CAA-PPI: A Computational Feature Design to Predict Protein-Protein Interaction using different Encoding Strategies

Bhawna Mewara (✉ mewara.bhawna2203@gmail.com)

Career Point University - Kota Campus: Career Point University

Gunjan Sahni (✉ gunjan22aug@gmail.com)

Career Point University - Kota Campus: Career Point University

Soniya Lalwani (✉ slalwani.math@gmail.com)

Career Point University, Kota (Raj.) <https://orcid.org/0000-0001-5757-9744>

Rajesh Kumar (✉ rkumar.ee@mnit.ac.in)

Malaviya National Institute of Technology

Research Article

Keywords: Machine Learning, Protein-Protein Interactions, Encoding Strategy, Feature representation

DOI: <https://doi.org/10.21203/rs.3.rs-443944/v2>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

CAA-PPI: A Computational Feature Design to Predict Protein-Protein Interaction using different Encoding Strategies

Bhawna Mewara · Gunjan Sahni · Soniya Lalwani ·
Rajesh Kumar

Received: date / Accepted: date

Abstract Protein-protein interactions (PPIs) carry out an extensive variety of biological procedures, containing cell-to-cell interactions, metabolic and developmental control. PPI is becoming one of the most important aims of system biology. PPI act as a fundamental part in predicting the protein function of the target protein and drug ability of molecules. Abundant work has been done to nurture methods to predict PPIs computationally as this supplements laboratory trials and offers cost-effective way of predicting the most likely set of interactions at the entire proteome scale. This article presents an innovative feature representation method (CAA-PPI) to extract features from protein sequence using two different encoding strategies and then an ensemble learning method, the Random Forest method is used as a classifier for PPIs prediction. CAA-PPI considers the role of trigram and bond of given amino acid with its nearby ones. The proposed PPI model achieves more than 98% prediction accuracies with one encoding scheme and more than 95% prediction accuracies with another encoding scheme respectively for the two diverse PPI datasets *i.e.* *H. Pylori* and *Yeast*. Further investigations are made to compare the CAA-PPI approach with existing sequence-based methods and reveals the proficiency of the proposed method with both encoding strategies. To further assess the practical prediction competence, a blind test has implemented on five other species' datasets independent of the training set, and obtained result ascertains the productivity of CAA-PPI with both encoding schemes.

Keywords Machine Learning · Protein-Protein Interactions · Encoding Strategy · Feature representation

1 Introduction

Proteins word is spawned from the Greek word 'protos', where its significance is the first element *Reeds (2000)*, undoubtedly protein is the fundamental of life. Proteins are complexes made with 20 types of amino acids that are linked through bonds called α -peptide bond. Merely protein can't act any function, therefore it accomplishes its roles by interacting with other molecules, like DNA, RNA, or other proteins, which catalyse different biological functions at a system or cellular stage. PPI can create a novel binding site for small effector molecules in accordance with different researches, PPI is the consequence of hydrophobic and electrostatic interactions and hydrogen bonds, all together contribute to the binding interaction nonetheless significance of hydrophobic forces have been proved *Maleki et al. (2013)*. There are some important properties of PPIs that are noteworthy. Firstly, the changes in kinematic properties of enzymes due to PPIs might cause delicate fluctuations in substrate binding or allosteric effects. Secondly, PPIs can form a new binding site for small effector molecules. Next, they can deactivate or suppress a protein. Interacting with different binding partners, PPI can alter the precision of protein for its substrate. PPI types are varied can be categorized

Bhawna Mewara and Gunjan Sahni
Department of Computer Science & Engineering, Career Point University, Kota, India
E-mail: mewara.bhawna2203@gmail.com, gunjan22aug@gmail.com

Soniya Lalwani
Department of Computer Science & Engineering, Career Point University, Kota, India
E-mail: slalwani.math@gmail.com

Rajesh Kumar
Department of Electrical Engineering, Malaviya National Institute of Technology, Jaipur, India
E-mail: rkumar.ee@gmail.com

Acronyms

AAC	Amino Acid Contact	AC	Auto Covariance
ACC	Auto Cross Covariance	CAA-PPI	Connecting Amino Acids Feature based PPI approach
CT	Conjoint Triad	DIP	Database of Interacting Proteins
DCI	Discrete Cosine Transform	E-ELM	Ensemble Extreme Learning Machine
DT	Decision Tree	FN	False Negatives
ES	Encoding Strategy	HKNN	K-local hyperplane distance nearest neighbor
FP	False Positives	LCPSSMAB	Local Coding PSSM Average Bigram
HOG	Histogram of Oriented Gradient	LCPSSMMF	Multifeatures fusion based on local coding PSSM matrix
KNN	K-Nearest Neighbour	LPQ	Local Phase Quantization
LCPSSMBG	Local Coding PSSM Bigram Group	MAC	Moran autocorrelation
LDA	Latent Dirichlet Allocation	ML	Machine Learning
LD	Local Descriptors	MLD	Multi-Scale Local Descriptor
LR	Logistic Regression	NPV	Negative Predictive Value
MCD	Multi-scale Continuous and Discontinuous	PCA	Principal Component Analysis
MCC	Matthew's correlation coefficient	PIR	Protein Information Resource
PC	Principal Components	PSSM	Position - Specific Scoring Matrix
PCVM	Probabilistic Classification Vector Machine	RF	Random Forest
PR	Property Response	RNN	Recurrent Neural Networks
PSSMMF	Multifeatures fusion based on original protein sequence PSSM matrix	SVD	Singular Value Decomposition
RoF	Rotation forest	SVM	Support Vector Machine
SMR	Substitution Matrix Representation	TP	True Positives
TN	True Negatives		

centred on stability, affinity, and composition of the consequential complex *Keskin et al. (2016)*, in transient or permanent, nonobligate interactions or obligate interactions respectively, similarly can be characterised in Homo-Oligomeric and Hetero-Oligomeric Complexes on the basis of the likeness of protein pair involved in the interaction. The interactions are obligate if the complex of PPI are unsteady on their own in vivo whereas resultant complexes of nonobligate interactions can occur autonomously. Nonobligate interactions exist either in transient or permanent mode. Transient interactions clasp and unclasp provisionally in vivo, though in permanent interaction complexes remain established once interaction of proteins. As PPIs implies various effects alike *Golemis and Adams (2002)*:

- Permit for substrate channelling.
- Formation of a novel binding site.
- Deactivate or abolish a protein.
- Alteration of the specificity of a protein.
- Oblige an adjusting role in an upstream or a downstream event.

