CoRNeA: A pipeline to decrypt the inter protein interfaces from amino acid sequence information

Kriti Chopra[1], Bhawna Burdak[1], Kaushal Sharma[2], Ajit Kembavi[2], Shekhar C. Mande[3], and Radha Chauhan[1*]

1- National Centre for Cell Science, Pune, Maharashtra, India.

2- Inter University Centre for Astronomy and Astrophysics, Pune, Maharashtra, India

3- Council of Scientific and Industrial Research (CSIR), New Delhi, India

***Corresponding Author**:

Dr. Radha Chauhan, Scientist 'E', National Centre for Cell Science, S.P. Pune University Campus, Ganeshkhind, Pune 411007, Maharashtra, India.

Email: radha.chauhan@nccs.res.in

Phone: +91-20-25708255

Abstract

Computational methods have been devised in the past to predict the interface residues using amino acid sequence information but have been majorly applied to predict for prokaryotic protein complexes. Since the composition and rate of evolution of the primary sequence are different between prokaryotes and eukaryotes, it is important to develop a method specifically for eukaryotic complexes. Here we report a new hybrid pipeline for the prediction of protein-protein interaction interfaces from the amino acid sequence information alone based on the framework of Co-evolution, machine learning (Random forest) and Network Analysis named CoRNeA trained specifically on eukaryotic protein complexes. We incorporate the intra contact information of the individual proteins to eliminate false positives from the predictions as the amino acid sequence also holds information for its own folding along with the interface propensities. Our prediction on various case studies shows that CoRNeA can successfully identify minimal interacting regions of two partner proteins with higher precision and recall.

Introduction

The biological machinery performs its cellular functions when its basic units such as DNA, RNA, and proteins interact with each other. To understand the overall functioning of the cell, it is important to delineate the pairwise interactions of these basic units such as DNA-protein, RNA-protein, and protein-protein. Of these, the inter protein interactions that a cell possesses play a very crucial role in understanding the various cellular processes and hence also their functioning or misfunctioning in the disease models. There are various experimental methods known for examining these interactions such as yeast two hybrid (Y2H)[1], co-immunoprecipitation (co-IP)[2], mass spectrometry [3], etc. which provide information only about the domains necessary for maintaining the interaction or the proximity of the interactions. Moreover, these methods are labor, cost and time intensive. Deciphering the PPII (Protein-Protein Interaction Interfaces) at the highest resolution through x-ray crystallography or cryo-electron microscopy methods is even more challenging due to their intrinsic technical difficulties.

A number of *in-silico* methods have been described earlier to predict these PPII based on available data such as 1) homology 2) machine learning and 3) co-evolution based. Homology based methods are generally applied when confident homologs of both the interacting proteins are available, followed by protein-protein docking for visualizing the protein interaction interfaces such as PredUS [4], PS-HomPPI [5], PriSE [6], etc. The machine learning (ML) methods which have been described till date are either structure-based or sequence-based. The structure-based ML methods (such as SPPIDER[7], PINUP[8], PAIRpred[9], PIER[10], ProMate[11], Cons-PPISP[12], Meta-PPISP[13], CPort[14], WHISCY[15], InterProSurf[16], VORFFIP[17], eFindSite[18], etc.) require three-dimensional information of the interacting proteins which can be either experimental or homology driven to incorporate the geometrical complementarities of amino acids as training features. Only a few sequence-based ML methods are known such as BIPSPI[19], PSIVER [20], and ComplexContact [21] which derive features based on conservation, physicochemical properties of amino acids, etc. However, the predictability of these ML methods is affected by the prevalence of high false-positive rates due to limitation of small number of protein-complex structures in the protein structure database (PDB) which restrict the training of these machine learning algorithms in terms of variability.

The third class, co-evolution-based methods which were originally formulated to predict contact forming residues within a single protein and therefore for the prediction of the structure of the protein. These methods have been extrapolated to also predict the inter-protein interaction interfaces based on the multiple sequence alignments (MSA) of the proteins. Concatenating the MSA of an interacting pair and using the same statistical formulae as described for intra pairs have been implemented to predict the co-evolving contact forming pairs by various methods such as DCA[22], EvComplex[23], etc. However, there are two main caveats known for these methods. Firstly, they use different downstream methods to filter out their results by using homology-based models and docking predictions in combination with their results. Secondly, most of these methods have been tested on prokaryotic proteins and have a limitation of predicting only for a maximum combined length of 1500 residues per protein pair. Almost all co-evolution-based methods have been only tested on prokaryotic lineage probably due to availability of huge number of sequences for generating variable multiple sequence alignments. Recently a hybrid method (co-evolution and machine learning based- ComplexContact [21]) was reported, however, its performance was also the tested on prokaryotic datasets. Overall these methods could not perform with similar accuracy when applied to eukaryotic complexes.

The low predictability of these methods for eukaryotic protein complexes can be attributed to the differences in the rate of evolution of the proteins in the two lineages. It has been reported that there is a difference in the composition of the type of amino acids present in prokaryotic versus eukaryotic proteins and also in the radius of gyration and planarity in the interaction interface. Since the eukaryotic proteins are not exclusive to only one set of function, it has been perceived that most of the eukaryotic protein interactions are transient, having smaller interaction hotspot zones and have more planar binding sites consisting of more polar and aromatic residues. These properties of the eukaryotic protein interactions make them essential part of cell signaling pathways [24].

Hence to delineate the vast PPII network of eukaryote lineage, e.g. human protein interaction network, which contains about 1,50,000 interactions (with only about 10% of known structures of these protein complexes)[25], it is important to develop a method specific for eukaryotic predictions. In this report, we present a new hybrid pipeline based on the framework of Co-evolution, Random forest (ML method) and Network Analysis (CoRNeA) for predicting the pairwise residues of the PPII from the protein sequence information of two interacting proteins (Figure 1). We also developed a new hybrid method for calculating co-evolving positions in the interacting pairs based on mutual information and Statistical Coupling Analysis (SCA)[26]. Owing to high signal to noise ratio, this method in consensus with the other co-evolution-based method does not perform well independently to extract the precise interacting pair of residues especially for eukaryotic proteins. Hence, we used this method as one of the features for machine learning pipeline. The other features derived for the random forest classifier are based on the physicochemical properties of the amino acids which depend on their side chain structure such as charge, size and hydrophobe compatibility, secondary structure information and relative solvent accessibility, were also derived using amino acid sequence information. To include the energetics of interactions, contact potentials were also included as features. Similar to other machine learning classifiers, our pipeline also predicted a number of false positives. In order to reduce them we employed network analysis by incorporating the intra contact information to generate residual networks for PPII. In summary, the major highlight of this method as compared to other methods developed on the similar lines are 1) use of eukaryotic protein structure database for training the classifier. 2) use of co-evolution information as conservation-based feature. 3) use of intra contact pairs to eliminate false positive pairs through network analysis. Thus, we present a holistic approach to this complex problem of identifying pair of residues forming the interaction interface in the heterodimers from the amino acid sequence information.

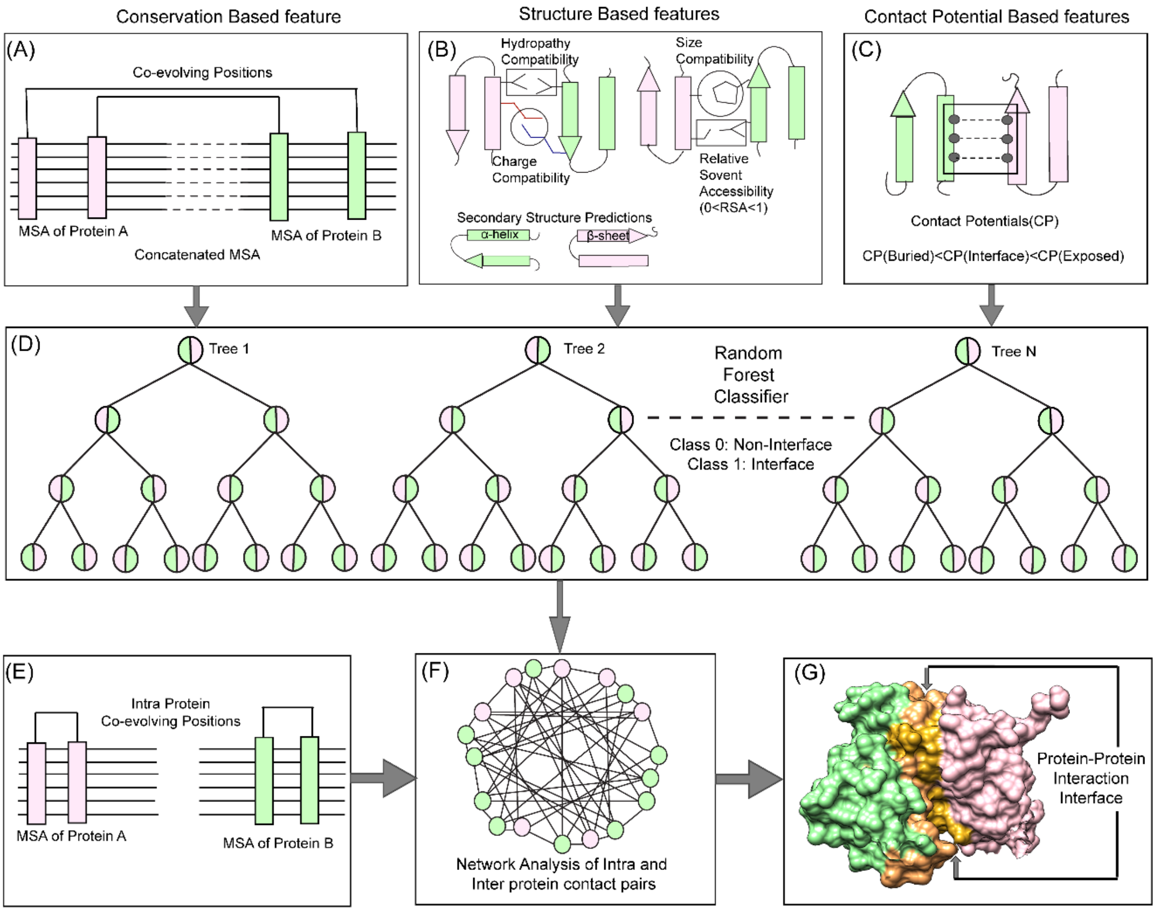

Figure 1: CoRNeA pipeline for predicting co-evolving contact forming residues in an interacting pair of proteins. The method for predicting the protein-protein interaction interface consists of three levels. The top panel depicts the features used for machine learning pipeline. (A). Conservation based (co-evolution) (B) Structure-based (Charge, Size, Hydropathy, Secondary structure, and Relative solvent accessibility) and (C) contact potential- based features (both for buried and exposed residues). (D) Random forest classification where pairwise values for both proteins are considered depicted in half green and pink circles for binary classification (Class 1: protein interface, Class 0: non-interface). The bottom panel depicts the application of network analysis by combining intra and inter protein contact predictions for reducing the false positives. (E) Prediction of intra contacts of Protein A and B. (F) Combined network analysis of inter and intra predicted contacts. (G) Interface prediction for PDB ID: 1H9D.

