Published with Open Access at Journal BiNET
Vol. 26, Issue 02: 2177-2184
Journal of Bioscience and Agriculture Research
Journal Home: www.journalbinet.com/jbar-journal.html

Modified naive Bayes classifier for classification of protein-protein interaction sites

Mohammad Ahsan Uddin[1] and Md. Shakil Ahmed[2]

[1]Department of Statistics, University of Dhaka, Dhaka-1000, Bangladesh.
[2]Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh.

✉ For any information: munna_stat@yahoo.com (Uddin, MA)
Article received: 22.10.2020; Revised: 17.11.2020; First published online: 10 December 2020.

ABSTRACT

The prediction of protein-protein interaction sites (PPIs) is a vital importance in biology for understanding the physical and functional interactions between molecules in living systems. There are several classification approaches for the prediction of PPI sites; the naïve Bayes classifier is one of the most popular candidates. But the ordinary naïve Bayes classifier is sensitive to unusual protein sequence profiling feature dataset and sometimes it gives ambiguous prediction results. To overcome this problem we have been modified the naïve Bayes classifier by radial basis function (RBF) kernel for the prediction of PPI sites. We investigate the performance of our proposed method compared with the popular classifiers like linear discriminant analysis (LDA), naïve Bayes classifier (NBC), support vector machine (SVM), AdaBoost and k-nearest neighbor (KNN) by the protein sequence profiling data analysis. The mNBC method showed sensitivity (86%), specificity (81%), accuracy (83%) and MCC (65%) for prediction of PPI sites.

Key Words*: Protein Sequences Profiling, PPI sites, Relative Solvent Accessibility (rSA), RBF Kernel and Naïve Bayes Classifier.*

Cite Article: Uddin, M. A. and Ahmed, M. S. (2020). Modified naive Bayes classifier for classification of protein-protein interaction sites. Journal of Bioscience and Agriculture Research, 26(02), 2177-2184.
Crossref: https://doi.org/10.18801/jbar.260220.266

I. Introduction

Every biological organism is controlled by proteins in a cell for several biological functions. Many proteins perform their functions independently; the overwhelming majority of proteins interact with others for correct biological activity. The interface residues are necessary for understanding the protein function and therefore, the mechanism of interaction is usually desired drug design. To identify protein-protein interaction sites can be treated as a classification problem, that is every amino acid residue is assigned of two classes such as interacting or non-interacting residues. The analysis of protein-protein interaction site prediction is such a lot essential for comprehending organic process is the function of one protein by interacting to others (Esmaielbeiki et al., 2014). It is important to identify individual protein-protein interactions and selectively working them through targeted residues (Sowa et al., 2001; Madabushi et al., 2004; Du X. et al., 2014). Proteins perform and control

Published with open access at journalbinet.com
EISSN: 2312-7945, ©2020 The Authors, Research paper

many processes within the cell through interactions with other proteins. The PPI sites help understand biological functions and develop new drugs (Fariselli et al., 2002; Esmaielbeiki et al., 2014; Li Hui et al., 2014). The machine learning algorithms are designed to find out by example during a multi-parameter space. Several kinds of literature have recently inaugurated to use them to envisage interacting surface residues, by neural networks and support vector machines (SVM) (Zhou et al., 2001; Fariselli et al., 2002). Investigated the alignment of residues and their structural neighbors used the neural networks for classification surface residues into interacting and non-interacting ones. This displayed the prominence of bearing in mind structural neighbors while construction of classifier (Yan et al., 2004a; Asadabadi et al., 2013) have trained in SVM to predict whether or not a surface residue is an interface residue. They have accomplished extraordinary sensitivity (82.3 and 78.5%) and specificity (81.0 and 77.6%) on two diverse datasets. Similar methods may be able to be applied to the proteins of unknown structures prediction is questionable issues. In that case, the knowledge on residue composition remains available, but the knowledge on neighboring residues and surface accessibility isn't (Ofran et al., 2003; Yan et al., 2004b) have independently shown that the interface residues tend to make clusters in sequence. Based on this observation have developed a two-stage classifier. It combines both SVM and Bayesian classifiers to predict which surface residues form interface and it achieves the accuracy of 72% and a coefficient of correlation of 0.30. However, they did not try to classify all residues in a protein but only those on its surface (which were determined by using the structure). In contrast attempted to classify residues from protein sequences into interacting and non-interacting ones (Ofran et al., 2003; Walia et al., 2014). Their method uses neural networks based on the sequence clustering of interface residues and interface composition. They report an accuracy of 70%, with 20% sensitivity. Identification of interacting residues from the sequence in a study by (Gallet et al., 2000; Bhaskara et al., 2014) where the authors have recommended that the identification of interacting residues is possible based on their hydrophobic moments (Yan et al., 2004b). However, tested this method on their dataset and obtained a negative correlation coefficient. The study of A non-redundant data set of heterodimers entailing 69 protein chains and accomplished a sensitivity of 66.3%, a specificity of 49.7%, an accuracy of 0.654 and MCC 0.297 (Wang et al., 2006; White et al., 2008; Su Z. et al., 2011).