All of the above placed excessive influence on plenty of biochemical events, consequently the exploration of PPIs can comfort scholars divulge tissue functions and structures and detect the maturation of syndromes and drug targets of gene remedy. In past years, numerous experimental methods have been employed to detect PPI resulted in high-productivity, comprising immunoprecipitation, Yeast two-hybrid system, affinity purification-mass spectrometry (AP-MS), and protein microarrays. Conversely, biological experiments are mostly expensive and onerous. Also, both the FN and FP rates of these methods are very high *Prieto et al. (2014)*. Thus, to develop reliable computing models for PPI prediction has great practical impact.

As per Galileo 'Book of Nature is written in mathematical language', therefore to detect the possibilities of interaction can be mapped by using different mathematical approaches by using properties and associated data of proteins as input to different computational models. Heretofore multiple research model has been introduced for predicting PPI which are categorize into approaches *Rai and Bhatnagar (2017)* like Gene data-Based methods, Network Topology-Based Methods, Structural profile-based methods, and ML-based methods as shown in Figure 1. Genetic linkage, genetic fusion, polygenetic profile, and in silico two-hybrid systems for PPI prediction are used in the genetic approach. Protein three-dimensional information is used in the structural approach whereas in the network topology based approach a confidence score matrix is generated for prediction purposes. ML based methods train the prediction model by exhausting diverse features of the interacting protein, far ahead train model predict the interaction of proteins.

In the proposed approach, ML-based model named Connecting Amino Acids Feature based PPI (CAA-PPI) approach is introduced to predict PPI. A Major contribution of the given model is the novel feature generation method using the hypothesis of association of different amino acid with a residue in a given trigram. CAA-PPI is using PCA to eliminate irrelevant and redundant features from the dataset. Also, CAA based feature extraction approach is implemented with two different encoding schemes named *ES1* and *ES2*, trailed by an

RF classifier to train the model. CAA-PPI with RF classifier model’s performance is verified with two different PPI datasets that are *Yeast* and *H. Pylori*, accomplished the average accuracies of 98.25% and 98.25% with one encoding scheme; 98.69% and 95.49% with another encoding scheme respectively and similarly, the comparison result of CAA-PPI with competitive approaches proved it more accurate. The proposed model is also tested with a random dataset (five spices dataset) define as an independent dataset due to its independence from the training dataset. Overall outcomes of the proposed method prove this approach more efficient in the prediction of PPI with both encoding strategies.

Fig. 1: Computational Approaches for predicting PPI.

The structure of this research article is: Firstly, the importance and challenges of PPI and the need for PPI prediction are discussed. Next, the studies about PPI that were published previously are deliberated including the researches related to encoding strategy and feature extraction approach as well. In the next section, details of additional materials that are used to carry out experimental work are presented. Tracked by the detailed description of the proposed approach including systematic workflow and pseudo code with an example for further understanding of CAA-PPI. After then, the performance of CAA-PPI with *ES1* and *ES2* on two diverse datasets is presented with seven standard measures followed by their respective Bonferroni post-hoc analysis comparisons with the state-of-the-art models. In the end, the research work is concluded with possible opportunities in CAA-PPI.

2 Documentary Research

The knowledge to build a PPI prediction model using sequence is primarily dependent on three factors:

- Select an appropriate manner to cover the possible essential information about PPI.
- Strategy to develop protein sequence feature extraction, and
- Apply a favourable classification algorithm.

This article mainly concentrates on the first two factors and therefore this section concisely considers their related study as briefed in table 1. Several investigations have been done in the development of an encoding scheme for fully capturing biological sequence information. *Stupar (2010)* suggested 7 classes of amino acids centered on their dipole and side-chain volumes, and according to those classes, the features of the protein pairs were extracted. *Al-Daoud (2011)* used three different encoding strategies based on amino acids’ chemical properties, polarity, and structure with three newly created feature sets. In 2017, Zhou encoded a protein sequence at multi-scale using seven properties, covering their qualitative and a quantitative explanation of amino acids. These encodings were then used for representing each protein sequence in terms of five different

Table 1: **Brief overview of previously suggested feature extraction methods for PPI prediction.**

