bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Kriti Chopra[1], Kaushal Sharma[2], Ajit 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

Abstract

Motivation

Decrypting the interface residues of the protein complexes provide insight into the functions of the proteins and hence the overall cellular machinery. Computational methods have been devised in the past to predict the interface residues using amino acid sequence information but all these methods have been majorly applied to predict for prokaryotic protein complexes. Since the composition and rate of evolution of the primary sequence is different between prokaryotes and eukaryotes, it is important to develop a method specifically for eukaryotic complexes.

Results

Here we report a new hybrid pipeline for the prediction of protein-protein interaction interfaces from the amino acid sequence information which is based on the framework of co-evolution, machine learning (random forest) and network analysis named CoRNeA trained specifically on eukaryotic protein complexes. We use conservation, structural and contact potential as major group of features to train the random forest classifier. We also incorporate the intra contact information of the individual proteins to eliminate false positives from the predictions keeping in mind that the amino acid sequence also holds information for its own folding and not only the interface propensities. Our prediction on example datasets shows that CoRNeA not only enhances the prediction of true interface residues but also reduces false positive rates significantly.

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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)(Godwin *et al.* 2000), co-immunoprecipitation (co-IP)(Masters 2004), mass spectrometry (Sobott and Robinson 2002) etc. which are labor, cost and time intensive. Deciphering the PPI (Protein-Protein Interaction) 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 PPI 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 (Zhang *et al.* 2011), PS-HomPPI (Xue, Dobbs and Honavar 2011), PriSE (Honavar *et al.* 2012) etc. The machine learning (ML) methods which have been described till date are either structure based or sequence based. The structure-based ML methods (SPPIDER(Porollo and Meller 2007), PINUP(Liang *et al.* 2006), PAIRpred(Afsar Minhas, Geiss and Ben-Hur 2014), PIER(Kufareva *et al.* 2007), ProMate(Neuvirth, Raz and Schreiber 2004), Cons-PPISP(Chen and Zhou 2005), Meta-PPISP(Qin and Zhou 2007), CPort(de Vries and Bonvin 2011), WHISCY(Vries, Dijk and Bonvin 2006), InterProSurf(Negi, S.S.; Catherine, H.S.; Oezguen, N.; Power 2007), VORFFIP(Segura, Jones and Fernandez-Fuentes 2011), eFindSite(Maheshwari and Brylinski 2016) 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 (Ruben Sanchez-Garcia *et al.* 2019), PSIVER (Murakami and Mizuguchi 2010), and ComplexContact (Zeng *et al.* 2018) 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

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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(Weigt *et al.* 2009), EvComplex(Green *et al.* 2014) 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 (Zeng *et al.* 2018)) 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 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 (Goncearenco *et al.* 2015).

Hence to delineate the vast PPI 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)(Rodriguez-Rivas *et al.* 2016), it is important to

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 PPI interface 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)(Lockless 2002). 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 specially 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 such as charge, size and hydrophobe compatibility, secondary structure information and relative solvent accessibility, which 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 PPI interface. 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.

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

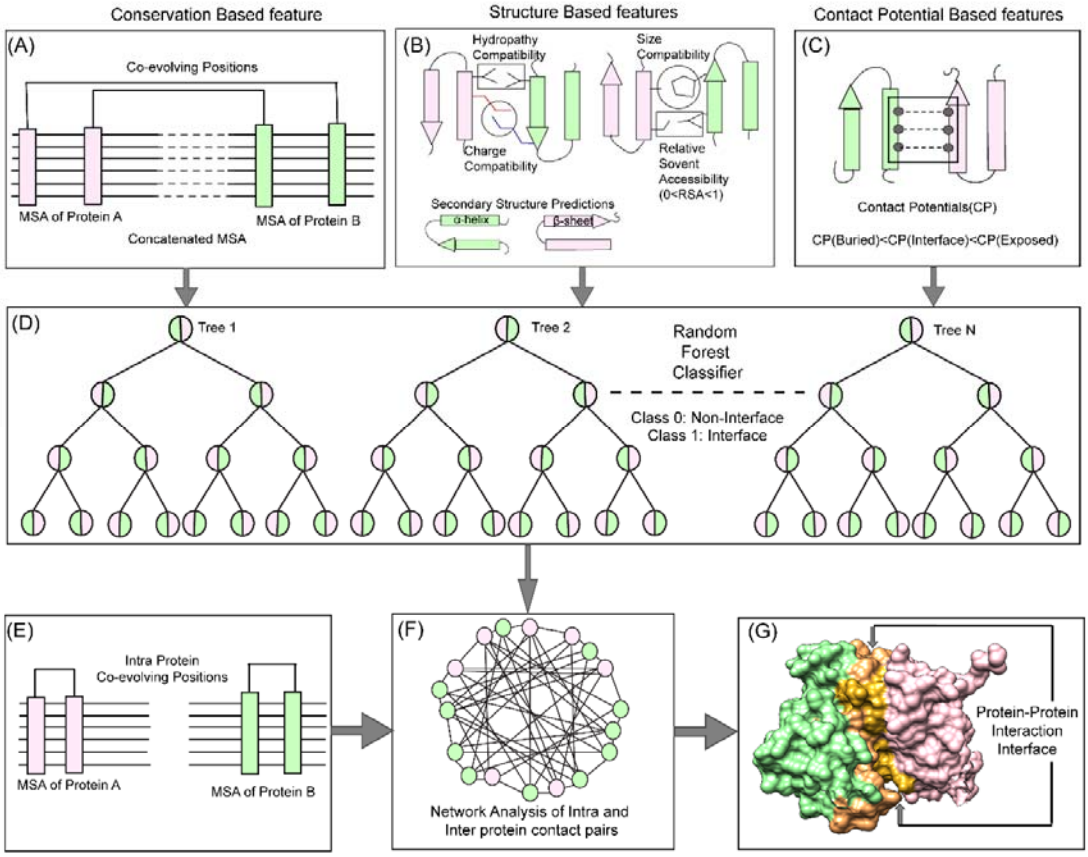

Figure 1: CoRNeA pipeline for predicting co-evolving contact forming residues in interacting pair of proteins. The method for predicting the protein-protein interaction interface consists of three levels. The top panel depicts the features used for machine learning pipeline. (A). Conservation based (coevolution) (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 intra 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 amino acid sequence of the two interacting proteins. 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. The third step is to combine the intra protein contact forming residues

