Review of the Recent Protein-Protein Interaction Techniques

Maad Shatnawi
College of Information Technology, United Arab Emirates University, Al-Ain, Abu Dhabi, UAE
E-mail: shatnawi@uaeu.ac.ae

Abstract— *Protein-protein interactions (PPI) play a crucial role in cellular functions and biological processes in all organisms. The identification of protein interactions can lead to a better understanding of infection mechanisms and the development of several medication drugs and treatment optimization. Several physiochemical experimental techniques have been applied to identify PPIs. However, these techniques are computationally expensive, significantly time consuming, and have covered only a small portion of the complete PPI networks. As a result, the need for computational techniques has been increased to validate experimental results and to predict non-discovered PPIs. This review investigates and compares most of the recent computational PPI prediction techniques, and discusses the technical challenges and open issues in this domain.*

Keywords: Protein-protein interaction prediction, PPI, protein sequences, computational techniques

1. Introduction

Proteins are the building blocks of all living organisms. The primary structure of a protein is the linear sequence of its amino acid (AA) units starting from the amino-terminal residue (N-terminal) to the carboxyl-terminal residue (C-terminal). Amino acids consist of carbon, hydrogen, oxygen, and nitrogen atoms that are clustered into functional groups. All amino acids have the same general structure, but each has a different R-group. The carbon atom to which the R group is connected is called the alpha carbon. There are twenty amino acids in proteins and are connected by a chemical reaction in which a molecule of water is removed, leaving two amino acids residues connected by a peptide bond. These twenty amino acids are Alanine, Arginine, Asparagine, Aspartic acid, Cysteine, Glutamic acid, Glutamine, Glycine, Histidine, Isoleucine, Leucine, Lysine, Methionine, Phenylalanine, Proline, Serine, Threonine, Tryptophan, Tyrosine, and Valine. These amino acids are represented by one-letter abbreviation as A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, and V [1].

The secondary structure of a protein is the general three-dimensional form of its local parts. The most common secondary structures are alpha (α) helices and beta (β) sheets. The α-helix is a right-handed spiral array while the β sheet is made up of beta strands connected crosswise by two or more hydrogen bonds, forming a twisted pleated sheet. These secondary structures are linked together by tight turns and loose flexible loops [1]. Protein domains are the basic functional units of protein tertiary structures. A protein domain is a conserved part of a protein sequence that can evolve, function, and exist independently. Each domain forms a three-dimensional structure and can be stable and folded independently. Several domains are joined together in different combinations forming multi-domain protein sequences [2], [3].

A protein interacts with other proteins in order to perform certain tasks. Protein-protein interaction (PPI) occurs at almost every level of cell functions. The identification of interactions among proteins provides a global picture of cellular functions and biological processes. Since most biological processes involve one or more protein-protein interactions, the accurate identification of the set of interacting proteins in an organism is very useful for deciphering the molecular mechanisms underlying given biological functions and for assigning functions to unknown proteins based on their interacting partners [4]–[6]. Protein interaction prediction is also a fundamental step in the construction of PPI networks for human and other organisms. The identification of possible viral-host protein interactions can lead to a better understanding of infection mechanisms and, in turn, to the development of several medication drugs and treatment optimization. Abnormal PPIs have implications in several neurological disorders such as Creutzfeld-Jacob and Alzheimer [7]–[9]. Therefore, the development of accurate and reliable methods for identifying PPIs has very important impacts in several protein research areas.

This review, as an extension of our paper shatnawi (2014) [10], provides a comprehensive comparative study and categorization of the existing computational approaches in PPI prediction and discuss the technical challenges and open issues in this field. The rest of this review is organized as follows. Next section addresses the key technical challenges that face PPI prediction and the open issues in this field. Section 3 discusses the performance measures that are typically used in PPI prediction. Section 4 provides a comprehensive description and comparison of the most current computational PPI predictors. Concluding remarks are presented in Section 5.