S. No.	Reference, Year	Feature Variant	Classifier	Approach
1.	<i>Yang et al. (2010)</i>	LD	SVM	Considered the residues’ interactions in both continuous and discontinuous regions and extracted more information of PPI from the protein sequence.
2.	<i>Stupar (2010)</i>	CT	SVM	k-mer based assembly algorithm that divides three successively occurred nearby amino acids into one collective entity and computes the frequency of every combination in the whole sequence.
3.	<i>Guo et al. (2008)</i>	AC	SVM	Considered the protein sequence as a set of signals which is then transformed in digitized form using suitable physicochemical properties which are promoted to scrutinize protein features.
4.	<i>Martin et al. (2005)</i>	Signature Descriptor	Signature kernel SVMs	Derived signature product that is a product of sub-sequences of protein and covers the signature descriptor from compound facts which is used like feature for PPI prediction using SVM.
5.	<i>You et al. (2013)</i>	AC+CT+LD+MAC	Ensemble of ELM	Used distinctive properties of ELM classifier comprising quick learning, well generalization performance, modest and onerous parameter tuning to predict protein interactions.
6.	<i>You et al. (2014)</i>	MCD	SVM	Employed the interfaces between serially remote but spatially near residues of amino acid to appropriately cover many overlying continuous and discontinuous segments present in sequence.
7.	<i>You et al. (2015)</i>	MLD	RF	Used a decomposition technique at multi-scale level to split sequence of protein into many variable size fragments which contain information of coinciding local sections.
8.	<i>Wong et al. (2015)</i>	PR+LPQ	RF	Generated PR matrix using the amino acids’ physicochemical property which was united to an LPQ descriptor to generate Eigenvalues for protein.
9.	<i>Ding et al. (2016)</i>	HOG+SVD	RF	Proposed SVD and HOG algorithms for feature vector generation.
10.	<i>An et al. (2019)</i>	LCPSSMMF	SVM	Proposed a feature extraction method which considered residues’ interactions of both continuous and discontinuous sections present in sequences.

protein descriptors *i.e.* AC, composition, frequency, transformation, and distribution *Zhou et al. (2017)*. *ElAbd et al. (2020)* compared two counterpart amino acid encoding schemes using CNN, RNN, and a hybrid CNN-RNN architecture, applied to two challenging problems. In 2020, a broad review was projected by Jing about various encoding methods for amino acids followed by systematic analysis of encoding methods and discussed the comparison of the performance of 16 different representatives of encoding methods which were classified in five categories by the author (*Le and Nguyen (2019)*).

Numerous computational methods proposed by several publishers for extracting sequence features mostly deal with evolutionary information of proteins, physiochemical information, or structure information. One of the popular feature extraction work published by Chou (*Kuo-Chen (2009)*) remarkably reflects the amino acids’ composition and progressions the locus information of amino acid. *Yang et al. (2010)* deliberated the residues’ interaction in both continuous and discontinuous regions and extracted more information of PPI present in the sequence of the protein and proposed LD and a KNN model.

Guo et al. (2008) done another excellent work by taking into account the discontinuous amino acid fragments of protein sequence by using an AC-based method. The process considered physicochemical properties, a descriptor ‘signature product’ was developed to determine PPIs (*Martin et al. (2005)*). The research work by (*You et al. (2013)*) proposed a new hierarchical model by first extracting the information that causes interaction of protein sequence using CT, AC, MAC, and LD and then use PCA and finally employed an E-ELM classifier to predict PPI. Another great research by *You et al. (2014)* considered the interfaces between serially remote but spatially near amino acid residues. Again in 2015, (*You et al. (2015)*) suggested another innovative feature representation approach with a postulation that the interaction between protein pairs could be possible in unceasing amino acid fragments having different segment lengths. The notable work was proposed by *Wong et al. (2015)* using image processing methods for feature extraction using Physicochemical PR Matrix and then employed LPQ for mining complex and essential coefficients from obtained features. To predict PPIs favorably, the RoF classifier was used and showed efficient performance by comparing with existing approaches. *Ding et al. (2016)* improved the precision of prediction using AAC matrix to obtain an SMR matrix followed by SVD and HOG algorithms that generate a feature vector. Another brilliant work was published by *An et al. (2019)* used both local and global features by using PSSM based local encoding approach to create a novel multi-features fusion matrix (CPSSM) and then employed Local Average Group (LAG) and Bigram Probability (BP) to extracted key features from the obtained matrix. In the current research article, an innovative feature

representation method CAA-PPI is projected to extract key information present in protein sequence which takes into consideration the association of different amino acids with a residue in a given trigram. These novel features are extracted using two different encoding schemes for representing amino acids. Then, RF is used as a classifier to prove the efficacy of the approach for predicting interaction between protein pairs. The proposed PPI model achieves prediction accuracies of 98.25% and 98.25% with one encoding scheme; 98.69% and 95.49% with another encoding scheme respectively when applied on two diverse PPI datasets including *Yeast* and *H. Pylori*. Further investigations are made to compare the proposed approach with existing sequence-based methods and reveals outstanding results which prove the proficiency of the proposed method. Further, to evaluate the practical prediction competence, a blind test has been implemented on five other species' datasets which are autonomous to the training set, and obtained results ascertain the productivity of CAA-PPI.

3 AUXILIARY MATERIALS

3.1 Dataset

The data is collected from DIP *Xenarios et al. (2001)* and PIR to validate the CAA-PPI approach. Evaluation is performed on *Yeast* and *H. Pylori*, all having different numbers of interacting protein pairs. *S. cerevisiae* which is basically *Yeast* protein, their PPI datasets are taken from DIP taking reference of An's work *An et al. (2019)* in the prediction of PPI. Replication of protein pair is done by scrutiny of a dataset having similarity lesser than 40% and 5594 datasets of interacting pairs is obtained. For effective testing of the performance of the model non-interacting pairs need to involve in the dataset to train the model. Consequently, 5594 datasets of non-interacting or negative pairs are selected consisting of diverse subcellular localizations. Finally total of 11,188 protein pair of *Yeast* dataset need to evaluate. *H. Pylori* are the next considered PPI dataset which is a collection of 2916 pairs of protein containing 1458 pairs of interacting and non-interacting each as Martin used in their work *Martin et al. (2005)*. Besides these, the PPI dataset of the following five species *M. musculus*, *H. Pylori*, *C. elegans*, *E. coli*, *H. sapien* are also used to test the performance of CAA-PPI, respectively comprising of 313, 1420, 4013, 6954, and 1412 number of interacting pairs also used in PPI related research work by *Zhou et al. (2011)*.