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

2.1 Datasets

The Affinity Database version 2.0(Kastritis *et al.* 2011) was used to select the protein complex structures for training. The amino acid sequences of the complex structures were extracted from www.rcsb.org and used as query to search for homologs. PHMMER(Finn *et al.* 2015) 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(Pei, Kim and Grishin 2008). The alignments were then analyzed/edited in JalView(Waterhouse *et al.* 2009) and then concatenated (Last residue of Protein A followed by first residue of Protein B) in R using package seqinr(Gouy *et al.* 1984). 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 MXN matrix (M=length of Protein A and N= length of Protein B). All the feature values were scaled between 0 and 1. (Figure S1)

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

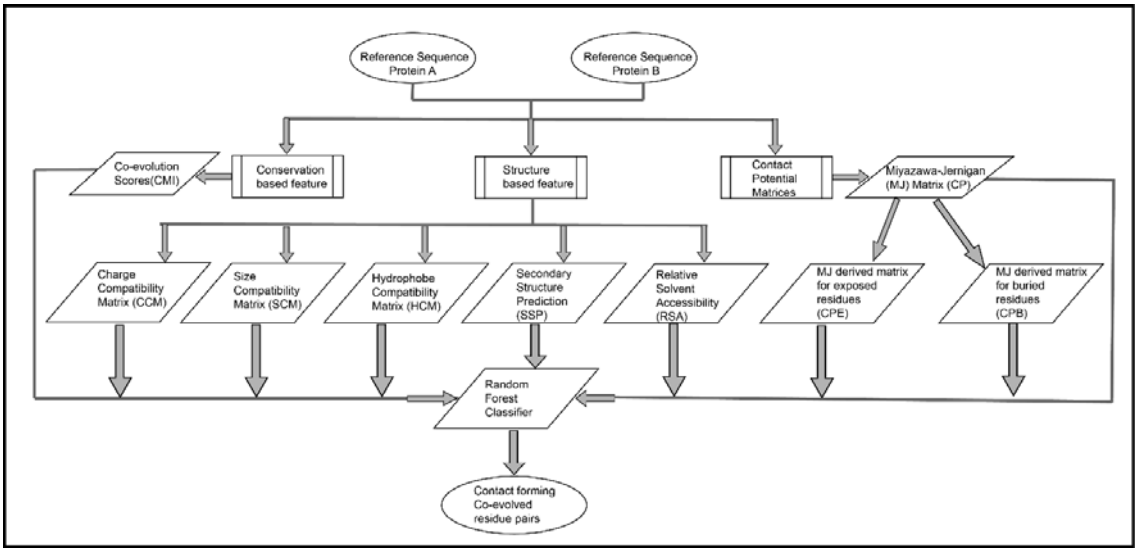

Figure S1: Flowchart depicting the feature generation for predicting pair of protein-protein interaction interface residues

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)(Lockless 2002). 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 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(Wyner 1978). 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 MXN matrix.

