Journal of Theoretical Biology 409 (2016) 27–37

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Classify vertebrate hemoglobin proteins by incorporating the evolutionary information into the general PseAAC with the hybrid approach

Muthu Krishnan S.

CSIR – Institute of Microbial Technology (IMTECH), Sector-39A, Chandigarh, India

HIGHLIGHTS

- Determine Vertebrate Hb protein, their animal classifications using general PseAAC's.
- Prediction performance was further investigated by ROC and prediction score graphs.
- ACC, SN, SP was examined to find the accurate predictions on the threshold level.
- Examined the newly developed models on an independent dataset (blind dataset).
- An analysis was tried to find relationship between sub-classes using developed models.

ARTICLE INFO

Article history:
Received 17 June 2016
Received in revised form
11 August 2016
Accepted 16 August 2016
Available online 27 August 2016

Keywords:
Hemoglobin
Vertebrate hemoglobin
Support vector machines
Confusion matrix
ROC analysis
Position specific scoring matrix
Hybrid approaches
PSSM
SVM
Fish hemoglobin
Amphibians hemoglobin
Reptiles hemoglobin
Aves hemoglobin
Mammalian hemoglobin

ABSTRACT

Hemoglobin is an oxygen-binding protein widely present in all kingdoms of life from prokaryotic to eukaryotic, but well established in the vertebrate system. An attempt was made to determine the Vertebrate hemoglobin (VerHb) protein on their animal classifications, based on general pseudo amino acid composition (PseAAC)'s evolutionary profiles and hybrid approach. The support vector machine (SVM) has been applied to develop all models, the prediction results further compared according to their animal classification. The performance of the approaches estimated using five-fold cross-validation techniques. The prediction performance was further investigated by receiver operating characteristic (ROC) and prediction score graphs. The prediction accuracy (ACC), sensitivity (SN) and specificity (SP) were examined to find the accurate predictions on the threshold level. Based on the approach, a web-tool has been developed for identifying the VerHb proteins.

1. Introduction

Hemoglobin is one of the oxygen binding proteins that plays an important role in oxygen storage and transport. Hemoglobin (Hb) are widespread in animal and plant kingdoms. Different species of vertebrate animals can carry different types of Hb in their blood (Perutz, 1983; Brittain 2005; Hardison 1998; Hardison 1996). It is made up of four globin chains (two alpha and two beta chains) that connect together. Each chain contains an important molecule at a central position called heme. Hemoglobin has many different functions, such as storage and transport of oxygen, enzymes, and signal transduction. It also plays an important role in maintaining the shape of red blood cells (Giardina et al., 1995). In the animal classifications, the VerHb has been classified mainly as fishes, amphibians, reptiles, aves and mammals. The molecular structure of the hemoglobin in the living organism analysis study reflects the relationship of protein and their evolution (Mylvaganam et al., 1996).

E-mail address: muthu@imtech.res.in

http://dx.doi.org/10.1016/j.jtbi.2016.08.027
0022-5193/

In the past nearly 15 years, a huge number of computational approaches have been developed to determine the functional protein predictions. Mainly, these prediction methods can be divided into three categories; protein structure based, protein sequence based and a hybrid approach that the combination of both structural and sequence information (Kumar et al., 2009; Cai and Lin, 2003; Yousef and Charkari, 2015; Panwar and Raghava, 2014). Most of the prediction methods were developed using SVM and it has been applied to many functional protein predictions as glycosylation site prediction (Caragea et al., 2007), subcellular localization (Xie et al., 2005), DNA-binding proteins (Ahmad et al., 2004), gene prediction (Chen et al., 2015), O-GlcNAcylation (Zhao et al., 2015) and Pupylation sites prediction (Hasan et al., 2015).

In the past studies, Oxypred and BacHbpred were developed for identifying the oxygen binding proteins using SVM (Muthukrishnan et al., 2007; Selvaraj et al., 2016), but it does not categorize the hemoglobin according to vertebrate animal classifications. Oxypred method carries only simple amino acid composition (AAC) and dipeptide compositions (DPC) and BacHbpred for identifying bacterial hemoglobin-like proteins developed with various approaches such as AAC, DPC, Hybrid and position-specific scoring matrix (PSSM).

The concept of Chou's pseudo amino acid components (PseAAC) was proposed in 2001, it entered almost all the area of computational proteomics, such as predicting bacterial virulent proteins (Nanni et al., 2012), predicting membrane protein types (Chen and Li, 2013; Huang and Yuan, 2013),discriminating outer membrane proteins (Hayat and Khan, 2012), identifying allergenic proteins (Mohabatkar et al., 2013), predicting metalloproteinase family (Beigi et al., 2011), predicting protein structural class (Sahu and Panda, 2010), identifying GPCRs (G protein-coupled receptors) and their types (Rehman and Khan, 2012; Xie et al., 2013),

However, most of the aforementioned prediction tools are predicting different functional proteins based on general amino acid properties. The PSSM and different features combination of hybrid methods are widely used by many researchers for protein predictions (Xiao et al., 2011; Chou and Cai, 2003; Cai et al., 2005; Qiu et al., 2014, 2016; Shen and Chou, 2009; Liu et al., 2014). But, using the combination of AAC and DPC as hybrid approach based predictions are low (Hamp and Rost, 2015; Kumar et al., 2007). According to my knowledge, no computational method has been developed using general pseudo amino acid composition (PseAAC)'s hybrid and evolutionary information of PSSM profiles for predicting the VerHb based on their animal classification.

The present study was an attempt to incorporate the general PseAAC such as hybrid and PSSM method based prediction for VerHb protein on their animal classification. I also tried the individual profile of AAC and DPC for VerHb prediction, which did not show the significant difference, so that, a systematic attempt was tried to improve the prediction accuracy. I developed a general-PseAAC based hybrid approach by integrating the AAC and DPC that increased the accuracy, which is significantly better than the individual approach (Garg and Raghava, 2008). Besides, this hybrid approach, a SVM model using evolutionary information of PSSM profiles received from Position-Specific Iterated BLAST (PSI-BLAST) (Rashid et al., 2007) was developed. All the models were developed for this study using five-fold cross-validation techniques.

As demonstrated by a series of recent publications (Chen et al., 2016; Jia et al. 2016.; Z. Liu et al., 2015, 2016; Qiu et al., 2016) in compliance with the 5-step rule (Chou, 2011), to establish a really useful sequence-based statistical predictor for a biological system, I should follow the following five guidelines: (a) construct or select a valid benchmark dataset to train and test the predictor; (b) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (c) introduce or develop a powerful algorithm (or engine) to operate the prediction; (d) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (e) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let me describe how to deal with these steps one-by-one.

Basically, the two-step prediction was applied, firstly, all classes of Hb proteins labeled as positive, the non-Hb sequence labeled as negative, almost nearly the same number of sequences used in both datasets. The dataset were equally divided into five sets, four sets used for training and one set used for testing and it has been repeated to all sets. Using the five-fold cross-validation, two SVM models were developed according to hybrid and PSSM approaches (Kumar et al., 2011). Secondly, all five animal classifications (fishes, amphibians, reptiles, aves and mammals) SVM models were developed. One sub-class sequence labeled as positive and all other classes labeled as negative for running the SVM to develop models. It has been repeated to all classes in the development of classification models. The same five-fold cross-validation technique was applied to develop each five SVM models in both approaches. All sequences in the dataset were tested with the newly developed models for recognizing the classes shown in the prediction score graphs. The blind dataset also used to identify the Hb and non-Hb. The prediction results of accuracy (ACC), Sensitivity (SN), Specificity (SP) were compared within the classes at the threshold level for the accurate prediction.

In summary, a number of large-scale protein analysis studies have been performed to determine the Hb protein on their amino acid sequences. The developed computational approaches have achieved success at different levels in the vertebrate animal classification Hb protein prediction. Based on this analysis and approach, online web-server has been developed for understanding the VerHbproteins available at http://bioinfo.imtech.res.in/servers/muthu/ver_hbpred/home.html.

