Stacked framework of machine learning classifiers for protein family prediction using protein characteristics

T. Idhaya[1,*], A. Suruliandi[1] and S. P. Raja[2]

[1]Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Abhishekapatti, Tirunelveli 627 012, India
[2]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore 632 014, India

A protein family must be identified, so that the protein can be modified and controlled for using it in the identification of drug target interactions, structure prediction, etc. Protein families are identified using the similarity between protein sequences. Alignment-free approaches use machine learning (ML) techniques for protein family prediction. In this study, two novel ML-based models, viz. a stacked framework of random forest, and a stacked framework of random forest, decision tree and naive Bayes for protein family prediction have been developed for a better identification of protein families. Both the models outperform state-of-the-art methods with an accuracy of 98.21% and 98.49% respectively. The proposed models give better results for twilight zone protein datasets as well.

Keywords: Alignment free method, machine learning, protein family prediction, stacked framework, twilight-zone proteins.

PROTEINS are molecules comprised of amino acids. These amino acids are linked by peptide bonds to forms protein sequences. Each protein sequences has a unique structure. Proteins are used in the fields like drug discovery, protein engineering, and identification of protein–protein interactions, protein 3D structure. So there is a need for studying proteins and their related families. The basis for protein identification is finding the similarity between sequences. This can be done with a part of the protein sequences, thus grouping similar sequences under a particular family. Experimental methods are costly and time-consuming. So, computational approaches have been introduced[1].

A large volume of data can be handled by computers today. This has lead to the rapid generation of raw data which are converted into information by data analytics. Based on historical data, predictive analysis is done to predict the future outcomes using machine learning (ML) techniques[2]. At present, the field of genetics is growing rapidly resulting a large amount of sequences of many cellular organisms. As the new generation technologies are cheaper and faster, thousands of protein sequences are obtained in a short time. In order to consider the various applications of proteins, there is need for identifying the protein family. Identifying a protein demands significant resource allocation, yet the predictive success rate remains notably minimal[3]. So computing techniques have been developed to identify the protein families. These methods predict the family of new proteins by comparing them with the existing proteins in databases. In this manner 2–10 sequences can be compared and the family identified based on similarity. The alignment based methods work for proteins with high similarities but it will not work for proteins which has different bio molecular functions (i.e. proteins with high dissimilarities). In order to overcome this drawback, several ML models have been developed[4]. INGA is a tool for protein family prediction only for proteins with 40% sequence similarity[5]. LOMETS[6] and HHSEARCH[7] work with twilight-zone proteins, i.e. proteins with very low similarity. The performance of these methods was poor when compared to other methods. The QUAST method is used to identify protein families with low similarity, but its performance is poor[8]. SVM Prot is a similarity-based ML method for predicting protein families it shows only average performance[9]. Thus protein family prediction is still a difficult task.

The literature reveals that combining two or more classifiers increases the efficiency of a model. In the present study, stacked framework models have been developed for protein family prediction. In this study we used data from the KEGG database for protein family prediction. The dataset was pre-processed for prediction. The pre-processed data fed to five different classifiers, viz. decision tree (DT), random forest (RF), *k*-nearest neighbors (*k*-NN), multilayer perceptron (MLP) and naïve Bayes (NB). RF, DT and NB were the top three best performing classifiers. The main goal of this study is to propose two different models for protein family prediction by combining the above classifiers. One is the stacked framework of the RF classifier (SFRF) and the other is the stacked framework of RF, DT and NB ML classifiers (SFRDN). The models have been evaluated with other ensemble techniques and other non-protein benchmark datasets to check their compatibility.

Motivation and justification

Human life is supported by the complex and coordinated interaction of proteins. Thus an understanding of the protein

*For correspondence. (e-mail: idhayathomas003@gmail.com)

structure and functions is important. Protein family prediction is a vital tool for learning about the function of proteins and their evolutionary relationships. By correctly predicting protein families, we can better understand the working of proteins in biological processes, and how they have evolved over time. This information can be used to design experiments to probe the function of proteins, and to develop new drugs and therapies. Hence, there is a need to study proteins in detail. Proteins can be used in various fields like genomics, proteomics, pharmacology, etc. This motivation fostered a drive to classify protein families into discrete categories.

Some alignment-free methods use ML classifiers to develop models for predicting protein families. However, most of the models perform poorly because of sequence similarity. So the ML based methods take protein sequence characteristics for prediction. Thus, there is a need for developing better models for protein family prediction. This motivated us to compare five different classifiers and their performance was evaluated for the KEGG dataset. The best performing classifiers were identified and two different models were developed using them for protein family prediction. In this study, we employed stacked framework of ensembles of ML classifiers for protein family prediction. The two proposed models are expected to perform better than the state-of-the-art methods.

Overview of the proposed work

Figure 1 provides an overview of the present study where the KEGG protein dataset was used[10]. A model for protein family prediction was developed by combining the best ML classification algorithms through majority voting. The models were evaluated using the performance evaluation metrics like accuracy and ROC.