3.2 Cross Validation

A cross-validation system is a typical procedure for circumventing any cross section prejudices as well as corroborating the reliability of the model *Stone (1974)*. In this article, a five-fold cross-validation technique is performed for assessing the classifier's performance. In the X -fold cross-validation technique (X is any valid number), the complete dataset is randomly fragmented into X equal fragments say folds; out of X folds, $X-1$ are used for training and the lingering one is used as a test set in each fold of the cross-validation. Similarly, this practise is recurred X times to achieve X distinct models. Lastly, the outcomes of X distinct trials are averaged to contribute to an inclusive assessment.

3.3 Performance Evaluation

With the purpose of quantitatively appraise the efficacy and constancy of a classifier, a number of broadly performed statistical measures are considered in this articles, namely accuracy (A), sensitivity/recall (Se), specificity (Sp), positive predictive value/precision (Pr), NPV, Fscore (F_s) and MCC . These measures are expressed as follows:

$$A = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Se = \frac{TP}{TP + FN} \quad (2)$$

$$Sp = \frac{TN}{TN + FP} \quad (3)$$

$$Pr = \frac{TP}{TP + FP} \quad (4)$$

$$NPV = \frac{TN}{TN + FN} \quad (5)$$

$$F_s = 2 \times \frac{SN \times PPV}{SN + PPV} \quad (6)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \quad (7)$$

Here, TP is the measure of true PPI pairs that are predicted appropriately as interacting pairs. TN defines the quantity of true non-interacting pairs predicted properly. FP is the amount of true non-interacting pairs that are wrongly predicted as interacting ones. FN states the number of falsely predicted true interacting pairs as non-interacting pairs. Though A is a simple assessment measure, it may lead to a very biased evaluation in case of a discrepancy dataset. Pr confirms the total predicted pairs that are allied to the PPIs. As Pr and Se contradict each other, F_s is evaluated as the weighted harmonic mean of Pr and Se to inclusively reveal the prediction performance of PPI (*Hripcsak and Rothschild (2005)*). Higher the value of F_s , so is the Pr and Se . MCC is a different objective index to imitate the whole method performance that considers under prediction as well as over-predictions (*Matthews (1975)*).

3.4 Principal Component Analysis (PCA)

PCA (*Vipsita et al. 2013*) is an unsupervised linear dimensionality reduction method, which is used for the projection of a data space into smaller dimensional space by using orthogonal transformation. It is a widespread method for eliminating redundant and noise data, and to extract relevant features. The objective of PCA is to condense the big feature set to small without dropping suitable information of the original set. The process used for reduction using PCA is given below in six basic steps:

- Transform the entire dataset into matrix of dimension $i \times j$ and ignore the class label.
- Calculate the mean vector of the matrix.
- Calculate the covariance of entire magnitudes.
- Compute eigenvalues and analogous eigenvectors.
- Arrange eigenvectors by declining eigenvalues and select any p eigenvector with highest eigenvalues result in a matrix of dimension $j \times p$.
- The resultant $j \times p$ matrix is used to convert sample space to new subspace.

3.5 Random Forest Classifier

RF is a booming classifier in the area of machine learning. It is a procedure of an ensemble classification that appoints a set of DTs to diminish the resultant variance of distinct trees to develop the constancy and exactitude of classification. RF judiciously take advantage of two influential ML techniques:

- for each tree, the election of training samples;
- the random feature selection to fragment the data set.

The selection of training samples is implemented by using a bootstrap sample from original data (termed bagging). The outcomes of bagging lead to two dismember bags, one holding the training data of around 63.2% instances and the other one holding the remaining samples generally stand for out-of-bag (OOB) samples. Usually, in-bag samples are used to build the RF classifier and OOB samples are used for the assessment of prediction. The next powerful ML technique selects a features' subgroup at every single node in respective classification tree *i.e.*, RF handpicked a fixed amount of features randomly at every node of a tree and the one with the thoroughgoing decrease in Gini index (*Qi and Y 2012*) is selected for the split when emerging the tree.

Naturally, a forest is made up of trees and more trees mean a more robust forest. In the same way, the RF algorithm generates DTs on data samples and acquires the prediction from each of them and lastly chooses the finest result through voting. This ensemble scheme is superior to a solitary DT as this condenses the over-fitting by averaging the outcome.

The RF algorithm can be understood with the help of the following steps:

- Firstly, select random samples from the assumed dataset.
- Then, a DT will be generated by the algorithm for every sample and prediction outcome from each DT will be achieved.
- Then, for every predicted outcome, voting will be implemented.
- In the end, the final prediction result will be the maximum voted prediction result.

4 Working of CAA-PPI

The proposed approach is based on the fact that interaction possibility-related information exists in the sequence of protein pairs. This information can be generated by deriving different features from the sequence by applying a varied feature extraction approach. The count of generated features can't vary with change in length of protein sequence; hence feature is generated in respect of amino acid.

CAA-PPI can generate these features by using the ratio of combinations of trigram presence in the protein sequence and count of central amino acid present in the same trigram. Here trigram (3-mer) represents a set of 3 consecutive symbols in the protein sequence. Since a protein sequence contain 20 amino acid so a trigram contains 20^3 tri-peptide combinations, so total 8000 feature values need to generate for each combination of tri-peptide of a protein, hence a protein pair have about 16000 values in its feature sets. It will be challenging to work with such huge feature sets.

Encoding is proposed to reduce the count of a protein's features by aggregating 20 amino acids in 7 classes, which can reduce the number of features from 20^3 to 7^3 . Encoding enhances the speed of the proposed model, but still had the issue of the performance that is analyzed by taking multiple encoding schemes and select the best one. The proposed work is evaluating results based on the encoding scheme proposed by *Stupar (2010) (ES1)* and *Talwar (2015) (ES2)*. Both encoding schemes are aggregating 20 amino acids in 7 classes nonetheless combination benchmarks are dissimilar as shown in Figures 2 and 3. In *ES1*, amino acids are categorized based on chemical properties like dipole scale and volume scale; whereas *ES2*, encoding is based on the structure of the side chain influenced by the significance of the amino acid's side chain in deciding properties of amino acid.