2. Methods

2.1. Benchmark datasets

The final dataset contains 791 of vertebrate hemoglobins (VerHb) which include hemoglobins in fishes, amphibians, reptiles, aves and mammals of 108, 63, 58, 75 and 487 respectively. The selected sequences were retrieved from Uniprot/Swissprot databases, removed all ‘fragments’, ‘isoforms’, ‘potentials’, ‘similarity’, or ‘probables’. Here, I applied 90% cutoff to all sub-class sequences , it means no two sequences are having more than 90% similar in the datasets. The non-Hb protein dataset containing 855 sequences were retrieved with the same cutoff. These proteins are different from VerHb proteins, but the lengths are nearly same. The average sequence identity between different classes was calculated using Percent Identity Matrix of Kalign program. The average identity between positive and negative dataset was 14.87% (similarity ranges 7.34% to 24.50%). The sequence identity of mammalian sequences with other classes was 32.35%, 28.63%, 28.22%, 30.82%, to reptiles, fishes, aves and amphibians. The reptiles sequence identity with fishes, aves and amphibians were 34.33%, 35.97% and 36.63% respectively. The fishes sequences with aves and amphibians, the average sequence identity was 30.22% and 34.24%. Finally, aves class of sequences with amphibians was 31.83%.

2.2. General pseudo amino acid composition

The successful application of Chou's pseudo amino acid composition has been applied to many computational proteomics

(Chou, 2011). It can be generated by a very powerful web-server called Pse-in-one, published recently (Liu et al., 2015b). Here, I proposed two general PseAAC approaches. 1. Hybrid and 2. Evolutionary profile in the form of PSSM for the prediction of VerHb proteins.

2.3. Hybrid models

The hybrid model is the combination of two or more profiles used for to improve the prediction accuracy (Kumar and Raghava, 2013; Verma and Melcher, 2012). In this study, a total of 420 vector length was used to develop hybrid models which were the combination of AAC andDPC. Firstly, calculated the AAC, the fraction of each amino acid in a protein was divided by the total length of amino acids in a protein. The final output results for AAC profile were 20. The DPC were calculated with the pattern length of 400 (20×20), the fraction of each dipeptide in a protein were divided by the number of all possible dipeptides (Mbah, 2014). Secondly, the AAC and DPC profile were merged to make a hybrid profile by the col_add program of GPSR_1.0 package.

2.4. Evolutionary models

PSSM profiles were developed using the gpsr_1.0 package, which is freely available for Linux/Windows (http://www.imtech.res.in/raghava/gpsr/) running against the non-redundant (nr) database downloaded through NCBI (ftp://ftp.ncbi.nih.gov/blast/db/). The position-specific scoring matrix was calculated using the suite (GPSR) programs. Initially, seq2PSSM_imp was used to calculate the PSSM matrix in column format without any normalization, by performing PSI-BLAST searches against the non-redundant protein database using different iterations (e.g. 3) with a cutoff e-value 0.001. For a sequence of length N, a $N \times 20$ position-specific substitution matrix (m) was computed from the PSI-BLAST alignment output where m [i, j], provided information on the evolutionary conservation of residue type (j) at sequence position (i). The values of PSSM matrix vary within a large range, which makes difficult to run SVM. Thus, every PSSM element $X(i)$ at position (i) is normalized using the program PSSM_n2 based on the following formula;

$$X(i) = \big(n(i) - l(i)\big)/\big(m(i) - l(i)\big) \tag{1}$$

where $X(i)$, $n(i)$, $l(i)$ and $m(i)$ are respectively defined as: the normalization value $X(i)$, the residue actual position score $n(i)$, the minimum score $l(i)$, and the maximum score $m(i)$ of the PSSM outputs for a single residue position (Mishra and Raghava, 2010; Mishra et al., 2014). The values are now normalized between 0 and 1, so that the minimum scores receive"0" and the maximum scores is set to "1". Finally, PSSM_comp and col2svm programs were used to generate the SVM_light input format (a 400 point vector representing the substitution rate of each amino acid into any other).

2.5. Support vector machine (SVM)

The SVM is a supervised learning method that can be used for numerous applications in the area of Bioinformatics. Mainly SVM used to solve the problem such as micro array gene expression data and protein secondary structure prediction. In the present study, SVM_light has been used to predict VerHb proteins. This package is freely downloadable at http://www.cs.cornell.edu/People/tj/svm_light/, which allows using a number of parameters and kernels such as linear, polynomial, radial basis function or any user-defined kernel (Joachims, 1999). The SVM has been applied for a number of predictions, including subcellular location (Chou and Shen, 2007), cancer prognosis prediction (Kourou et al., 2014; Cruz and Wishart, 2007), progesterone receptor (J.L. Liu et al., 2015), 14-3-3-binding phosphopeptides (Madeira et al., 2015), DNA-binding proteins prediction (Xu et al., 2015), Cyclin-Dependent Inhibitors prediction (Saha et al., 2015), Breast Cancer Resistant Protein Inhibitors and Prediction of antimicrobial peptides (Belekar et al., 2015; Ng et al., 2015).

2.6. Evolution of performance

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, sub-sampling test, and jackknife test. However, of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset as elaborated in (Chou and Zhang, 1995). Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (see, e.g. Liu et al., 2015a; Ali and Hayat, 2015; Khan et al., 2015; Kumar et al., 2015; Jia et al., 2016). However, to reduce the computational time, I adopted the five-fold cross-validation in this study as done by many investigators with SVM as the prediction engine. In a five-fold cross validation technique, the positive and the negative dataset were randomly divided into five equal sets. In the five sets, four sets used as training and the remaining one set for testing. In the classifications, one class was used as positive and all other classes consider as negative sets. It has been repeated to all classes. Using this technique, ACC, SN, SP and Matthews correlation coefficient (MCC) was calculated for correct predictions. Accuracy were calculated the correct prediction of positive and negative examples. Sensitivity calculated for positive examples, in which correctly predicted as positively. The calculation of specificity was used for negative example, which correctly predicted as was used in the SVM-based machine learning as a measure of the quality of binary classifications (Kaundal and Raghava, 2009; Ramana and Gupta, 2009). The following equation was used to measure the ACC, SN, SP, and MCC

$$Accuracy\ (ACC) = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

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

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

$$MCC\ = \frac{TPXTN - FPXFN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

where TP, TN, FP, and FN as true positive, true negative, false positive and false negative respectively. To most of the biologists, the above mentioned four formulas are not quite easy to understand, specially for MCC. Here, let me adopt the formulation proposed by a series of studies published very recently (Liu et al., 2014; Chen et al., 2013; Qiu et al., 2014; Lin et al., 2014; Guo et al., 2014), based on Chou's symbol and definition (K.-C. Chou, 2001). According to the formulation, the four formulas can be expressed as

$$\begin{cases} \mathrm{Sn}=1-\dfrac{N_-^+}{N_-^+}, & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sn}=1-\dfrac{N_+^-}{N^-}, & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc}=1-\dfrac{N_-^+ + N_+^-}{N^+ + N^-}, & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC}=\dfrac{1-\left(\frac{N_-^+}{N^+}+\frac{N_+^-}{N^-}\right)}{\sqrt{\left(1+\frac{N_+^- - N^+}{N^+}\right)\left(1+\frac{N_-^+ - N_+^-}{N^-}\right)}}, & -1 \le \mathrm{MCC} \le 1 \end{cases} \quad (6)$$

where is $N+$ is the total number of VerHb proteins investigated while $N\pm$ the number of VerHb predicted wrongly as non-Hb; N^- is the total number of non-Hb investigated while $N\mp$ is the total number of non-Hb incorrectly predicted as VerHb (C. Chou, 2001).