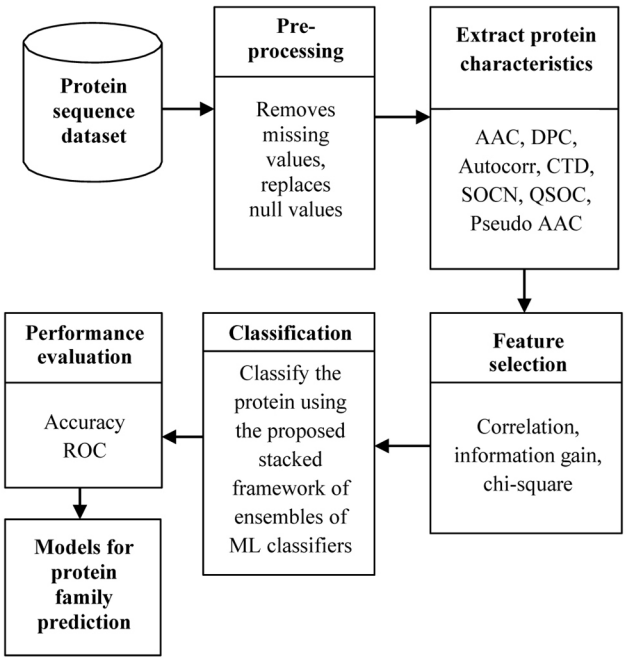

Figure 1. Proposed methodology.

Methodology

Protein sequences

A protein sequence consists of amino acids connected by peptide bonds[11]. The primary structure is the linear sequence of amino acids. A protein sequence is used to study its structure and functions[12].

Pre-processing

ML inputs raw data. Hence the dataset may contain missing, redundant or inaccurate information. Pre-processing raw data before classification improves the dataset quality. The standard values replaced in the missing places to make the input data classifiable. The extracted features were grouped with labels to provide a better predictive dataset.

Characteristics and features of proteins

Proteins possess several features that are important for protein family prediction, binding site prediction, identification of protein–protein interactions and drug–target interactions, and other applications. Table 1 shows seven sets of structural and physio-chemical properties called sequence-based features, which include 51 descriptors and 1497 descriptor values.

Protein family

In this study, four protein families have been considered. These include enzyme (E), G-protein coupled receptor (GPCR), ion channel (IC) and nuclear receptor (NR)[13].

We used the KEGG dataset in the study. Table 2 describes protein sequences collected from the KEGG database with 20,518 human genes. From the database, four different classes were chosen for study, which had around 1497 attributes.

Feature selection

Feature selection is an important step in ML[14,15]. From the study, it was found that the filter-based methods performed better than other feature selection methods. So only filter-based methods were considered. The most significant features were selected using the filter-based feature selection methods, viz. correlation[16], information gain[17] and chi square[18]. According to the study, the filter-based methods

Table 1. Characteristics and features of proteins

Protein characteristics	Features	Feature count
Amino acid composition (AAC)	A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V	20
Dipeptide composition (DC)	AA, AR, AN, AD, RN, RD, RQ, NN, ND, NR, NQ,QD, QN, QR, etc.	400
Normalized Moreau–Broto autocorrelation (MB)	Eight amino acid properties	240
Geary autocorrelation (GC)		240
Moron autocorrelation (MC)		240
Composition (C)	Seven amino acid properties	21
Transition (T)		21
Distribution (D)		105
Sequence order coupling number (SOCN)	Two amino acid properties and side chain volumes	60
Quasi-sequence order number (QSON)		100
Pseudo amino acid composition (PseudoAAC)	Two amino acid properties and side chain mass	50

Table 2. Protein sequence dataset

Protein class	No. of sequences
E	664
GPCR	204
IC	95
NR	26

E, Enzyme; GPCR, G-Protein coupled receptor; IC, Ion channel; NR, Nuclear receptor.

work on the basis of statistical measures and the most important feature selection methods are information gain, correlated features and chi-square. The most highly correlated features selected were K, D, A, G, I, E, R, Y, N, hydrophobicity.Group 1 polarity, charge, solvent access. The procedure for filter-based feature selection method is given below.

Machine learning classifiers

ML is a subfield of artificial intelligence (AI)[19]. It uses algorithms and statistical models to analyse data patterns and form inferences, advancing computer systems that can learn and adapt without being explicitly instructed. There are two types of learners in ML – lazy learners (*k*-NN, case-based learners, etc.) and eager learners (DT, NB, ANN, etc.). In this study we used eager learners to form a stacked framework of classifiers.

Proposed models

The two proposed classifiers SFRF and SFRDN were not chosen in a random manner. On performing doing experiments with five classifiers, the top three best-performing ones were chosen. Though a number of ensemble models are available for various applications, they have not yet been used in the bioinformatics domain. Hence, the novelty of this study lies in applying an ensemble of classifiers for protein family prediction, which is the need of the hour for proteomics researchers and pharmacologists.