2. Technical Challenges and Open Issues

There are several technical challenges that face computational analysis of protein sequences in general and

PPI prediction in particular. First, there have been a huge amount of newly discovered protein sequences in the past genomic era. Second, protein chains are typically long which are difficult, time-consuming, and expensive to characterize by experimental methods. Third, the availability of large, comprehensive, and accurate benchmark datasets is required for the training and evaluation of prediction methods. Fourth, appropriate performance measures to evaluate the significance of the predictors should be developed to minimize the number of results that give false positives and false negatives. Fifth, it is difficult to distinguish between novel interactions and false positives. Sixth, computational PPI methods are based on experimentally collected data, and therefore, any error in the experimental data will effect the computational PPI predictions.

One of the challenges of protein prediction methods is protein representation. Protein prediction methods vary in protein representation and feature extraction in order to build their classification models. There are two kinds of models that were generally used to represent protein samples; the sequential model and the discrete model. The most and simplest sequential model for a protein is its entire AA sequence. However, this representation does not work well when the query protein does not have high sequence similarity to any attribute-known proteins. Several non-sequential models, or discrete models, have been proposed. The simplest discrete model is AA composition which is the normalized occurrence frequencies of the twenty native amino acids in a protein. However, all the sequence-order knowledge will be lost using this representation which, in turn, will negatively affect the prediction quality [11]. Some approaches use AA physiochemical properties. Other approaches use pairwise similarity. Some approaches are template-based while others are statistical-based or machine learning-based.

There are various challenges that face machine-learning (ML) protein interaction prediction methods. Selecting the best ML approach is a great challenge. There is a variety of techniques that diverse in accuracy, robustness, complexity, computational cost, data diversity, over-fitting, and ability to deal with missing attributes and different features. Most ML approaches of protein sequences are computationally expensive and often suffer from low prediction accuracy. They are further susceptible to overfitting [12].

Most PPI prediction approaches have achieved reasonable performance on balanced datasets containing equal number of interacting and non-interacting protein pairs. However, this ratio is highly unbalanced in nature and these approaches have not been comprehensively assessed with respect to the effect of the large number of non-interacting pairs in realistic datasets. In addition, since highly unbalanced distributions usually lead to large datasets, more efficient prediction methods, algorithmic optimizations and continued improvements in hardware performance are required to handle such challenging tasks.

3. Performance Measures

There are several performance measures that are used to evaluate a PPI predictor and compare it with other approaches. The most frequently used evaluation measures in this field are accuracy, sensitivity, specificity, precision, $F1, MCC, ROC$, and AUC.

Accuracy (Ac) is the proportion of correctly predicted interacting and non-interacting protein pairs to all of the protein pairs listed in the dataset. Sensitivity, or recall (R), is the proportion of correctly predicted interacting protein pairs to all of the interacting protein pairs listed in the dataset. Precision (P) is the proportion of correctly predicted interacting protein pairs to all of the predicted interacting protein pairs. Specificity (Sp) is the proportion of correctly predicted non-interacting protein pairs to all the non-interacting protein pairs listed in the dataset. These metrics can be represented mathematically as follows:

$$Ac = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$Sp = \frac{TN}{TN + FP} \tag{4}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

The F-measure ($F1$) is an evaluation metric that combines precision and recall into a single value and is defined as the harmonic mean of precision and recall [13], [14].

$$F1 = \frac{2PR}{P + R} \tag{5}$$

Mathew correlation coefficient (MCC) is a measure that balances prediction sensitivity and specificity. MCC ranges from -1, indicating an inverse prediction, through 0, which corresponds to a random classifier, to +1 for perfect prediction.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{6}$$

The Receiver Operating Characteristic (ROC) curve is created by plotting the true positive rate (Recall) against the false positive rate (1- Specificity) at various threshold settings. AUC is the area under the ROC curve. AUC represents the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. The AUC can also be interpreted as the average recall over the entire range of possible specificity, or the average specificity over the entire range of possible recalls [15]–[17].