Now it is clear from the set of formula (6) that when $N\pm=0$, it means none of the verHb predicted to be a non-Hb, the sensitivity $Sn=1$. When $N\pm=N+$, it means that all VerHb were incorrectly predicted as non-Hb, the sensitivity $Sn=0$. Likewise, $N\mp=0$ meaning that none of the non-Hb was incorrectly predicted as VerHb, in this the specificity $Sp=1$; Whereas $N\mp=N^-$ meaning that all the non-Hb were incorrectly predicted as VerHb , the specificity $Sp=0$; When the $N\pm=N\mp=0$ meaning that none of the VerHb in the positive (+ve) dataset and none of the non-Hb in the negative (−ve) dataset was incorrectly predicted, now the overall accuracy Acc=1 and MCC=1; When $N\pm=N+=N\mp=N^-$ meaning that all VerHb in the positive ($+ve$) dataset and all non-Hb in the negative ($-ve$) dataset were incorrectly predicted, i.e all positive sequences predicted to be negative and all negative sequences predicted as positive. Now, the overall accuracy Acc=0 and MCC=−1; whereas when $N\pm=N+/2=$ and $N\mp=N^-/2$ means, the accuracy Acc=0.5 and MCC=0 no better than random prediction. As per the above mentioned discussion based on the Equations-6 one can easily understand the meaning of sensitivity, specificity, overall accuracy, and MCC.

Howerver, the formulas (2)–(5) and a set of formulas -(6) are valid only for the single -label systems. For the multiple-label systems, needs to have, such as the subcellular localization of multiplex proteins (Chou et al., 2012, 2011a) where the protein may have two or more locations, a completely different set of metrics is needed as defined in (Chou, 2013).

2.7. Web-server

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful methods (Lin and Lapointe, 2013; Chou and Shen, 2009), a web-server for the method presented in this paper has been established at http://bioinfo.imtech.res.in/servers/muthu/ver_hbpred/home.html. In this studies, all the scripts were written in Perl, CGI-Perl and the web page was designed in HTML format. The developed web-tool is a user-friendly interface, which allows the user to submit their query sequence in a plain text without any format. The results are displayed in a tabular format which includes sequence length, input sequence and chosen approach.

3. Results

3.1. Amino acids profile analysis

In the amino acid profile analysis, the average amino acids were calculated for VerHb and non-Hb proteins, found the residues “A” and “L” are higher in VerHb protein. Residues “F”, “H”, “K”, and “V” are higher in VerHb than the non-Hb proteins. However, the residues “I”, “P”, “Q”, “R” and “S” are higher in non-Hb than the verHb proteins. Residues “C”, “M”, and “W” are present less than 2% in both datasets. The sequence length profile also calculated, displayed in the histogram and compared both VerHb and non-Hb, nearly 550 of VerHb sequences are belong to the range between 101 to 200 residues length. The non-Hb sequence length also similar to VerHb. The complete analysis results are shown in Fig. 1a and b.

In the classification of VerHb proteins, the residue profiles don't show the differences, mostly similar range shown in Fig. 1c. I compared further the residues profile at the median level, mostly the differences are between −0.1 and 1.6 (Fig. 1d). In addition, I sorted the residues from maximum percentage present to a minimum level, the residue order more or less similar in all classes. Interestingly, the residues “A”, “L” and “V” are mostly present in a similar position at the maximum residue level in all classes. In the minimum residue level “M”, “W” and “C” residues remain same in all classes. The complete results are displayed in Fig. 1 f-1–5. The sequence length also calculated for VerHb proteins separately, most of the sequences belong to 101–200 length range (Fig. 1e).

3.2. Prediction based on general PseAAC-hybrid approach

An attempt was tried to predict the VerHb proteins by the individual composition approach such as AAC and DPC. The individual AAC prediction scores were 83.83, 90.98, 77.23 and 0.77 as ACC, SN,SP, and MCC. The DPC prediction approach shows 85.22, 92.96, 78.07 and 0.80 as ACC, SN, SP, and MCC.

The Hybrid approach was developed to show the better prediction than the simple amino and dipeptide composition. It is basically a combination of both AAC and DPC profiles, with the effect of 420 vector dimensions, comprising 20 of AAC and 400 of DPC. Through this approach, achieved the maximum accuracy was 84.25% in VerHb with non-Hb combined dataset along with the sensitivity, specificity, and MCC as 94.23%, 75.04% and 0.80 respectively. In the classification of VerHb proteins achieved the maximum accuracy was 85.21%, 86.87%, 87.30%, 84.53% and 84.38% along with MCC 0.86, 0.75, 0.83, 0.74 and 0.74 as fishes, amphibians, reptiles, aves and mammals respectively. All the prediction results of hybrid approach ACC, SN, SP and MCC are shown in Table 1.

3.3. Prediction based on general PseAAC-evolutionary profile

Evolutionary profile in the form of PSSM has been used successfully for predicting the VerHb animal classification of hemoglobin proteins. The maximum accuracy was achieved 89.56% with MCC 0.88 in VerHb and non-Hb datasets. The evolutionary profile generation flow chart has been shown in Fig. 2. In the classification models, achieved the maximum accuracy was 86.23%, 87.38%, 88.30%, 86.98% and 86.74% with MCC 0.85, 0.81, 0.84, 0.73 and 0.79 as fishes, amphibians, reptiles, aves and mammals. The PSSM profile based complete prediction results ACC, SN, SP, and MCC are shown in Table 1.

3.4. Prediction score analysis using confusion matrix methods

Using graphical approaches to study biological problems can provide an intuitive picture or useful insights for helping analyzing complicated relations in these systems (Lin and Lapointe, 2013), as demonstrated by many previous studies on a series of important biological topics, such as enzyme-catalyzed reactions (Chou and Forsen, 1980; Zhou and Deng, 1984), inhibition of HIV-1 reverse transcriptase (Althaus et al., 1993a., 1993b), inhibition kinetics of processive nucleic acid polymerases and nucleases (Chou et al., 1994), drug metabolism systems (Chou, 2010), and using wenxiang

Fig. 1. vertebrate animal classification amino acid distribution chart; (a) aminoacid distribution chart between VerHb and non-Hb proteins; (b) sequence length profile of VerHb and non-Hb; (c) aminoacid distribution in VerHb according to animal classifications. (d) Amino acid distribution difference between VerHb proteins, the difference calculated based on median scores; (e) sequence length profile of VerHb proteins, *X*-axis used for sequence length range and *Y*-axis for a number of sequences; (f 1–5) maximum to the minimum residues profile of VerHb proteins according to their animal classifications.

Table 1
The performance of the SVM models using PSSM and Hybrid profiles on a original datasets.

		ACC	SEN	SEP	MCC	AUC
Hb/non-Hb	PSSM	89.55793	97.73089	82.05409	0.87985	0.982
	Hybrid	84.25152	94.22468	75.03655	0.79589	0.96
Fishes	PSSM	86.22611	89.58333	85.70772	0.8513	0.979
	Hybrid	85.2057	91.96429	84.16971	0.85909	0.963
Amphibians	PSSM	87.38057	80.72917	87.93103	0.80764	0.955
	Hybrid	86.86709	72.39583	88.05651	0.74757	0.948
Reptiles	PSSM	88.29618	84.65909	88.57021	0.84021	0.942
	Hybrid	87.30222	84.09091	87.54252	0.82928	0.923
Aves	PSSM	86.98248	70	88.77641	0.73264	0.876
	Hybrid	84.53323	74.16667	85.62063	0.73867	0.912
Mammals	PSSM	86.74363	96.74479	71.0041	0.78944	0.958
	Hybrid	84.375	95.74742	66.29098	0.74178	0.951

diagram or graph (Chou et al., 2011b) to study protein-protein interactions (Zhou, 2011).

Confusion Matrix (CM) is used to describe the performance of a developed classification model on a testing set data for which the true values are known. In order to test the developed models, a confusion matrix was generated that shows no confusion occurred in VerHb PSSM models. All sequences are recognized as positive and negative as per present in the datasets. In the performance of hybrid models as shown in Fig. 3b, all positive sequences were correctly predicted as positive, but one of the negative sequences were predicted wrongly. The individual class model's performance shown that two of aves class sequences were predicted wrongly. In that, one sequence was identified by mammals model, but the another sequence was not recognized by any other class models in the PSSM based approach. At the same time, one of the reptile sequences were predicted negatively in PSSM approach, but it has not been recognized by any other sub-class models.