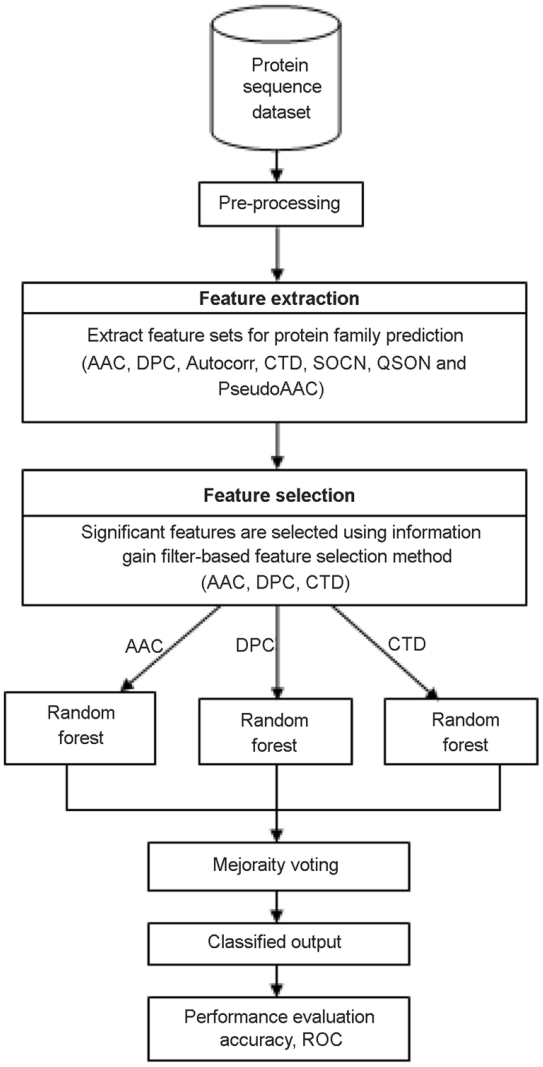

Figure 2. Working of the stacked framework of random forest (SFRF).

Proposed stacked framework of RF classifiers

This framework uses only RF for prediction. Figure 2 shows the working of this framework. Here three different protein features are given as in put to three RF classifiers and

the results are combined using majority voting. Then the performance of the proposed model is evaluated using the performance metrics accuracy and ROC.

Algorithm of the proposed stacked framework of the RF model

Parameters:
B: Base classifier for prediction.
E: Prediction of classifier.
S: Number of classifiers used.
Input of the prediction: Training dataset $(T) = \{x_i, y_i\}_{i=1}^{m}$ with class label *C*.
Output of the prediction: Proposed classifier's prediction E.
Step 1: Base classifiers learns
for L =1 to S do
Extract three different feature sets as AAC, DPC and CTD and give as input to each base classifier (RF classifier).
Training set (T) is used for learning
end for
Step 2: Compare the prediction using majority voting
for L = 1 to S do
Call B with T_m and receive the classifier S_m.
Compare the prediction E with C_m generated from S_m, update vote.
Aggregate vote
end for
Step 3: return E.

Proposed stacked framework of ML classifiers

This framework considers three ML classifiers, viz. RF (R), DT (D) and NB (N). Three different features of proteins are given as input to the classifiers. Then the prediction results are combined using majority voting. The above model has been developed using SFRF as the base model. Figure 3 shows the working of this model.

Algorithm of the proposed stacked framework of the RDN model

Parameters:
B: Base classifier for prediction.
E: Prediction of classifier.
S: Number of classifiers used.
Input of the prediction: Training dataset $(T) = \{x_i, y_i\}_{i=1}^{m}$ with class label *C*.
Output of the prediction: Proposed classifier's prediction E.
Step 1: Base classifiers learns
for L=1 to S do
Extract three different feature sets as AAC, DPC and CTD and give as input to each base classifier (RF, DT and NB).
Training set (T) is used for learning
end for
Step 2: Compare the prediction using majority voting
for L = 1 to S do
Call B with T_m and receive the classifier S_m.
Compare the prediction E with C_m generated from S_m, update vote.
Aggregate vote
end for
Step 3: Compare the voting of the classifiers.
for L = 1 to S do
Compare the voting generated from S_m,
end for
Step 4: return E.

Procedure for protein family prediction

Figure 4 shows the procedure for protein family prediction. The Homo sapiens (HSA) protein is found in the KEGG database. Before feature selection, the collected dataset was pre-processed. After identifying the predictive features, the dataset was separated into training and testing sets. The proposed stacked frameworks were used to train the classifier with the data. After training a classifier, the testing dataset was used to predict protein families. The performance of the classifier was evaluated using different measures.

