2012 International Conference on Bioinformatics and Computational Biology (ICBCB 2012)
IPCSIT vol.34 (2012) © (2012) IACSIT Press, Singapore

Prediction of FAD Binding Residues with Combined Features from Primary Sequence

Chun Fang[1+], Tamotsu Noguchi[2] , Hayato Yamana[1]
[1]Department of Computer Science and Engineering of Waseda University, Tokyo, Japan
[2]Computational Biology Research Center(CBRC), Tokyo Japan

Abstract. In order to analyze the impact of different characteristics of protein sequence on small nucleotide binding site prediction, this paper proposed four sequence-based methods for identifying FAD binding residues of flavin binding proteins (FBP) by means of support vector machine (SVM). We used the different combined features obtained from primary sequence as input for the prediction: evolutionary conservation, predicated solvent accessibility, physicochemical characteristics, residue neighbor list and so on. Our result shows that, the three methods which combined the evolutionary information performed much better than the predictor which did not adopt the evolutionary information. The predictor which combined the feature of residue neighbor list, evolutionary information and physicochemical properties performs the best, achieved accuracy of 87.7326% which is 4.87% higher than the previous methods.

Keywords: FAD, binding prediction, sequence-based, SVM, combined features

1. Introduction

The function of proteins depends on their interaction with other molecules (proteins, DNA, ligands and so on). Flavin adenine dinucleotide (FAD), also known as active-type vitamin B2, is one such important small molecule which plays critical role as a cofactor in the functionality of flavin binding proteins (FBP). It plays critical role as a coenzyme in high-energy electron transfer, signaling transduction, metabolism and so on.

FAD proteins have attracted the interest of many researchers. Orly Dym et al [1] analyzed the protein sequence and structural features of FAD-containing proteins. Mariana Babor et al [2] analyzed the conserved positions and water bridging interactions among FAD-protein complexes. Thus identification of FAD interacting residues in a protein is important for understanding their function and mechanism.

Now, there are many sequence-based tools for predicting residues interaction sites between polynucleotide (DNA/RNA) and proteins [4-5] or protein-protein interactions [6-8], but few for small molecules and proteins. Previous prediction methods mainly focused on the sequence conservation information or amino acid composition preference [9], without considering the structure information or residues flexibility. Since there are some proteins that have only a few homologous protein sequences, the methods only using the evolutionary information for prediction may be ineffective. Thus, more effective feature extraction method from sequence is indispensable.

In order to illustrate the impact of different features on the performance of the predictor. we proposed four methods, each of them used the support vector machines (SVM) and the combined different features obtained from primary sequence for the prediction of FAD binding residues: such as conservation, predicated structural information, physicochemical characteristics, residue neighbor list and so on. We found that there were significant differences between the predictors which used different combinations of features as input.

2. Materials and Methods

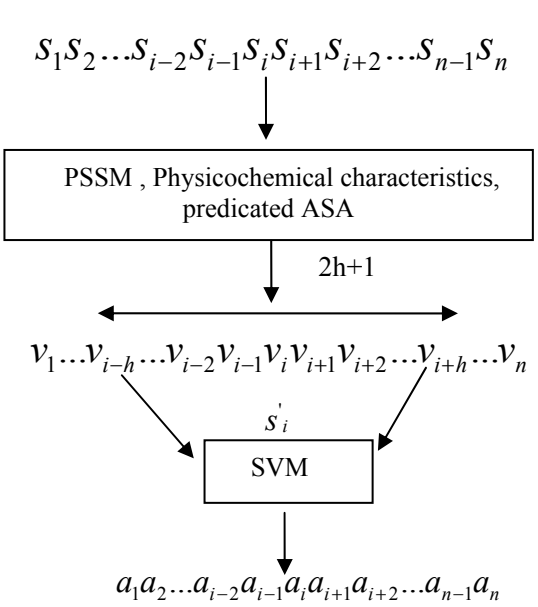
[+] Corresponding author.
E-mail address: fangchun@yama.info.waseda.ac.jp