In the hybrid approach, two of aves sequences were predicted wrongly by their own model. Interestingly, these two sequences were identified by reptiles and mammalian models respectively. According to our prediction results, no confusion was found in fishes, amphibians, and mammals in both approaches. But the confusion found in aves and reptiles in PSSM models and hybrid reptiles model does not show any confusion. However, the confusions happened due to their close relationship or similarity of one sub-class with others. The complete prediction score based analysis graph was shown in Fig. 3.

3.5. *Prediction performances*

In order to find the accurate prediction on the threshold level,

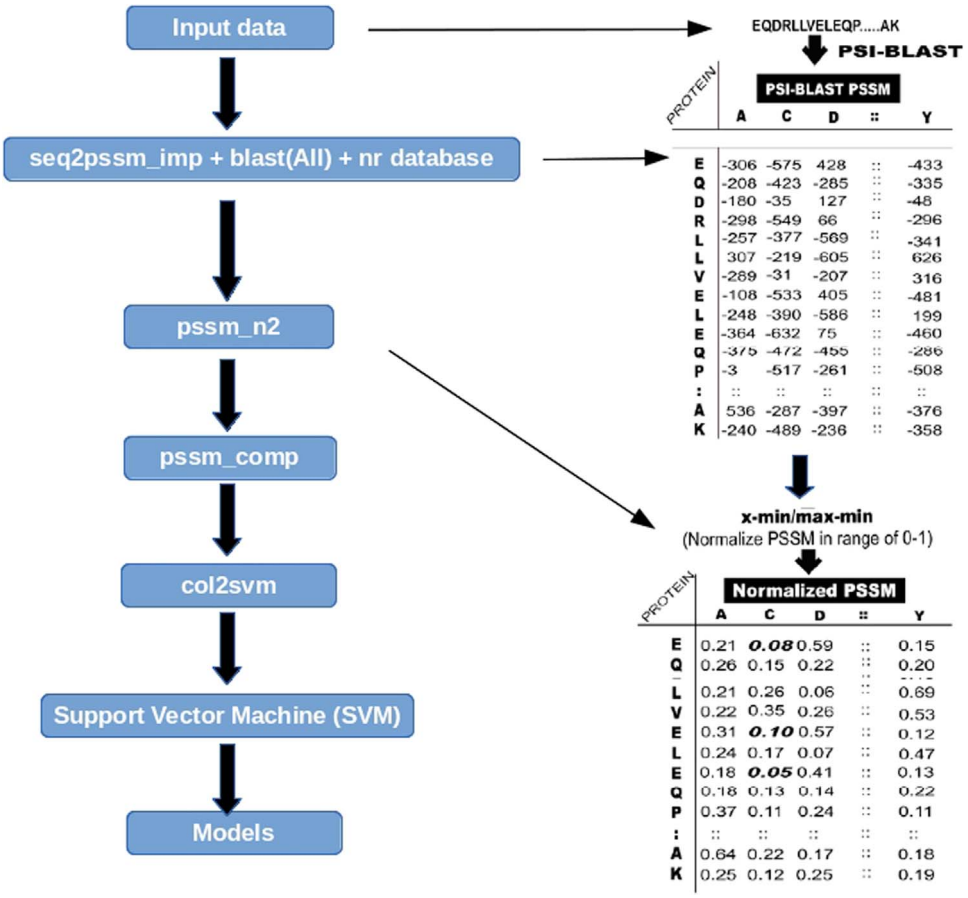

Fig. 2. Evolutionary profile in the form of PSSM flow chart. This profile was generated by various GPSR 1.0 programs, which makes the profiles and normalization of PSSM value for running SVM 400 vector.

where the ACC, SN and SP's prediction scores are same on a particular level of threshold. In this analysis, sub-classes of fishes, amphibians, reptiles prediction scores are met in negative thresholds in both Hybrid and PSSM approaches. The aves subclass was found in the positive thresholds in both approaches. The PSSM of mammalian class appears in the positive threshold, but not in the hybrid approach. Interestingly, the accuracy and specificity prediction scores are same over the -1.5 to $+1.5$ thresholds, but the parallel scores do not shown in the mammalian class of Hb proteins. The details of the results are shown in Fig. 4.

Each class of SVM model's performance was measured by receiver operating characteristic (ROC) plotting, curve created by plotting the sensitivity against the specificity at various threshold settings. Sensitivity is the percentage of true positive and the specificity is the percentage of true negative. The ROC provides clear information about the performance of all developed modules optimized with best parameters (Barman et al., 2014). The area under curve (AUC) was calculated for PSSM (0.982) and Hybrid (0.960) approaches. The PSSM based models to be the highest among all models in the classifications. The complete AUC details are available in Table 1 along with prediction scores and the details of the plots shown in Fig. 5.

Further, I examined the newly developed models on an independent dataset (blind dataset), which consist of 247 Hb and 11218 nonHb sequences obtained from UniProt database. The obtained sequences are not in our original datasets. The performance of individual AAC and DPC approaches on blind-data set, the both approach recognized the blind Hb proteins 408 out of 523 and in NonHb blind data 10906 recognized correctly as negative out of 11218. But, the prediction performance on blind dataset was increased in the Hybrid and PSSM based approaches, it recognized 243 Hb sequences as positively and 4 sequences predicted negatively. The search information shows on non-predicted sequences mostly belong to fragments that are less than 30 residues. At the same time, I tried the blind dataset of non-Hb sequences; a total of 11,218 and 137 predicted false positively, the false positive rate (FPR) was nearly 1.22%. Moreover, the results are mostly recognized by PSSM methods. This demonstration shows that some of them were recognized by hybrid, but completely refused by PSSM models. It means that models developed using evolutionary information in the form of PSSM perform better than the hybrid method. The complete prediction on blind-data results are shown in the supplementary file.

The prediction performance of individual sub-class models output results in both approaches are available in Supplementary Tables 1–6. Highlighted in red shows default threshold cut-off 0.0, maximum score of accuracy, sensitivity, specificity, and MCC over the thresholds.

3.6. Separation and the relationship of the models

An analysis was tried to find the relationship between subclasses using the developed models. The question behind that how one class model recognizes the other classes in the prediction and how it separates?. The measurement has been calculated based on the average prediction scores. According to the Fig. 6, the PSSM-fishes model in order to separate as fishes, amphibians, aves, mammalian and reptiles. The PSSM-Amphibians model separate in order as amphibians, fishes, mammals, reptiles, and aves. The PSSM-reptiles model shows that aves sequence performance next