4. Computational Approaches

PPI prediction has been studied extensively by several researchers and a large number of approaches have been proposed. These approaches can be classified into physiochemical experimental and computational approaches. Physiochemical experimental techniques identify the physiochemical interactions between proteins which, in turn, are used to predict the functional relationships between them. These techniques include yeast two-hybrid based methods [18], mass spectrometry [19], Tandem Affinity Purification [20], protein chips [21], and hybrid approaches [22]. Although these techniques have succeeded in identifying several important interacting proteins in several species such as Yeast, Drosophila, and Helicobacter-pylori [23], they are computationally expensive and significantly time consuming, and so far the predicted PPIs have covered only a small portion of the complete PPI network. As a result, the need for computational tools has been increased in order to validate physiochemical experimental results and to predict non-discovered PPIs [4], [24].

Several computational methods have been proposed for PPI prediction and can be classified according to the used protein features into sequence-based and structure-based methods. Sequence-based methods utilize AA features and can be further categorized into statistical and Machine Learning (ML)-based methods. The structure-based methods use three-dimensional structural features [25] and can be categorized into template-based, statistical and ML-based methods. This section provides an overview and discussion of some of the current computational sequence-based and structure-based PPI prediction approaches.

4.1 Sequence-Based Approaches

Sequence-based PPI prediction methods utilize AA features such as hydrophobicity, physiochemical properties, evolutionary profiles, AA composition, AA mean, or weighted average over a sliding window [25]. Sequence-based methods can be categorized into statistical and Machine Learning (ML)-based methods. This section presents and evaluates some of the existing sequence-based approaches.

4.1.1 Statistical Sequence-Based Approaches

This section presents and describes several existing statistical sequence-based PPI prediction approaches.

Mirror Tree Method:

Pazos and Valencia [26] introduced the Mirror Tree Method based on the comparison of the evolutionary distances between the sequences of the associated protein families and using topological similarity of phylogenetic trees to predict PPIs. These distances were calculated as the average value of the residue similarities taken from the McLachlan amino acid homology matrix [27]. The similarity between trees was calculated as the correlation between the distance matrices used to build the trees.

The Mirror Tree Method does not require the creation of the phylogenetic trees but only the underlying distance matrices are analyzed, and therefore, this approach is independent of any given tree-construction method. Although the mirror tree method does not require the presence of fully sequenced genomes, it requires the presence of the orthologous proteins in all the species under consideration. As a result, when more species genomes become available, fewer proteins could be applied. In addition to that, the method is restricted to cases where at least eleven sequences were collected from the same species for both proteins. This minimum limit was set empirically as a compromise between being sufficiently small to provide enough cases and large enough for the matrices to contain sufficient information. The approach can be improved by increasing the number of possible interactions by collecting sequences from a larger number of genomes. Further, since the distance matrices are not a perfect representation of the corresponding phylogenetic trees, it is possible that some inaccuracies are introduced by comparing distance matrices instead of the real phylogenetic trees.

PIPE:

Pitre *et al.* [28] introduced PIPE (Protein-protein Interaction Prediction Engine) to estimate the likelihood of interactions between pairs of the yeast Saccharomyces cerevisiae proteins using protein primary structure information. PIPE is based on the assumption that interactions between proteins occur by a finite number of short polypeptide sequences observed in a database of known interacting protein pairs. These sequences are typically shorter than the classical domains and reoccur in different proteins within the cell. PIPE estimates the likelihood of a PPI by measuring the reoccurrence of these short polypeptides within known interacting proteins pairs. To determine whether two proteins A and B interact, the two query proteins are scanned for similarity to a database of known interacting proteins pairs. For each known interacting pair (X, Y), PIPE uses sliding windows to compares the AA residues in protein A against that in X and protein B against Y, and then measures how many times a window of protein A finds a match in X and at the same time a window in protein B matches a window in Y. These matches are counted and added up in a 2D matrix. A positive protein interaction is predicted when the reoccurrence count in certain cells of the matrix exceed a predefined threshold value. PIPE was evaluated on a randomly selected set of 100 interacting yeast protein pairs and 100 non-interacting proteins from the database of interacting proteins (DIP) (http://dip.doe-mbi.ucla.edu) [29] and MIPS [30] databases. PIPE showed a prediction sensitivity of 0.61 and specificity of 0.89.