2. Methodology

The overall pipeline to predict pairwise contact forming residues from sequence derived data can be divided into three distinct parts as depicted in Figure 1. The first step is to generate pairwise features (conservation, structural and contact potential based) from the amino acid sequence of the two interacting proteins (Figure 1(A)-(C)). The second step is to feed these pairwise features in a random forest classifier and hence optimize its various hyperparameters to obtain the best evaluation statistics (Figure 1(D)). The third step is to combine the intra protein contact forming residues from co-evolution-based method and inter-protein contact forming residues from random forest classifier and perform network analysis to predict the exclusive pair of residues forming the interface of the two interacting proteins (Figure 1(E)-(G)).

2.1 Datasets

The Affinity Database version 2.0[27] was used to select the protein complex structures for training (42 complexes were selected for training). The amino acid sequences of the complex structures were extracted from www.rcsb.org and used as a query to search for homologs. PHMMER[28] was used to fetch maximum homologs of the query sequence which were then manually curated to remove redundant sequences. The sequences having less than 25% sequence identity were removed. The final dataset for each of the interacting protein consisted of identical species.

2.2 Multiple Sequence Alignments

The datasets for each interacting pair of proteins having identical species were subjected to structure-guided multiple sequence alignments using PROMALS3D[29]. The alignments were then analyzed/edited in JalView[30] and then concatenated (Last residue of Protein A followed by first residue of Protein B) in R using package seqinr[31]. These concatenated MSA datasets were used for co-evolution matrix calculations.

2.3 Features

For calculating sequence-based features, the sequences were extracted from the protein databank (www.rcsb.org) and any missing regions reported in the structure were removed from the sequence data. All the features for training and testing were compiled as all versus all residue pairs between sequence of the interacting pair of protein (Protein A and Protein B) in form of M*N matrix (M=length of Protein A and N= length of Protein B). All the feature values were scaled between 0 and 1. (Figure S1)

2.3.1 Evolution based features

Co-evolution matrices (CMI)

The co-evolution scores between the pair of residues of the interacting proteins were calculated based on Conditional Mutual Information as depicted in Figure 2. The concatenated MSA's were subjected to perturbation experiment similar to that used in Statistical Coupling Analysis (SCA)[26]. The amino acids were converted from alphabetic nomenclature to numeric for the ease of calculation (table S1). For each column in the MSA of Protein A and B, a condition pertaining to the presence of one of the 20 amino acid was given to subset the concatenated MSA. For example, position 1 in concatenated MSA, a condition given to subset the MSA for the presence of valine (V). A subset of sequences was selected which had only valine at position 1 of MSA. Frequencies of the amino acid present in the subset were calculated and subjected to the conditional mutual information formula[32]. It resulted in 20 such conditions for each column in the MSA of Protein A which were summed up to obtain the final co-evolution M*N matrix.

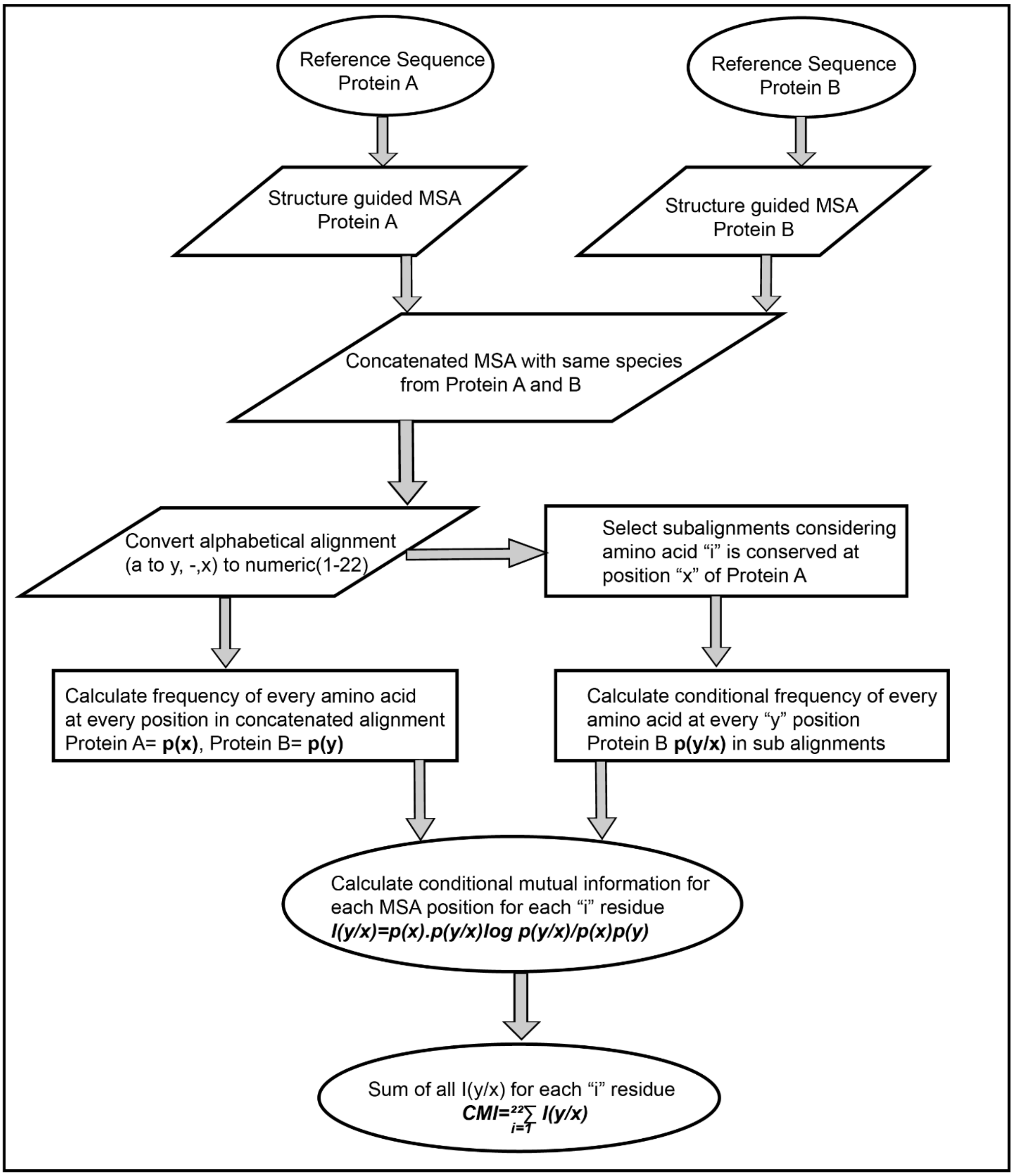
Reference Sequence Protein A
Reference Sequence Protein B
Structure guided MSA Protein A
Structure guided MSA Protein B
Concatenated MSA with same species from Protein A and B
Convert alphabetical alignment (a to y, -,x) to numeric(1-22)
Select subalignments considering amino acid "i" is conserved at position "x" of Protein A
Calculate frequency of every amino acid at every position in concatenated alignment Protein A= **p(x)**, Protein B= **p(y)**
Calculate conditional frequency of every amino acid at every "y" position Protein B **p(y/x)** in sub alignments
Calculate conditional mutual information for each MSA position for each "i" residue
$I(y/x)=p(x).p(y/x)\log p(y/x)/p(x)p(y)$
Sum of all I(y/x) for each "i" residue
$CMI=\sum_{i=1}^{22} I(y/x)$

Figure 2: Flow chart representing an algorithm for calculating inter protein co-evolving positions from multiple sequence alignments.

2.3.2 Structure based features

Charge, Hydrophobe and size compatibility matrices

The physicochemical properties of the residue determined by the composition and chemical structure were used to derive the structure-based features. These features can be derived from sequence information but to derive pair wise values for these properties, we employed the 20X20 residue matrices which were described to aid in *ab initio* modeling of single protein[33]. These matrices were used to derive an all versus all residue matrix (M*N) for the interacting pair of proteins as features i.e. hydropathy compatibility (HCM), charge compatibility (CCM) and size compatibility matrices (SCM)

Relative Solvent Accessibility (RSA)

To calculate the pairwise RSA values, RSA of independent proteins were calculated using SPIDER3[34] and multiplied to form an all versus all (M*N) matrix of the pair of interacting proteins.

Secondary Structure Predictions (SSP)

The secondary structure of the proteins was predicted using PSIPRED[35] and all residues were assigned numbers (i.e. 1= α-helix, 2=β-sheet and 3=l-loop). Simple multiplication and scaling of these numbers between 0 and 1 would yield in a combination where α-helix to α-helix instance will be ranked lowest. To avoid this mis scaling, the training dataset was inspected for the nature of residue-residue combinations in terms of secondary structures and the 6 possible combinations (i.e. α-α, α-β, α-l, β-β, β-l and l-l) were ranked in order of occurrence. These values were then used as standard to fill in all M*N matrices of the two interacting proteins.

2.3.3. Contact Potential based features

Three different approximations of contact potentials were used to generate contact potential-based features. The first approximation was the original matrix (MJ matrix) [36] where the effective inter-residue contact energies for all amino acid pairs were calculated based on the statistical analysis of protein structures. The other two approximations were derived from the MJ matrix, where a 2-body correction was applied on this matrix to generate two separate matrices [37]. One of them was specific for capturing the interactions between exposed residues and the other one for buried residues. Thus, all three possible combinations were used to derive three contact potential (M*N) matrices namely, **CP**: original MJ matrix, **CPE**: MJ matrix derived for exposed residues and **CPB**: MJ matric derived for buried residues, for the pair of interacting proteins.

2.4. Environment features

To include residue environment information for training the machine learning algorithm, a kernel matrix of size 5*5 was defined and convolved over the nine feature matrices as described above. The convoluted features were generated by using OpenImageR (https://github.com/mlampros/OpenImageR) package in R and the size of the matrices were kept same to avoid any loss of information. Additionally, various other kernel matrices were also used to train and test different datasets varying from 3*3 to 7*7 with varying percentage decrease in the weights from 10% to 25%. Hence, for each independent training/testing cycle, 18 feature matrices were used for each pair of interacting protein for training the random forest classifier (9 original features and 9 derived features).

2.5 Interface residue labeling

The interface residues for the protein complexes were extracted using PISA[38]. The number of residue pairs present in the interface (500 pairs for 42 complexes) was far less than all possible residue pairs of the two interacting proteins (20,00,000 for 42 complexes). To increase the search space and take into consideration the environment of the contact forming residues, a distance cut off of 10Å was used to search for possible pair of residues flanking -2 to +2 positions of the interface residues extracted from PISA. This yielded ten times more positive labels (5000 pairs for 42 complexes) for training the classifier.

2.6 Data Imbalance Problem

Although increasing the search space as explained above yielded 10 times more data points, still the complete protein complex database exhibited highly imbalanced data. 5000 pairs were labeled as positive out of the total 20,00,000 pairs. In order to address this imbalance class problem, the majority class, which was the negative data labels (non-interface residues pairs) was down sampled. A number of ratios for negative to positive samples were tested iteratively (e.g. 2:1, 5:1, 10:1 and 20:1) and best evaluation statistics were obtained when the negative sample size was five times that of positive samples (5:1). This was used as training set for the supervised classification model.