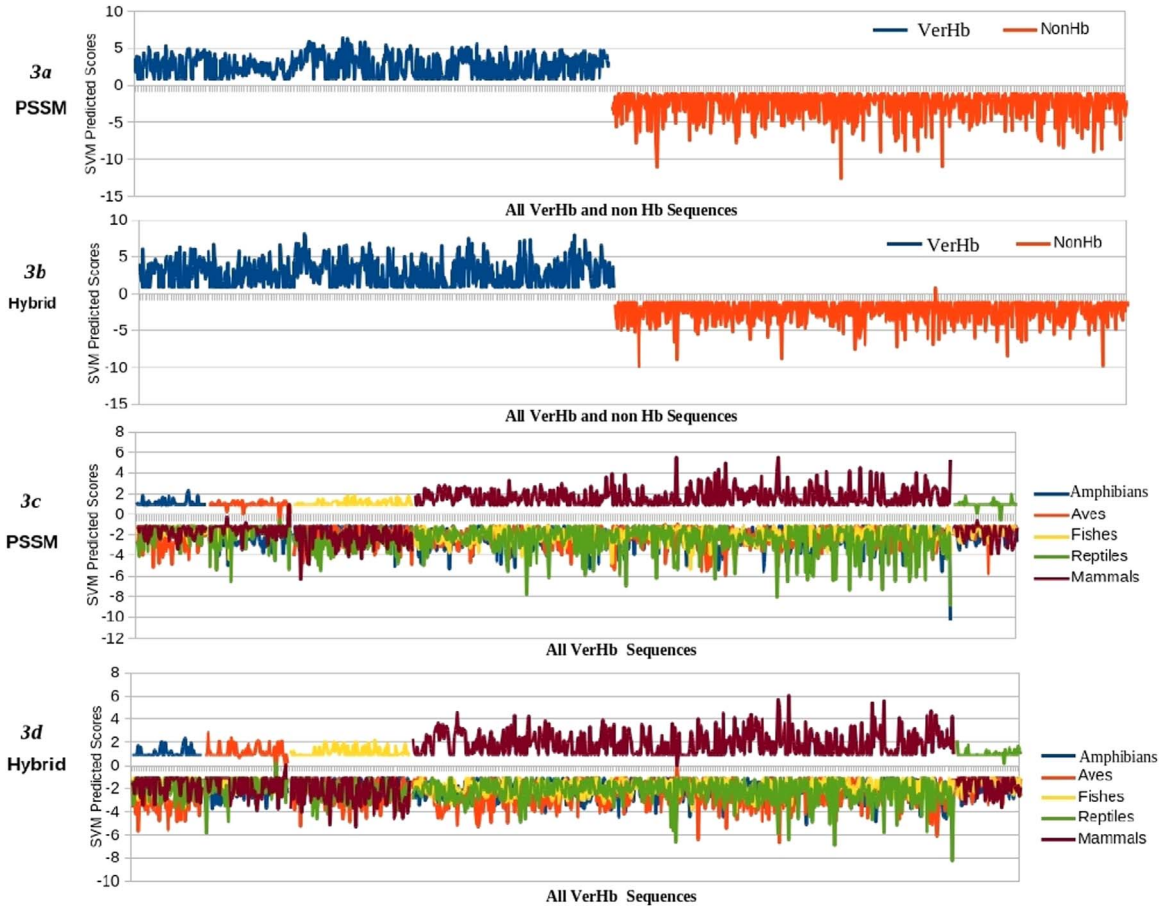

Fig. 3. Developed VerHb model performance on original datasets. (a) Performance based on the evolutionary profile in the form of PSSM developed model separation of positive and negative sequences. (b) A Hybrid model performance, which separating the VerHb and non-Hb sequences in the datasets. (c) VerHb classification models of PSSM profile, recognizing the original datasets of own and other classes of the same family. (d) Hybrid of VerHb classification model's performance, separating the positive and negative class sequences.

to the reptiles, but the fishes, mammals, and amphibians sequences followed by the aves. In PSSM-aves model separates in order as mammals, fishes, reptiles, and amphibians. The PSSM-mammalian model separates in order as fishes, aves, amphibians, and reptiles. The hybrid model's performance is also shown in Fig. 6. Interestingly, the amphibians model performances are exactly the same in both PSSM and hybrid approaches. According to the performance of the individual models are showing the relationship in order from nearby to the distant one.

4. Discussion

Hemoglobin is one of the main proteins in all living organisms and difficult to identify the Hb protein by in-vitro analysis; there is a need of a computational method to identify the Hb protein and their animal classification on the basis of the amino acid sequence of a protein. Most of the prediction methods are available for structural information, but these methods are limited scope because these structures are unknown for many proteins. In the current studies, a highly accurate method has been developed for predicting the VerHb proteins from their protein sequences. The similarity-based annotation data may produce accurate when the experimentally annotated homologous proteins present in the dataset. But, in the case of more accurate, it has to be used in the absence of the significant similarities. So that 90% cutoff has been applied to retrieve the sequences from the UniProt databases. The condition was extended to negative sequences with different length size, which is similar to positive set sequences. In the prediction studies, two types of approaches were applied for prediction, i.e Hybrid and evolutionary information of protein from PSSM profiles. A Hybrid approach was developed in the combination of both AAC and DPC profiles. The evolutionary information of PSSM was obtained from PSI-BLAST search against 'nr' database (Gupta et al., 2014; Li et al., 2013; Tao et al., 2015). The overall accuracy in Hybrid method was 84.25% at 0.80 (-g 05 -c 1700) parameters. But the accuracy was significantly improved with the same parameters in PSSM which show 89.56%. In the vertebrate animal classification based, the reptiles Hb class accuracy was the maximum in both approaches among all the other classes. Interestingly, overall the prediction accuracy was significantly improved in general-PseAAC's PSSM profile than the Hybrid approach in all classes. According to the performance on blind data sets, developed models are performing equally well in both approaches, but PSSM models more correctly identifying positive and negative sequences.

The main objectives of this study were to develop a method for identification of new Hb protein which is similar to VerHb proteins. In order to develop these methods, SVM-based models were developed using Hybrid and PSSM profiles by the fixed parameter according to the dataset size. The analysis results suggest that both methods are important for predicting the Hb proteins and PSSM based approach was better than the Hybrid approach. Based on these studies, a web-server has been developed for the scientific community to provide direct access, which allows the user to predict Hb on their vertebrate animal classification proteins.

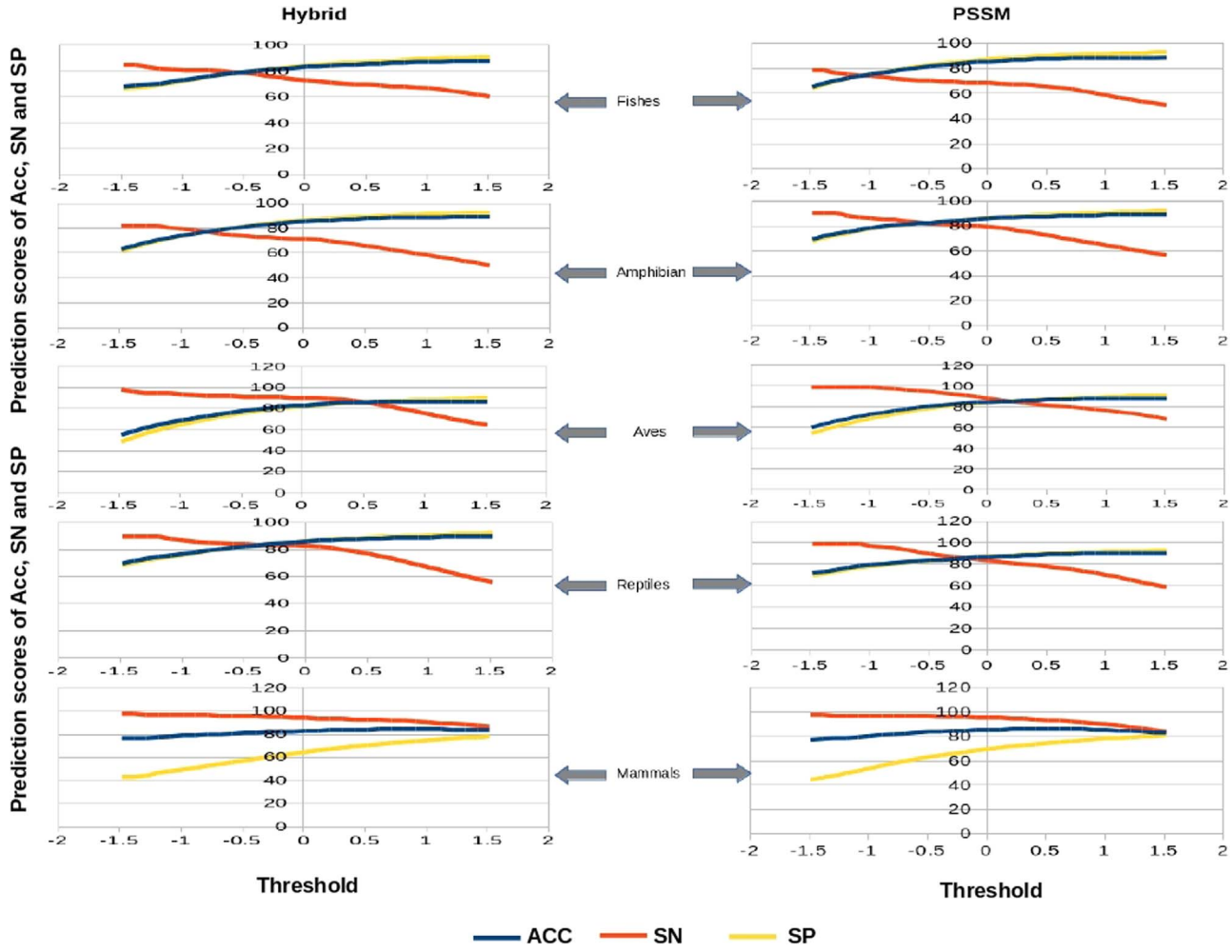

Fig. 4. The performance of accuracy, sensitivity, and specificity based on the threshold value used in VerHb classifications in the approach of PSSM and Hybrid.