Results and discussion

Dataset

Table 3 depicts the features extracted from the protein sequence based on its characteristics.

Performance metrics

Table 4 shows the confusion matrix for protein family prediction. Table 5 lists the performance metrics used.

Protein family prediction output

Table 6 shows the predicted output of the existing and proposed classifiers.

Choosing the best feature selection method

We employed the RF classifier as the base classifier. The feature selection approach chooses the most important features for quick algorithm training and therefore reduces computational complexity.

Table 7 shows that the filter-based feature selection method, viz. information gain selects the best subset of features. It selects the most significant features – AAC, DPC

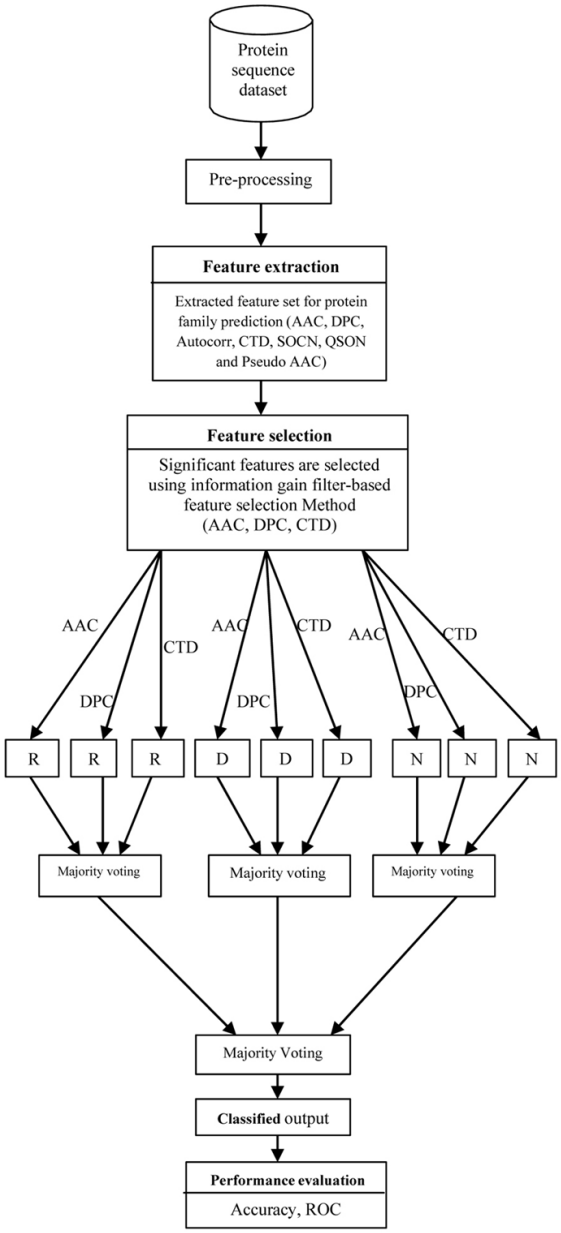

Figure 3. Working of the stacked framework of RDN (SFRDN). R, Random forest, D, decision tree and N, Naïve Bayes.

Table 3. Description of the dataset

Features	Description	Data type	Feature count
A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V	Composition of sequence	Numeric	20
AA, AR, AN, AD, RN, RD, RQ, NN, ND, NR, NQ, QD, QN, QR, etc.	Two-letter composition of sequence	Numeric	400
Auto correlation – eight amino acid properties	Correlation of physio-chemical properties	Numeric	720
CTD – seven amino acid properties	Physio-chemical properties – distributions and variants	Numeric	147
SOCN – two amino acid properties and side chain volume	Physio-chemical properties and combination of sequences	Numeric	60
QSON – three amino acid properties and side chain volume	Physio-chemical properties and combination of sequences	Numeric	100
Pseudo amino acid composition	Physio-chemical properties – combination of sequences and square correlation	Numeric	50
Total no. of features			1497