II. Materials and Methods

Dataset

The Protein Data Bank (PDB) web server is available at this web address (Berman et al., 2000). The PDB (http://www.rcsb.org/pdb/home/home.do) archive is the single worldwide repository of information for protein or nucleic acids 3D structures of large biological molecules. From the PDB webserver, we collected 115 protein sequences in FASTA format for creating protein sequence profile (Murakami et al., 2010).

Protein sequence profiling

The protein sequence profile predicted accessibility (pA) of a residue was obtained using SABLE which is available by this web address http://sable.cchmc.org/. The pA represents the rSA of each residue and is expressed on a scale of0 (fully buried) to 100 (fully exposed).

Relative Solvent Accessibility (rSA)

Solvent accessibility is calculated from the 'solvent accessible' surface we saw in the last section. The area described by the center of the probe as it travels across the protein surface is assigned to the adjacent atoms giving atomic solvent accessibility. From the atomic solvent accessibility, a residue solvent accessibility can be calculated, by summing the component atomic accessibilities. Standard accessibility values can be calculated for each amino acid type in an extended conformation as part of an Ala-XAla tripeptide. By expressing residue accessibility as a percentage of the standard accessibility, one can obtain relative accessibility.

Modified Naïve Bayes Classifier

A naive Bayes classifier is a probabilistic classifier depends on the Bayes theorem and assumes the independence of features for given a class. The standard rule is to select the most probable hypothesis; this is often referred to as the utmost a posteriori (MAP) decision rule. The radial basis function (RBF) kernel was used for modification of mNBC method and it gives the better performance than the

Published with open access at journalbinet.com
EISSN: 2312-7945, © 2020 The Authors, Research paper

popular classifiers by 5-fold cross validation datasets. Prediction of PIs the RBF kernel and naïve Bayes classifier was used to calculate the posterior probability using protein sequence profile features.

Step-1: From the extracted protein sequence profiling feature data matrix as input values.

Step-2: Select the optimum window size based on 5-fold cross validation, here the optimum window size is 37 (Figure 01).

Step-3: Calculate the probability using RBF kernel function

Step-4: Classification of protein-protein interaction (PPI) sites using modified Naïve Bayes classifier. It is binary classification +ve (interaction sites or interface residue) and –ve (non-interacting sites or non-interface residue). All the calculations implemented by R v.3.2.0 programing language for all numerical data analysis. The probability calculation using RBF kernel of rbf(x,...) function under RSNNS v0.4-12 (https://www.rdocumentation.org/packages/RSNNS/versions/0.4-12/topics/rbf) package in R programing language. Also the Naïve Bayes classification was calculated using naive Bayes (formula, data, laplace = 0, ..., subset, na.action = na.pass) function under e1071 v1.7-3 packages (https://www.rdocumentation.org/packages/e1071/versions/1.7-3/topics/naiveBayes) using R Programming language.