Fig. 2: Classification of Amino Acids based on *ES1*.

Fig. 3: Classification of Amino Acids based on *ES2*.

The first step of CAA-PPI is encoding of protein sequence, this process converts amino acid characters into symbols ('1', '2', '3', '4', '5', '6', '7'), features values need to generate for 7^3 combinations of trigrams so simul-

taneously we need to generate 7^3 features. For which combinations will generate. Now for each combination feature value is generated, for example, if the sequence is 'AWGVWEGIAVGWAWG'.

Then for combination 'AWG', the feature value will be:

- Count of 'AWG' in given sequence/Count of W in given sequence = $2/4=0.5$

For instance, the input dataset has positive (interacting) and negative (non-interacting) protein pairs, so labeling is done with 1 and -1 receptively. Now PCA will apply to the resultant feature for the feature selection process, followed by a fivefold cross-validation process of the dataset as shown in figure 4. Five-fold process divide dataset into 5 equal parts from, which one part is used at the place of test dataset and remaining are used like train dataset, hence it represents 1-4 partition of data. The selected train dataset is used to train a model using the random forest classification method. Trained model predict the class of the test dataset, which support in calculating performance measure as A , Pr , Se , MCC , Fs , Sp , and NPV of results.

Fig. 4: Flowchart of CAA-PPI Approach.

The proposed feature generation algorithm is depicted in Algorithm 1-3, representing CAA-PPI. Since the main procedure of the model, taking protein sequence (seq), encoding pattern style ($encoding_pattern$), and numbers of elements in combination (p_gram) as arguments. The first step is encoding, perform by function $encoding_PPI$ using a sequence of protein and encoding pattern like an input argument and generate encoded protein sequence either according to $ES1$ or $ES2$ depend on the value of the encoding pattern. The next step

of the procedure is to generate features of CAA-PPI that is performed by function *Generate_Feature_CAA* by counting trigrams, and central amino acid in a given sequence and generate all features for all combinations of a given sequence and return value *CAA_featureset*. To filter out the correlated data from feature sets PCA function is used that return *CAA_filtered_featureset*, that is further labeled by *Add_label* function after testing that whether protein pairs are interacting or non-interacting.

```

Initialize()
Procedure CAA_PPI()
    Input : Protein sequence (seq), Size of combinational p-gram = 3
    seq_encoded = Encoding_PPI(seq, encoding_pattern)
    CAA_featureset = Generate_Feature_CAA(seq_encoded, p-gram)
    CAA_filtered_featureset = PCA(CAA_featureset)
    If y in CAA_filtered_featureset is Negative_PPI
        Add_label -1 to y
    else
        Add_label 1 to y
    end
end Procedure

```

Algorithm 1: Initialization of CAA-PPI

5 Results and Discussions

5.1 Performance of PPI Prediction

This section specifies the performance of the neighbourhood based feature representation approach to predict PPIs *via* two diverse PPI datasets with two encoding strategies discussed in previous sections. The outcomes are then compared with numerous existing approaches that were suggested in already published works. Subsequently, a blind test has implemented on five other species' datasets (*M. musculus*, *H. sapiens*, *C. elegans*, *H. Pylori* and *E. coli*) which are autonomous to the training set to prove the productivity of CAA-PPI.

Performance of CAA-PPI model using ES1 and ES2 on the Yeast Dataset. CAA representation of protein sequence having RF predictor is tested using five-fold cross-validation with the dataset of *Yeast* as shown in Table 2 using *ES1* and in Table 3 using *ES2*. It is noteworthy from Table 2 that great prediction accuracy of 98.25% is attained for the CAA-PPI approach with *ES1*. The values of the other six standard measures are also assessed for the proposed model to well discover the prediction capability, and achieve decent performance with both the encoding schemes but *ES1* performs comparatively better than *ES2*.

Furthermore, prediction model using CAA approach with *ES1* and *ES2* is compared against the methodologies proposed by various publishers as follows: *An et al. (2019)*, *You et al. (2014)*, *Wong et al. (2015)*, *Zhou et al. (2011)*, *Guo et al. (2008)* and presents the Bonferroni post-hoc analysis in Table 4. These comparing approaches has discussed in previous sections, they distinctly used LCPSSMMF, PSSMMF, LCPSSMAB, LCPSSMBG, AC+CT+LD+MAC, MCD, PR-LPQ, LD, ACC, AC for encoding amino acid sequence and predict PPI using SVM, RF, RoF, E-ELM classifier. It is notable from Table 4 that CAA-PPI using *ES1* and *ES2* beats all competitive methods *i.e.* it generally has a significant difference in prediction accuracy than these state-of-art PPI predictors for the *Yeast* dataset also depicted by Figure 5.