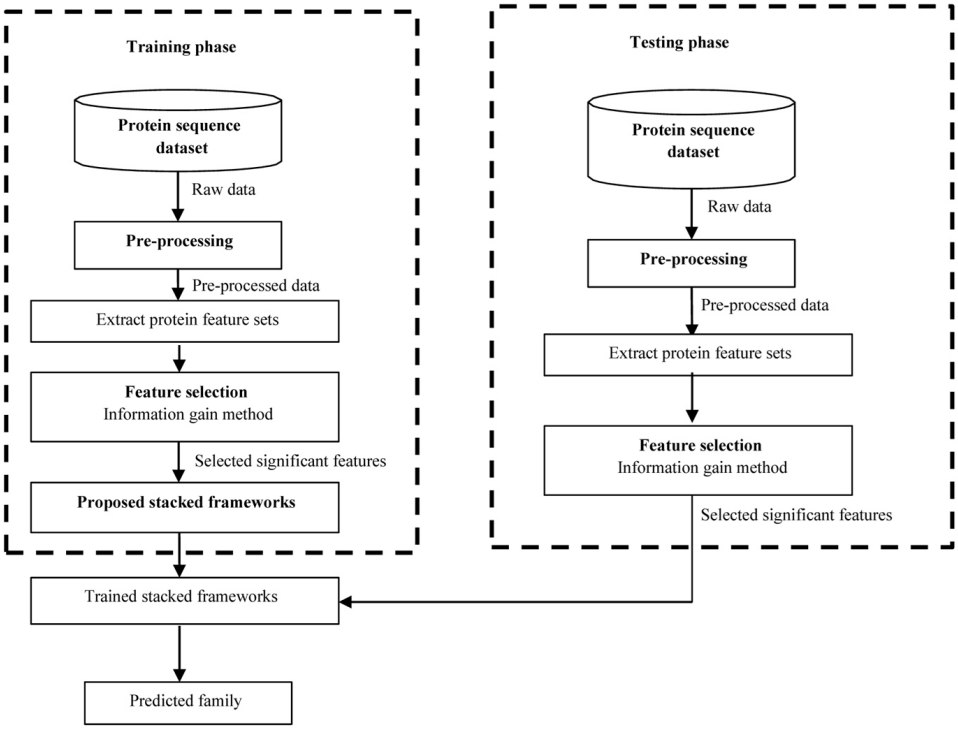

Figure 4. Procedure for protein family prediction.

Table 4. Confusion matrix for protein family prediction

	E	IC	GPCR	NR
E	TP	FN	FN	FN
IC	FP	TP	FN	FN
GPCR	FP	FP	TP	FN
NR	FP	FP	FP	TP

TP, True positive; FN, False negative; FP, False positive; TN, True negative.

Table 5. Performance metrics

Metrics	Formula	Description
Accuracy (ACC)	(TP + TN)/(TP + TN + FP + FN)	Ratio of correct predictions to total predictions
Precision (Prec)	TP/(TP + FP)	Measure of exactness
Recall (Rec)	TP/(TP + FN)	Measure of completeness
F-measure (FM)	$2\left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\right)$	Harmonic average of precision and recall
ROC	True positives versus false positives	Relationship between true positives and false positives
Error rate (ER)	1-Accuracy	Error of prediction

and CTD – among the seven feature sets of protein characteristics.

Finding the best base classifier for the stacked framework

The five different classifiers were evaluated and the best chosen for the proposed stacked framework construction. Table 8 shows the confusion matrix for the RF classifier.

Table 9 reveals that the RF classifier has the best performance of 95% accuracy, followed by DT and NB.

Optimal parameters for the classifiers of the proposed stacked framework

The performance of the classifiers can be improved by optimizing their parameters. To determine the best classifier parameters, experiments were conducted. Tables 10–12 show the results.