Naïve Bayes Classifier

Let the vector $X = (x_{1k}, x_{2k}, \dots\dots, x_{pk})$ representing some p features, it assigns to this instance probabilities $f(C_k|x_{1k}, x_{2k}, \dots\dots, x_{pk}, \theta_k)$ for each of K populations or classes. The conditional probability based on the Bayes theorem can be written as:

$$f(C_k|\boldsymbol{X}, \boldsymbol{\theta}_k) = \frac{f(C_k)\, f(\boldsymbol{X}|C_k, \boldsymbol{\theta}_k)}{f(\boldsymbol{X}|\boldsymbol{\theta})}$$

Where, $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots\dots, \theta_k\}$and $f(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{k=1}^{K} f(C_k)\, f(\boldsymbol{X}|C_k, \boldsymbol{\theta}_k)$

Now the "naive" conditional independence assumptions come into play: assume that each feature $x_j$is conditionally independent of every other feature x_i for j≠i, given the category C. The conditional distribution over the class variable C under the independence assumptions can be decomposed as:

$$f\left(C_k\middle|x_{1k}, x_{2k}, \dots\dots, x_{pk}, \boldsymbol{\theta}_k\right) = \frac{1}{Z} f(C_k) \prod_{j=1}^{p} f\left(x_{jk}\middle|C_k, \boldsymbol{\theta}_k\right)$$

Where, the evidence $Z=\sum_{k=1}^{K} f(C_k|x_{1k}, x_{2k}, \dots\dots, x_{pk}, \boldsymbol{\theta}_k)$ is a scaling factor dependent only on $x_{1k}, x_{2k}, \dots\dots, x_{pk}$ that is, a constant if the values of the feature variables are known.

Radial basis function (RBF) kernel

The radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms in machine learning. Especially it is frequently used in classification in the Bioinformatics dataset. Let us consider two samples **x** and **x'**, represented as feature vectors in some input space then the RBF kernel can be defined as,

$$K\,(x, x^T) = \exp\left(-\frac{\|x - x^T\|^2}{2\sigma^2}\right)$$

Where,

$\|x - x^T\|$ is the squared Euclidean distance between the two feature vectors and σ is a free parameter.

Published with open access at journalbinet.com
EISSN: 2312-7945, ©2020 The Authors, Research paper

Evaluation measures and validation

The following measures were calculated to assess the performance of Bayes and Gaussian Naïve Bayes classifier, using counts of true positives (TP; residues correctly predicted as interface), false positives (FP; residues incorrectly predicted as interface or Type-I Error), true negatives (TN; residues correctly predicted as non-interface) and false negatives (FN; residues incorrectly predicted as non-interface or Type-II Error).

ACC: Accuracy (ACC) is the proportion of the known residues that are correctly predicted in all prediction and is defined as:

$$ACC = \frac{(TP + TN)}{(TP + FN + TN + FP)}$$

MCC: MCC (Matthews Correlation Coefficient) indicates the degree of the correlation between the actual and predicted classes of the residues. MCC values range between $-1 \leq MCC \leq +1$, +1 means all the predictions are correct, and -1 none are correct. The MCC can be defined as:

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

Sensitivity or TPR: The sensitivity that is True Positive Rate (TPR) measures the proportion of the known interface residues that are correctly predicted as interface residues. The sensitivity is defined as:

$$Sensitivity\ or\ TPR = \frac{TP}{(TP + FN)}$$

Specificity or TNR: The specificity or True Negative Rate (TNR) measures the proportion of the known non-interface residues that are correctly predicted as non-interface residues. The specificity can be defined as:

$$Specificity\ or\ TNR = \frac{TN}{(TN + FP)}$$

Positive Predictive Value (PPV) or Precision: The PPV measures the proportion of the residues predicted as interface that are known interface residues. The PPV can be defined as:

$$PPV = \frac{TP}{(TP + FP)}$$

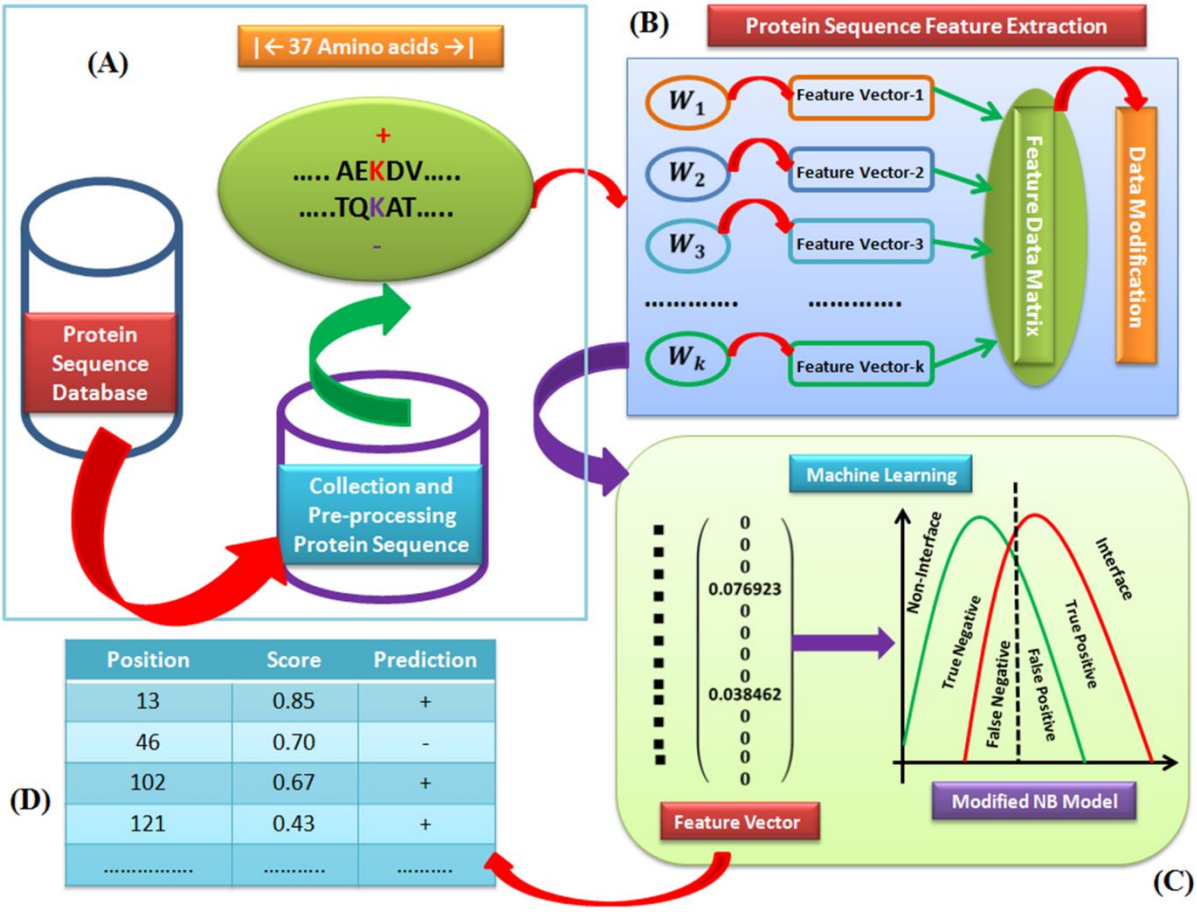
Negative Predictive Value (NPV): The NPV measures the proportion of the residues predicted as non-interface that are known non-interface residues. The NPV can be defined as:

$$NPV = \frac{TN}{(TN + FN)}$$

III. Results and Discussion

The dataset was analyzed by four steps such that, firstly, collect the protein sequences from the protein databases, preprocessing the protein sequences and select the optimum window size. In this paper the optimum window size 37 amino acids. Secondly, the feature extraction from the protein sequences based on the selected optimum window size. The statistical non-parametric method was used for feature selection and fit the machine learning method. Thirdly, fit the machine learning method and finally calculate the prediction score of PPI sites (Figure 01).

Published with open access at journalbinet.com
EISSN: 2312-7945, © 2020 The Authors, Research paper



Position	Score	Prediction
13	0.85	+
46	0.70	-
102	0.67	+
121	0.43	+
.....	.....	.....

Figure 01. Schematic outline of our study (A) Protein sequence collection from protein database and preprocessing of the collected sequences, (B) Protein sequence feature extraction using statistical methods, (C) Fit the machine learning method modified naïve Bayes model, and (D) Protein-protein interaction sites prediction score.