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

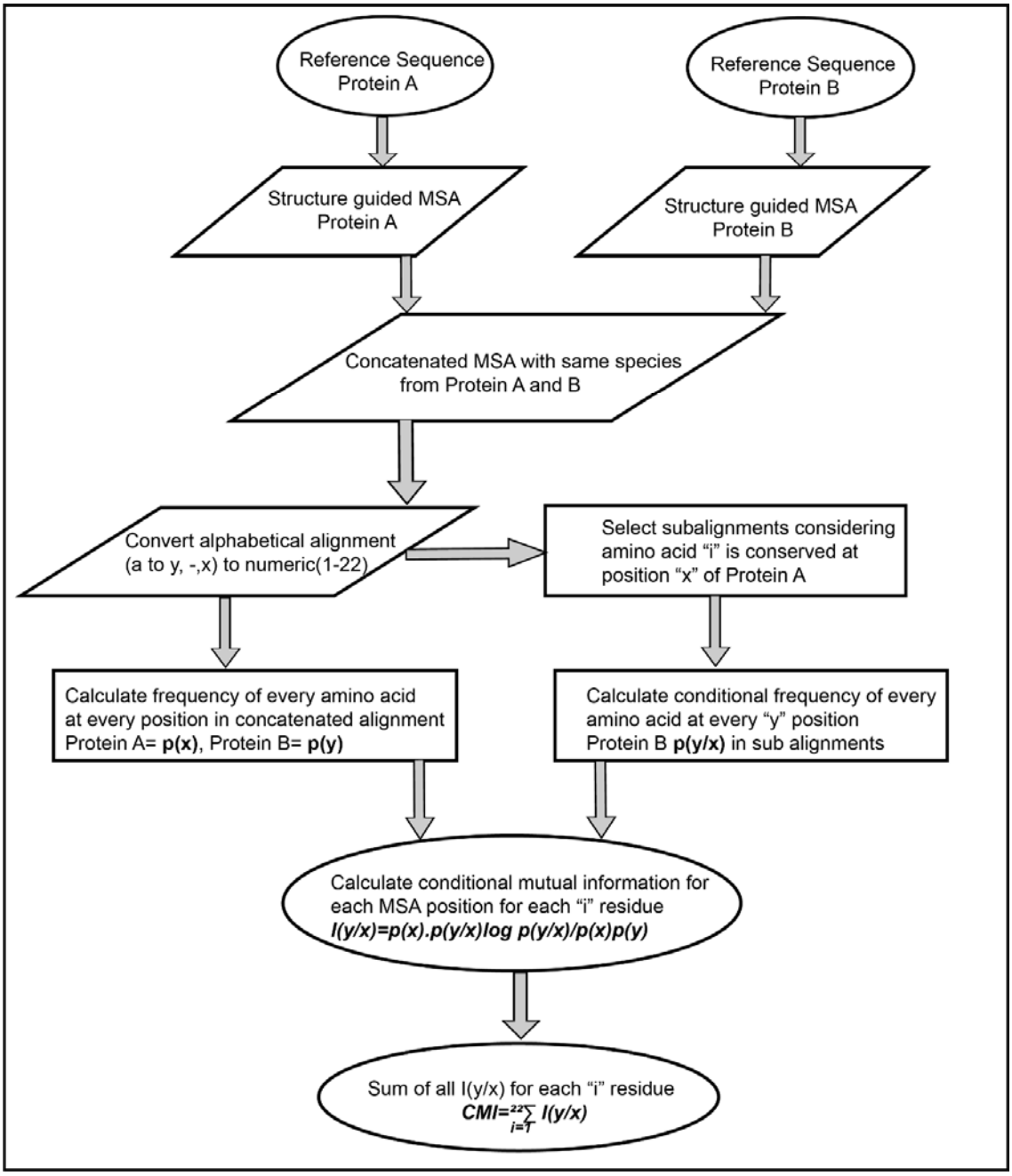

Figure 2: Flow chart representing algorithm for calculating inter protein co-evolving positions from multiple sequence alignments.

2.3.2 Structure based features

Charge, Hydrophobe and size compatibility matrices

The physicochemical properties of the residue 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* modelling of single protein(Biro 2006). These

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

matrices were used to derive an all versus all residue matrix (MXN) 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(Heffernan *et al.* 2017) and multiplied to form an all versus all (MXN) matrix of the pair of interacting proteins.

Secondary Structure Predictions (SSP)

The secondary structure of the proteins was predicted using PSIPRED(Jones 1999) and all residues were assigned numbers (i.e. 1= α-helix, 2=β-sheet and 3=l-loop). A 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 MXN 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) (Miyazawa and Jernigan 1996) where the effective inter-residue contact energies for all amino acid pairs were calculated based on 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 (Zeng, Liu and Zheng 2012). 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 (MXN) 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

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. Hence, 18 feature matrices were used for each pair of interacting protein for training the random forest classifier.

2.5 Interface residue labelling

The interface residues for the protein complexes were extracted using PISA(Krissinel and Henrick 2007). The number of residue pairs present in the interface (500 pairs for 42 complexes) was far less than all possible residues 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 datapoints, still the complete protein complex database exhibited highly imbalance data. 5000 pairs were labelled 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 was trained first using 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 a 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(Pedregosa *et al.* 2011) 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

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 3B). 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(Csárdi and Nepusz 2006).

The predictions from the random forest classifier were used to plot 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(Heffernan *et al.* 2017) for both the proteins independently. A residual network was hence computed for the inter-protein contact predications by first eliminating the nodes representing RSA=0 and then the intra-protein contacts from Protein A and B (figure 3C and 3D. This residual network was then analysed for the false positives and true positives on a protein complex with known 3D structure of the protein of interest.

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

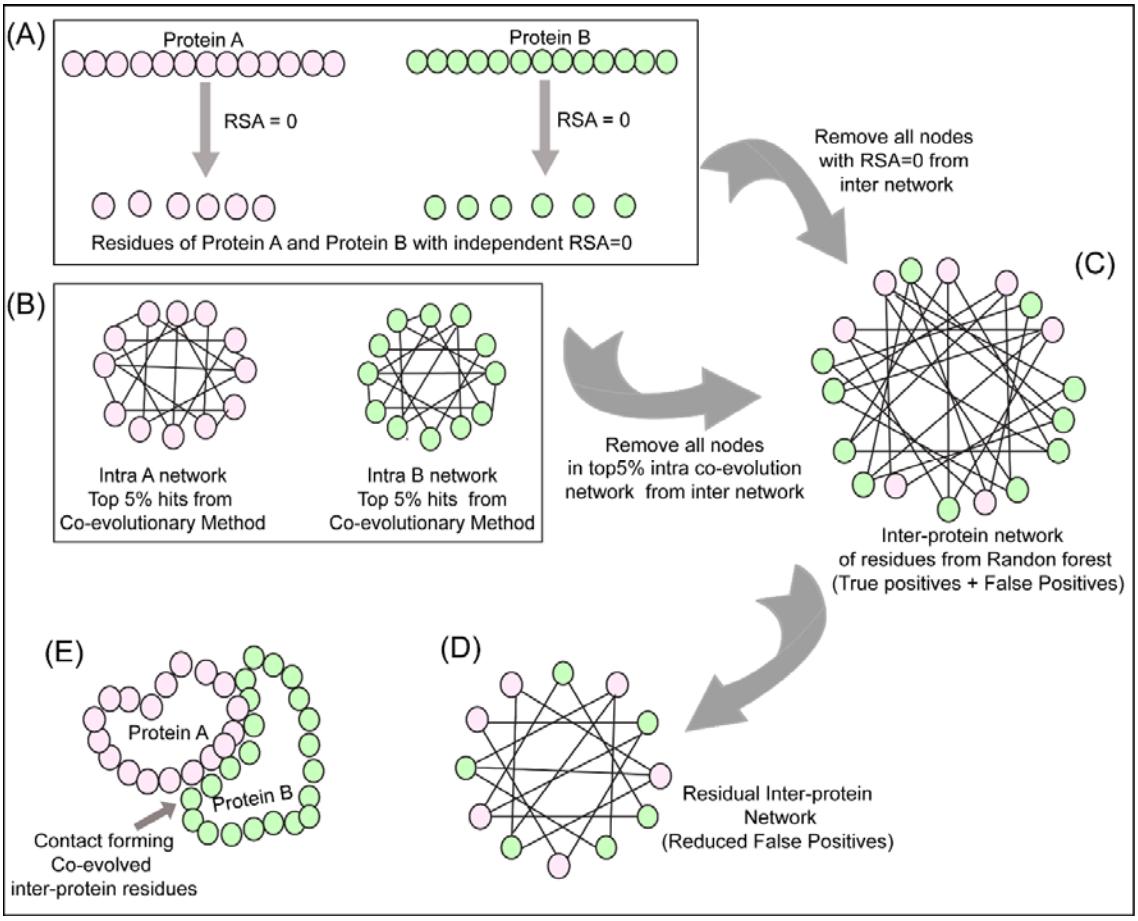

Figure 3: Network analysis of intra and inter protein contacts. (A) Extraction of residues with RSA=0 for Protein A and B. (B) Intra contact prediction for Protein A and B (top 5% co-evolving residue pairs). (C) Predicted inter protein network from random forest classifier. (D) The false positive inter 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.