2.7 Random Forest Classifier

The random forest classifier[39] was trained first using a grid search to optimize the hyperparameters for the model yielding the best evaluation statistics through cross-validation. The hyperparameters obtained from the grid search were then used to train the classifier with training to test sample split to 75:25. The scoring function used for optimizing the hyperparameters was chosen as F1 score owing to imbalanced nature of the dataset used for training. Scikit-learn[40] was used to import the random forest classifier base algorithm. Training was performed on the same data sets both with and without environment features. All the data sets were compiled using R and Rstudio(*http://www.rstudio.com/*) and machine learning was performed using python3.7 via anaconda-navigator (https://anaconda.com).

2.8 Network Analysis

To reduce the number of false positives obtained from the random forest classifier, a holistic approach was adopted as described in Figure 3 to include the intra protein predictions. To determine the intra contacts, we used the co-evolution method as described in 2.3.1 by concatenating Protein A with itself (similarly for Protein B) (Figure 3(B)). To determine the contact forming intra-protein residue pairs, the residues present at a sequential distance less than 5 residues were eliminated and only top 5% of the coevolution values were taken as positive. The residue pairs obtained from this analysis for both proteins were used to plot the intra-protein residue networks in Rstudio using igraph package[41].

The predictions from the random forest classifier were used to plot the inter-protein residue network as a bipartite graph using the igraph package in Rstudio. Since the RSA for residues present in the core of the protein should be 0, these residues were extracted from SPIDER3[34] for both the proteins independently. A residual network was hence computed for the inter-protein contact predictions by first eliminating the nodes representing RSA=0 and then the intra-protein contacts from Protein A and B (Fsigure 3(C) and 3(D)).

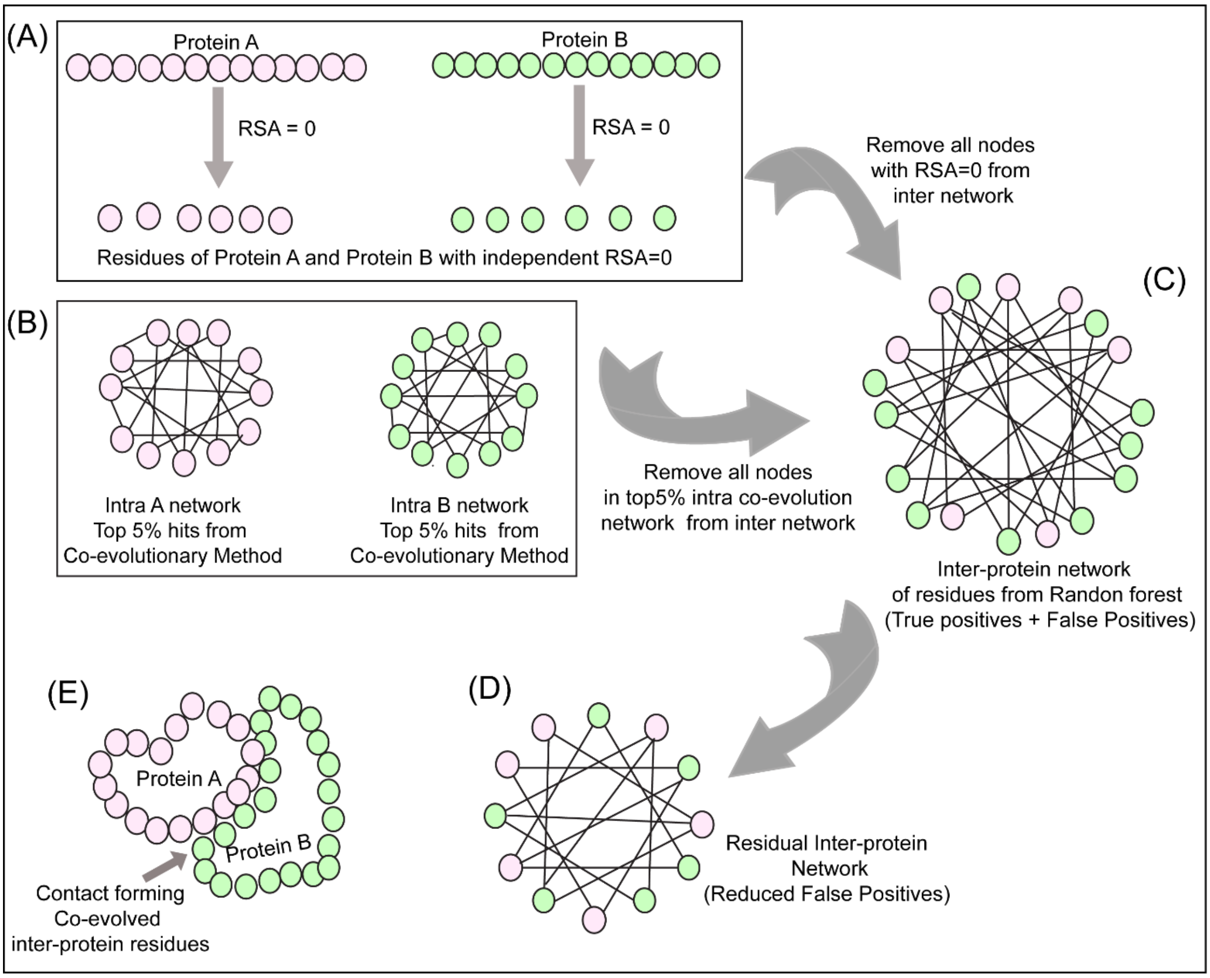
(A)
Protein A
Protein B
RSA = 0
RSA = 0
Residues of Protein A and Protein B with independent RSA=0
Remove all nodes with RSA=0 from inter network
(B)
Intra A network Top 5% hits from Co-evolutionary Method
Intra B network Top 5% hits from Co-evolutionary Method
Remove all nodes in top5% intra co-evolution network from inter network
(C)
Inter-protein network of residues from Randon forest (True positives + False Positives)
(E)
Protein A
Protein B
Contact forming Co-evolved inter-protein residues
(D)
Residual Inter-protein Network (Reduced False Positives)

Figure 3: Network analysis of intra and inter protein contacts. (A) Extraction of residues with RSA=0 for Protein A and B. (B) Intra contact prediction for Protein A and B (top 5% co-evolving residue pairs). (C) Predicted inter protein network from random forest classifier. (D) The false-positive inter protein residue pairs obtained from the random forest classifier are reduced by removing nodes having RSA=0 for Protein A and B as well as top 5% co-evolving intra protein residues of Protein A and B. (E) Analysis of the inter-contact from residual network onto the structure of Protein A and B.

2.9 Scoring of positive pairs using convolution feature matrix

The residual inter-protein network obtained were then plotted as a binary matrix of Protein A versus Protein B where 0 represented predicted non interface pairs and 1 represented predicted interface pairs. To identify the most probable interaction interfaces, cluster of 1's was identified by convolving a unitary matrix of size equal to that of kernel matrix used for deriving environmental features (i.e. 3*3 or 5*5) over the prediction matrix. Sub sections having the maximum number of 1's hence obtained the highest score (score of 9 for 3*3 matrix and 25 for 5*5 matrix). A cut off value of 2 for 3*3 matrix and 6 for 5*5 matrix was selected to sort the high scoring pairs considering that at least 25% of the 3*3 or 5*5 subsections of the prediction matrix are populated with 1's. These high scoring pairs were then extracted and mapped onto the test dataset structures to identify the true positives such that they also occur in the group of 3 residues at a stretch in both the proteins.

2.10 Immunoprecipitation for validating interface residues

Human Nup93 (KIAA0095) fragments (full length (1-819), 1-150, 1-82, 96-150) were cloned in pEGFP-C1 expression vector (Clontech) fused with GFP at N-terminus. HEK293F cells (Invitrogen) cultured in freestyle media (Gibco) in a humidified incubator maintained with 8% CO_2, 37°C at 110 rpm, were transfected with plasmid DNA using Polyethylenimine (Polysciences). Cells were harvested after 60 hours and lysed with lysis buffer (1X DPBS (Gibco), 0.2% tween 20, protease inhibitor cocktail, 1mM PMSF) by incubating the cells on ice for 30 minutes followed sonication and centrifugation. 1 mg of supernatant was incubated with glutathione beads (Pierce) pre-bound with GST tagged Anti-GFP nanobody (Addgene ID # 61838)[42] for 4 hours and 5% lysate was taken as input. The beads were then washed with lysis buffer thrice and the pulled fractions were eluted by incubating with elution buffer (1X DPBS, 50 mM Tris Cl pH 8, 150 mM NaCl, 0.5 mM EDTA, 5 mM β-mercaptoethanol, 10 mM reduced glutathione. Eluted fractions were separated on 10 % SDS PAGE, and transferred onto PVDF membrane (Millipore). Blots were then probed with primary antibody Anti-Nup205 at 1:4000 (Sigma HPA024574), Anti-GFP 1:3000 (Sigma G1546) followed by secondary HRP conjugate. Blots were developed using Quant HRP substrate (Takara) and images were acquired on Amersham Imager 600 (GE).

3. Result and Discussion

3. 1 Feature Derivation

The predictability of any supervised machine learning method is dependent on the nature of features used for training. Random forest classifier is a tree-structure based algorithm where the classification rules are learned based on the feature values and their target class provided while training. Various features generated for training the random forest classifier were divided into three categories viz conservational, structure-based and contact potential-based features. For the conservation-based feature, a new co-evolution algorithm was derived as explained in 2.3.1 and figure 2. The new method as described in section 2.3.1 provided better scores for the interface residues as opposed to other co-evolution methods (table S2). Another important difference was generation of only a single non-symmetric M*N matrix from this method as opposed to LXL (where L= M+N) from other methods which result in higher signal to noise ratios. Thus, the conditional mutual information (CMI) based method was able to provide more confidence to the co-evolving pair of residues and decreasing the noise by generating the M*N matrices. Moreover, the co-evolving pair of residues in the interacting proteins maintain the homeostasis of the interaction across species hence using them as a feature as opposed to the standard PSSM based conservation methods(such as PAIRpred[9], eFindSite[18], Cons-PPISP[12], PSIVER[20] , BIPSPI[19], etc.) provided better predictability.

The nature of physicochemical properties of the residue interaction in the protein interface is somewhere in between their properties when present in the core or on the surface of the protein. It has been reported that the interface environment is closer to that exhibited on the outside in contact with the solvent as opposed to that present in the core of the protein[43]. For example, relative solvent accessibility of a residue which defines its possible position in the protein i.e. whether it will be present in the core of the protein (relative solvent accessibility of 0) or is solvent-exposed (relative solvent accessibility >0). For the residues which lie in the PPI interface should have value as 0<RSA<1 if the value is scaled between 0 and 1. Due to lack of specific standard matrices for inter-protein residue contacts, those derived for intra-protein contacts were used for feature generation in this method which includes charge, hydrophobe and size compatibilities, relative solvent accessibility and secondary structure predictions.

The knowledge-based statistical potentials have also been used previously to mimic the interactions between the amino acids in a protein. One of such knowledge-based potential is the contact potential derived by Miyazawa and Jernigan based on statistical analysis of the protein structures. These contact potentials are widely used in the computational prediction for protein folding. The contact potentials for the residue lying in the PPI interface should ideally lie in between those of buried and exposed residues. To assess their applicability in identifying interface residues of the interacting proteins three approximations of these contact potentials were used as features.

The contacts between two residues of the interacting proteins also depend on its neighboring residues by creating a favorable niche for the interaction to take place. Hence the properties governing the interaction (as described above) of the neighboring residues will also have an impact on the overall predictability of the random forest classifier. To address this, the random forest classifier was trained in two different modes i.e. with and without environment features, the results of which are explained below.