Classification of protein-protein interaction (PPI) site prediction is important for any biological functional process. Prediction of PPI sites LDA (Linear Discriminant Analysis) was shown that the ~77% sensitivity, ~73% specificity, ~75% accuracy (ACC) and ~59% of MCC respectively for full dataset. In case of NBC (naïve Bayes classifier) shown that sensitivity, specificity, ACC, MCC were ~78%, ~74%, ~76%, and ~59% respectively for the prediction of PPI sites. The SVM (support vector machine) tool was used to predict PPI sites the prediction output 82%, 77%, 79%, and 62% sensitivity, specificity, ACC, and MCC respectively. On the other hand, the AdaBoost, KNN method performance shown that sensitivity 81% and 80%, specificity 76% and 76%, ACC 76% and 78%, and MCC 61% and ~61% respectively for classification of PPI sites based on the full dataset.

The proposed modified naïve Bayes classifier (mNBC) shown that the better performance for the prediction of PPI sites the sensitivity, specificity, ACC, and MCC 86%, 81%, 83%, and 65% respectively. Also compare the popular machine learning tools were used for prediction of PPI sites for 5-fold datasets, for all datasets shown the better performance of mNBC method (Table 01). The ROC curves were plotted for 5-fold datasets and calculate the optimum AUC for different machine learning methods. Figure 02 shown that for all the datasets the mNBC method better prediction performance with comparison SVM, NBC, LDA, AdaBoost, and KNN machine learning methods. All the calculated AUC was plotted using box plot for different datasets. The box plot showed that the mNBC method better prediction performance than the other machine learning methods (Figure 03).

Published with open access at journalbinet.com
EISSN: 2312-7945, ©2020 The Authors, Research paper