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

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 MXN 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 MXN 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(Afsar Minhas, Geiss and Ben-Hur 2014), eFindSite(Maheshwari and Brylinski 2016), Cons-PPISP(Chen and Zhou 2005), PSIVER(Murakami and Mizuguchi 2010) etc) provided better predictability.

The nature of physicochemical properties of the residue interaction in the protein interface are 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(Jones and Thornton 1995). 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 access their applicability in identifying interface residues of the interacting proteins three approximations of these contact potentials were used as features.

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The contacts between two residues of the interacting proteins also depends on its neighbouring residues by creating a favourable niche for the interaction to take place. Hence the properties governing the interaction (as described above) of the neighbouring 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 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.

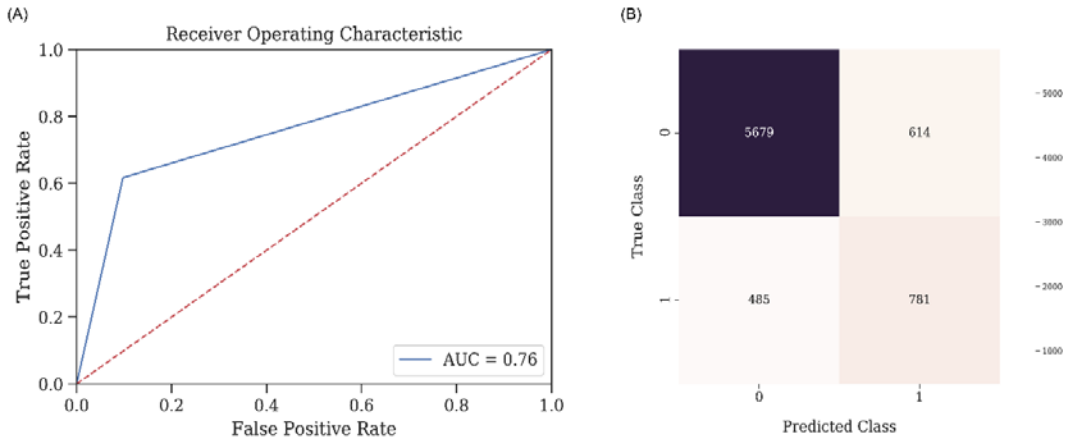

Figure 4: Statistics for the Random Forest Classifier Model for predicting contact forming residue pairs. (A) Receiver-operator curve (ROC) depicting Area under the curve (AUC) as 0.76 when the model is tested on the 75:25 data split. (B) Confusion matrix for the tested model on 75:25 data split with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

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

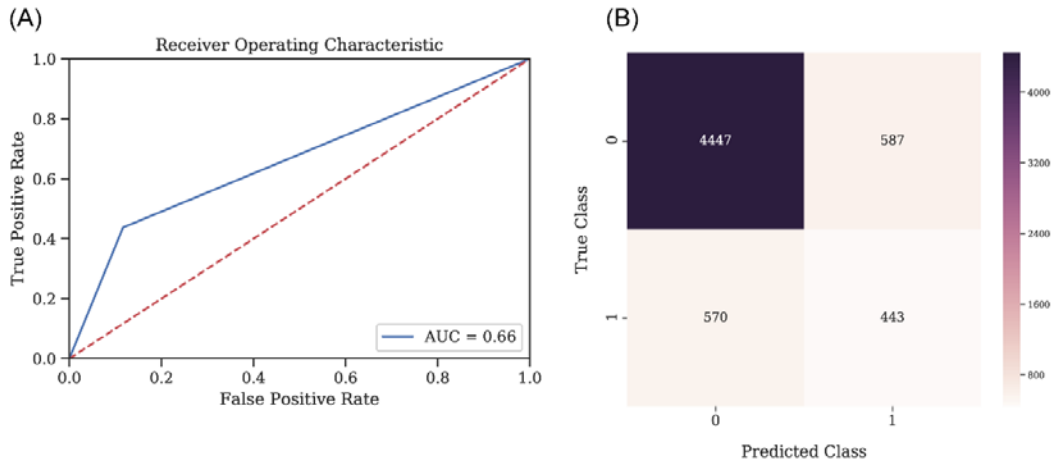

Figure S2: Statistics for the Random Forest Classifier Model for predicting contact forming residue pairs without environmental features. (A) Receiver-operator curve (ROC) depicting Area under the curve (AUC) as 0.66 when the model is tested on the 75:25 data split. (B) Confusion matrix for the tested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

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 features along with RSA, CMI and CP.

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Thus, it is evident that all these features play a crucial role for the prediction of protein interaction interfaces.

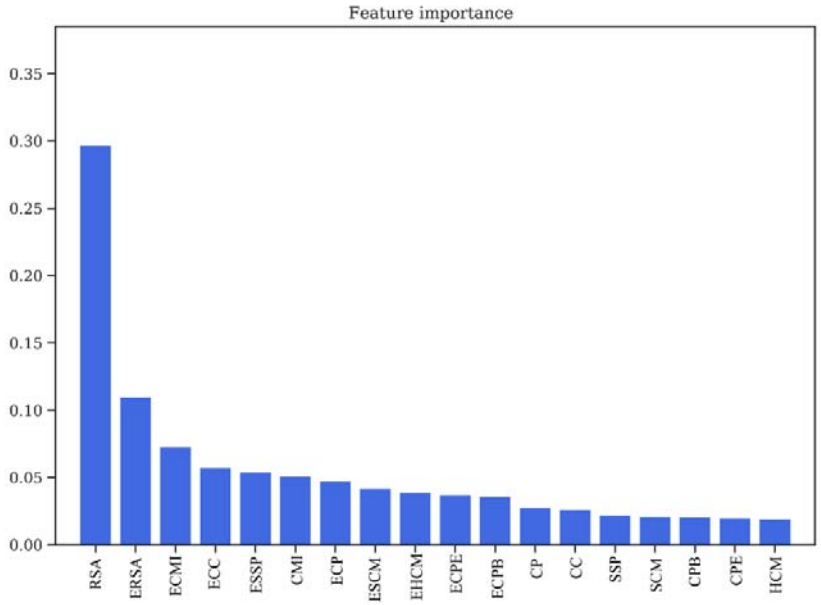

Figure 5: Feature Importance obtained from Random Forest Classifier.