Table 6. Predicted output of the existing and proposed classifiers

Protein sequence	Expected output	Predicted output of classifiers						
		DT	RF	KNN	NB	MLP	Proposed SFRF	Proposed SFRDN
MSTTGQVIRCKAAILWKPGAPFSIEEVEVAPPKAKEVRIKVVATGLCGTEMKVLGSKHLDLLYPTILGHEGAGIVESIGEGV STVKPGDKVITLFLPQCGECTSCLNSEGNFCIQFKQSKTQLMSDGTSRFTCKGKSIYHFGNTSTFCEYTVIKEISVAKIDAV APLKVCLISCGFSTGFGAAINTAKVTPGSTCAVFGLGGVGLSVVMGCKAAGAARIIGVDVNKEKFKKAQELGATECLNP QDLKKPIQVLFDMTDAGIDFCFEAIGNLDVLAAALASCNESYGVCVVVGVLPASVQLKISGQLFFSGRSLKGSVFGGWK SRQHIPKLVADYMEKLNLDPLITHTLNLDKINEAVELMKTGKW	E	E	E	IC	E	G	E	E
MEDDSLYLRGEWQFNHFSKLTSSRPDAAFAEIQRTSLPEKSPLSCETRVDLCDDLAPVARQLAPREKLPLSSRRPAAVGAGL QNMGNTCYVNASLQCLTYTPPLANYMLSREHSQTCHRHKGCMLCTMQAHITRALHNPGHVIQPSQALAAGFHRGKQE DAHEFLMFTVDAMKKACLPGHKQVDHHSKDTTLIHQIFGGYWRSQIKCLHCHGISDTFDPYLDIALDIQAAQSVQQALE QLVKPEELNGENAYHCGVCLQRAPASKTLTLHTSAKVLILVLKRFSDVTGNKIAKNVQYPECLDMQPYMSQPNTGPLV YVLYAVLVHAGWSCHNGHYFSYVKAQEGQWYKMDDAEVTASSITSVLSQQAYVLFYIQKSEWERHSESVSRGREPRA LGAEDTDRRATQGELKRDHPCLQAPELDEHLVERATQESTLDHWKFLQEQNKTKPEFNVRKVEGTLPPDVLVIHQSKY KCGMKNHHPEQQSSLLNLSSSTPTHQESMNTGTLASLRGRARRSKGKNKHSKRALLVCQ	E	E	G	E	IC	E	E	E
MPIMGSSVYITVELAIAVLAILGNVLVCWAVWLNSNLQNVTNYFVVSLAAADIAVGVLAIPFAITISTGFCAACHGCLFIACF VLVLTQSSIFSLLAIAIDRYIAIRIPLRYNGLVTGTRAKGIIAICWVLSFAIGLTPMLGWNNCGQPKEGKNHSQGCGEGQV ACLFEDVVPMNYMVYFNFFACVVPLLLMLGVYLRIFLAARRQLKQMESQPLPGERARSLQKEVHAAKSLAIIVGLFAL CWLPLHIINCFTFFCPDCSAPLWLMYLAIVLSHTNSVVNPFIYAYRIREFRQTFRKIIRSHVLRQQEPFKAAGTSARVLAA HGSDGEQVSLRLNGHPPGVWANGSAPHPERRPNGYALGLVSGGSAQESQGNTGLPDVELLSHELKGVCPEPPGLDDPL AQDGAGVS	G	G	G	G	G	G	G	G
MVNENTRMYIPEENHQGSNYGSPRPAHANMNANAAAGLAPEHIPTPGAALSWQAAIDAARQAKLMGSGNATISTVSSTQR KRQQYGKPKKQGSTTATRPPRALLCLTLKNPIRRACISIVEWKPFEIIILLTIFANCVALAIYIPFPEDDSNATNSNLERVEY LFLIIFTVEAFLKVIAYGLLFHPNAYLRNGWNLLDFIIVVVGLFSAILEQATKADGANALGGKGAGFDVKALRAFRVLRP LRLVSGVPSLQVVLNSIIKAMVPLLHIALLVLFVIIIYAIIGLELFMGKMHKTCYNQEGIADVPAEDDPSPCALETGHGRQ CQNGTVCKPGWDGPKHGITNFDNFAFAML	IC	IC	IC	IC	G	E	IC	IC
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
MEQKPSKVECGSDPEENSARSPDGKRKRKNGQCSLKTSMSGYIPSYLDKDEQCVVCGDKATGYHYRCITCEGCKGFFRRTI QKNLHPTYSCKYDSCCVIDKITRNQCQLCRFKKCIAVGMAMDLVLDDSKRVAKRKLIEQNRERRRKEEMIRSLQQRPEP TPEEWDLIHIATEAHRSTNAQGSHWKQRRKFLPDDIGQSPIVSMPDGDKVDLEAFSEFTKII	NR	NR	NR	NR	E	NR	NR	NR

Table 7. Choosing the best feature selection method

Feature selection method	Evaluator	No. of features selected	Performance metrics		
			Acc	Prec	FM
Without feature selection	–	1497	0.90	0.89	0.93
Filter method	Correlation	493	0.70	0.69	0.82
	Information gain	617	0.93	0.92	0.96
	Chi-square	598	0.91	0.89	0.92

Table 8. Confusion matrix of random forest (RF) classifier

Classified data	E	IC	GPCR	NR	Row total
E	80	0	1	1	80
IC	5	21	0	0	26
GPCR	0	1	11	0	12
NR	1	0	0	2	3
Column total	84	22	12	3	121

Table 9. Finding the best base classifier for the stacked framework

Classifier	Performance metrics				
	Acc	Prec	Rec	FM	ROC
DT	0.8234	0.70	0.76	0.72	0.66
RF	0.95	0.76	0.83	0.77	0.97
KNN	0.81	0.85	0.62	0.68	0.55
MLP	0.73	0.63	0.68	0.64	0.58
NB	0.82	0.87	0.80	0.82	0.58

Table 10. Optimizing the parameters of the RF classifier

Trial no.	Hyper parameter experimental set	Performance metrics				
		Acc	ER	Prec	Rec	FM
1	Entropy, 700, 2, Sqrt	0.96	0.04	0.85	0.84	0.81
2	Gini, 200, 3, Auto	0.88	0.12	0.91	0.71	0.76
3	Gini, 400, 5, Sqrt	0.87	0.13	0.68	0.62	0.64
4	Gini, 600, 7, Log 2	0.82	0.18	0.65	0.54	0.57

Performance evaluation of the proposed models with the RF classifier for all three feature sets

The performance of the proposed models was evaluated against the RF classifier with the selected feature sets (Table 13).

Table 13 reveals that SFRDN performs better than the RF classifier with three different features. Thus it outperforms both SFRF and the RF classifier.