Since PIPE is based on protein primary structure information without any previous knowledge about the higher structure, domain composition, evolutionary conservation or the function of the target proteins. It can identify interactions of protein pairs for which limited structural information is available. The limitations of PIPE are as follows. PIPE is computationally intensive and requires hours of computation per protein pair as it scans the interaction library repeatedly every time. Second, PIPE shows weakness in detecting novel interactions among genome wide large-scale datasets as it reported a large number of false positives. Third, PIPE was evaluated on uncertain data of interactions that were determined using several methods, each having a limited accuracy.

Pitre *et al.* [31] then developed PIPE2 as an improved and more efficient version of PIPE which showed a specificity of 0.999. PIPE2 represents AA sequences in a binary code which speeds up searching the similarity matrix. Unlike the original PIPE that scans the interaction database repeatedly every time, PIPE2 pre-computes all window comparisons in advance and stores them on a local disk.

Although PIPE2 achieves a high specificity, it has a large number of false positives with a sensitivity of 0.146 only. False positives rate can be reduced by incorporating other information about the target protein pairs including sub-cellular localization or functional annotation. A major limitation of PIPE2 is that it relies exclusively on a database of pre-existing interaction pairs for the identification of re-occurring short polypeptide sequences and in the absence of sufficient data, PIPE2 will be ineffective. PIPE2 is also less effective for motifs that span discontinuous primary sequence as it does not account for gaps within the short polypeptide sequences.

Co-evolutionary Divergence:

Liu *et al.* [32] introduced a sequence-based co-evolution PPI prediction method in the human proteins. The authors defined the co-evolutionary divergence (CD) based on two assumptions. First, PPI pairs may have similar substitution rates. Second, protein interaction is more likely to conserve across related species. CD is defined as the absolute value of the substitution rate difference between two proteins. CD can be used to predict PPIs as the CD values of interacting protein pairs are expected to be smaller than those of non-interacting pairs. The method was evaluated using 172,338 protein sequences obtained from Evola database [33] for Homo sapiens and their orthologous protein sequences in thirteen different vertebrates. The PPI dataset was downloaded from the Human Protein Reference Database [34]. Pairwise alignment of the orthologous proteins was made with ClustalW2 software. The absolute value of substitution rate difference between two proteins was used to measure the CDs of protein pairs which were then used to construct the likelihood ratio table of interacting protein pairs.

The CD method combines co-evolutionary information of interacting protein pairs from many species. The method does not use multiple alignments, thus taking less time than other alignment methods such as the mirror tree method. The method is not limited to proteins with orthologous across all species under consideration. However, increasing the number of species will provide more information to improve the accuracy of the co-evolutionary divergence method. Although this method could rank the likelihood of interaction for a given pair of proteins, it did not infer specific features of interaction such as the interacting residues in the interfaces.

Table 1 summarizes these statistical sequence-based approaches including the features that are used, the technique and/or the tools applied, and the validation datasets used.

4.1.2 Machine Learning Sequence-Based Approaches

This section describes several existing ML sequence-based PPI prediction approaches.