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.

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

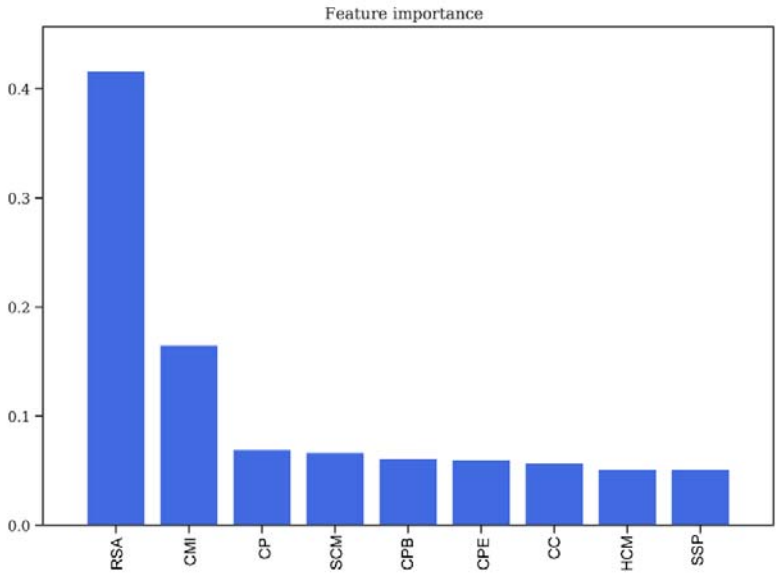

Figure S3: Feature Importance obtained from Random Forest Classifier without environmental features.

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.

4. Validation of prediction onto test dataset

The pipeline CoRNeA was used to test its predictability on a protein complexes with a known crystal structure. One of them was the crystal structure of Vav and Grb2 Sh2 domain (PDB ID: 1GCQ)(Nishida *et al.* 2001) 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 model. 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

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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)) to reduce the total pairs to 371 of which 52 were true positives and 319 were false positives.

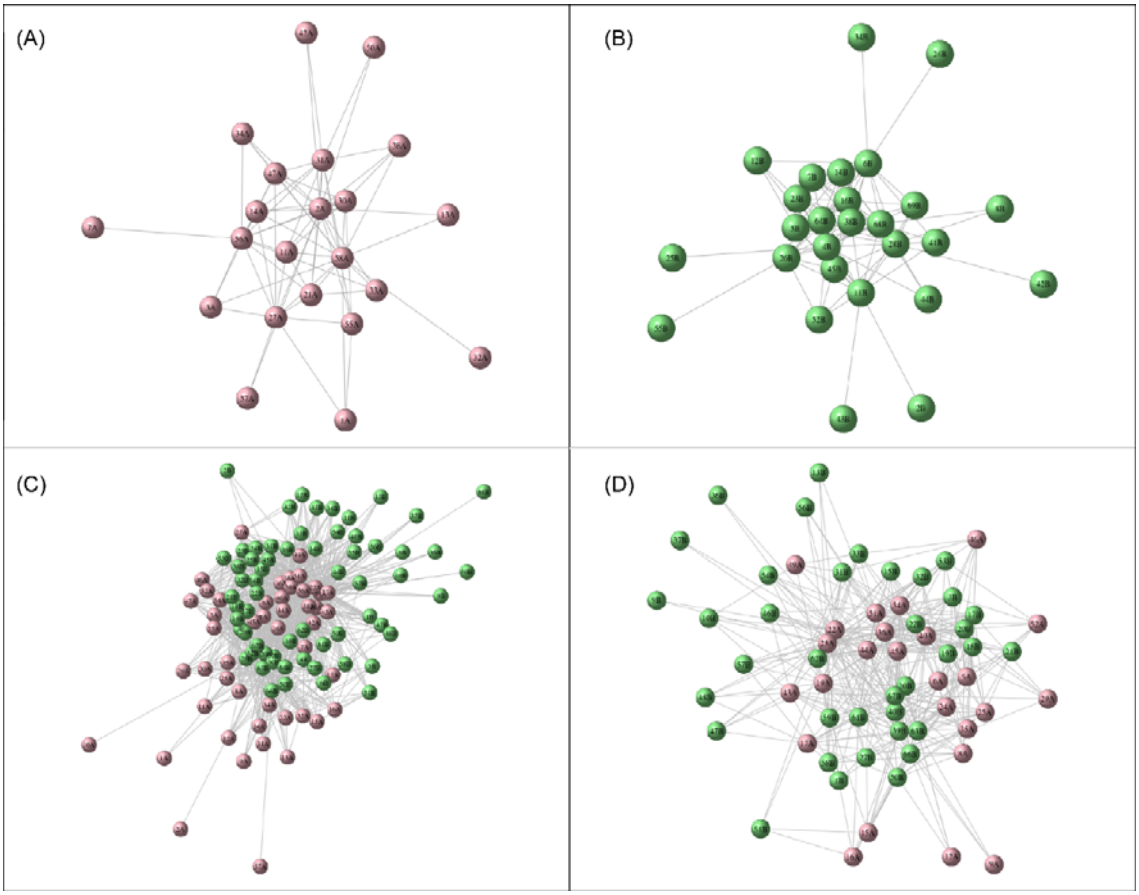

Figure S4: Network analysis for PDB ID 1GCQ. (A) Intra-protein network for Chain A/B of 1GCQ obtained from top 5% co-evolving intra residue pairs. (B) Intra-protein network for Chain C of 1GCQ obtained from top 5% co-evolving intra residue pairs. (C) Inter-protein network for 1GCQ obtained from random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

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The results obtained from the network are shown onto the structure of VAV and GRB2 SH3 domains (PDB ID 1GCQ) (Figure 6A). Interestingly, the data labels provided while testing were 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 6B) as well as Chain B and Chain C (Figure 6C) (details in supplementary table S4) within 10Å distance. 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 S5.