Performance of CAA-PPI model using ES1 and ES2 on the H. Pylori Dataset. Further, to evaluate the efficacy of the proposed method, the CAA-PPI model with *ES1* and *ES2* has been tested on the *H. Pylori* dataset using five-fold cross-validation which reports the result as shown in Table 5 and Table 5. It can be observed from Table 5 and 6 that the average accuracy of the proposed model is 98.25% with *ES1* and 98.69% with *ES2*. Moreover, the performance of CAA-PPI is computed comprehensively with other evaluation metrics including *Se*, *Sp*, *Pr*, *NPV*, *Fs*, *MCC* as shown in Table 5 and 6. Likewise, the performance of CAA-PPI with both *ES1* and *ES2* are compared with the approaches suggested in previous literary works and presents the Bonferroni post-hoc analysis in Table 7. These comparing approaches independently used HOG+SVD (*Ding et al. (2016)*), AC+CT+LD+MAC (*You et al. (2013)*), MCD (*You et al. (2014)*), DCT+SMR (*Huang et*

```

Function Seq_encoded = Encoding_PPI(seq, encoding_pattern)
  If encoding_pattern is ES1
    Aggregation of amino acid in 7 classes in seq_encoded
    {'I', 'L', 'F', 'P'} as 1, {'H', 'N', 'Q', 'W'} as 2, {'C'} as 7
    {'Y', 'M', 'T', 'S'} as 3, {'A', 'G', 'V'} as 4, {'D', 'E'} as 5
    {'R', 'K'} as 6
  else If encoding_pattern is ES2
    Aggregation of amino acid in 7 classes in seq_encoded
    {'A', 'G', 'V', 'I', 'L'} as 1, {'Q', 'N', 'T', 'S'} as 2,
    {'Y', 'F', 'W'} as 3, {'R', 'K', 'H'} as 4, {'D', 'E'} as 5,
    {'C', 'M'} as 6, {'P'} as 7
  end
end Function

```

Algorithm 2: Encoding of Given Protein Sequence

```

Function feature_set = Generate_Feature_CAA(seq_encoded, p-gram)
  Set list_code to {'1', '2', '3', '4', '5', '6', '7'}
  trigram_set = Generate_Combination(list_code, p-gram)
  forall x in trigram_set
    Set k to Central amino acid in x
    Set a to Count of x in seq_encoded
    Set b to Count of k in seq_encoded
    Add a ratio b in feature_list
  end for
  Add feature_list in feature_set
end Function

```

Algorithm 3: Generate feature CAA

Table 2: Result of Five-fold cross validation for proposed approach on *Yeast* dataset using *ES1*

Testing Set	A (%)	Se (%)	Sp (%)	Pr (%)	NPV (%)	Fs (%)	MCC (%)
1	98.54	100	96.97	97.26	100	98.61	97.11
2	98.54	100	97.22	97.01	100	98.48	97.12
3	99.27	100	98.44	98.65	100	99.32	98.54
4	98.54	100	97.30	96.92	100	98.44	97.11
5	96.38	98.59	94.03	94.59	98.44	96.55	92.83
Average	98.25	99.72	96.79	96.89	99.69	98.28	96.54

Table 3: Result of Five-fold cross validation for proposed approach on *Yeast* dataset using *ES2*

Testing Set	A (%)	Se (%)	Sp (%)	Pr (%)	NPV (%)	Fs (%)	MCC (%)
1	97.08	95.89	98.44	98.59	95.45	95.45	94.19
2	94.89	90.14	100	100	90.41	94.81	90.28
3	95.62	90.77	100	100	92.31	95.16	91.54
4	98.54	100	97.47	96.67	100	98.31	97.07
5	91.30	86.84	96.77	97.06	85.71	91.67	83.19
Average	95.49	92.73	98.54	98.46	92.78	95.43	91.25

al. 2015), LD (Zhou *et al.* (2011)), Phylogenetic bootstrap (Bock and Gough (2003)), HKNN Nanni (2005), Ensemble of HKNN Nanni and Lumini (2006), Signature products (Martin *et al.* (2005)), Boosting (Liu *et al.* (2013)) to express amino acid sequence and use favourable classifier to predict PPIs. From Table 7, it is worth noting that CAA-PPI is more effective than other competitive methods with both *ES1* and *ES2* as well i.e. it generally has a significant difference than these state-of-art PPI predictors for the *H. Pylori* dataset also represented by Figure 6.

Outcomes on five species datasets To assess the hands-on prediction aptitude, initially, CAA-PPI is trained with PPIs of *Yeast* dataset using *ES1* and *ES2* separately and used five independent species' datasets to test,

Table 4: Bonferroni post-hoc analysis result of CAA-PPI using *ES2* and *ES2* compared with existing approaches for Overall prediction accuracy for *Yeast* dataset.

Approaches	A	B	C	D	E	F	G	H	I	J	K	L	M
A	-	○	○	●	●	●	●	●	○	●	●	●	●
B		-	○	○	●	●	●	○	○	●	○	●	●
C			-	○	○	○	○	○	○	○	○	○	●
D				-	○	○	○	○	○	○	○	○	○
E					-	○	○	○	●	○	○	○	○
F						-	○	○	○	○	○	○	○
G							-	○	●	○	○	○	○
H								-	○	○	○	○	○
I									-	○	○	●	●
J										-	○	○	○
K											-	○	○
L												-	○
M													-

A: CAA_ES1+RF, B: CAA_ES2+RF, C: LCPSSMMF + SVM, D: PSSMMF+SVM, E: LCPSSMAB+SVM, F: LCPSSMBG+SVM, G: (AC+CT+LD+MAC) + E-ELM, H: MCD+SVM, I: PR-LPQ+ROF, J: LD +SVM, K: ACC+SVM, L: AC +SVM, M: LD +KNN, ●: Significant Difference, ○: Non-significant Difference

Fig. 5: Comparison of CAA-PPI applied on *Yeast* dataset using *ES1* and *ES2* with existing approaches (Ref: Table 4).

Table 5: Five-fold cross validation result of CAA-PPI using *ES1* on *H. Pylori* dataset.