Auto Covariance:

Guo *et al.* [36] proposed a sequence-based method using Auto Covariance (AC) and Support Vector Machines (SVM). AA residues were represented by seven physicochemical properties. These properties are hydrophobicity, hydrophilicity, volumes of side chains, polarity, polarizability, solvent-accessible surface area, and net charge index of AA side chains. AC counts for the interactions between residues a certain distance apart in the sequence. AA physicochemical properties were analyzed by AC based on the calculation of covariance. A protein sequence was characterized by a series of ACs that covered the information of interactions between each AA residue and its 30 vicinal residues in the sequence. Finally, a SVM model with a Radial Basis Function (RBF) kernel was constructed using the vectors of AC variables as input. The optimization experiment demonstrated that the interactions of one AA residue and its 30 vicinal AAs would contribute to characterizing the PPI information. The software and datasets are available at http://www.scucic.cn/Predict_PPI/index.htm. A dataset of 11,474 yeast PPIs extracted from DIP [37] was used to evaluate the model and the average prediction accuracy, sensitivity, and precision achieved are respectively 0.86, 0.85, and 0.87.

One of the advantages of this approach is that AC includes long-range interaction information of AA residues which are important in PPI identification. The use of SVM as a predictor is another advantage. SVM is the state of the art ML technique and has many benefits and overcomes many limitations of other techniques. SVM has strong foundations in statistical learning theory [38] and has been successfully applied in various classification problems [39]. SVM offers several related computational advantages such as the lack of local minima in the optimization [40].

Table 1: Statistical Sequence-based PPI prediction approaches.

Approach	Extracted Features	Technique/Tool	Datasets
Mirror Tree (Pazos and Valencia 2001)	Similarity of phylogenetic trees	Evolutionary distance, McLachlan AA homology matrix	*Escherichia coli* protein (Dandekar *et al.* 1998) [35]
PIPE (Pitre *et al.* 2006), PIPE2 (Pitre *et al.* 2008)	Short AA polypeptides	Similarity measure	Yeast protein (DIP and MIPS)
Co-evolutionary Divergence (Liu *et al.* 2013)	Co-evolutionary information,	Pairwise alignment, ClustalW2	Human protein (Matsuya *et al.* 2008, Prasad *et al.* 2009)

Pairwise Similarity:

Zaki *et al.* [4] proposed a PPI predictor based on pairwise similarity of protein primary structure. Each protein sequence was represented by a vector of pairwise similarities against large AA subsequences created by a sliding window which passes over concatenated protein training sequences. Each coordinate of this vector is the E-value of the Smith-Waterman (SW) score [41]. These vectors were then used to compute the kernel matrix which was exploited in conjunction with a RBF-kernel SVM. Two proteins may interact by the means of the scores similarities they produce [42], [43]. Each sequence in the testing set was aligned against each sequence in the training set, counted the number of positions that have identical residues, and then divided by the total length of the alignment.

The method was evaluated on a dataset of yeast *Saccharomyces cerevisiae* proteins created by Chen and Liu [44] and contains 4917 interacting protein pairs and 4000 non-interacting pairs. The method achieved an accuracy of 0.78, a sensitivity of 0.81, a specificity of 0.744, and a ROC of 0.85.

SW alignment score provides a relevant measure of similarity between proteins. Therefore protein sequence similarity typically implies homology, which in turn may imply structural and functional similarity [45]. SW scores parameters have been optimized over the past two decades to provide relevant measures of similarity between sequences and they now represent core tools in computational biology [46]. The use of SVM as a predictor is another advantage. This work can be improved by combining knowledge about gene ontology, inter-domain linker regions, and interacting sites to achieve more accurate prediction.