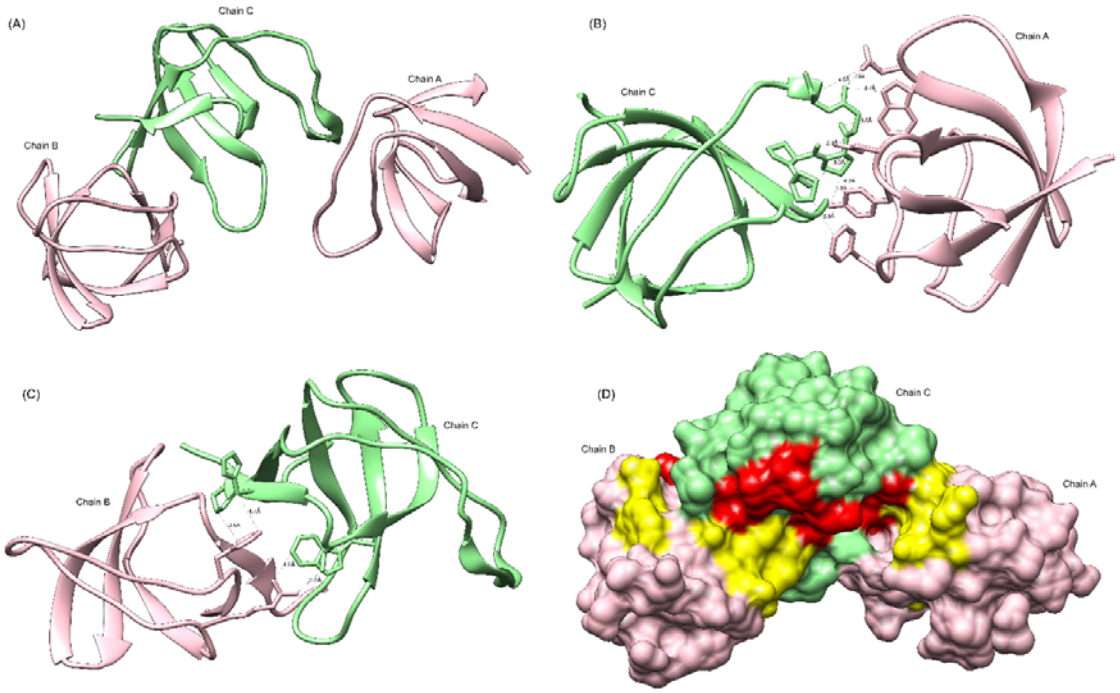

Figure 6: PDB ID 1GCQ evaluated by CoRNeA.

(A) Cartoon representation of 1GCQ. (B) Interface residues predicted by this method between Chain A (pink) and Chain C (green) within 5Å distance. (C) Interface residues predicted by this method between Chain B (pink) and Chain C (green) within 5Å distance. (D) Surface representation of 1GCQ depicting interface residues. Chain A and B in pink and their respective interface residues are shown in yellow. Chain C in green and its interface residues are depicted in red.

Table S4: 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	Residue number	Distance (Å)	Residue number	Residue number	Distance (Å)

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Chain A)	(Chain C)		(Chain B)	(Chain C)	
208	**609**	**3**	**179**	**652**	**3.3**
208	**608**	**3.3**	**165**	**657**	**3.6**
209	**610**	**3.5**	**179**	**637**	**4**
192	**611**	**3.6**	**165**	**656**	**5.3**
208	**611**	**3.6**	211	629	5.9
193	**610**	**4**	179	653	6.6
193	**611**	**4**	165	653	7.25
208	**612**	**4.3**	179	651	7.7
192	**612**	**4.4**	179	636	8
165	**608**	**4.8**	179	656	8
209	**611**	**4.9**	179	657	8
208	610	5.2	209	612	8.3
193	612	5.6	163	657	8.3
206	612	6	179	630	8.7
193	609	7.3	182	630	8.8
208	607	7.7	179	627	9
192	609	7.7	180	637	9
166	653	7.8	208	593	9.3
179	607	8.5	211	593	9.3
165	609	8.7	179	629	9.5
193	608	8.8	179	600	10
165	610	8.9	180	630	10
209	653	9.3	211	652	10
192	608	9.6	192	657	10
165	651	9.6	211	657	10
179	608	9.8			
174	612	10			

Table S5: 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**	

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

After Network Analysis	**True Class 0**	**True Negatives= 3575**	**False Positives = 319**
	True Class 1	**False Negatives= 56**	**True Positives= 52**
		0	**Predicted Class 1**

To test the applicability of the pipeline on larger protein complexes, the structure of the alpha gamma heterodimer of human IDH3 (PDB ID: 5YVT)(Liu *et al.* 2018) (Figure S5A) 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). The random forest classifier was able to predict 64 out of 164 contacts with precision. 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 of which 24 pairs formed the actual contacts when mapped onto the structure. In terms of the interface residues covered amongst these 24 pairs, 50% of the pairs where correctly identified by CoRNeA as shown in figure 6A and 6B. 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.