3.2 Evaluation of environment features in random forest classifier

To validate the effect of the environment features on the random forest classifier, the classifier was trained both with and without the environment features. The evaluation metrics obtained for both the cases are listed in supplementary table S3. The overall accuracy obtained for the dataset trained with the environment features was 85.3% as opposed to that for without environment features was 80%. The Receiver-Operator Curve and confusion matrix for five-fold cross-validation for the dataset with environment features is shown in figure 4 and that without environment is depicted in supplementary figure S2. As observed through all the evaluation statistics, the classifier predicts with better precision and recall and hence F1 measure, especially for the class label 1, when the environment features are used for training. Thus, validating that these derived features (environment features) are important in predicting the contact forming residue pairs for the interacting proteins.

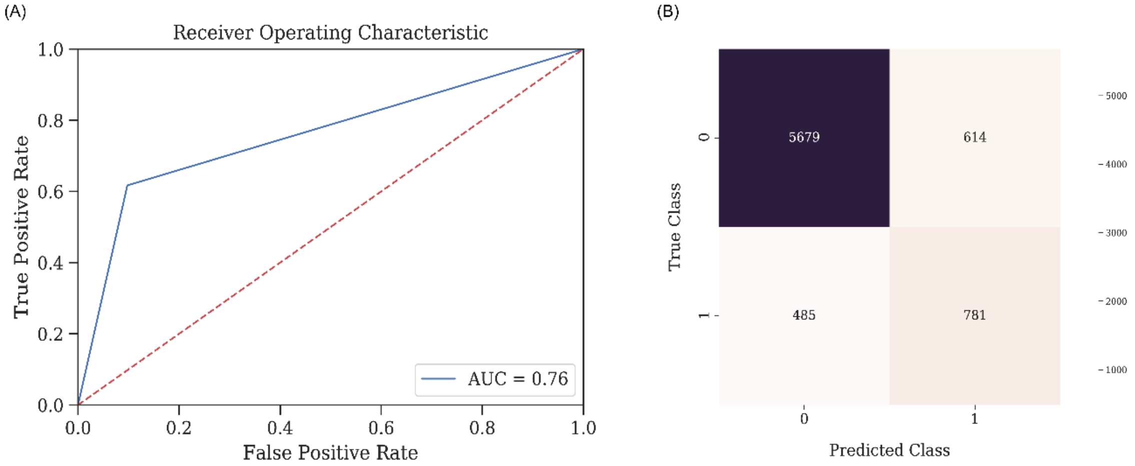

Figure 4: Statistics for the Random Forest Classifier Model for predicting contact forming residue pairs. (A) Receiver-operator curve (ROC) depicting Area under the curve (AUC) as 0.76 when the model is tested on the 75:25 data split. (B) Confusion matrix for the tested model on 75:25 data split with a final accuracy of 85.33%

3.3 Feature importance evaluation

One of the marked features of random forest classifier is that it is able to decipher the importance of every feature used for training which can be used to determine the over-fitting of a model as well as to gain insights about the physical relevance of the features in predicting the PPI interface. The feature importance plot for the dataset without the environment features (supplementary figure S3) depicts that the three most important features are relative solvent accessibility (RSA), co-evolution scores (CMI) and the contact potentials (CP). However, the feature importance plot for the dataset with environment features (18 features in all) (figure 5), depicts the importance of these derived features. Of the 18 features, used for training, top 12 positions have all 9 derived/environment features along with RSA, CMI, and CP. Thus, it is evident that all these features play a crucial role in the prediction of protein interaction interfaces.

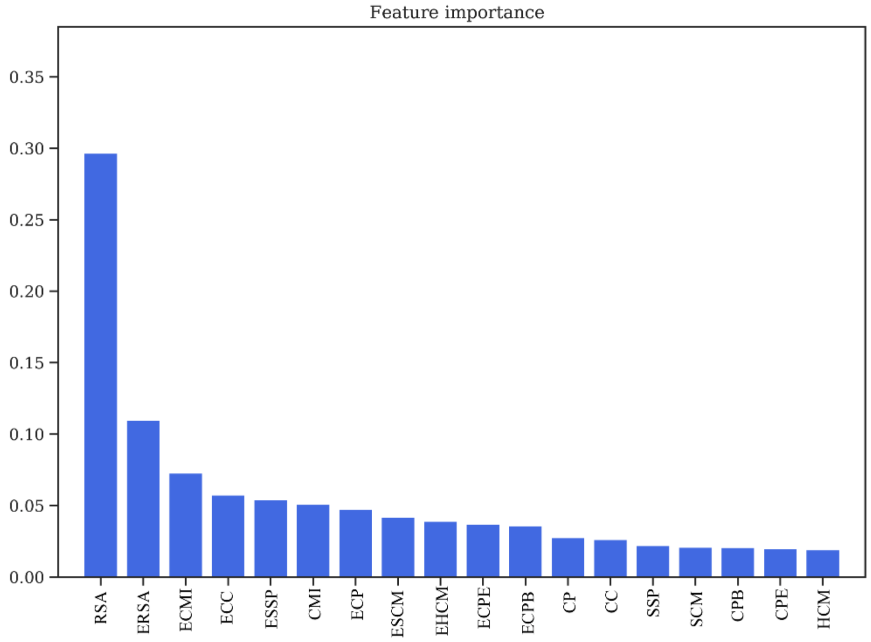

Figure 5: Feature Importance obtained from Random Forest Classifier.

Relative Solvent Accessibility (RSA/ERSA) and Co-evolution Scores (ECMI/CMI) as two of the most important features in training the model. **RSA:** Relative Solvent Accessibility. **ERSA:** Environment Relative Solvent Accessibility. **ECMI:** Environment Conditional Mutual Information. **ECC:** Environment Charge Compatibility. **ESSP:** Environment Secondary Structure Prediction. **CMI:** Conditional Mutual Information. **ECP:** Environment Contact Potential. **ESCM:** Environment Structure Compatibility Matrix. **EHCM:** Environment Hydropathy Compatibility Matrix. **ECPE:** Environment Contact Potential for Exposed residues. **ECPB:** Environment Contact Potential for Buried residues. **CP:** Contact Potential. **CC:** Charge Compatibility. **SSP:** Secondary Structure Prediction. **SCM:** Structure Compatibility Matrix. **CPB:** Contact Potential for Buried residues. **CPE:** Contact Potential for Exposed residues. **HCM:** Hydropathy Compatibility Matrix.

3.4 Relationship between the size of feature kernel matrix and type of secondary structures in the interaction hotspots

The interaction interfaces of the proteins can be classified into 6 possible categories based on the secondary structure compositions of the interface hotspot regions, such as α-α, α-β, α-l, β-β, β-l and l-l (where α denotes helices, β denoted sheets, and l denoted loops). Since the residue environment features were identified as the most critical features in the training of random forest classifier model, it is important to consider the role of the size of kernel matrix used for training the classifier. The residue environment for any protein can range from n-1 to n+1 position and up to n-3 to n+3 positions, thus all such variations were tested by training different classifiers. For every different size and weight of the feature kernel matrix, the derived features were generated and used to train different random forest models. For each of the test dataset, all these different models were tested to determine a relationship between the nature of interaction in terms of secondary structure pairs and the size and weight of feature kernel matrices. The optimized models were then utilized to test for pair of interacting proteins with known crystal structure which were not a part of the training dataset to validate the predictability of the method. As observed from table S4, for interface hotspots consisting of loop-loop or loop-sheet interactions were predicted better using 5*5 kernel matrix derived model and those consisting of helix-helix interfaces were predicted better using the 3*3 kernel matrix derived model.

3.5. Validation of prediction onto test datasets

The pipeline CoRNeA was used to test its predictability on four eukaryotic protein complexes with known crystal structures. These protein complexes were not a part of the training dataset. The combined amino acid length of the two proteins in these hetero dimers ranged from 127 amino acids to 986 amino acids. Additionally, variability in terms of secondary structure combinations in the interface were also considered while selecting these test datasets. The features for each dataset were generated as for the training dataset and different kernel matrix derived environmental feature-based models were used for predicting the interface residues for each test case. The model which predicted with the best evaluation statistics was considered for the downstream network analysis and final prediction matrix processing. Moreover, CoRNeA was used to predict the interaction interface of a known interacting pair of protein from the inner ring of the nuclear pore complex to access the applicability of the pipeline to filter high scoring pairs in absence of structural information.

3.5.1 Vav and Grb2 Sh3 domain heterodimer (PDB ID: 1GCQ)

One of them was the crystal structure of Vav and Grb2 Sh3 domain (PDB ID: 1GCQ)[44] which consists of three chains. One of Vav proto-oncogene (Chain C) and the other two of growth factor receptor-bound protein 2 (Chain A and Chain B). The dataset was compiled for this protein pair using Chain A and Chain C of 1GCQ as query. The features were calculated as described above and used as test dataset for evaluating the trained random forest models with environment features. The total size of the dataset created by these two chains amounted to 4002 pairs of residues. The random forest classifier predicted 25 pairs correctly as true positives and 967 pairs were predicted as false positives.

To further reduce the number of false-positive pairs, network analysis was performed. The intra protein contact forming residue pairs for Chain A (Protein A) and Chain C (Protein B) of 1GCQ were obtained from co-evolution analysis where only top 5% pairwise values were considered to be true cases. The length of Chain A is 56 amino acids which would lead to 3,136 intra pairs. The highest scoring 157 pairs were considered while constructing the intra protein contact forming residue network of Chain A of 1GCQ as depicted in supplementary figure S4 (A). The length of Chain C is 69 amino acids which would lead to 4,761 intra protein pairs. The highest scoring 238 pairs were considered while constructing the intra protein contact forming network of Chain C of 1GCQ as depicted in figure S4(B). The inter protein contact forming residue pair network of Chain A and Chain C as obtained from random forest classifier is shown in figure S4(C) which consisted to 992 predicted pairs of which 967 were false positives. A residual network was calculated from the three networks mentioned above (as shown in Figure S4(D)) and the final pairs were plotted as a matrix of Protein A versus Protein B. Since a 5*5 matrix was used to derive the environmental features, a unitary matrix of 5*5 was convolved onto the resultant interface prediction matrix. Pairs having convolved value more than 6 were selected which reduced the total pairs to 359 of which 42 were true positives and 317 were false positives. The results obtained from the pipeline are shown onto the structure of Vav and Grb2 Sh3 domains (PDB ID 1GCQ) (Figure 6A(i-ii)). Interestingly, the data labels provided while testing was only for Chain A and Chain C but the labels obtained after prediction were for both the pairs i.e. Chain A and Chain C (Figure 6A(i-ii)) as well as Chain B and Chain C (Figure 6A(i-ii)) (table S5) within 10Å distance. In comparison to the interface predicted by PISA using the structural information, CoRNeA was able to predict at least 50% of true pairs as depicted in figure 6A(iii). Thus, the overall pipeline to predict the PPI interface is fair in predicting the probable pairs of interacting residues as well as separate out the residue which might reside on the surface of the protein from those present in the core of the individual proteins only from amino acid sequence information. The confusion matrix before and after the network analysis is provided in supplementary table S6.