AA Composition:

Roy *et al.* [47] examined the role of amino acid composition (AAC) in PPI prediction and its performance against well-known features such as domains, tuple feature, and signature product feature. Every protein pair was represented by AAC and domain features. AAC was represented by monomer and dimer features. Monomer features capture composition of individual amino acids, whereas dimer features capture composition of pairs of consecutive AAs. To generate the monomer features, a 20-dimensional vector representing the normalized proportion of the 20 AAs in a protein was created. The real-valued composition was then discretized into 25 bits producing a set of 500 binary features. To generate the dimer features, a 400-dimensional vector of all possible AA pairs were extracted from the protein sequence and discretized into 10 bits producing a set of 4000 binary features. The domains were represented as binary features with each feature identified by a domain name. To compare AAC against other non-domain sequence-based features, tuple features [48] and signature products [49] were obtained. The tuple features were created by grouping AAs into six categories based on their biochemical properties, and then creating all possible strings of length 4 using these categories. The signature products were obtained by first extracting signatures of length 3 from the individual protein sequences. Each signature consists of a middle letter and two flanking AAs represented in alphabetical order. Thus two 3-tuples with the first and third amino acid letter permuted have the same signature. The signatures were used to construct a signature kernel specifying the inner product between two proteins.

The proposed approach was examined using three machine learning classifiers (logistic regression, SVM, and the Naive Bayes) on PPI datasets from yeast, worm and fly. Three datasets for yeast *S. cerevisiae* were extracted from the General Repository for Interaction Datasets (GRID) database [50], TWOHYB (Yeast Two-hybrid), AFFMS (Affinity pull down with mass spectrometry), and PCA (protein complementation assay). In addition to that, a dataset each for worm, *C. elegans* (Biogrid dataset) [51] and fly, *D. melanogaster* [50] were used. The authors reported that AAC features performed almost equivalent contribution as domain knowledge across different datasets and classifiers which indicated that AAC captures significant information for identifying PPIs. AAC is a simple feature, computationally cheep, applicable to any protein sequence, and can be used when there is lack of domain information. AAC can be combined with other features to enhance PPI prediction.

AA Triad:

Yu *et al.* [52] proposed a probability-based approach of estimating triad significance to alleviate the effect of AA distribution in nature. The relaxed variable kernel density estimator (RVKDE) [53] was employed to predict PPIs based on AA triad information. The method is summarized as follows. Each protein sequence was represented as AA triads by considering every three continuous residues in the protein sequence as a unit. To reduce feature dimensionality vector, the 20 AA types were categorized into seven groups based on their dipole strength and side chain volumes [23]. The triads were then scanned one by one along the sequence, and each scanned triad is counted in an occurrence vector, O. Subsequently, a significance vector, S, was proposed to represent a protein sequence by estimating the probability of observing less occurrences of each triad than the one that is actually observed in O. Each PPI pair was then encoded as a feature vector by concatenating the two significance vectors of the two individual proteins. Finally, the feature vector was used to train a RVKDE PPI predictor. The method was evaluated on 37,044 interacting pairs within 9,441 proteins from the Human Protein Reference Database (HPRD) [54], [55]. Datasets with different positive-to-negative ratios (from 1:1 to 1:15) were generated with the same positive instances and distinct negative sets, which are obtained by randomly sampling from the negative instances. The authors concluded that the degree of dataset imbalance is important to PPI predictor behavior. With 1:1 positive-to-negative ratio, the proposed method achieves 0.81 sensitivity, 0.79 specificity, 0.79 precision, and 0.8 F-measure. These evaluation measures drop as the data gets more imbalanced to reach 0.39 sensitivity, 0.97 specificity, 0.495 precision, and 0.44 F-measure with 1:15 positive-to-negative ratio.

RVKDE is a ML algorithm that constructs a RBF neural network to approximate the probability density function of each class of objects in the training dataset. One main distinct feature of RVKDE is that it takes an average time complexity of $O(nlogn)$ for the model training process, where n is the number of instances in the training set. In order to improve the prediction efficiency, RVKDE considers only a limited number of nearest instances within the training dataset to compute the kernel density estimator of each class. One important advantage of RVKDE, in comparison with SVM, is that the learning algorithm generally takes far less training time with an optimized parameter setting. In addition to that, the number of training samples remaining after a data reduction mechanism is applied is quite close to the number of support vectors of SVM algorithm. Unlike SVM, RVKDE is capable of classifying data with more than two classes in one single run [53].