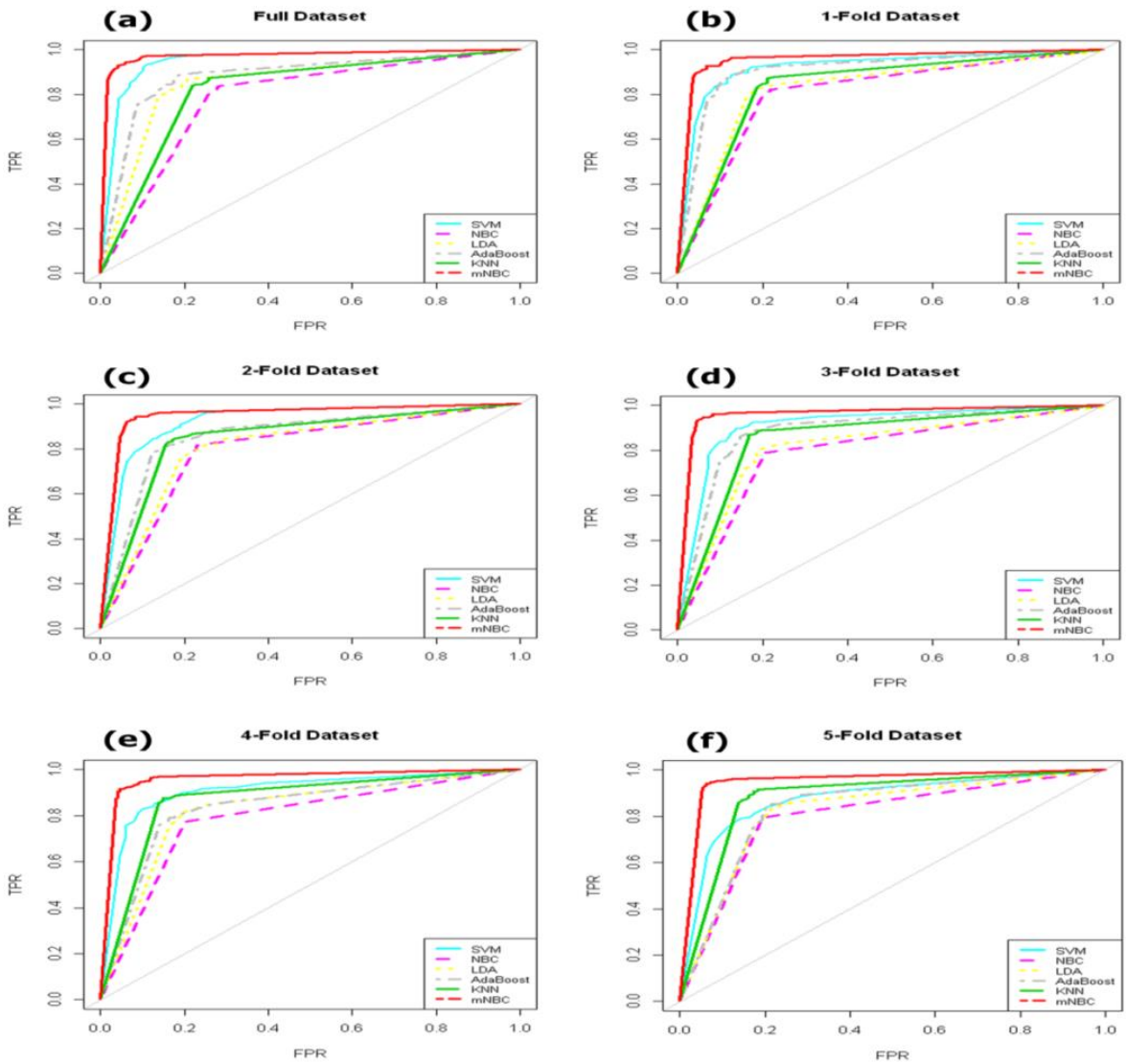


Figure 02. ROC curve for the classification of PPI sites using (a) Full dataset, (b) 1-Fold dataset, (c) 2-Fold dataset, (d) 3-Fold dataset, (e) 4-Fold dataset, and (f) 5-Fold cross validation datasets for comparison of popular classifiers and mNBC method.

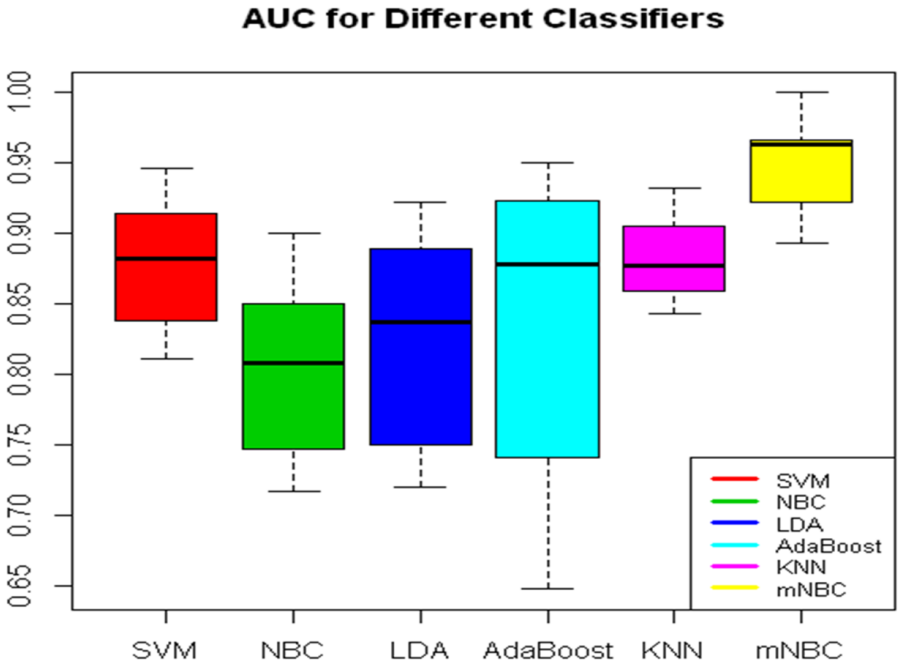


Figure 03. AUC box plot for different classifiers including our mNBC (proposed) methods based on the 5-fold cross validation datasets

Published with open access at journalbinet.com
EISSN: 2312-7945, © 2020 The Authors, Research paper

Table 01. Selection of best classifiers and comparison with other different datasets for classification of PPI sites using 5-fold cross validation.