2.1 Data sets

To facilitate the comparison, we used the same data sets with Mishra's experimental [10]. All the data are downloaded from PDB database and have below 40% residue identity. Mishra's experimental data contains 198 protein chains which interact with FAD. But two of them are too short to find homologous sequences that results in preventing the generate of the PSSM scoring matrix by PSI-blast [11]. On the other hand, 9 protein chains of them are too long for the relevant predictor to predict solvent accessibility of single residues. Thus, these 2 and 9 protein chains are removed. The remaining 187 protein chains are used for our experiments. The 187 protein chains contain 4972 FIRs (FAD interacting residue) and 73425 non-FIRs (non-FAD interacting residue). All the FIRs and 5200 non-FIRS that are selected randomly are used for developing our model.

2.2 Prediction model

Our prediction model is shown in Fig.1. Detailed description of each part is explained later.



Output: sequence of binding site and non-binding site

Fig.1 prediction model

2.3 Definition of FAD-binding pattern

We use slide-windows to generate the positive data sets and negative data sets. The length of windows is 2h+1. If the middle site residue v_i is a binding site, the pattern $v_{i-h}\ldots v_{i-2}v_{i-1}v_iv_{i+1}v_{i+2}\ldots v_{i+h}$ is considered to be a positive data, otherwise, it is a negative data. In dealing with residues at beginning and end of the sequence, h 0's are added to both sides of the sequence as follows:

$$\underbrace{00\ldots00}_{h}s_1s_2\ldots s_{i-2}s_{i-1}s_is_{i+1}s_{i+2}\ldots s_{n-1}s_n\underbrace{00\ldots00}_{h}$$

2.4 Evolutionary information (PSSM)

Evolutionary information is obtained from position specific scoring matrixes (PSSM), which are generated by PSI-BLAST search against NCBI non-redundant (nr) database [12] by three times iteration with an e-value of 0.001. The evolutionary information for each amino acid is encapsulated in a vector of 20 dimensions where the size of PSSM matrix of a protein with N residue is 20 * N. 20 dimensions are standard amino acid. N is the length of a protein, each value of the PSSM matrix is normalized into the range of [-1, 1] by he following function:

$$f(v)=\begin{cases}1, f(v)/10>1.0\\ f(v)/10, -1.0<f(v)/10<1.0\\ -1, f(v)/10<-1.0\end{cases}$$

2.5 Predicated accessible surface area (ASA)

We use the predicted solvent accessibility of residues as a simplified form to represent the tertiary structural features, that is "buried" and "exposed". If a residue is predicted to be buried, we code it as 0. Otherwise it is coded by 1. We get Solvent Accessibility by the server of RVP-NET which is designed by Shandar Ahmad[13]. It provides Real Values of solvent accessibility in a protein using neural network algorithm. Since it is worth noting that the largest sequence size for predication is limited to 700, part of the sequence larger than 700 will be ignored during prediction.

2.6 Physicochemical features

Ten kinds of physicochemical features of residue are considered in our study, they are hydrophobic, polar, small, proline, tiny, aliphatic, aromatic, positive, negative, charged. Each amino acid was represented by a vector of 10 length (e.g. Ala by 1 0 1 0 1 0 0 0 0 0).

2.7 Support vector machines(SVM)

Many studies have shown that SVM is powerful in dealing with high dimensional data, especially for binary classification. Identification of FIRs can be addressed as a two-classification problem: determining whether a given residue is interacting with FAD or not. In our study, there are 73425 non-FIRs and only 4972 FIRs, in order to maintain a balance between positive dataset and negative dataset, 5200 non-FIRs which are selected from the non-FIRs dataset randomly and all the 4972 FIRs are used for training model. The prediction model is trained by the libSVM software package which is written by in Chih-Jen [14]. Here, the Radial Basis Function (RBF kernel) was adopted to construct the SVM classifiers in our model, the grid search method was used to search for the best parameters c and g, 5-fold cross-validation was used to evaluate the performance of the developed modules.