Testing Set	A (%)	Se (%)	Sp (%)	Pr (%)	NPV (%)	Fs (%)	MCC (%)
1	94.89	100	90.41	90.14	100	94.81	90.28
2	98.54	100	96.97	97.26	100	98.61	97.11
3	100	100	100	100	100	100	100
4	98.54	100	96.88	97.33	100	98.65	97.10
5	99.28	100	98.61	98.51	100	99.25	96.56
Average	98.25	100	96.57	96.65	100	98.26	96.61

consisting of 6954 interacting pairs of *E. coli*, 4013 interacting pairs of *C. elegans*, 1420 interacting pairs of *H. Pylori*, 1412 interacting pairs of *H. sapiens* and *M. musculus* dataset of 313 interacting pairs. The postulation behind this procedure is that ortholog protein have alike functional natures and so is their interacting nature (*Shi et al. (2010)*). Henceforth, in this section, above stated and experimentally demonstrated the interaction of any one species *Yeast* dataset (with 11,188 samples in this case) is employed to predict the interactions of other ones. Then, a blind test has implemented on five other species' datasets which are autonomous to

Table 6: Five-fold cross validation result of CAA-PPI using *ES2* on *H. Pylori* dataset.

Testing Set	A (%)	Se (%)	Sp (%)	Pr (%)	NPV (%)	Fs (%)	MCC (%)
1	98.54	100	97.14	97.10	100	98.53	97.12
2	99.27	98.51	100	100	98.59	99.25	98.55
3	99.27	98.59	100	100	98.51	99.29	98.55
4	96.35	98.63	93.75	94.74	98.36	96.64	92.74
5	100	100	100	100	100	100	100
Average	98.69	99.15	98.18	98.37	99.09	98.74	97.39

Table 7: Bonferroni post-hoc analysis result of CAA-PPI using *ES1* and *ES2* compared with existing approaches for Overall prediction accuracy for *H. Pylori* dataset.

Approaches	A	B	C	D	E	F	G	H	I	J	K	L
A	-	○	●	●	●	●	●	●	●	●	●	●
B		-	●	●	●	●	●	●	●	●	●	●
C			-	○	○	○	○	●	○	○	○	●
D				-	○	○	○	●	○	○	○	●
E					-	○	○	●	○	○	○	○
F						-	○	●	○	○	○	○
G							-	●	○	○	○	○
H								-	●	●	●	○
I									-	○	○	○
J										-	○	○
K											-	○
L												-

A: CAA_ES1, B: CAA_ES2, C: HOG+SVD, D: AC+CT+LD+MAC, E: MCD, F: DCT+SMR, G: LD, H: Phylogenetic bootstrap, I: HKNN, J: Signature products, K: Ensemble of HKNN, L: Boosting, ●: Significant Difference, ○: Non-significant Difference

Fig. 6: Comparison of CAA-PPI applied on *H. Pylori* dataset using *ES1* and *ES2* with existing approaches (Ref: Table 7).

the training set using the same proposed approach. The resultant performances are shown in Table 8. The outstanding result of the proposed method using both *ES1* and *ES2* ascertains the significant proficiency ($p < 0.05$) of the CAA-PPI compared to the existing published works from Table 8.

Table 8: Performance of PPI Prediction on five species datasets taking Yeast dataset for training (in terms of Accuracy).

Species	Testing Pairs	CAA_ES1 (%)	CAA_ES2 (%)	HOG+SVD (<i>Ding et al. 2016</i>) (%)	MLD (<i>c You et al. 2015</i>) (%)	DCT+SMR (<i>Huang et al. 2015</i>) (%)	LD (<i>a Zhou et al. 2011</i>) (%)
<i>C.elegans</i>	4013	95.94	96.01	90.28	87.71	81.19	75.73
<i>E.coli</i>	6954	93.36	93.72	93.18	89.30	66.08	71.24
<i>H.sapiens</i>	1412	95.89	96.90	94.58	94.19	82.22	76.27
<i>H.Pylori</i>	1420	93.59	92.32	92.03	90.09	82.18	N/A
<i>M.musculus</i>	313	97.76	97.12	92.25	91.96	79.87	76.68

6 Conclusion and Future Scope

With the growing number of PPI calculation methods, the codification of numerous amino acid feature vectors are also evolving. Even though considerable advancement has been achieved by now, further operational approaches are required to deal with the precincts. This research presents a ML based model (CAA-PPI) to predict PPI using two distinct encoding strategies. Major contribution of the given model is novel feature generation method using association of different amino acid with a residue in a given trigram. CAA based feature extraction approach is implemented with different encoding scheme trailed by random forest classifier, to train the model. Proposed CAA-PPI with RF classifier model’s performance is then verified with two diverse PPI datasets that are *Yeast* and *H. Pylori* and attain a favourable outcomes with both the encoding scheme. Additionally, it is worth noting that both encoding strategies are equally effective with CAA-PPI approach to better predict new protein interactions.

It is said that there is always room for improvement or change, the only challenge is to discover the same. The next step could be to discover more interacting protein pairs and to generate a new set of features using the proposed approach. Moreover, CAA-PPI can be assessed by using other encoding strategies and apply them to other organisms. Likewise, a new encoding scheme can be developed with the systematic categorization of amino acids. The proposed approach can also be extended in the direction of the interaction of the protein with other possible molecules.

Acknowledgements The authors would like to express deep gratitude to the editors and referees for their favourable suggestions which led to a finer presentation of this paper.

Compliance with Ethical Standards

Conflict of interest

The authors declare that they have no conflict of interest.

Ethical approval

This article work is tested on the already available data in research community.

Authorship Contributions

All the authors contributed equally to this research paper.

References

- Essam Al-Daoud. Improving protein-protein interaction prediction by using encoding strategies and random indices. *interaction*, 1:2, 2011.
- Ji-Yong An, Yong Zhou, Yu-Jun Zhao, and Zi-Ji Yan. An efficient feature extraction technique based on local coding pssm and multifeatures fusion for predicting protein-protein interactions. *Evolutionary Bioinformatics*, 15:1176934319879920, 2019.

