding residues prediction by boosting

SVMs with random under-sampling, Neurocomputing 104 (2013) 180–190.
[15] D.J. Yu, et al., TargetATPsite: a template-free method for ATP-binding sites prediction with residue evolution image sparse representation and classifier ensemble, J. Comput. Chem. 34 (11) (2013) 974–985.
[16] Y.-N. Zhang, et al., Predicting protein-ATP binding sites from primary sequence through fusing bi-profile sampling of multi-view features, BMC Bioinf. 13 (1) (2012) 118.
[17] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.
[18] B. Alipanahi, et al., Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning, Nat. Biotechnol. 33 (8) (2015) 831.
[19] H.R. Hassanzadeh, M.D. Wang, DeeperBind: enhancing prediction of sequence specificities of DNA binding proteins, in: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2016.
[20] J. Zhou, et al., CNNsite: prediction of DNA-binding residues in proteins using Convolutional Neural Network with sequence features, in: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2016.
[21] A. Morrow, et al., Convolutional Kitchen Sinks for Transcription Factor Binding Site Prediction, 2017 arXiv preprint arXiv:1706.00125.
[22] X. Gao, et al., DeepPolyA: a convolutional neural network approach for polyadenylation site prediction, IEEE Access 6 (2018) 24340–24349.
[23] J. Jiménez, et al., DeepSite: protein-binding site predictor using 3D-convolutional neural networks, Bioinformatics 33 (19) (2017) 3036–3042.
[24] H. Zeng, et al., Convolutional neural network architectures for predicting DNA–protein binding, Bioinformatics 32 (12) (2016) i121–i127.
[25] Y.S. Vang, X. Xie, HLA class I binding prediction via convolutional neural networks, Bioinformatics 33 (17) (2017) 2658–2665.
[26] J. Zuallaert, et al., Interpretable convolutional neural networks for effective translation initiation site prediction, in: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2017.
[27] J. Jiménez, et al., K DEEP: protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks, J. Chem. Inf. Model. 58 (2) (2018) 287–296.
[28] X. Pan, H.-B. Shen, Learning distributed representations of RNA sequences and its application for predicting RNA-protein binding sites with a convolutional neural network, Neurocomputing 305 (2018) 51–58.
[29] D. Wang, et al., MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction, Bioinformatics 33 (24) (2017) 3909–3916.
[30] X. Pan, H.-B. Shen, Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional neural networks, Bioinformatics 34 (20) (2018) 3427–3436.
[31] N.Q.K. Le, Q.-T. Ho, Y.-Y. Ou, Using two-dimensional convolutional neural networks for identifying GTP binding sites in Rab proteins, J. Bioinform. Comput. Biol. 17 (1) (2019), 1950005-1950005.
[32] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012.
[33] Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (0) (1997) 17.
[34] Y. Sun, A.K. Wong, M.S. Kamel, Classification of imbalanced data: a review, Int. J. Pattern Recognit. Artif. Intell. 23 (04) (2009) 687–719.
[35] S.W. Taju, N.-Q.-K. Le, Y.-Y. Ou, Using Deep Learning with Position Specific Scoring Matrices to Identify Efflux Proteins in Membrane and Transport Proteins, 2016, pp. 101–108.
[36] F. Chollet, Deep Learning with Python, Manning Publications Co, 2017.
[37] M. Abadi, et al., Tensorflow: a system for large-scale machine learning, OSDI (2016) 265–283.
[38] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2014 arXiv preprint arXiv:1412.6980.
[39] M.D. Zeiler, ADADELTA: an Adaptive Learning Rate Method, 2012 arXiv preprint arXiv:1212.5701.
[40] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, J. Mach. Learn. Res. 12 (Jul) (2011) 2121–2159.
[41] L. Bottou, Large-scale machine learning with stochastic gradient descent, in: Proceedings of COMPSTAT'2010, Springer, 2010, pp. 177–186.
[42] Bengio, Y., RMSProp and Equilibrated Adaptive Learning Rates for Non-convex Optimization.
[43] M.C. Thomsen, M. Nielsen, Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion, Nucleic Acids Res. 40 (2012) W281–W287 (Web Server issue).
[44] M. Hong, Biochemical studies on the structure-function relationship of major drug transporters in the ATP-binding cassette family and solute carrier family, Adv. Drug Deliv. Rev. 116 (2017) 3–20.
[45] X.-Q. Yu, et al., Multidrug resistance associated proteins as determining factors of pharmacokinetics and pharmacodynamics of drugs, Curr. Drug Metabol. 8 (8) (2007) 787–802.
[46] W. Mo, J.-T. Zhang, Human ABCG2: structure, function, and its role in multidrug resistance, Int. J. Biochem. Mol. Biol. 3 (1) (2012) 1.