Performance evaluation of the proposed models for other protein datasets

The proposed models were evaluated for the cluster of orthologous (COG) dataset[20]. COG also called as twilight zone protein, Twilight zone proteins show 20–35% sequence similarity, while most other protein shows more than 40% sequence similarity. The twilight zone proteins are collected from the databases PDB[21], GPCR-COG[22] and Pfam[23] the dataset has different protein family. Table 14 shows the performance of the models for other protein datasets.

Table 14 reveals that SFRF performs well for the Pfam dataset while SFRDN performs well for the COG dataset.

Performance of the proposed models with other classifiers for varied non-protein datasets based on accuracy

Table 15 provides a description of the non-protein datasets. Table 16 shows the performance of proposed models in evaluating the non-protein datasets with the other classifiers.

Table 11. Optimizing the parameters of the decision tree (DT) classifier

Trial no.	Hyper parameter experimental set	Performance metrics				
		Acc	ER	Pre	Rec	FM
1	Entropy, 10, 2, Sqrt	0.80	0.20	0.75	0.74	0.60
2	Entropy, 32, 2, Auto	0.824	0.18	0.76	0.78	0.77
3	Gini, 64, 3, Sqrt	0.81	0.19	0.64	0.69	0.63
4	Gini, 32, 2, Auto	0.79	0.21	0.55	0.52	0.59

Table 12. Optimizing the parameters of the Naïve Bayes (NB) classifier

Trial no.	Hyper parameter experimental set	Performance metrics				
		Acc	ER	Pre	Rec	FM
1	1, True, none	0.8254	0.17	0.83	0.84	0.65
2	0, False, none	0.80	0.20	0.54	0.59	0.53

Table 13. Performance of the proposed models with the RF classifier

Classifier	Feature set	Performance metrics (%)				
		Acc	Prec	Rec	FM	ROC
RF classifier	AAC	91.75	91.75	90.85	89.75	90.50
	DPC	95.87	95.57	95.70	95.80	94.87
	CTD	96.10	95.04	96.04	96.54	96.04
Proposed SFRF	All three features	98.21	97.45	96.55	95.87	96.89
Proposed SFRDN	All three features	98.49	96.78	95.76	94.79	97.90

Table 14. Performance evaluation of the proposed models for other protein datasets

Proposed models	Protein dataset	No. of instances	No. of classes	Performance metrics (%)				
				Acc	Prec	Recall	FM	ROC
SFRF	COG dataset	1,389,595	30	94.92	93.23	93.56	94.35	93.05
	PDB dataset	1,234	2	90.87	91.23	92.45	93.55	93.78
	GPCR-COG	8,345	5	92.34	92.44	93.45	94.12	93.44
	Pfam dataset	19,632	5	95.87	90.45	93.45	93.55	96.65
SFRDN	COG dataset	1,389,595	30	96.54	85.65	90.45	93.44	95.89
	PDB dataset	1,234	2	91.73	91.43	90.45	92.55	94.78
	GPCR-COG	8,345	5	90.14	91.34	93.25	93.02	92.41
	Pfam dataset	19,632	5	93.89	92.50	93.45	90.67	92.09

Table 15. Description of the various non-protein datasets

Dataset	Number of instances	Number of attributes	Number of classes
Breast cancer[24]	498	10	2
Prisoner[25]	463	31	3
Hypothyroid[26]	3772	30	3
Advertisement[27]	3279	1559	2

Table 16. Performance evaluation of the proposed models using metric accuracy for other non-protein datasets

Dataset	Proposed model SFRF	Proposed model SFRDN	SVM	DT	NB
Breast cancer	0.93	0.96	0.92	0.90	0.78
Prisoner	0.92	0.95	0.88	0.89	0.76
Hypothyroid	0.86	0.90	0.88	0.90	0.80
Advertisement	0.75	0.82	0.73	0.75	0.79

Table 17. Performance evaluation of the proposed models with other ensemble models

Stacked framework	Performance metrics (%)				
	Acc	Prec	Recall	FM	ROC
Bagging	95.00	96.34	97.8	91.8	97.45
Adaboost	95.20	95.32	96.5	92.56	96.50
Proposed model SFRF	98.21	97.45	96.55	95.87	96.89
Proposed model SFRDN	98.49	96.78	95.76	94.79	97.90

Table 16 reveals that SFRDN performs better than SFRF and the other classifiers with greater accuracy.

Comparing performance evaluation of the proposed models with ensemble models

The performance of the proposed models was evaluated with other ensemble models (Table 17).

From Table 17, it can be inferred that the proposed models outperform the other ensemble models.

Conclusion

Protein sequence dataset under four different classes from the KEGG database was considered for the present study. The protein characteristics were extracted from the protein sequences. The extracted feature sets were pre-processed. Information gain was chosen as the best feature selection method. The selected features were given as input to the proposed stacked framework models for protein family prediction. The cross-validation (CV) method used was *k*-fold CV. Here, ten fold CV was used for selecting the subset of the dataset. Then the chosen fold was divided into 70% for training and 30% for testing. The proposed SFRF model achieved an accuracy of 98.21% and ROC of 96.89%, while SFRDN achieved an accuracy of 98.49% and ROC of 97.90% for the KEGG dataset. In future, these models can be used to identify proteins in the Pfam dataset, which has a specific section for unidentified proteins.