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

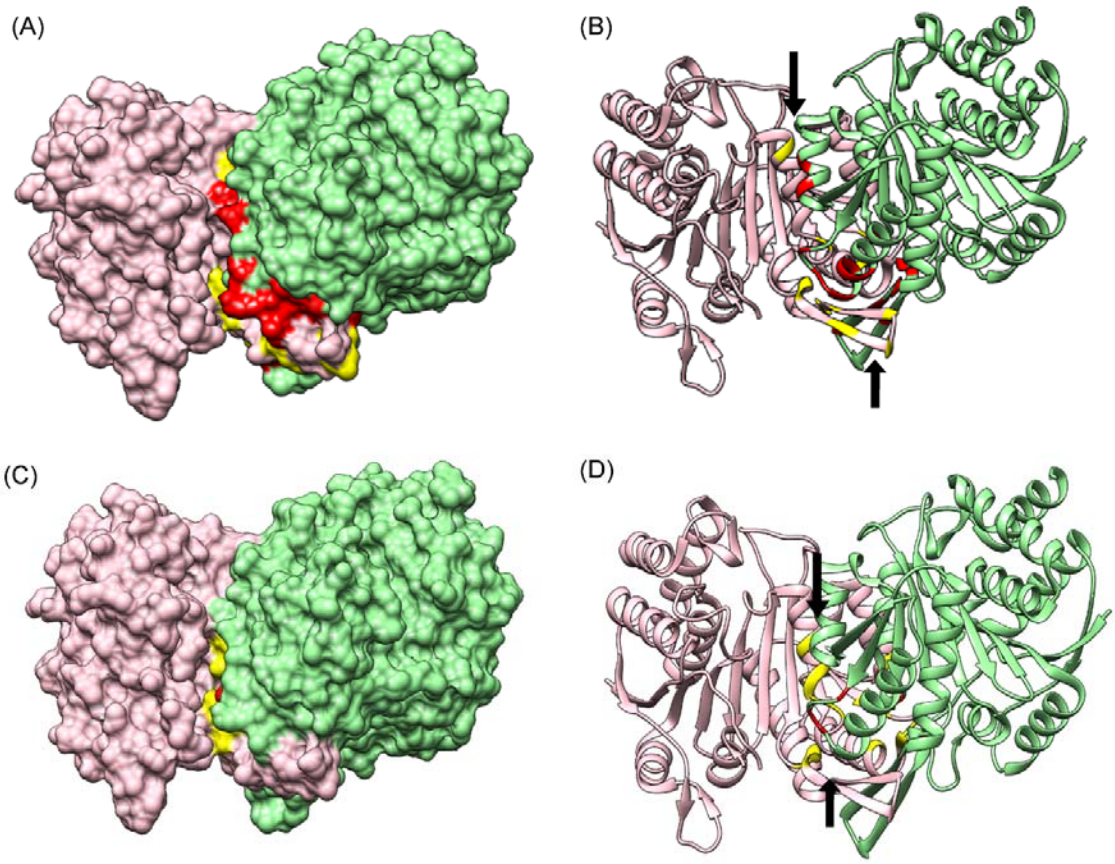

Figure 6: PDB ID 5YVT evaluated using CoRNeA and BIPSPI

A. Surface representation of 5YVT depicting interface residues predicted by CoRNeA. B. Cartoon representation of interface residues predicted by CoRNeA. C. Surface representation of 5YVT depicting interface residues predicted by BIPSPI. D. Cartoon representation of interface residues predicted by BIPSPI Chain A in pink and their respective interface residues are shown in yellow. Chain B in green and its interface residues are depicted in red. The black arrows indicate the regions of interface predicted by CoRNeA/BIPSPI.

Comparison with other methods

To access 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(Ruben Sanchez-Garcia *et al.* 2019) which is closest to our implementation. The sequence mode of prediction on BIPSPI server was employed for predicting the interface residues of 1GCQ and 5YVT. In case of 1GCQ, none of the predicted pairs had a prediction score more than 0.5 which is the threshold for any machine learning based method. Of the top 20 pairwise predictions obtained, only two pairs were found to be in the interface zone when mapped onto the structure. For 5YVT, 1234 pairs were reported by BIPSPI, above the threshold prediction value of 0.5, of which 24 were true interface forming pairs. The results

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

obtained were mapped onto the structure of 5YVT as shown in figure 6C and 6D. It was observed that the regions which spanned most of these predictions were smaller as compared to that predicted by CoRNeA (figure 6B). Moreover, 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 S6).

Table S6: Comparison of predictions from CoRNeA with BIPSPI

	Method	Expected no of residues within 10Å	Number of True positives with probability more than 0.5	Number of False Positives
PDB ID: 1GCQ	**BIPSPI**	**108**	**0**	**N/A**
	CoRNeA		**52**	**56**
PDB ID: 5YVT	**BIPSPI**	**164**	**24**	**1210**
	CoRNeA		**24**	**968**

The numbers depicted for CoRNeA are post network analysis. 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. It is evident that the environmental features play a very important role in training the classifier and thus tweaking around the size and weights of the kernel matrix can be performed to generate the derived features and yield in better and specific results.

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 rated 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 specially for large multimeric complexes. CoRNeA can be utilized to identify the minimal interacting regions in the heterodimers

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 case 3D structure models (experimental or homology based) are available.

Author Contributions

References

Afsar Minhas F ul A, Geiss BJ, Ben-Hur A. PAIRpred: Partner-specific prediction of interacting residues from sequence and structure. *Proteins Struct Funct Bioinforma* 2014;**82**:1142–55.

Biro JC. Amino acid size, charge, hydropathy indices and matrices for protein structure analysis. *Theor Biol Med Model* 2006;**3**:1–12.

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**:21–35.

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.

Finn RD, Clements J, Arndt W *et al.* HMMER web server: 2015 Update. *Nucleic Acids Res* 2015;**43**:W30–8.

Godwin B, Uetz P, Johnston M *et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 2000;**403**:623–7.

Goncearenco A, Shaytan AK, Shoemaker BA *et al.* Structural Perspectives on the Evolutionary Expansion of Unique Protein-Protein Binding Sites. *Biophys J* 2015;**109**:1295–306.

Gouy M, Milleret F, Mugnier C *et al.* ACNUC: a nucleic acid sequence data base and analysis system. *Nucleic Acids Res* 1984;**12**:121–7.

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

518 Green AG, Marks DS, Bonvin AMJJ *et al.* Sequence co-evolution gives 3D contacts and
519 structures of protein complexes. *Elife* 2014;**3**:1–45.

520 Heffernan R, Yang Y, Paliwal K *et al.* Capturing non-local interactions by long short-term
521 memory bidirectional recurrent neural networks for improving prediction of protein
522 secondary structure, backbone angles, contact numbers and solvent accessibility.
523 *Bioinformatics* 2017;**33**:2842–9.

524 Honavar V, Jordan RA, EL-Manzalawy Y *et al.* Predicting protein-protein interface residues
525 using local surface structural similarity. *BMC Bioinformatics* 2012;**13**:41.

526 Jones DT. Protein secondary structure prediction based on position-specific scoring matrices.
527 *J Mol Biol* 1999;**292**:195–202.