Datasets	Prediction methods	LDA	NBC	SVM	AdaBoost	KNN	mNBC (Proposed)
Full Dataset	*Sn (%)*	77.382	78.155	81.681	80.864	80.047	85.98
	Sp (%)	73.17	73.901	77.235	76.462	75.690	81.30
	ACC (%)	75.132	75.883	79.306	78.512	77.719	83.48
	MCC (%)	58.698	59.284	61.959	61.339	60.719	65.22
1-fold Dataset	*Sn (%)*	78.057	78.837	82.393	81.57	80.745	86.73
	Sp (%)	74.079	74.819	78.194	77.413	76.630	82.31
	ACC (%)	76.068	76.828	80.294	79.491	78.688	84.52
	MCC (%)	60.39	60.993	63.745	63.108	62.470	67.10
2-fold Dataset	*Sn (%)*	79.587	80.382	84.008	83.168	82.328	88.43
	Sp (%)	72.207	72.929	76.218	75.456	74.694	80.23
	ACC (%)	77.886	78.664	82.213	81.391	80.568	86.54
	MCC (%)	62.001	62.621	65.445	64.791	64.136	68.89
3-fold Dataset	*Sn (%)*	81.189	82.000	85.699	84.843	83.985	90.21
	Sp (%)	76.707	77.474	80.968	80.159	79.349	85.23
	ACC (%)	78.903	79.692	83.286	82.454	81.620	87.67
	MCC (%)	63.981	64.620	67.535	66.86	66.184	71.09
4-fold Dataset	*Sn (%)*	74.385	75.128	78.517	77.732	76.947	82.65
	Sp (%)	72.018	72.738	76.019	75.258	74.498	80.02
	ACC (%)	72.153	72.874	76.161	75.399	74.638	80.17
	MCC (%)	61.101	61.712	64.495	63.850	63.205	67.89
5-fold Dataset	*Sn (%)*	85.707	86.564	90.468	89.564	88.659	95.23
	Sp (%)	83.286	84.118	87.913	87.034	86.154	92.54
	ACC (%)	82.134	82.955	86.697	85.83	84.963	91.26
	MCC (%)	72.585	73.310	76.617	75.851	75.085	80.65

V. Conclusion

The relative solvent accessibility (rSA) features from the profiling of protein sequences used in this paper as training dataset for scrutinizing the performance of our proposed method (mNBC). The conditional probability was calculated using RBF kernel. The mNBC method showed better performance for the prediction of PPI sites than the other machine learning methods. The 5-fold cross validation datasets were used for the exploration of the performance of mNBC. The mNBC method was shown high sensitivity, specificity, accuracy, and MCC 86%, 81%, 83%, and 65% respectively'.

VI. References

[1]. Asadabadi, E. B. and Abdolmaleki, P. (2013). Predictions of Protein-Protein Interfaces within Membrane Protein Complexes. Avicenna Journal of Medical Biotechnology, 5(3), 148-57.

[2]. Berman, Helen, M., John, Westbrook, Zukang, Feng, Gary Gilliland, Talapady, N. Bhat, Helge, Weissig, Ilya, N. Shindyalov, and Philip E. B. (2000). The protein data bank. Nucleic acids research, 28(1), 235-242. https://doi.org/10.1093/nar/28.1.235

[3]. Bhaskara, R. M., Padhi, A. and Srinivasan, N. (2014). Accurate prediction of interfacial residues in two-domain proteins using evolutionary information: implications for three-dimensional modeling. Proteins, 82(7), 1219-34. https://doi.org/10.1002/prot.24486

[4]. Du, X., Cheng, J., Zheng, T., Duan, Z. and Qian, F. (2014). A novel feature extraction scheme with ensemble coding for protein-protein interaction prediction. International Journal of Molecular Sciences, 15(7), 12731-49. https://doi.org/10.3390/ijms150712731

[5]. Esmaielbeiki, R., and Nebel, J. C. (2014). Scoring docking conformations using predicted protein interfaces. BMC Bioinformatics, 15, 171. https://doi.org/10.1186/1471-2105-15-171

[6]. Fariselli, P., Pazos, F., Valencia, A. and Casadio, R. (2002). Prediction of protein--protein interaction sites in heterocomplexes with neural networks. European Journal of Biochemistry, 269(5), 1356-61. https://doi.org/10.1046/j.1432-1033.2002.02767.x

Published with open access at journalbinet.com
EISSN: 2312-7945, ©2020 The Authors, Research paper

[7]. Gallet, X., Charloteaux, B., Thomas, A. and Brasseur, R. (2000). A fast method to predict protein interaction sites from sequences. Journal of Molecular Biology, 302(4), 917-26. https://doi.org/10.1006/jmbi.2000.4092

[8]. Li, Hui, Dechang, Pi, and Chishe, W. (2014). The prediction of protein-protein interaction sites based on RBF classifier improved by SMOTE. Mathematical Problems in Engineering (2014). https://doi.org/10.1155/2014/528767

