An Improvement on Identification of ATP Binding Residues of Proteins from Primary Sequences

Aditya Adhikary *(2015007),* Mayank Pal *(2015147)*

I. Introduction

We have carried out a brief implementation of the paper "**Identification of ATP binding residues of a protein from its primary sequence**"[1] and attempted to improve on the existing methods by using different machine learning techniques, an extended dataset and optimized parameters on the models.

A. Motivation

ATP is an important ligand that plays a critical role as a coenzyme in the functionality of many proteins. Hence, it is essential to develop novel methods for identifying ATP interacting residues in ATP binding proteins in order to understand the mechanism of protein-ligands interaction. This can be done via simple statistical methods or supervised machine learning techniques.

B. Methods previously explored

The given paper discussed a Support Vector Machine (SVM)-based method to perform classification. It compared the amino acid composition of ATP interacting and non-interacting regions of proteins and concluded that certain residues are preferred for interaction with ATP. It trained and tested the model (carried out cross validation) on 168 non-redundant ABP chains. The SVM based model using primary sequence of proteins obtained a maximum MCC of 0.33 with validation accuracy of 66.25 %.

II. Our methods

We at first tried a simple implementation of the paper using the linked *dataset* on Python, using the *scikit-learn* library and the svm SVC module, and carried it out on different window sizes. We obtained maximum cross-validation accuracies of around **64%** on a balanced dataset, with window size 17. We afterwards trained a sequential neural network model using the *keras* library, but did not obtain satisfactory results. Afterwards, we fell back on expanding the size of the training dataset and optimizing the parameters of the existing svm model.

III. Implementation Details

A. Preprocessing of Data

The given dataset already had 168 non-redundant ATP binding protein chains, preprocessed with the help of CD-HIT and Ligand-Protein Contact (LPC). We loaded the amino acid sequences from the same, and appended *(window_size-1)*/2 X's or unknown residues at the end of each amino acid sequence to account for the terminal residues. We then broke each sequence into *window_size* length patterns and generated *n* patterns, *n* being the total number of residues over all the sequences.

We then converted the above patterns into (*window_size**21) length vectors for the input to the machine learning model. First, we labeled each pattern as ATP interacting(1) or non-interacting(0) depending on whether the central residue was interacting or not. Then, we went through each pattern and replaced each residue by a 21-length binary vector with the position of a 1 representing the presence of the residue. We then flattened the [*window_size X* 21] matrix to a single (*window_size**21) length vector.

We also carried out balancing of data since there were too many negative samples. We chose to undersample the number of negative (ATP non-interacting) patterns by randomly choosing and keeping only *x* negative patterns, where *x* was the number of positive patterns.

TABLE I. Performance Evaluation for Improved SVM Model

Window Size	Accuracy	Precision	Recall	Specificity	FScore	MCC	FPR
24	0.6958	0.7162	0.6486	0.7430	0.6887	0.3934	0.2569

B. Models and Cross Validation

We then performed cross validation using *StratifiedKFold*, in which each fold has an approximately equal representation of each class, with $K = 5$, and the preliminary svm model. We used *cross_val_predict* to carry out predictions during validation, which keeps the prediction obtained on a sample when it was in the test set during its validation iteration.

C. Performance Evaluation

We evaluated several performance measures on the predictions, including *Classification Accuracy*, *Precision*, *Recall*, *Specificity*, *F1-Score*, *MCC*, and *False Positive Rate*, using either library functions or from the values obtained from the *Confusion Matrix*. This was essential to test the different characteristics of the predictions and to improve upon the existing results.

IV. Improvements

A. Deep Neural Networks

Since we were getting results similar to the paper, we attempted to improve them by using a deep neural network model. We used the *keras* library to train a *Sequential* model, with the initial layers using $tanh$ activation functions, the final layers using *softmax*, and the loss function as *cross-entropy*. However, the neural network failed to give a comparable accuracy (around 50%) or MCC as the training data was not adequately large for training it.

B. Expanding the Training Dataset

We derived more ATP binding protiens using a lower redundancy threshold for CD-HIT. We now had 223 primary sequences, stored in a "generated.txt" file. By running the same svm model as before, the accuracy did not improve further beyond 66%.

C. Optimizing SVM parameters

We tried a number of different parameters for SVM via an exhaustive grid search and found that a value of 'C' (learning rate) of 100 and a polynomial kernel with *gamma* as 0.1 gave best results.

V. Results

The improved results obtained are shown in Table I.

VI. Conclusion

We managed to obtain an improved cross validation accuracy of 69.58 % and MCC score 0.393 using a larger training set and an optimized SVM model.

Acknowledgment

We would like to thank Prof. G.P.S Raghava, who guided us during the initial phase of understanding.

References

[1] J.S Chauhan, N.K Mishra & G.P.S Raghava, *Identification of ATP binding residues of a protein from its primary sequence*, BMC Bioinformatics, 2009