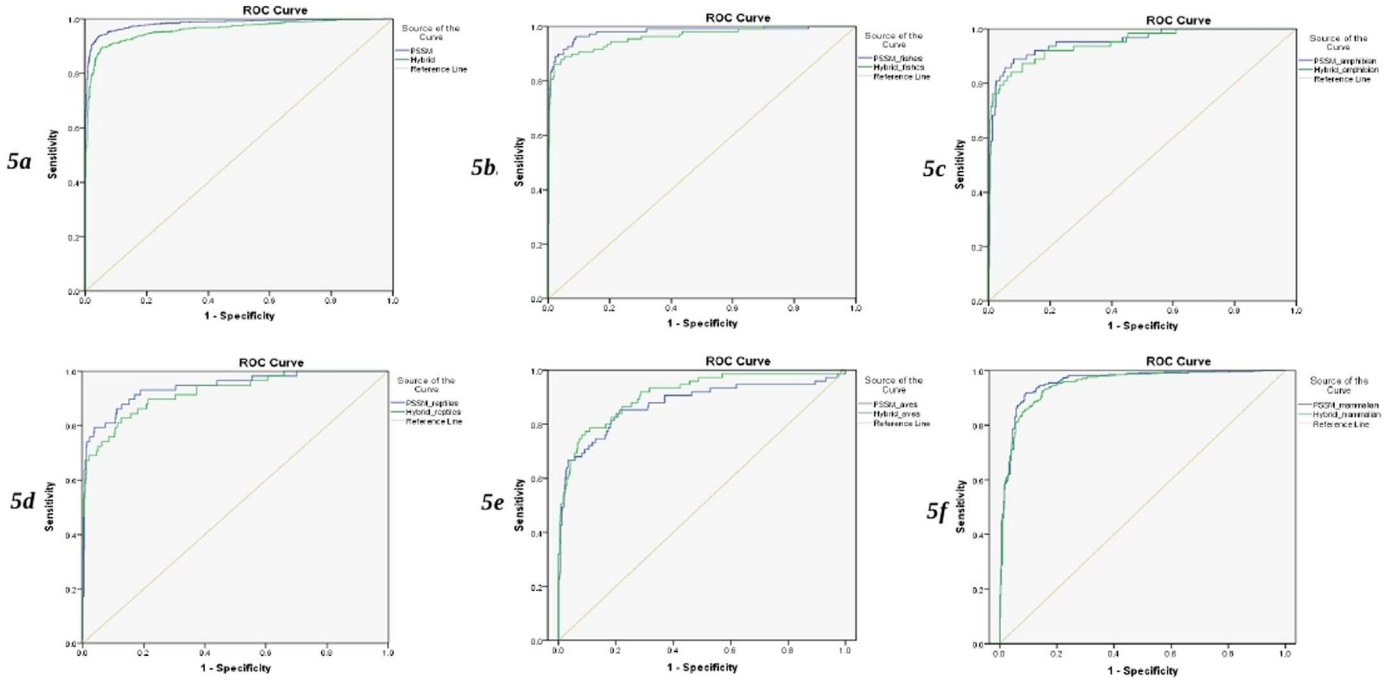

Fig. 5. ROC curve plot. The performance of developed VerHb models by the receiver operating characteristic (ROC) plots in Hybrid and PSSM approaches. The area under curve (AUC) was measured for all developed models. It is mainly to show the relationship between sensitivity and 1-specificity for each threshold of the real value out-puts. (a) VerHb performance, (b) fishes, (c) amphibians; (d) reptiles, (e) aves, (f) mammals.

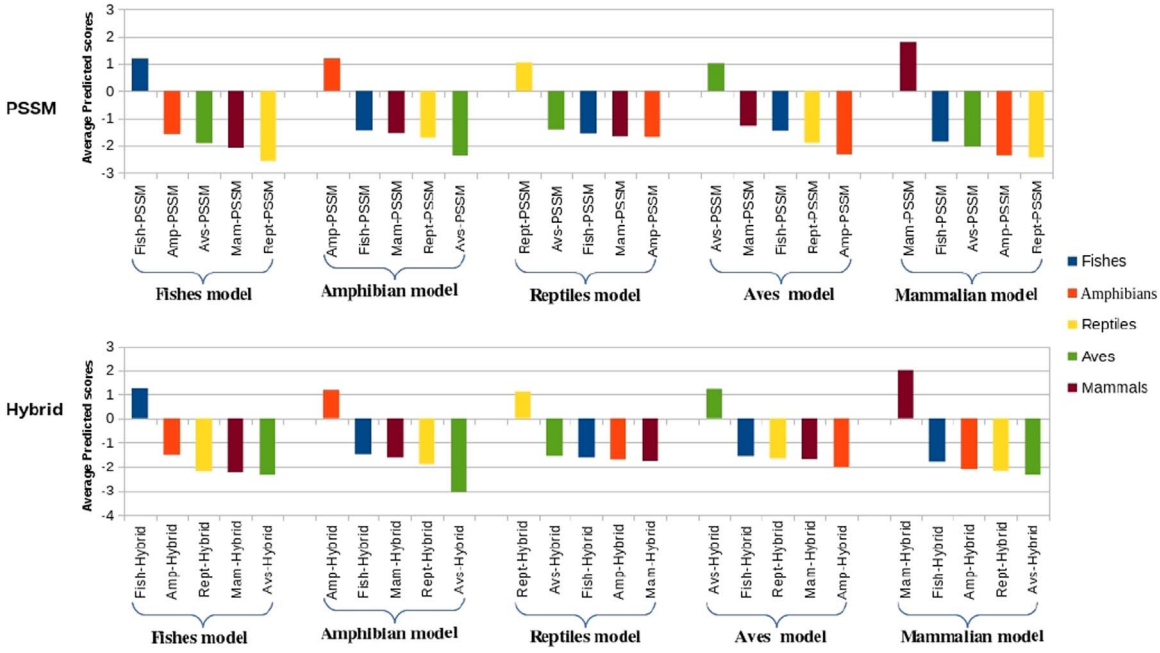

Fig. 6. Separation and the relationship of the models. Individual model performance for the separation of sub-classes based on the prediction scores data.

5. Conclusion

I developed a highly accurate method for identifying the VerHb proteins reported to their animal classifications. According to the literature studies, these approaches were applied for the first time for Hb proteins on the animal classifications. The online prediction server is freely available publicly, which allows the user to identify VerHb protein. I hope this study will assist and support the biologist in the annotation of genomes.

Conflict of interests

The author declares that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

I sincerely thank to Dr. Grish Sahni, Director, CSIR-IMTECH for his support. I would like to thank the anonymous reviewers for their valuable comments and suggestion to improve the quality of paper. I am thankful to Dr. K. L. Dikshit and Dr. Sri Krishna Subramanian of CSIR-IMTECH for their valuable suggestion to this manuscript. CSIR-IMTECH communication number for this manuscript is 0101/2015.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2016.08.027.