2.8 Four combination models

In order to analyze the impact of different characteristics on prediction, four predictors using different combinations of features as input are designed respectively.

Predictor-1(SVM$_{\text{-PSSM}}$): this model used residue neighbor list and evolutionary information (PSSM) as input. Each residue was encoded as a feature vector with 17×20 dimensions, i.e. (the surface residue to be predicted +16 neighbors) $\times$(20 amino acids).

Predictor-2 (SVM$_{\text{-PSSM_ASA}}$): this model used residue neighbor list , evolutionary information and predicted Solvent Accessibility as input. Each residue was encoded as a feature vector with 17×21 dimensions, i.e. (the surface residue to be predicted +16 neighbors) $\times$(20 amino acids+ predicted solvent accessibility).

Predictor-3 SVM$_{\text{-PSSM_Physicalchemical}}$: this model used residue neighbor list , evolutionary information and ten kinds of physicochemical characteristics of residue as input. Each residue was encoded as a feature vector with 17×30 dimensions, i.e. (the surface residue to be predicted +16 neighbors) $\times$(20 amino acids+ 10 Physicochemical features).

Predictor-4 (SVM$_{\text{_ ASA_Physicalchemical}}$):this model used residue neighbor list , Solvent Accessibility and physicochemical information as input. Each residue was encoded as a feature vector with 17×11 dimensions, i.e. (the surface residue to be predicted +16 neighbors) $\times$(10 Physicochemical features + predicted solvent accessibility).

3. Evaluation Criteria

The performance of the SVM was measured by the sensitivity, specifity, accuracy and MCC. Where TP, TN, FP and FN represents true positive, true negative, false positive and false negative respectively.

$$senstivity = \frac{TP}{TP + FN} \times 100\% \quad (1)$$

$$specifity = \frac{TN}{TN + FP} \times 100\% \quad (2)$$

$$precision = \frac{TP}{TP + FP} \times 100\% \quad (3)$$

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}\times 100\% \quad (4)$$

$$MCC=\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}\times 100\% \quad (5)$$

4. Results and Discussion

In order to analyze the relationship among the amount of training data, running time and the prediction accuracy, the SVM_PSSM method was analyzed as an example, different sizes of datasets from 500 to 6000 are used for training the prediction model. The result is shown in Figure 2 and Figure 3.

Figure 2 shows the relationship between the size of training-data and training time. Figure 3 shows the relationship between the size of traning-data and the prediction accuracy.