3.5.2 Alpha gamma heterodimer of human Isocitrate dehydrogenase (IDH3) (PDB ID: 5YVT)

To test the applicability of the pipeline on larger protein complexes, the structure of the alpha gamma heterodimer of human IDH3 (PDB ID: 5YVT)[45] (Figure 6B) was used as a test dataset. This protein complex is from mitochondrial origin and its length (M+N) is larger (693 amino acids) as compared to the previous example (PDB ID: 1GCQ, 127 amino acids). Network analysis was performed for this dataset by calculating the intra contacts of both chains A and B. The residual network resulted in 992 edges which were then mapped back in the form of the matrix of Protein A versus Protein B. A unitary matrix of 5*5 was convolved onto the predicted matrix and 537 pairs having value more than 6 were selected for analysis. Of these, 30 pairs formed the actual contacts when mapped onto the structure having distance within 10Å as shown in figure 6B (i-ii). Hence this new pipeline can be used for proteins from eukaryotic origin as well as the length of the pair of proteins in consideration is not a limiting factor.

3.5.3 Ubiquitin like activating enzyme E1A and E1B (PDB ID: 1Y8R)

The crystal structure of ubiquitin-like activating enzyme E1A and E1B (PDB ID: 1Y8R[46]) having a combined length of 986 amino acids (Protein A: 346 amino acids and Protein B: 640 amino acids) was used as another test dataset. Network analysis was performed for this dataset by calculating the intra contacts of both chains A and B. The residual network resulted in 1166 edges which were then mapped back in the form of the matrix of Protein A versus Protein B. A unitary matrix of 3*3 was convolved onto the predicted matrix owing to the occurrence of α helical structure of the pair of proteins under consideration resulting in total number of 898 positives pairs of which 18 were true positives and remaining 880 were false positives (Figure 6C).

3.5.4 Nup107-Nup133 heterodimer of the outer ring of the Nuclear Pore Complex (PDB ID: 3CQC)

The crystal structure of Nup107-Nup133 complex (Nup107: 270 amino acids, Nup133: 227 amino acids, combined length of 497 amino acids) consists of the C-terminal region of both the proteins was used as another test dataset. The residual network consisting of 540 pairs was generated after removing the nodes which are a part of the intra network in either of the proteins. The total number of points were further reduced to 240 after performing convolution on the final prediction matrix using a unitary 3*3 matrix and keeping a cut off of more than 2.

Of the 240 pairs, 6 pairs were identified as true positives within the distance of 10Å (Figure 6D).

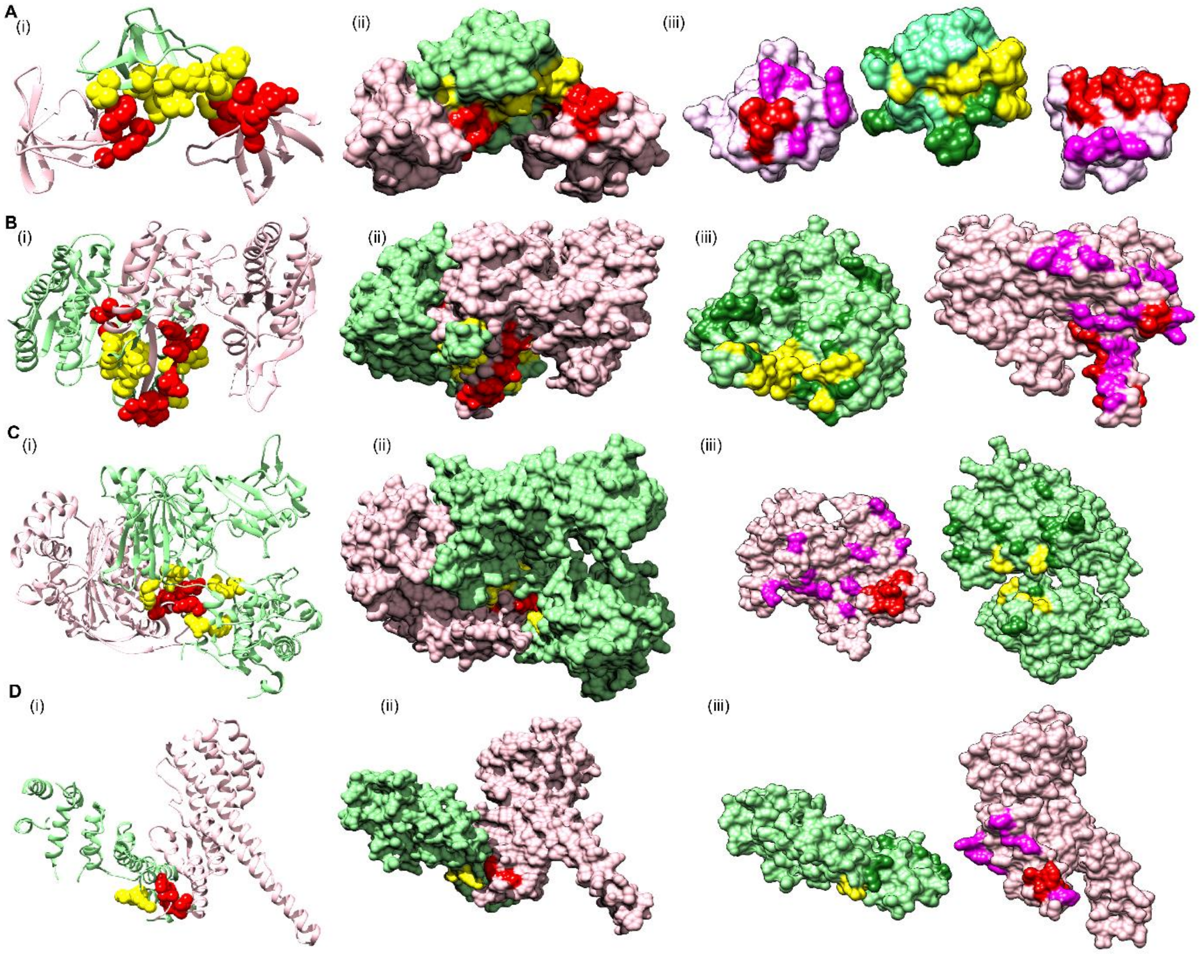

Figure 6: Prediction of interface hotspots on test datasets using CoRNeA.

Predictions of the interface residues for 4 test datasets were mapped onto their crystal structures, A. PDB ID: 1GCQ B. PDB ID: 5YVT, C. PDB ID: 1Y8R, D. PDB ID: 3CQC. The first column (i) for all four datasets depict ribbon representation where Protein A is colored in pink and Protein B in light green; interface residues predicted using CoRNeA for Protein A (red) and Protein B (yellow) are depicted as spheres. The second column (ii) depicts surface representation of the same. The third column (iii) depicts open book representation of the interface residues where the interface hotspots predicted by PISA and not by CoRNeA are colored as purple for Protein A and forest green for Protein B.

3.5.5 Nup93-Nup205 complex of the adapter ring of the Nuclear Pore Complex (NPC)

To test the applicability of the pipeline on the dataset without known structural information, hNup93-hNup205 interaction interface was explored. Nup93 is a linker protein of the Nup93-subcomplex of the NPC. It is known to connect the adaptor/ inner ring of the spoke region with the central channel pore of the NPC[47]. The adaptor region consists of the four proteins viz., Nup188, Nup205, Nup35, and Nup155. In terms of the known interactions of the specific domains of the Nup93, its R1 region which spans the first 82 amino acids is known to interact with the Nup62 of the central channel[48]. Nup93 is specifically known to form mutually exclusive complexes with either Nup188 or Nup205 of the adapter ring[49,50]. The interaction interface information for these pair of proteins is not known specifically from mammalian origin owing to difficuties in biochemical reconstitution of these complexes. However, for hNup93-hNup205, proximity information for this pair of proteins is known through crosslinking based mass spectrometry analysis[51]. The cross-linking data suggests three different regions of Nup93 to be in proximity of Nup205 (i.e. N-terminal, middle and C-terminal) but the most prominent hits are seen between the R2 (96-150) region at the N-terminal of Nup93 with the C-terminal of Nup205 (Figure 7A).

CoRNeA was employed to identify the interaction interface of Nup93-Nup205 complex by utilizing full length sequence information of both the proteins (Nup93: 819 amino acids and Nup205: 2012 amino acids). Since, the secondary structure prediction of both these proteins depicts α- helices, hence the 3*3 kernel matrix derived random forest model was utilized to predict the interface pairs. The resultant high scoring pairs, which pertained to specifically the R2 region of Nup93 (96-150) with the C-terminal region of Nup205 obtained from CoRNeA (Figure 7B), are in consensus with cross-linking mass spectrometry analysis (table S7). However various low scoring pairs were also identified for Nup93 middle and C-terminal region but they did not span more than three continuous pairs (such as 89-91 of Nup93 with 1201-1205 of Nup205) between the two proteins.

Further, validation of the interacting interface between Nup93 and Nup205 predicted with CoRNeA analysis was done by *in-vitro* pull-down experiment using Nup93 deletion constructs (Figure 7C). Upon pull down with GST tagged anti-GFP nanobody, N-terminal region of Nup93(1-150) was able to pull endogenous Nup205 efficiently. Further mapping the minimal interaction region, R2 fragment of Nup93 (96-150) was found to interact with endogenous Nup205 thus validating the *in-silico* prediction by CoRNeA. A diminished interaction of the Nup93 region (176-819) was also observed through this pull-down experiment which is also consistent with the identification of low scoring regions identified by CoRNeA. This experimental validation depicts that CoRNeA is able to predict the short stretches of interaction hotspots between known pair of interacting proteins from only their sequence information and hence can be used to decipher the minimal interacting regions of pair of large proteins. Thus, aiding in their biochemical reconstitution followed by structural elucidation.

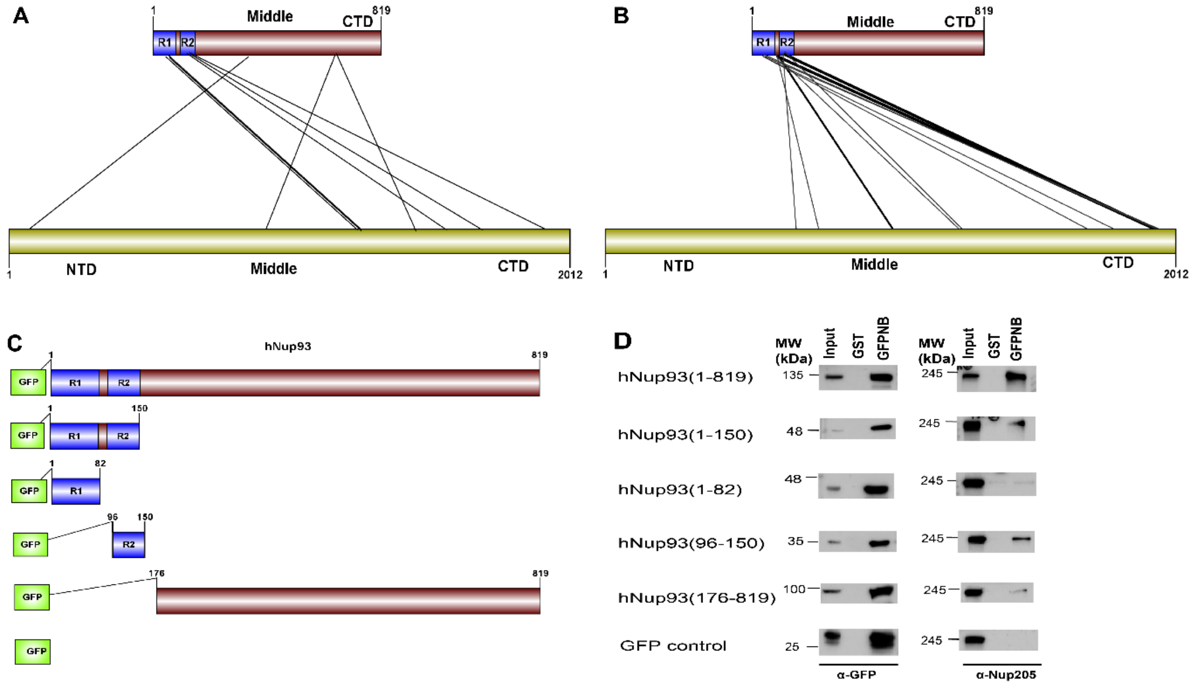

Figure 7: Prediction and validation of interface regions for Nup93-Nup205

A. Cross-linking based mass spectrometry defined proximity regions between Nup93-Nup205 (adapted from Jan Kosinski, et.al, Science, 2016). B. Top 10% regions predicted by CoRNeA. Edges in bold depict three most significant regions (N-terminal of Nup93 with C-terminal of Nup205) (details in table S7). C. GFP-fused deletion constructs for Nup93 for validating the predictions. D. Immunoprecipitation results depicting N-terminal region (1-150) and R2 regions (96-150) of Nup93 specifically interact with endogenous Nup205. GFPNB: GST-anti-GFP-nanobody.