References

Ahmad, S., Gromiha, M.M., Sarai, A., 2004. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics. 20 (4), 477–486.
Ali, F., Hayat, M., 2015. Classification of membrane protein types using Voting Feature Interval in combination with Chou's pseudo amino acid composition. J. Theor. Biol. 384, 78–83.
Althaus, W., Gonzales, A.J., Chou, J.J., Romero, D.L., Deibel, M.R., Chou, K.C., Kezdy, F. J., Resnick, L., Busso, M.E., So, A.G., 1993a. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J. Biol. Chem. 268 (20), 14875–14880.
Althaus, W., Chou, J.J., Gonzales, A.J., Deibel, M.R., Chou, K.C., Kezdy, F.J., Romero, D. L., Palmer, J.R., Thomas, R.C., 1993b. Kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32 (26), 6548–6554.
Barman, R.K., Saha, S., Das, S., 2014. Prediction of interactions between viral and host proteins using supervised machine learning methods. PLoS One 9 (11), e112034. http://dx.doi.org/10.1371/journal.pone.0112034, eCollection 2014.
Beigi, M.M., Behjati, M., Mohabatkar, H., 2011. Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. J. Struct. Funct. Genom. 12 (4), 191–197.
Belekar, V., Lingineni, K., Garg, P., 2015. Classification of breast cancer resistant protein (BCRP) inhibitors and non-inhibitors using machine learning approaches. Comb. Chem. High Throughput Screen. 18 (5), 476–485.
Brittain, T., 2005. Root effect hemoglobins. J. Inorg. Biochem. 99 (1), 120–129.
Cai, Y.D., Lin, S.L., 2003. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. Biochim. Biophys. Acta 1648 (1–2), 127–133.
Cai, Y.-D., Zhou, G.-P., Chou, K.-C., 2005. Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. J. Theor. Biol. 234 (1), 145–149.
Caragea, J., Sinapov, A., Silvescu, D., Dobbs, Honavar, V., 2007. Glycosylation site prediction using ensembles of Support Vector Machine classifiers. BMC Bioinform. 8, 438.
Chen, G., Han, N., Li, G., Li, X., Liu, Y., Wu, W., Wang, Y., Chen, Y., Sun, G., Li, Z., Li, Q., 2015. Prediction of feature genes in trauma patients with the TNF rs1800629 A allele using support vector machine. Comput. Biol. Med. 64, 24–29. http://dx.doi.org/10.1016/j.compbiomed.2015.06.002, Epub 2015 Jun 9.
Chen, W., Feng, P.-M., Lin, H., Chou, K.-C., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res., gks1450
Chen, W., Feng, P., Ding, H., Lin, H., Chou, K.-C., 2016. Using deformation energy to analyze nucleosome positioning in genomes. Genomics 107 (2), 69–75.
Chen, Y.-K., Li, K.-B., 2013. Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. J. Theor. Biol. 318, 1–12.
Chou, C., 2001. Prediction of protein signal sequences and their cleavage sites. Proteins: Struct. Funct. Bioinform. 42 (1), 136–139.
Chou, K.C., Forsen, S., 1980. Graphical rules for enzyme-catalysed rate laws. Biochem. J. 187 (3), 829–835.
Chou, K.C., Shen, H.B., 2007. Large-scale plant protein subcellular location prediction. J. Cell Biochem. 100 (3), 665–678.
Chou, K.-C., 2001. Using subsite coupling to predict signal peptides. Protein Eng. 14 (2), 75–79.
Chou, K.-C., 2010. Graphic rule for drug metabolism systems. Curr. Drug Metab. 11 (4), 369–378.
Chou, K.-C., 2011. Some remarks on protein attribute prediction and pseudo amino