1. https://www.dnastar.com/blog/structural-biology/why-structure-prediction-matters (accessed on 12 September 2022).
2. Ranjini, K., Suruliandi, A. and Raja, S. P., A stacked framework of heterogeneous incremental classifiers for assisted reproductive technology outcome prediction. *IEEE Trans. Comput. Soc. Syst.*, 2021, **8**(3), 557–567.
3. Cao, R. *et al*., DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC Bioinform*., 2016, **17**, 495; https://doi.org/10.1186/s12859-016-1405-y.
4. Mukherjee, S. *et al.*, Genomes OnLine database (GOLD) v.7: updates and new features. *Nucleic Acids Res.*, 2019, **47**(D1), D649–D659; https://doi.org/10.1093/nar/gky977.
5. Piovesan, D., Giollo, M., Leonardi, E., Ferrari, C. and Tosatto, S. C., INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity. *Nucleic Acids Res.*, 2015, **43**(W1), W134–W140; https://doi.org/10.1093/nar/gkv523.
6. Wu, S. and Zhang, Y., LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res.*, 2007, **35**(10), 3375–3382; https://doi.org/10.1093/nar/gkm251.
7. Söding, J., Protein homology detection by HMM–HMM comparison. *Bioinformatics*, 2005, **21**(7), 951–960; https://doi.org/10.1093/bioinformatics/bti125.
8. Smaili, F. Z. *et al.*, QAUST: protein function prediction using structure similarity, protein interaction, and functional motifs. *Genom. Proteom. Bioinformat.*, 2021, **19**(6), 998–1011; https://doi.org/10.1016/j.gpb.2021.02.001.
9. Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, X. and Chen, Y. Z., SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.*, 2003, **31**(13), 3692–3697; https://doi.org/10.1093/nar/gkg600.
10. Kanehisa, M. and Goto, S., KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., 2000, **28**, 27–30.
11. https://www.sciencelearn.org.nz/resources/1901-proteins-what-they-are-and-how-they-re-made (accessed on 12 September 2022).
12. https://bio.libretexts.org/Bookshelves/Microbiology (accessed on 12 September 2022).
13. https://www.guidetopharmacology.org/targets.jsp (accessed on 12 September 2022).
14. Saeys, Y., Inza, I. and Larrañaga, P., A review of feature selection techniques in bioinformatics. *Bioinformatics*, 2007, **23**(19), 2507–2517.
15. Chandrashekar, G. and Sahin, F., A survey on feature selection methods. *Comput. Electr. Eng.*, 2014, **40**(1), 16–28.
16. Yu, L. and Liu, H., Feature selection for high-dimensional data: a fast correlation-based filter solution. In Proceedings, Twentieth International Conference on Machine Learning (eds Fawcett, T. and Mishra, N.), 2003, vol. 2, pp. 856–863.
17. Jaynes, E. T., Information theory and statistical mechanics II. *Phys. Rev*., 1957, **108**(2), 171–190; Bibcode:1957PhRv.108.171J; doi: 10.1103/physrev.108.171.
18. https://machinelearningmastery.com/feature-selection-machine-learning-python/Chi-square (accessed on 12 September 2022).
19. Wagstaff, K., Machine learning that matters, 2012; arXiv:1206.4656.
20. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. and Koonin, E. V., Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, 2015, **43**(D1), D261–D269.
21. Xu, Y. *et al.*, Deep dive into machine learning models for protein engineering. *J. Chem. Inf. Model*., 2020, **60**(6), 2773–2790; doi: 10.1021/acs.jcim.0c00073.
22. Yusuf, S. M., Zhang, F., Zeng, M. and Li, M., Deep PPF: a deep learning framework for predicting protein family. *Neurocomputing*, 2021, **428**, 19–29; doi:10.1016/j.neucom.2020.11.062.
23. Blum, M. *et al.*, The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, 2020, **49**(D1), D344–D354.
24. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data (accessed on 2 November 2022).
25. David, H., Fredrick, B., Suruliandi, A. and Raja, S. P., Preventing crimes ahead of time by predicting crime propensity in released prisoners using data mining techniques. *Int. J. Appl. Decis. Sci.*, 2010, **12**(3), 307–336; https://github.com/Benjamindavid03/Crime-PropensityPredictionDataset
26. https://www.kaggle.com/datasets/yasserhessein/thyroid-disease-data-set (accessed on 2 November 2022).
27. https://www.kaggle.com/datasets/bumba5341/advertisingcsv (accessed on 2 November 2022).

Received 17 December 2022; revised accepted 7 June 2023

doi: 10.18520/cs/v125/i5/508-517