3.6. Comparison with other methods/BIPSPI

To assess the predictability of CoRNeA, the results obtained from it for the two test cases described above were compared to the predictions of recently published method BIPSPI[19] which is closest to our implementation and the only available method to predict the interface residues using only amino acid sequence information. BIPSPI also utilizes similar physiochemical properties as well as residue environment information through hot encodings. Although the major point of difference between BIPSPI and CoRNeA lies choice of conservation-based feature (PSSM in BIPSPI versus co-evolution in CoRNeA) and derivation of the environmental features (hot encoding in BIPSPI versus convolution averaging in CoRNeA). Moreover, the network analysis post processing of the results to remove the intra contacts is one of the unique attributes of the pipeline CoRNeA which is not present with other machine learning based methods known for predicting the interaction interfaces. Since CoRNeA utilizes only the amino acid sequence information, the sequence mode of prediction on BIPSPI server was employed for predicting the interface residues of the four test datasets (PDB ID: 1GCQ, 5YVT, 1Y8R and 3CQC). The Nup93-Nup205 dataset could not be processed using BIPSPI owing to its limitation to consider proteins larger than 1500 amino acids in length. The results obtained for these datasets depicted that the final predictions from CoRNeA yielded in fewer false positives than BIPSPI hence validating the overall improvement in the accuracy of the prediction of PPI interface residues (Table 1).

Table 1: Comparison of predictions from CoRNeA with BIPSPI

Test Dataset	**Method**	**Expected no of residues within 10Å**	**Number of True positives with probability more than 0.5**	**Number of False Positives with probability more than 0.5**
PDB ID: 1GCQ	BIPSPI	108	0	N/A
	CoRNeA		42	317
PDB ID: 5YVT	BIPSPI	164	24	1210
	CoRNeA		30	537
PDB ID: 1Y8R	BIPSPI	157	1	57
	CoRNeA		18	880
PDB ID: 3CQC	BIPSPI	48	0	1
	CoRNeA		6	240

The numbers depicted for CoRNeA are post convolution of prediction matrix. For 1GCQ the total number of expected contacts and true positives are for both chain combinations i.e. Chain A and C; Chain B and C

CoRNeA can, however, be further optimized to reduce the false-positive rates as well as improve the true positive predictions by increasing the training dataset. As it is evident that the environmental features play a very important role in training the classifier and there is a correlation between the type of secondary structures and kernel matrices used to derive these environmental features, different training sub-datasets can be used to train specifically on various combinations of secondary structures to decrease the false positive prediction by random forest classifiers and hence increase the specificity of the overall pipeline.

Conclusions

Predicting the pairwise interacting residues for any two-given pair of proteins from only the amino acid sequence still remains a challenging problem. In this study, the newly designed pipeline CoRNeA addresses some of the challenges for predicting the PPI interfaces such as applicability to eukaryotic PPI and high false-positive rates, by incorporating co-evolution information and intra contacts for improving the precision and recall of the pipeline. This pipeline can be utilized to predict the interface residues as a pairwise entity and also to understand folding of the individual proteins though intra contact predictions. Obtaining the structural information of proteins individually as well as in complex with their interacting partners is a tremendously challenging problem especially for large multimeric complexes. CoRNeA can be utilized to identify the minimal interacting regions in the heterodimers for its biochemical reconstitution, which can then be utilized in structure elucidation studies. The information obtained from CoRNeA can also be used as a starting point for protein docking studies in cases where 3D structure models (experimental or homology-based) are available. The web server is currently under development and the R codes along with the trained models are available on github.

Author Contributions

KC and RC conceived the project. KC performed all computational analysis. BB performed the pull-down experiment. SCM contributed for intellectual suggestions for the project. KS and AKK helped in the optimization of machine learning algorithm. The manuscript was written by KC and RC. All authors read and approved the manuscript.

Acknowledgments

This work is supported by Department of Science and Technology-Science and Engineering Research Board grant (SERB/EMR/2017/000272) and Department of Biotechnology grant

(DBT/PR26398/BRB/10/1637/2017) to RC, DST-Inspire SRF fellowship to KC and DBT-SRF fellowship to BB. AKK and KS acknowledge financial support from a Raja Ramanna Fellowship awarded by Department of Atomic Energy (10/1(16)/2016/RRF-R&D-II/630). We would like to thank all members of Lab of Structural Biology at NCCS for help and support through discussions on this manuscript. We would also like to thank Professor Ninan Sajeetth Philip (Department of Physics, St. Thomas College, Kozhencherry, Kerala, India)**,** Dr Sheelu Abrahim (Marthoma College, Chungathara, Nilambur, Kerala, India) and Dr Kaustabh Vaghmare (Inter University Centre for Astronomy and Astrophysics, Pune, Maharashtra, India) for their inputs for machine learning part.

References

1. Uetz P, Giot L, Cagney G, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*. 2000;403(6770):623-627. doi:10.1038/35001009

2. Masters SC. Co-Immunoprecipitation from Transfected Cells. In: Fu H, ed. *Methods Mol Biol*. Totowa, NJ: Humana Press; 2004:337-350. doi:10.1385/1-59259-762-9:337

3. Sobott F, Robinson C V. Protein complexes gain momentum. *Curr Opin Struct Biol*. 2002;12(6):729-734. doi:10.1016/S0959-440X(02)00400-1

4. Zhang QC, Deng L, Fisher M, Guan J, Honig B, Petrey D. PredUs: A web server for predicting protein interfaces using structural neighbors. *Nucleic Acids Res*. 2011;39(SUPPL. 2):283-287. doi:10.1093/nar/gkr311

5. Xue LC, Dobbs D, Honavar V. HomPPI: A class of sequence homology based protein-protein interface prediction methods. *BMC Bioinformatics*. 2011;12. doi:10.1186/1471-2105-12-244

6. Jordan RA, EL-Manzalawy Y, Dobbs D, Honavar V. Predicting protein-protein interface residues using local surface structural similarity. *BMC Bioinformatics*. 2012;13(1):41. doi:10.1186/1471-2105-13-41

7. Porollo A, Meller J. Prediction-Based Fingerprints of Protein–Protein Interactions. *PROTEINS Struct Funct Bioinforma*. 2007;66(2006):630-645. doi:10.1002/prot.21248

8. Liang S, Zhang C, Liu S, Zhou Y. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res*. 2006;34(13):3698-3707. doi:10.1093/nar/gkl454

9. Minhas F ul AA, Geiss BJ, Ben-Hur A. PAIRpred: Partner-specific prediction of interacting residues from sequence and structure. *Proteins Struct Funct Bioinforma*. 2014;82(7):1142-1155. doi:10.1002/prot.24479

10. Kufareva I, Budagyan L, Raush E, Totrov M, Abagyan R. PIER: Protein Interface Recognition for Structural Proteomics. *PROTEINS Struct Funct Bioinforma*. 2007;67:400-417. doi:DOI: 10.1002/prot.21233

11. Neuvirth H, Raz R, Schreiber G. ProMate: A structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol*. 2004;338(1):181-199. doi:10.1016/j.jmb.2004.02.040

12. Chen H, Zhou HX. Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data. *Proteins Struct Funct Genet*. 2005;61(1):21-35. doi:10.1002/prot.20514

13. Qin S, Zhou HX. Meta-PPISP: A meta web server for protein-protein interaction site prediction. *Bioinformatics*. 2007;23(24):3386-3387. doi:10.1093/bioinformatics/btm434

14. de Vries SJ, Bonvin AMJJ. Cport: A consensus interface predictor and its performance in prediction-driven docking with HADDOCK. *PLoS One*. 2011;6(3). doi:10.1371/journal.pone.0017695

15. Vries SJ de, Dijk ADJ van, Bonvin AMJJ. WHISCY: What Information Does Surface Conservation Yield? Application to Data-Driven Docking. *PROTEINS Struct Funct Bioinforma*. 2006;63:479-489. doi:DOI: 10.1002/prot.20842

16. Negi SS, Schein CH, Oezguen N, Power TD, Braun W. InterProSurf: a web server for predicting interacting sites on protein Surfaces. *Bioinformatics*. 2007;23(24):3397-3399. doi:10.1093/bioinformatics/btm474.InterProSurf

17. Segura J, Jones PF, Fernandez-Fuentes N. Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams. *BMC Bioinformatics*. 2011;12. doi:10.1186/1471-2105-12-352

18. Maheshwari S, Brylinski M. Template-based identification of protein-protein interfaces using eFindSitePPI. *Methods*. 2016;93:64-71. doi:10.1016/j.ymeth.2015.07.017

19. Sanchez-Garcia R, Sorzano COS, Carazo JM and, Segura J. BIPSPI: a method for the prediction of Partner- Specific Protein-Protein Interfaces. *Bioinformatics*. 2019;35(3):470-477. doi:10.1093/bioinformatics/xxxxx

20. Murakami Y, Mizuguchi K. Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics*. 2010;26(15):1841-1848. doi:10.1093/bioinformatics/btq302

21. Zeng H, Wang S, Zhou T, et al. ComplexContact: A web server for inter-protein contact prediction using deep learning. *Nucleic Acids Res*. 2018;46(W1):W432-W437. doi:10.1093/nar/gky420

22. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci U S A*. 2009;106(1):67-72.