acid composition. J. Theor. Biol. 273 (1), 236–247.
Chou, K.-C., 2013. Some remarks on predicting multi-label attributes in molecular biosystems. Mol. Biosyst. 9 (6), 1092–1100.
Chou, K.-C., Zhang, C.-T., 1995. Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30 (4), 275–349.
Chou, K.-C., Cai, Y.-D., 2003. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. Biochem. Biophys. Res. Commun. 311 (3), 743–747.
Chou, K.-C., Shen, H.-B., 2009. Review: recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 1 (02), 63.
Chou, K.-C., Kézdy, F.J., Reusser, F., 1994. Kinetics of processive nucleic acid polymerases and nucleases. Anal. Biochem. 221 (2), 217–230.
Chou, K.-C., Wu, Z.-C., Xiao, X., 2011. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS One 6 (3), e18258.
Chou, K.-C., Lin, W.-Z., Xiao, X., 2011. Wenxiang: a web-server for drawing wenxiang diagrams. Nat. Sci. 3 (10), 862.
Chou, K.-C., Wu, Z.-C., Xiao, X., 2012. iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. 8 (2), 629–641.
Cruz, J.A., Wishart, D.S., 2007. Applications of machine learning in cancer prediction and prognosis. Cancer Inf. 2, 59–77.
Garg, A., Raghava, G.P., 2008. A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. In Silico Biol. 8 (2), 129–140.
Giardina, B., Messana, I., Scatena, R., Castagnola, M., 1995. The multiple functions of hemoglobin. Crit. Rev. Biochem. Mol. Biol. 30 (3), 165–196.
Guo, S.-H., Deng, E.-Z., Xu, L.-Q., Ding, H., Lin, H., Chen, W., Chou, K.-C., 2014. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics, Btu083.
Gupta, R. Kapil, Dhakan, D.B., Sharma, V.K., 2014. MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. PLoS One 9 (4), e93907. http://dx.doi.org/10.1371/journal.pone.0093907, eCollection 2014.
Hamp, T., Rost, B., 2015. Evolutionary profiles improve protein-protein interaction prediction from sequence. Bioinformatics 31 (12), 1945–1950. http://dx.doi.org/10.1093/bioinformatics/btv077, Epub 2015 Feb 4.
Hardison, R., 1998. Hemoglobins from bacteria to man: evolution of different patterns of gene expression. J. Exp. Biol. 201 (Pt 8), 1099–1117.
Hardison, R.C., 1996. A brief history of hemoglobins: plant, animal, protist, and bacteria. Proc. Natl. Acad. Sci. USA 93 (12), 5675–5679.
Hasan, M.M., Zhou, Y., Lu, X., Li, J., Song, J., Zhang, Z., 2015. Computational identification of protein pupylation sites by using profile-based composition of k-spaced amino acid pairs. PLoS One. 10 (6), e0129635. http://dx.doi.org/10.1371/journal.pone.0129635, ECollection 2015.
Hayat, M., Khan, A., 2012. Discriminating outer membrane proteins with Fuzzy K-nearest Neighbor algorithms based on the general form of Chou's PseAAC. Protein Pept. Lett. 19 (4), 411–421.
Huang, C., Yuan, J.-Q., 2013. A multilabel model based on Chou's pseudo-amino acid composition for identifying membrane proteins with both single and multiple functional types. J. Membr. Biol. 246 (4), 327–334.
Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.-C., 2016. pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. J. Theor. Biol. 394, 223–230.
Joachims, T., 1999. Making large-scale SVM learning practical. In: Scholkopf, B., Burges, C., Smola, A. (Eds.), Advances in Kernel Methods—Support Vector Learning. MIIT Press, Cambridge, MA;London, England.
Kaundal, R., Raghava, G.P., 2009. RSLpred: an integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information. Proteomics 9 (9), 2324–2342. http://dx.doi.org/10.1002/pmic.200700597.
Khan, Z.U., Hayat, M., Khan, M.A., 2015. Discrimination of acidic and alkaline enzyme using C Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. J. Theor. Biol. 365, 197–203.
Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I., 2014. Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. 13, 8–17. http://dx.doi.org/10.1016/j.csbj.2014.11.005, eCollection 2015.
Kumar, K.K., Pugalenthi, G., Suganthan, P.N., 2009. DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest. J Biomol. Struct. Dyn. 26 (6), 679–686.
Kumar, M., Gromiha, M.M., Raghava, G.P., 2007. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinform. 8, 463.
Kumar, R., Raghava, G.P., 2013. Hybrid approach for predicting coreceptor used by HIV-1 from its V3 loop amino acid sequence. PLoS One 8 (4), e61437. http://dx.doi.org/10.1371/journal.pone.0061437, Print 2013.
Kumar, R., Panwar, B., Chauhan, J.S., Raghava, G.P., 2011. Analysis and prediction of cancerlectins using evolutionary and domain information. BMC Res. Notes 4, 237. http://dx.doi.org/10.1186/1756-0500-4-237.
Kumar, R., Srivastava, A., Kumari, B., Kumar, M., 2015. Prediction of beta-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. J. Theor. Biol. 365, 96–103.
Li, W., Kondratowicz, B., McWilliam, H., Nauche, S., Lopez, R., 2013. The annotation-enriched non-redundant patent sequence databases. Database . http://dx.doi.org/10.1093/database/bat005, Print 2013.
Lin, H., Deng, E.-Z., Ding, H., Chen, W., Chou, K.-C., 2014. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. 42 (21), 12961–12972.
Lin, S.-X., Lapointe, J., 2013. Theoretical and experimental biology in one-a symposium in honour of Professor Kuo-Chen Chou's 50th anniversary and Professor Richard Giegé's 40th anniversary of their scientific careers. J. Biomed. Sci. Eng. 6 (4), 435.
Liu, B., Long, R., Chou, K.-C., 2016. iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. Bioinformatics, btw186.
Liu, B., Fang, L., Wang, S., Wang, X., Li, H., Chou, K.-C., 2015. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. J. Theor. Biol. 385, 153–159.
Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., Chou, K.-C., 2015. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 43 (W1), W65–W71.
Liu, B., Zhang, D., Xu, R., Xu, J., Chen, Q., Dong, Q., Chou, K.-C, 2014. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics 30 (4), 472–479.
Liu, J.L., Peng, Y., Fu, Y.S., 2015. Efficient prediction of progesterone receptor interactome using a support vector machine model. Int. J. Mol. Sci. 16 (3), 4774–4785. http://dx.doi.org/10.3390/ijms16034774.
Liu, Z., Xiao, X., Yu, D.-J., Jia, J., Qiu, W.-R., Chou, K.-C., 2015. pRNAm-PC: predicting N-methyladenosine sites in RNA sequences via physical-chemical properties. Anal. Biochem. 497, 60–67.
Madeira, F., Tinti, M., Murugesan, G., Berrett, E., Stafford, M., Toth, R., Cole, C., MacKintosh, C., Barton, G.J., 2015. 14−3-3-Pred: improved methods to predict 14-3-3-binding phosphopeptides. Bioinformatics 31 (14), 2276–2283. http://dx.doi.org/10.1093/bioinformatics/btv133, Epub 2015 Mar 3.
Mbah, N., 2014. Application of hybrid functional groups to predict ATP binding proteins. ISRN Comput. Biol. 2014, 581245.
Mishra, N.K., Raghava, G.P., 2010. Prediction of FAD interacting residues in a protein from its primary sequence using evolutionary information. BMC Bioinform. 11 (Suppl. 1), S48. http://dx.doi.org/10.1186/1471-2105-11-S1-S48.
Mishra, N.K., Chang, J., Zhao, P.X., 2014. Prediction of membrane transport proteins and their substrate specificities using primary sequence information. PLoS One 9 (6), e100278. http://dx.doi.org/10.1371/journal.pone.0100278, eCollection 2014.
Mohabatkar, H., Beigi, M.M., Abdolahi, K., Mohsenzadeh, S., 2013. Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. Med. Chem. 9 (1), 133–137.
Muthukrishnan, S., Garg, A., Raghava, G.P., 2007. Oxypred: prediction and classification of oxygen-binding proteins. Genom. Proteom. Bioinform. 5 (3–4), 250–252. http://dx.doi.org/10.1016/S1672-0229(08)60012-1.
Mylvaganam, S.E., Bonaventura, C., Bonaventura, J., Getzoff, E.D., 1996. Structural basis for the root effect in haemoglobin. Nat. Struct. Biol. 3 (3), 275–283.
Nanni, L., Lumini, A., Gupta, D., Garg, A., 2012. Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information. IEEE/ACM Trans. Comput. Biol. Bioinform. 9 (2), 467–475.
Ng, X.Y., Rosdi, B.A., Shahrudin, S., 2015. Prediction of antimicrobial peptides based on sequence alignment and support vector machine-pairwise algorithm utilizing LZ-complexity. Biomed. Res. Int. . http://dx.doi.org/10.1155/2015/212715, Epub 2015 Feb 23
Panwar, Raghava, G.P., 2014. Prediction of uridine modifications in tRNA sequences. BMC Bioinform. 15, 326. http://dx.doi.org/10.1186/1471-2105-15-326.
Perutz, M.F., 1983. Species adaptation in a protein molecule. Mol. Biol. Evol. 1 (1), 1–28.
Qiu, W.R., Sun, B.Q., Xiao, X., 2016. iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory. Mol. Inform. . http://dx.doi.org/10.1002/minf.201600010
Qiu, W.-R., Xiao, X., Chou, K.-C., 2014. iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. Int. J. Mol. Sci. 15 (2), 1746–1766.
Ramana, J., Gupta, D., 2009. LipocalinPred: a SVM-based method for prediction of lipocalins. BMC Bioinform. 10, 445. http://dx.doi.org/10.1186/1471-2105-10-445.
Rashid, M., Saha, S., Raghava, G.P., 2007. Support vector machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinform. 8, 337.
Rehman, Z.U., Khan, A., 2012. Identifying GPCRs and their types with Chou's pseudo amino acid composition: an approach from multi-scale energy representation and position specific scoring matrix. Protein Pept. Lett. 19 (8), 890–903.
Saha, B. Rak, Bhowmick, S.S., Maulik, U., Bhattacharjee, D., Koch, U., Lazniewski, M., Plewczynski, D., 2015. Binding activity prediction of cyclin-dependent inhibitors. J. Chem. Inf. Model. 55 (7), 1469–1482. http://dx.doi.org/10.1021/ci500633c, Epub 2015 Jul 10.
Sahu, S.S., Panda, G., 2010. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Comput. Biol. Chem. 34 (5–6), 320–327.
Selvaraj, M., Puri, M., Dikshit, K.L., Lefevre, C., 2016. BacHbpred: support vector machine methods for the prediction of bacterial hemoglobin-like proteins. Adv. Bioinform., Feb 29;2016

Shen, H.-B., Chou, K.-C., 2009. QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. J. Proteom. Res. 8 (3), 1577–1584.
Tao, P., Liu, T., Li, X., Chen, L., 2015. Prediction of protein structural class using tri-gram probabilities of position-specific scoring matrix and recursive feature elimination. Amino Acids 47 (3), 461–468. http://dx.doi.org/10.1007/s00726-014-1878-9, Epub 2015 Jan 13.
Verma, R., Melcher, U., 2012. A support vector machine based method to distinguish proteobacterial proteins from eukaryotic plant proteins. BMC Bioinform. 13 (Suppl. 15), S9. http://dx.doi.org/10.1186/1471-2105-13-S15-S9, Epub 2012 Sep 11.
Xiao, X., Wang, P., Chou, K.-C., 2011. GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. Mol. Biosyst. 7 (3), 911–919.
Xie, A., Li, M., Wang, Z., Fan, Feng, H., 2005. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. Nucleic Acids Res., W105–W110.
Xie, H.-L., Fu, L., Nie, X.D., 2013. Using ensemble SVM to identify human GPCRs N-linked glycosylation sites based on the general form of Chou's PseAAC. Protein Eng. Des. Sel. 26 (11), 735–742.
Xu, R., Zhou, J., Wang, H., He, Y., Wang, X., Liu, B., 2015. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. BMC Syst. Biol. 9 (Suppl. 1), S10. http://dx.doi.org/10.1186/1752-0509-9-S1-S10, Epub 2015 Feb 6.
Yousef, A., Charkari, N.M., 2015. A novel method based on physicochemical properties of amino acids and one class classification algorithm for disease gene identification. J. Biomed. Inform. 56, 300–306. http://dx.doi.org/10.1016/j.jbi.2015.06.018, Epub 2015 Jul 2.
Zhao, X., Ning, Q., Chai, H., Ai, M., Ma, Z., 2015. PGlcS: prediction of protein O-GlcNAcylation sites with multiple features and analysis. J. Theor. Biol. 380, 524–529. http://dx.doi.org/10.1016/j.jtbi.2015.06.026, Epub 2015 Jun 24.
Zhou, G.P., Deng, M.H., 1984. An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. Biochem. J. 222 (1), 169–176.
Zhou, G.-P., 2011. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. J. Theor. Biol. 284 (1), 142–148.