- Joel R Bock and David A Gough. Whole-proteome interaction mining. *Bioinformatics*, 19(1):125–134, 2003.
- Yijie Ding, Jijun Tang, and Fei Guo. Identification of protein–protein interactions via a novel matrix-based sequence representation model with amino acid contact information. *International journal of molecular sciences*, 17(10):1623, 2016.
- Hesham ElAbd, Yana Bromberg, Adrienne Hoarfrost, Tobias Lenz, Andre Franke, and Mareike Wendorff. Amino acid encoding for deep learning applications. *BMC bioinformatics*, 21(1):1–14, 2020.
- Erica Golemis and Peter David Adams. *Protein-protein interactions: a molecular cloning manual*. Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY, 2002.
- Yanzhi Guo, Lezheng Yu, Zhining Wen, and Menglong Li. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic acids research*, 36(9):3025–3030, 2008.
- George Hripcsak and Adam S Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298, 2005.
- Ozlem Keskin, Nurcan Tuncbag, and Attila Gursoy. Predicting protein–protein interactions from the molecular to the proteome level. *Chemical reviews*, 116(8):4884–4909, 2016.
- Chou Kuo-Chen. Pseudo amino acid composition and its applications in bioinformatics. *Proteomics and System Biology, Curr. Proteomics*, 6:262–274, 2009.
- NQ Le and BP Nguyen. Prediction of fmn binding sites in electron transport chains based on 2-d cnn and pssm profiles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.
- Bin Liu, Jimmy Yi, SV Aishwarya, Xun Lan, Yilin Ma, Tim HM Huang, Gustavo Leone, and Victor X Jin. Qchipat: a quantitative method to identify distinct binding patterns for two biological chip-seq samples in different experimental conditions. *BMC genomics*, 14(8):1–11, 2013.
- Mina Maleki, Gokul Vasudev, and Luis Rueda. The role of electrostatic energy in prediction of obligate protein–protein interactions. *Proteome science*, 11(1):1–12, 2013.
- Shawn Martin, Diana Roe, and Jean-Loup Faulon. Predicting protein–protein interactions using signature products. *Bioinformatics*, 21(2):218–226, 2005.
- Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.
- Loris Nanni. Hyperplanes for predicting protein–protein interactions. *Neurocomputing*, 69(1-3):257–263, 2005.
- Loris Nanni and Alessandra Lumini. An ensemble of k-local hyperplanes for predicting protein–protein interactions. *Bioinformatics*, 22(10):1207–1210, 2006.
- DaRue A Prieto, Donald J Johann Jr, Bih-Rong Wei, Xiaoying Ye, King C Chan, Dwight V Nissley, R Mark Simpson, Deborah E Citrin, Crystal L Mackall, W Marston Linehan, et al. Mass spectrometry in cancer biomarker research: a case for immunodepletion of abundant blood-derived proteins from clinical tissue specimens. *Biomarkers in medicine*, 8(2):269–286, 2014.
- Sneha Rai and Sonika Bhatnagar. Computational methods for prediction of protein-protein interactions: Ppi prediction methods. In *Materials Science and Engineering: Concepts, Methodologies, Tools, and Applications*, pages 309–340. IGI Global, 2017.
- Peter J Reeds. Dispensable and indispensable amino acids for humans. *The Journal of Nutrition*, 130(7):1835S–1840S, 2000.
- Ming-Guang Shi, Jun-Feng Xia, Xue-Ling Li, and De-Shuang Huang. Predicting protein–protein interactions from sequence using correlation coefficient and high-quality interaction dataset. *Amino acids*, 38(3):891–899, 2010.
- Mervyn Stone. Cross-validation and multinomial prediction. *Biometrika*, 61(3):509–515, 1974.
- Robert M Stupar. Into the wild: The soybean genome meets its undomesticated relative. *Proceedings of the National Academy of Sciences*, 107(51):21947–21948, 2010.
- GP Talwar. *Textbook of biochemistry, biotechnology, allied and molecular medicine*. PHI Learning Pvt. Ltd., 2015.
- Leon Wong, Zhu-Hong You, Shuai Li, Yu-An Huang, and Gang Liu. Detection of protein-protein interactions from amino acid sequences using a rotation forest model with a novel pr-lpq descriptor. In *International Conference on Intelligent Computing*, pages 713–720. Springer, 2015.
- Ioannis Xenarios, Esteban Fernandez, Lukasz Salwinski, Xiaoqun Joyce Duan, Michael J Thompson, Edward M Marcotte, and David Eisenberg. Dip: the database of interacting proteins: 2001 update. *Nucleic acids research*, 29(1):239–241, 2001.
- Lei Yang, Jun-Feng Xia, and Jie Gui. Prediction of protein-protein interactions from protein sequence using local descriptors. *Protein and Peptide Letters*, 17(9):1085–1090, 2010.
- Zhu-Hong You, Ying-Ke Lei, Lin Zhu, Junfeng Xia, and Bing Wang. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. In

- BMC bioinformatics*, volume 14, pages 1–11. Springer, 2013.
- Zhu-Hong You, Lin Zhu, Chun-Hou Zheng, Hong-Jie Yu, Su-Ping Deng, and Zhen Ji. Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. In *BMC bioinformatics*, volume 15, pages 1–9. Springer, 2014.
- Zhu-Hong You, Keith CC Chan, and Pengwei Hu. Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. *PloS one*, 10(5):e0125811, 2015.
- Chang Zhou, Hua Yu, Yijie Ding, Fei Guo, and Xiu-Jun Gong. Multi-scale encoding of amino acid sequences for predicting protein interactions using gradient boosting decision tree. *PLoS One*, 12(8):e0181426, 2017.
- Yu Zhen Zhou, Yun Gao, and Ying Ying Zheng. Prediction of protein-protein interactions using local description of amino acid sequence. In *Advances in computer science and education applications*, pages 254–262. Springer, 2011.