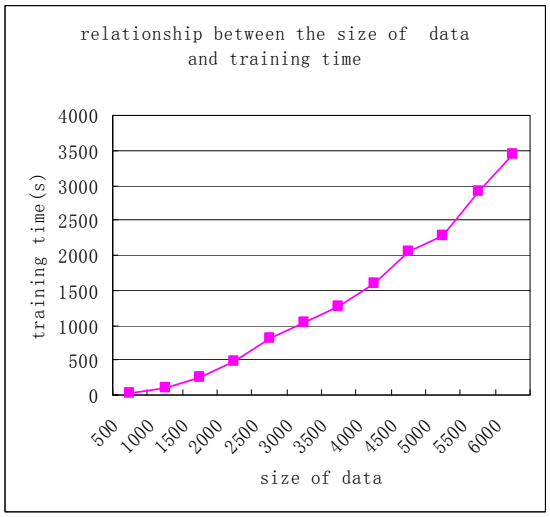


Fig.2 relationship between the size of data and training time

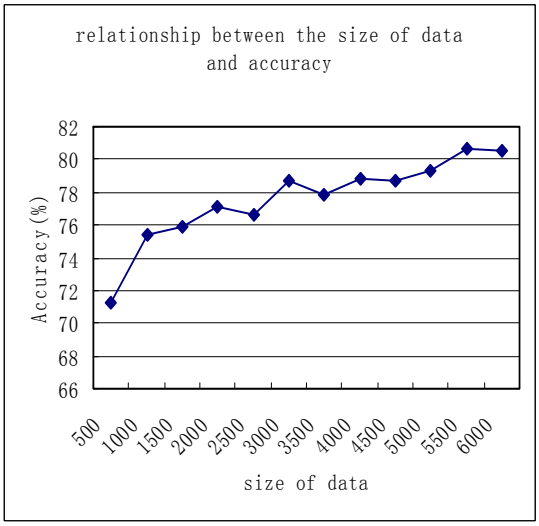


Fig.3 relationship between the size of data and accuracy

From the Figure3, we can find that, with the increasing of the amount of training data, the prediction accuracy is improving. Therefore, in our experiment, all the positive data and roughly the same amount of negative data were used, although some method took a long time to search for the best parameters c and g.

4.1 Comparison of performance in the four methods.

We use our four models to predict for the 187 protein chains which are downloaded from PDB. The highest, the lowest and the average accuracy are shown in table1.

Table 1**:** The highest, the lowest and the average accuracy

Predictor	the highest	the lowest	average accuracy
SVM$_{\text{-PSSM}}$	98.72%	72.76%	87.61%
SVM$_{\text{- PSSM\_ASA}}$	92.16%	74.28%	87.13
SVM$_{\text{-PSSM\_Physicalchemical}}$	97.62%	75.60%	87.73%
SVM$_{\text{-ASA-Physicalchemical}}$	92.75%	56.31%	75.31%

The average of sensitivity, specificity, precision and mcc of the four methods are shown in Figure 4.

The PDB_id of the best and the worst top 20 prediction by the four methods are shown in Table 2 and Table 3.

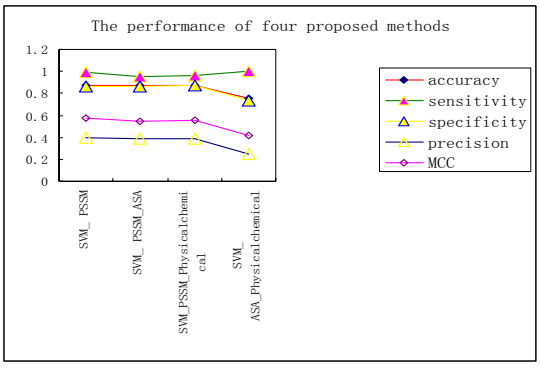


Fig.4 performance of the four proposed methods

Table 2: The PDB_id of the best top 20 prediction

SVM_ PSSM	SVM_{-ASA-} Physicalchemical	SVM_ PSSM_ASA	SVM_ PSSM_Physicalchemical
2IW5:B	2CZ8:H	2IW5:B	2GAG:C
2CZ8:H	1OQC:D	2HJ3:A	2HJ3:A
2HJ3:A	2HJ3:A	1JR8:B	2IW5:B
1RP4:A	2CIF:A	2GAG:C	1OQC:D
1OQC:D	1EBD:C	1OQC:D	1JR8:B
2BRY:B	2V3B:B	1KF6:O	2BRY:B
2GAG:C	2E0I:A	1RP4:A	1ZY8:M
1O96:A	1JR8:B	1E7P:L	1KF6:P
1JR8:B	2CFA:B	1O96:A	1E7P:L
1E7P:L	2G37:B	2CZ8:H	1RP4:A
1KF6:O	1KQ4:D	2BRY:B	1O96:A
2BUN:A	2YVJ:B	1ZY8:M	1YVV:B
1EFV:B	1EFV:B	1XOP:A	1ZXI:C
1ZY8:M	2E5V:B	1I19:A	1BF3:A
1F0X:B	1VRQ:D	1T3Q:F	1SEZ:B
3C4A:A	2CVJ:A	1RM6:E	1I19:A
1I19:A	2PD7:B	1SEZ:B	1RM6:E
1KF6:P	2E1M:B	1EFV:B	1EFV:B
1YVV:B	2IW5:B	3C4A:A	3C4A:A
1K87:A	2CUL:A	1ZXI:C	1T3Q:F