23. Hopf TA, Schärfe CPI, Rodrigues JPGLM, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife*. 2014;3:1-45. doi:10.7554/elife.03430

24. Goncearenco A, Shaytan AK, Shoemaker BA, Panchenko AR. Structural Perspectives on the Evolutionary Expansion of Unique Protein-Protein Binding Sites. *Biophys J*. 2015;109(6):1295-1306. doi:10.1016/j.bpj.2015.06.056

25. Rodriguez-Rivas J, Marsili S, Juan D, Valencia A. Conservation of coevolving protein interfaces bridges prokaryote–eukaryote homologies in the twilight zone. *Proc Natl Acad Sci*. 2016;113(52):15018-15023. doi:10.1073/pnas.1611861114

26. Lockless SW, Ranganathan R. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science (80-)*. 2002;286(5438):295-299. doi:10.1126/science.286.5438.295

27. Kastritis PL, Moal IH, Hwang H, et al. A structure-based benchmark for protein-protein binding affinity. *Protein Sci*. 2011;20(3):482-491. doi:10.1002/pro.580

28. Finn RD, Clements J, Arndt W, et al. HMMER web server: 2015 Update. *Nucleic Acids Res*. 2015;43(W1):W30-W38. doi:10.1093/nar/gkv397

29. Pei J, Kim BH, Grishin N V. PROMALS3D: A tool for multiple protein sequence and structure alignments. *Nucleic Acids Res*. 2008;36(7):2295-2300.

doi:10.1093/nar/gkn072

30. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics*. 2009;25(9):1189-1191. doi:10.1093/bioinformatics/btp033

31. Gouy M, Milleret F, Mugnier C, Jacobzone M, Gautier C. ACNUC: a nucleic acid sequence data base and analysis system. *Nucleic Acids Res*. 1984;12(1Part1):121-127. doi:10.1093/nar/12.1Part1.121

32. Wyner AD. A definition of conditional mutual information for arbitrary ensembles. *Inf Control*. 1978;38(1):51-59. doi:10.1016/S0019-9958(78)90026-8

33. Biro JC. Amino acid size, charge, hydropathy indices and matrices for protein structure analysis. *Theor Biol Med Model*. 2006;3(1):1-12. doi:10.1186/1742-4682-3-15

34. Heffernan R, Yang Y, Paliwal K, Zhou Y. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*. 2017;33(18):2842-2849. doi:10.1093/bioinformatics/btx218

35. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*. 1999;292:195-202. doi:10.1006/jmbi.1999.3091

36. Miyazawa S, Jernigan RL. Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading - 1-s2.0-S002228369690114X-main.pdf. *J Mol Biol*. 1996:623-644. http://ac.els-cdn.com.cuhsl.creighton.edu/S002228369690114X/1-s2.0-S002228369690114X-main.pdf?_tid=24355ac2-9cdb-11e2-9995-00000aab0f6c&acdnat=1365047759_ff446b9f5d285ed566bab28b5354da32.

37. Zeng H, Liu K-S, Zheng W-M. The Miyazawa-Jernigan Contact Energies Revisited. *Open Bioinforma J*. 2012;6(1):1-8. doi:10.2174/1875036201206010001

38. Krissinel E, Henrick K. Inference of Macromolecular Assemblies from Crystalline State. *J Mol Biol*. 2007;372(3):774-797. doi:10.1016/j.jmb.2007.05.022

39. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32.

doi:10.1023/A:1010933404324

40. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825-2830. doi:10.1007/s13398-014-0173-7.2

41. Csárdi G, Nepusz T. The igraph software package for complex network research. *InterJournal , Complex Syst*. 2006;1695. doi:10.3724/SP.J.1087.2009.02191

42. Katoh Y, Nozaki S, Hartanto D, Miyano R, Nakayama K. Architectures of multisubunit complexes revealed by a visible immunoprecipitation assay using fluorescent fusion proteins. *J Cell Sci*. 2015;128(12):2351-2362. doi:10.1242/jcs.168740

43. Jones S, Thornton JM. PROTEIN-PROTEIN INTERACTIONS: A REVIEW OF PROTEIN DIMER STRUCTURES. *Prog Biophys molec Biol*. 1995;63(94):31-65. doi:10.1016/0079-6107(94)00008-W

44. Nishida M, Nagata K, Hachimori Y, et al. Novel recognition mode between Vav and Grb2 SH3 domains. *EMBO J*. 2001;20(12):2995-3007. doi:10.1093/emboj/20.12.2995

45. Liu Y, Hu L, Ma T, Yang J, Ding J. Insights into the inhibitory mechanisms of NADH on the αγ heterodimer of human NAD-dependent isocitrate dehydrogenase. *Sci Rep*. 2018;8(1):1-12. doi:10.1038/s41598-018-21584-7

46. Lois LM, Lima CD. Structures of the SUMO E1 provide mechanistic insights into SUMO activation and E2 recruitment to E1. 2005;24(3):439-451. doi:10.1038/sj.emboj.7600552

47. Benjamin V, Wolfram A. The diverse roles of the Nup93/Nic96 complex proteins – structural scaffolds of the nuclear pore complex with additional cellular functions. *Biol Chem*. 2014;395:515. doi:10.1515/hsz-2013-0285

48. Sachdev R, Sieverding C, Flotenmeyer M, Antonin W. The C-terminal domain of Nup93 is essential for assembly of the structural backbone of nuclear pore complexes. *Mol Biol Cell*. 2011;23(4):740-749. doi:10.1091/mbc.e11-09-0761

49. Vincent Galy, Iain W. Mattaj and PA. Caenorhabditis elegans Nucleoporins Nup93 and Nup205 Determine the Limit of Nuclear Pore Complex Size Exclusion In Vivo. *Mol Biol Cell*. 2003;14(December):5104–5115. doi:10.1091/mbc.E03

50. Theerthagiri G, Eisenhardt N, Schwarz H, Antonin W. The nucleoporin Nup188 controls passage of membrane proteins across the nuclear pore complex. *J Cell Biol*. 2010;189(7):1129 LP - 1142. doi:10.1083/jcb.200912045

51. Kosinski J, Mosalaganti S, Von Appen A, et al. Molecular architecture of the inner ring scaffold of the human nuclear pore complex. *Science (80-)*. 2016;352(6283):363-365. doi:10.1126/science.aaf0643

Supplementary Material

CoRNeA: A pipeline to decrypt the inter protein interfaces from amino acid sequence information

Kriti Chopra[1], Bhawna Burdak[1], Kaushal Sharma[2], Ajit K. Kembavi[2], Shekhar C. Mande[3] and Radha Chauhan[1*]

1- National Centre for Cell Science, Pune.

2- Inter University Centre for Astronomy and Astrophysics, Pune

3- Council of Scientific and Industrial Research (CSIR), New Delhi

***Corresponding Author**:

Dr. Radha Chauhan, Scientist 'E', National Centre for Cell Science, S.P. Pune University Campus, Ganeshkhind, Pune 411007, Maharashtra, India.

Email: radha.chauhan@nccs.res.in

Phone: +91-20-25708255

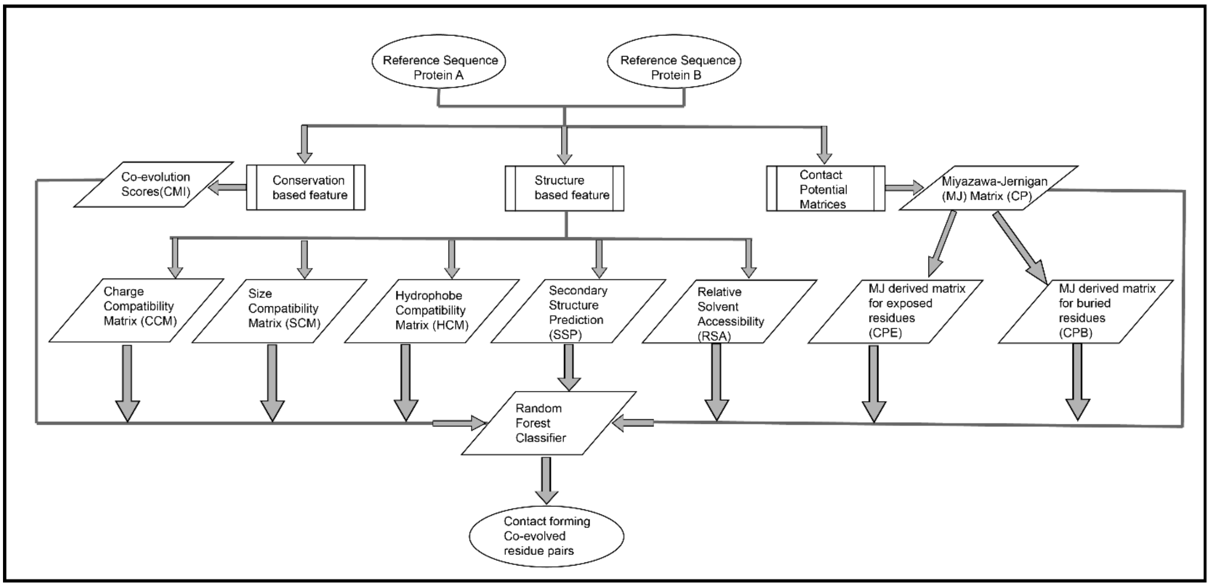
Reference Sequence Protein A
Reference Sequence Protein B
Co-evolution Scores(CMI)
Conservation based feature
Structure based feature
Contact Potential Matrices
Miyazawa-Jernigan (MJ) Matrix (CP)
Charge Compatibility Matrix (CCM)
Size Compatibility Matrix (SCM)
Hydrophobe Compatibility Matrix (HCM)
Secondary Structure Prediction (SSP)
Relative Solvent Accessibility (RSA)
MJ derived matrix for exposed residues (CPE)
MJ derived matrix for buried residues (CPB)
Random Forest Classifier
Contact forming Co-evolved residue pairs

Figure S1: Flowchart depicting the feature generation for predicting pair of protein-protein interaction interface residues

Table S1: Numeric Coding for amino acids used for co-evolution score calculations

Amino Acid	**Numeric Coding**
V (Valine)	1
I (Isoleucine)	2
L (Leucine)	3
M (Methionine)	4
F (Phenylalanine)	5
W (Tryptophan)	6
Y (Tyrosine)	7
S (Serine)	8
T (Threonine)	9
N (Asparagine)	10
Q (Glutamine)	11
H (Histidine)	12
K (Lysine)	13
R (Arginine)	14
D (Aspartic Acid)	15
E (Glutamic acid)	16
A (Alanine)	17
G (Glycine)	18
P (Proline)	19
C (Cysteine)	20
- (Gap)	21
X (Non-Standard Amino Acid)	22

Table S2: Comparison of known methods for PPI interface prediction with the new hybrid method

Interface residues (PISA)			**Various algorithms for finding contacts**				
Nup107	**Nup133**	**Distance(Å)**	**MI** **(2.03)**	**DCA** **(0.158)**	**Evfold** **(0.155)**	**SCA** **(3.86)**	**New Method (CMI)** **(1.00)**
D 879	**T 696**	3.37	0.4285	0.0022	0.0052	0.618	**0.804**
S 822	**K 975**	2.78	0.2379	0.0009	0.0023	0.1607	**0.591**
E 884	**K 975**	2.69	0.2379	0.0001	0.0021	0.339	**0.524**
D 917	**K 966**	2.53	0.0104	0.0005	0.0013	0.192	**0.642**
Y 921	**K 966**	3.37	0.225	0.0008	0.003	0.616	**0.364**
E 922	**R 962**	3.18	0.7898	0.0015	0.002	0.742	**0.342**
K 894	**D 982**	3.82	0.354	0.005	0.0005	0.223	**0.371**
R 898	**A 980**	3.28	0.179	0.001	0.0025	0.039	**0.233**
Q 902	**Q 944**	3.35	0.8474	0.002	0.001	1.46	**0.159**

The interface residues for a test case as predicted by PISA. The value under the name of the method represents the highest score calculated by the algorithm. MI: Mutual information, DCA: Direct Coupling Analysis, SCA: Statistical Coupling Analysis.