[9]. Madabushi, S., Gross, A. K., Philippi, A., Meng, E. C., Wensel, T. G. and Lichtarge, O. (2004). Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. Journal of Biological Chemistry, 279(9), 8126-32. https://doi.org/10.1074/jbc.M312671200

[10]. Murakami, Y. and Mizuguchi, K. (2010). Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. Bioinformatics, 26(15), 1841-8. https://doi.org/10.1093/bioinformatics/btq302

[11]. Ofran, Y. and Rost, B. (2003). Predicted protein-protein interaction sites from local sequence information. FEBS Letters, 544(1-3), 236-9. https://doi.org/10.1016/S0014-5793(03)00456-3

[12]. Sowa, M. E., He, W., Slep, K. C., Kercher, M. A., Lichtarge, O. and Wensel, T. G. (2001). Prediction and confirmation of a site critical for effector regulation of RGS domain activity. Nature Structural Biology, 8(3), 234-7. https://doi.org/10.1038/84974

[13]. Su, Z., Ning, B., Fang, H., Hong, H., Perkins, R., Tong, W. and Shi, L. (2011). Next-generation sequencing and its applications in molecular diagnostics. Expert Review of Molecular Diagnostics, 11(3), 333-43. https://doi.org/10.1586/erm.11.3

[14]. Walia, R. R., Xue, L. C., Wilkins, K., El-Manzalawy, Y., Dobbs, D. and Honavar, V. (2014). RNA Bind R Plus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. PLoS One., 9(5), e97725. https://doi.org/10.1371/journal.pone.0097725

[15]. Wang, B., Chen, P., Huang, D. S., Li, J. J., Lok, T. M. and Lyu, M. R. (2006). Predicting protein interaction sites from residue spatial sequence profile and evolution rate. FEBS Letters., 580(2), 380-4. https://doi.org/10.1016/j.febslet.2005.11.081

[16]. White, A. W., Westwell, A. D. and Brahemi, G. (2008). Protein-protein interactions as targets for small-molecule therapeutics in cancer. Expert Reviews in Molecular Medicine, 10, e8. https://doi.org/10.1017/S1462399408000641

[17]. Yan, C., Dobbs, D. and Honavar, V. (2004b). A two-stage classifier for identification of protein-protein interface residues. Bioinformatics, 20 Suppl 1, i371–i378. https://doi.org/10.1093/bioinformatics/bth920

[18]. Yan, C., Honavar, V. and Dobbs, D. (2004a). Identification of interface residues in protease-inhibitor and antigen antibody complexes: a support vector machine approach. Neural Computing and Applications, 13(2), 123-129. https://doi.org/10.1007/s00521-004-0414-3

[19]. Zhou, Huan-Xiang, and Yibing, S. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins: Structure, Function, and Bioinformatics, 44(3), 336-343. https://doi.org/10.1002/prot.1099

HOW TO CITE THIS ARTICLE?

Crossref: https://doi.org/10.18801/jbar.260220.266

MLA
Uddin and Ahmed. "Modified naive Bayes classifier for classification of protein-protein interaction sites". Journal of Bioscience and Agriculture Research, 26(02), (2020): 2177-2184.

APA
Uddin, M. A. and Ahmed, M. S. (2020). Modified naive Bayes classifier for classification of protein-protein interaction sites. *Journal of Bioscience and Agriculture Research*, 26(02), 2177-2184.

Chicago
Uddin, M. A. and Ahmed, M. S. "Modified naive Bayes classifier for classification of protein-protein interaction sites". Journal of Bioscience and Agriculture Research, 26(02), (2020): 2177-2184.

Harvard
Uddin, M. A. and Ahmed, M. S. 2020. Modified naive Bayes classifier for classification of protein-protein interaction sites. Journal of Bioscience and Agriculture Research, 26(02), pp. 2177-2184.

Vancouver
Uddin MA and Ahmed MS. Modified naive Bayes classifier for classification of protein-protein interaction sites. Journal of Bioscience and Agriculture Research, 2020 December 26(02): 2177-2184.

Published with open access at journalbinet.com
EISSN: 2312-7945, © 2020 The Authors, Research paper