Table 3: The PDB_id of the worst 20 prediction

SVM_ PSSM	SVM_{-ASA-} Physicalchemical	SVM_ PSSM_ASA	SVM_ PSSM_Physicalchemical
2GQT:A	1YQ3:D	1EP1:A	1T3Q:D
1RZ0:A	1KF6:O	1XHC:A	2GQT:A
1EP1:A	1NEK:D	2PD7:B	1XHC:A
2YVJ:B	1W1O:A	1NOH:A	1NOH:A
1YOA:A	1KF6:P	1YQ3:B	1GPE:A
1HSK:A	1JU2:A	1P3Y:1	1EP3:B
1EP3:B	1GPE:A	1T3Q:D	1EP1:A
2ED4:B	1T9G:R	2A1T:A	1T9G:R
1T3Q:D	2E1M:C	2GQT:A	2GJ3:B
1GPE:A	1P3Y:1	1VRQ:D	2GQF:A
1W4X:A	1NEK:C	2BS2:E	2A1T:A
2GQF:A	2VFR:A	1EP3:B	1VRQ:D
1DNP:B	1YQ3:C	1T9G:R	1JU2:A
1IQR:A	1COY:A	1GPE:A	2PD7:B
1OWL:A	1EP1:A	1KF6:N	2V6O:A
1JU2:A	1E7P:L	2GJ3:B	1HSK:A
1MBB:A	1O96:B	2V6O:A	1VQW:B
2PD7:B	1T3Q:F	1Y9D:A	1Y9D:A
1NOH:A	2IX6:F	1HSK:A	1W4X:A
1R2J:A	1XDI:B	1NEK:C	2CZ8:H

According to the above results, the following conclusions can be drawn:

(1) It can be seen from table 1 that the $SVM_{-PSSM-Physicalchemical}$ method which combines the evolutionary information and ten kinds of physicochemical characteristics of residues performed best, that got an average accuracy of 87.73%, 0.12% better than the second best method of SVM_{-PSSM} whose average accuracy is 87.61%. Although the increase was not very significant, but it can illustrate that considering the physicochemical properties of residues can improve the predictive performance. Moreover, there is a need to explain, we didn't analyze the unique impact of certain physicalchemical property to the result. This is a place worthy of further consideration.

(2) Many existing studies have suggested that, considering the solvent accessibility of residues can improve the performance of interaction prediction, however, in our experiment, from the Table 1 and Figure 4. We can see that, the performance of SVM_{-PSSM_ASA} method which considers the evolutionary information and solvent accessibility is 0.5% poor than the SVM_{-PSSM}.

(3) In the four methods, the prediction performance of SVM_{-PSSM},$SVM_{-PSSM-ASA}$,$SVM_{\_PSSM\_Physicalchemical}$ which combins evolutionary information are significantly better $SVM_{-ASA\_Physicalchemical}$ which does not use evolutionary information, this proves once again that evolution of information plays a key role in prediction of protein molecules. In order to maintain certain function, the functional sites of proteins must have a high degree of evolutionary conservation.

(4) From Table 2 and Table 3, we can find a noteworthy phenomenon that the performances of different predictors on different protein are very different. For example, protein 1VRQ: D, 2CZ8: H and 2PD7: B has got predictive accuracy all above 90% in the $ASA_{-Physicalchemical}$ method where but less than 70% in the $SVM_{-PSSM-Physicalchemical}$ method. So it is necessary to consider designing different predictor according to the types of proteins.

4.2 Comparison with other methods

In order make an objective comparison with the results of others, we use the same dataset with PS Raghave's methods[14]. PS Raghave's used two methods for the prediction of FAD Binding Residues, that is, using Binary pattern of Amino acids sequence and evolution of information respectively as input for SVM, they used SVM_light package and each value of the PSSM matrix was normalized into range [0, 1]. Result of the comparison is shown in Table 4.