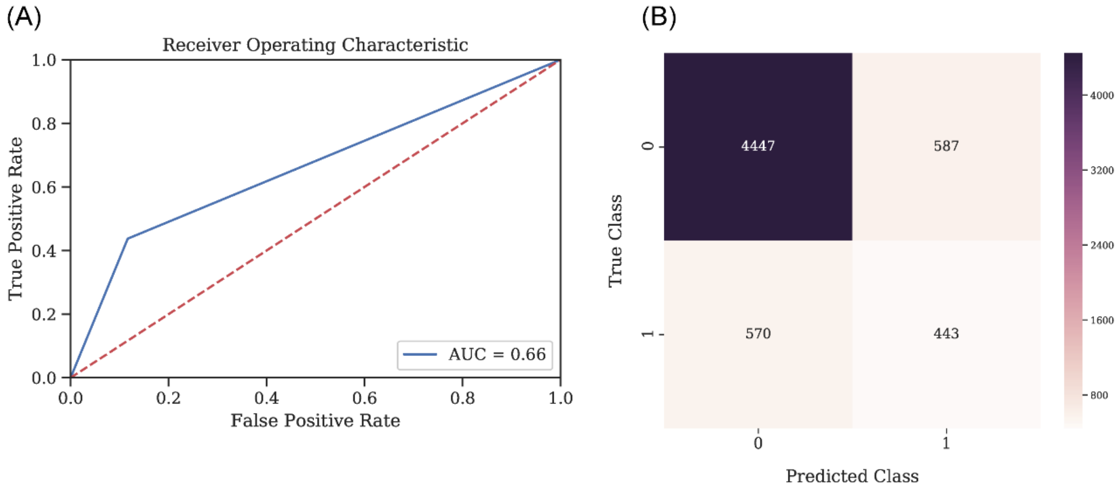

Figure S2: Statistics for the Random Forest Classifier Model for predicting contact forming residue pairs without environmental features. (A) Receiver-operator curve (ROC) depicting Area under the curve (AUC) as 0.66 when the model is tested on the 75:25 data split. (B) Confusion matrix for the tested model on 75:25 data split with a final accuracy of 80%

Table S3: Comparison of evaluation statistics, with and without environmental features.

	Class	Precision	Recall	F1-score
Without Environmental Features	**0**	**0.89**	**0.88**	**0.88**
	1	**0.43**	**0.44**	**0.43**
	Weighted Avg	**0.81**	**0.81**	**0.81**
With Environmental Features	**0**	**0.92**	**0.91**	**0.91**
	1	**0.56**	**0.59**	**0.58**
	Weighted Avg	**0.86**	**0.85**	**0.86**

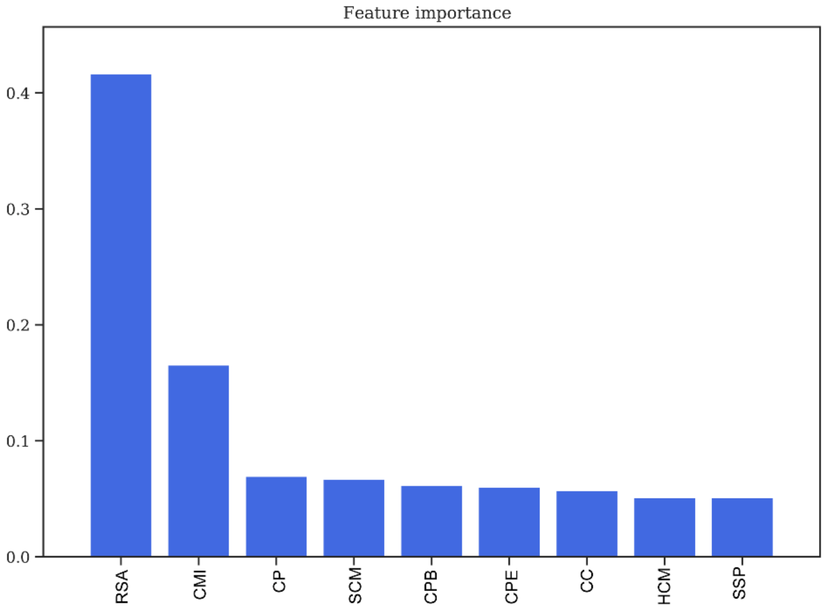

Figure S3: Feature Importance obtained from Random Forest Classifier without environmental features.

Relative Solvent Accessibility (RSA) and Co-evolution Scores (CMI) as two of the most important features in training the model. **RSA:** Relative Solvent Accessibility. **CMI:** Conditional Mutual Information. **CP:** Contact Potential. **SCM:** Structure Compatibility Matrix. **CPB:** Contact Potential for Buried residues. **CPE:** Contact Potential for Exposed residues. **CC:** Charge Compatibility. **HCM:** Hydropathy Compatibility Matrix. **SSP:** Secondary Structure Prediction.

Table S4: Evaluation of different kernel matrix derived random forest classifier on different test datasets

PDB ID	Type of secondary structure	Best Kernel Matrix	Number of true positive labelled	Actual true positives predicted with best kernel matrix
1GCQ	Loop:Loop Loop:Sheet	5*5	81	25
1Y8R	Helix:Helix Loop:Loop	3*3	157	23
4YDU	Helix:Helix	3*3	86	23
5YVT	Helix:Helix Sheet:Sheet Loop:Loop	5*5	164	64
3CQC	Helix:Helix	3*3	48	13

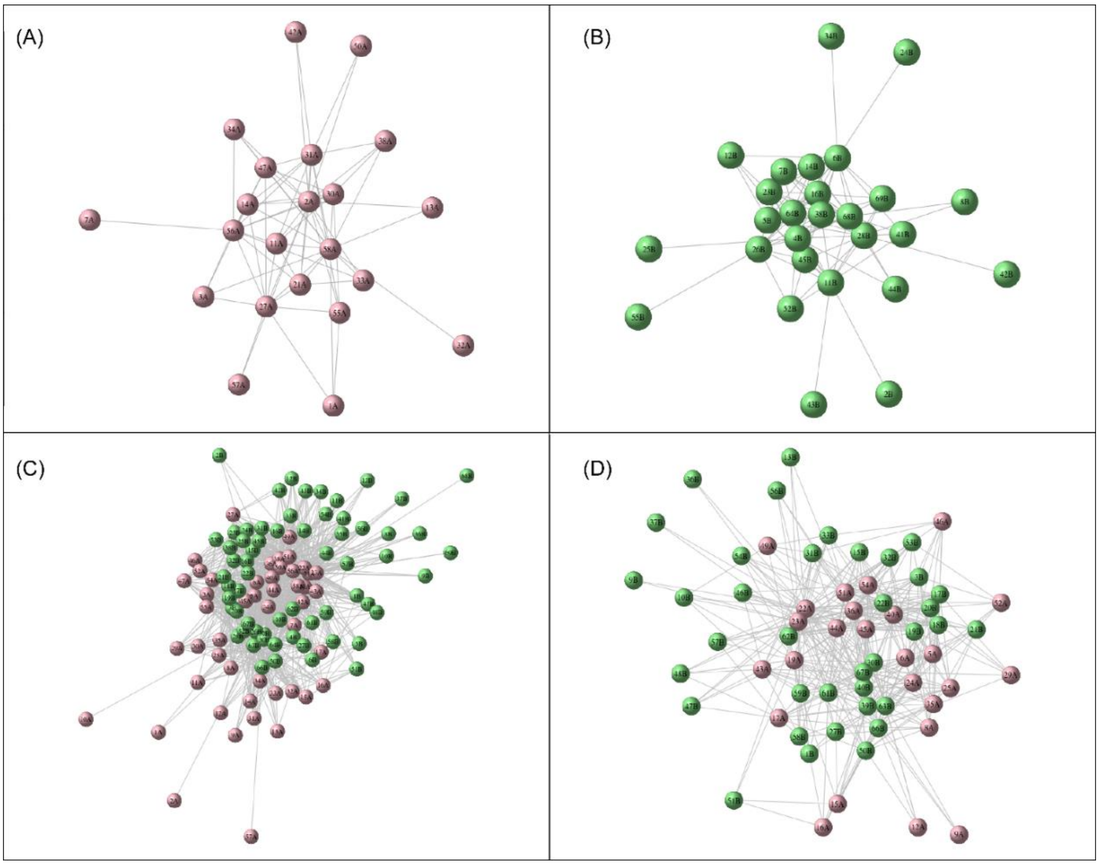

Figure S4: Network analysis for PDB ID 1GCQ. (A) Intra-protein network for Chain A/B of 1GCQ obtained from top 5% co-evolving intra residue pairs. (B) Intra-protein network for Chain C of 1GCQ obtained from top 5% co-evolving intra residue pairs. (C) Inter-protein network for 1GCQ obtained from random forest classifier. (D) Inter-protein network for 1GCQ after removing intra-protein network nodes and all nodes having relative solvent accessibility as 0.

Table S5: Pairwise true contacts predicted for PDB ID 1GCQ Chain A with Chain C and Chain B with Chain C within a distance cutoff of 10 Å.

Residue number (Chain A)	Residue number (Chain C)	Convolution Value	Distance (Å)	Residue number (Chain B)	Residue number (Chain C)	Convolution Value	Distance (Å)
208	612	7	3.53	179	652	7	3.3
192	611	7	3.6	165	655	8	4.66
208	611	8	3.62	179	655	9	6.7
194	608	7	3.7	164	657	7	7.2
209	607	8	3.7	179	653	7	7.5
209	610	11	3.9	179	654	8	8.9
193	610	9	4	179	629	8	9.8
193	611	7	4.17				
208	610	9	4.39				
209	609	11	4.78				
165	608	7	4.8				
209	611	9	4.9				
209	608	9	5.13				
207	611	8	5.2				
209	651	7	6.8				
164	607	9	7.15				
193	609	9	7.3				
207	610	9	7.47				
164	608	11	7.49				
179	606	9	7.6				
192	609	9	7.7				
209	612	7	7.8				
179	607	12	8.5				
165	609	8	8.7				
193	608	7	8.8				
165	610	7	8.9				
209	653	7	9.3				
192	608	7	9.6				
179	608	12	9.8				

Table S6: Confusion Matrix statistics for PDB ID 1GCQ before and after network analysis

Before Network Analysis	**True Class**	**0**	**True Negatives= 2954**	**False Positives = 967**
		1	**False Negatives= 56**	**True Positives= 25**
			0 Predicted Class 1	
After Network Analysis	**True Class**	**0**	**True Negatives= 3575**	**False Positives = 317**
		1	**False Negatives= 56**	**True Positives= 42**
			0 Predicted Class 1	

Table S7: Top 10% pairs predicted for Nup93-Nup205

Nup205	Nup93	Convolution Score	No of pairs in the predicted regions
1932-1936	86-99	272	57
1932-1936	101-117	234	54
1013-1014	86-109	100	30
1945-1948	44-48	82	16
1801-1805	44-48	71	15
749-751	86-97	66	18
1935-1939	448-452	65	16
1928-1930	87-94	65	17
682-684	109-115	63	21
1937-1940	44-48	63	14
1696-1700	44-48	59	15
1250-1252	87-93	55	17
1250-1252	109-113	45	15