528 Jones S, Thornton JM. PROTEIN-PROTEIN INTERACTIONS: A REVIEW OF PROTEIN
529 DIMER STRUCTURES. *Prog Biophys molec Biol* 1995;**63**:31–65.

530 Kastritis PL, Moal IH, Hwang H *et al.* A structure-based benchmark for protein-protein
531 binding affinity. *Protein Sci* 2011;**20**:482–91.

532 Krissinel E, Henrick K. Inference of Macromolecular Assemblies from Crystalline State. *J*
533 *Mol Biol* 2007;**372**:774–97.

534 Kufareva I, Budagyan L, Raush E *et al.* PIER: Protein Interface Recognition for Structural
535 Proteomics. *PROTEINS Struct Funct Bioinforma* 2007;**67**:400–17.

536 Liang S, Zhang C, Liu S *et al.* Protein binding site prediction using an empirical scoring
537 function. *Nucleic Acids Res* 2006;**34**:3698–707.

538 Liu Y, Hu L, Ma T *et al.* Insights into the inhibitory mechanisms of NADH on the αγ
539 heterodimer of human NAD-dependent isocitrate dehydrogenase. *Sci Rep* 2018;**8**:1–12.

540 Lockless SW. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein
541 Families. *Science (80-)* 2002;**286**:295–9.

542 Maheshwari S, Brylinski M. Template-based identification of protein-protein interfaces using
543 eFindSitePPI. *Methods* 2016;**93**:64–71.

544 Masters SC. Co-Immunoprecipitation from Transfected Cells BT - Protein-Protein
545 Interactions: Methods and Applications. In: Fu H (ed.). Totowa, NJ: Humana Press,
546 2004, 337–48.

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

547 Miyazawa S, Jernigan RL. Residue-Residue Potentials with a Favorable Contact Pair Term
548 and an Unfavorable High Packing Density Term, for Simulation and Threading - 1-s2.0-
549 S002228369690114X-main.pdf. *J Mol Biol* 1996:623–44.

550 Murakami Y, Mizuguchi K. Applying the Naïve Bayes classifier with kernel density
551 estimation to the prediction of protein-protein interaction sites. *Bioinformatics*
552 2010;**26**:1841–8.

553 Negi, S.S.; Catherine, H.S.; Oezguen, N.; Power TD. BW. InterProSurf: a web server for
554 predicting interacting sites on protein Surfaces. *Bioinformatics* 2007;**23**:3397–9.

555 Neuvirth H, Raz R, Schreiber G. ProMate: A structure based prediction program to identify
556 the location of protein-protein binding sites. *J Mol Biol* 2004;**338**:181–99.

557 Nishida M, Nagata K, Hachimori Y *et al.* Novel recognition mode between Vav and Grb2
558 SH3 domains. *EMBO J* 2001;**20**:2995–3007.

559 Pedregosa F, Varoquaux G, Gramfort A *et al.* Scikit-learn: Machine Learning in Python. *J*
560 *Mach Learn Res* 2011;**12**:2825–30.

561 Pei J, Kim BH, Grishin N V. PROMALS3D: A tool for multiple protein sequence and
562 structure alignments. *Nucleic Acids Res* 2008;**36**:2295–300.

563 Porollo A, Meller J. Prediction-Based Fingerprints ofProtein–Protein Interactions.
564 *PROTEINS Struct Funct Bioinforma* 2007;**66**:630–45.

565 Qin S, Zhou HX. Meta-PPISP: A meta web server for protein-protein interaction site
566 prediction. *Bioinformatics* 2007;**23**:3386–7.

567 Rodriguez-Rivas J, Marsili S, Juan D *et al.* Conservation of coevolving protein interfaces
568 bridges prokaryote–eukaryote homologies in the twilight zone. *Proc Natl Acad Sci*
569 2016;**113**:15018–23.

570 Ruben Sanchez-Garcia, Sorzano COS, Carazo JM and *et al.* BIPSPI: a method for the
571 prediction of Partner- Specific Protein-Protein Interfaces. *Bioinformatics* 2019;**35**:470–
572 7.

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

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

576 Sobott F, Robinson C V. Protein complexes gain momentum. *Curr Opin Struct Biol*
577 2002;**12**:729–34.

578 de Vries SJ, Bonvin AMJJ. Cport: A consensus interface predictor and its performance in
579 prediction-driven docking with HADDOCK. *PLoS One* 2011;**6**, DOI:
580 10.1371/journal.pone.0017695.

581 Vries SJ de, Dijk ADJ van, Bonvin AMJJ. WHISCY: What Information Does Surface
582 Conservation Yield? Application to Data-Driven Docking. *PROTEINS Struct Funct*
583 *Bioinforma* 2006;**63**:479–89.

584 Waterhouse AM, Procter JB, Martin DMA *et al.* Jalview Version 2-A multiple sequence
585 alignment editor and analysis workbench. *Bioinformatics* 2009;**25**:1189–91.

586 Weigt M, White RA, Szurmant H *et al.* Identification of direct residue contacts in protein–
587 protein interaction by message passing. *Proc Natl Acad Sci U S A* 2009;**106**:67–72.

588 Wyner AD. A definition of conditional mutual information for arbitrary ensembles. *Inf*
589 *Control* 1978;**38**:51–9.

590 Xue LC, Dobbs D, Honavar V. HomPPI: A class of sequence homology based protein-
591 protein interface prediction methods. *BMC Bioinformatics* 2011;**12**, DOI: 10.1186/1471-
592 2105-12-244.

593 Zeng H, Liu K-S, Zheng W-M. The Miyazawa-Jernigan Contact Energies Revisited. *Open*
594 *Bioinforma J* 2012;**6**:1–8.

595 Zeng H, Wang S, Zhou T *et al.* ComplexContact: A web server for inter-protein contact
596 prediction using deep learning. *Nucleic Acids Res* 2018;**46**:W432–7.

597 Zhang QC, Deng L, Fisher M *et al.* PredUs: A web server for predicting protein interfaces
598 using structural neighbors. *Nucleic Acids Res* 2011;**39**:283–7.

599 **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

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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**

bioRxiv preprint doi: https://doi.org/10.1101/741280; this version posted August 31, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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**

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