Table 4: comparison with PS Raghave's method

Predictor	Accuracy	MCC
PS Raghave's Binary pattern	69.65%	0.39
PS Raghave's PSSM	82.86%	0.66
$SVM_{\_PSSM}$	87.61%	0.57
$SVM_{\_PSSM\_ASA}$	87.13%	0.55
$SVM_{\_PSSM\_Physicalchemical}$	87.73%	0.56
$SVM_{\_ASA\_Physicalchemicall}$	75.31%	0.42

From Table 4, we can see that prediction accuracy of our three methods ($SVM_{\_PSSM}$, SVM_{-PSSM_ASA}, $SVM_{\_PSSM\_Physicalchemical}$) are all nearly 5% higher than PS Raghave's method although with 0.1 lower on the MCC.

5. Conclusions

In this paper, we developed four new methods to predict the FAD interacting residues from protein sequence based on support vector machine. The preliminary experiments indicate that using combined features could lead to better prediction performance. We proposed several different sequence encoding schemes and compared their resulting prediction performance. The purpose of this study was to find which kind of information input can lead to the best prediction result. The prediction accuracies were averaged by using 5-fold cross-validation. It was found that using residue neighbor list and evolutionary information could significantly improve the prediction performance, the prediction accuracy increased from 75.31% with single sequence to 87.61% and MCC from 0.42 to 0.57. Moreover, if coupled with the ten kinds of physicochemical information of residue, the prediction accuracy was further improved to 87.73% with MCC of 0.56. This study reinforced that SVM is a powerful prediction tool for extracting the relationship between FAD and primary protein sequence, all the result will provide helpful and complementary information in understanding FAD complex proteins and their function.

6. Acknowledgments

We would like to thank the protein function team in Computational Biological Research Center (CBRC) for helpful advice and discussion. We also thank an anonymous reviewer for his/her helpful comments, which improved the manuscript.

7. References

[1] Orly Dym, David Eisenberg, "Sequence-structure analysis of FAD-containing proteins". Protein Science 2001. Vol.10, Issue 9, p. 1712–1728.

[2] Babor M, Sobolev V. "Conserved positions for ribose recognition: importance of water bridging interactions among ATP, ADP and FAD-protein complexes". Journal of molecular biology, 2002

[3] Saito M, Go M and Shirai T: “An empirical approach for detecting nucleotide-binding sites on protein”. Protein Engineering Design Selection 2006, 19:67–75.

[4] Kumar M, Gromiha MM, and Raghava GPS: “Prediction of RNA binding sites in the protein using SVM and PSSM profile”. Proteins 2008, 71:189-194.

[5] Ofran Y, Mysore V, Rost B: “Prediction of DNA-binding residues from sequence”. Bioinformatics 2007, 23:i347-353.

[6] Tuncbag, N , Kar, G, Keskin, O, Nussinov, R. “A survey of available tools and web servers for analysis of protein-protein interactions and interfaces Brief ”, Bioinformatics (2009) 10(3): 217-232

[7] Xuewen C, Jong cheol J, “Sequence-based prediction of protein interaction sites with an integrative method”, Bioinformatics (2009) 25(5): 585-591

[8] Espadaler J, Romero-Isart O. “Prediction of protein–protein interactions using distant conservation of sequence patterns and structure relationships”. Bioinformatics .2005, pages 3360–3368

[9] Mishra, N.K. and Raghava, G. P. S..”Prediction of FAD interacting residues in a protein from its primary sequence using evolutionary information”. BMC Bioinformatics. 2010.

[10] http://www.imtech.res.in/raghava/fadpred/FADPred_data

[11] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”. Nucleic Acids Res 1997, 25:3389-3402

[12] NR ftp://ftp.ncbi.nih.gov/blast/db/fasta/nr.gz

[13] Shandar Ahmad M. , Akinori S, “ Real value prediction of solvent accessibility from amino acid sequence”, Proteins: Structure, Function, and Bioinformatics, p 629–635, 2003

[14] http://www.csie.ntu.edu.tw/~cjlin/libsvm/