UNISPPI:

Valente et al. [56] (2013) introduced UNISPPI (Universal In Silico Predictor of Protein-Protein Interactions). The authors examined both the frequency and composition of the physicochemical properties of the twenty protein AAs to train a decision tree PPI classifier. The frequency feature set includes the percentages of each of the 20 AA in the protein sequence. The composition feature set was obtained by grouping each AA of a protein into one of three different groups related to seven physicochemical properties and calculating the percentage of each group for each feature ending up by a total of 21 composition features. The seven physicochemical properties are hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, secondary structure, and solvent accessibility. When tested on a dataset of PPI pairs of twenty different eukaryotic species including eukaryotes, prokaryotes, viruses, and parasite-host associations, UNISPPI correctly classified 0.79 of known PPIs and 0.73 of non-PPIs. The authors concluded that using only the AA frequencies was sufficient to predict PPIs. They further concluded that the AA frequencies of Asparagines (N), Cysteine (C), and Isoleucine (I) are important features for distinguishing between interacting and non-interacting protein pairs.

The main advantages of UNISPPI are its simplicity and low computational cost as small amount of features were used to train the decision tree classifier. Decision tree classifier is fast to build and has few parameters to tune. Decision trees can be easily analyzed and the features can be ranked according to their capabilities of distinguishing PPIs from non-PPIs. However, decision tree classifiers normally suffer from overfitting.

ETB-Viterbi:

Kern et al. [57] proposed the Early Traceback Viterbi (ETB-Viterbi) as a decoding algorithm with an early traceback mechanism in ipHMMs (Interaction Profile Hidden Markov Models) [58] which was designed to optimally incorporate long-distance correlations between interacting AA residues in input sequences. The method was evaluated on real data from the 3DID database [59] along with simulated data generated from 3DID data containing different degrees of correlation and reversed sequence orientation. ETB-Viterbi was capable to capture the long-distance correlations for improved prediction accuracy and was not much affected by sequence orientation. Hidden Markov models (HMMs) are powerful probabilistic modeling tool for analyzing and simulating sequences of symbols that are emitted from underlying states and not directly observable [60]. The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states. However, the Viterbi algorithm is expensive in terms of memory and computing time. The HMM training involves repeated iterations of the Viterbi algorithm which makes it quite slow. HMM Model may not converge to a truly optimal parameter set for a given training set as it can be trapped in local maxima, and can

Table 2: Machine Learning Sequence-based PPI prediction approaches.

Approach	Extracted Features	Technique/Tool	Datasets
Auto Covariance (Guo *et al.* 2008)	AA physicochemical properties	Auto covariance, SVM	Yeast protein (DIP and MIPS)
Pairwise Similarity (Zaki et al 2009)	Pairwise similarity	SVM	Yeast protein
AA Composition (Roy *et al.* 2009)	AAC	Logistic regression, SVM, Naive Bayes	Yeast protein (GRID), worm protein (Li *et al.* 2004), fly protein (Biogrid)
AA Triad (Yu *et al.* 2010)	AA triad information	RVKDE	Human protein (HPRD)
UNISPPI (Valente *et al.* 2013)	Frequency and composition of AA physiochemical properties	Decision tree	Twenty different eukaryotic species
ETB-Viterbi (Kern *et al.* 2013)	AA residues	Hidden Markov models, Early Traceback Viterbi	3DID database

suffer from overfitting [61]–[64].

Table 2 summarizes these ML sequence-based approaches and compared them in terms of features, techniques, tools, and validation datasets.

4.2 Structure-Based Approaches

Structure-based PPI prediction methods use three-dimensional structural features such as domain information, solvent accessibility, secondary structure states, and hydrophobic and polar surface locations [25]. Structure-based PPI prediction methods can be categorized into template-based, statistical, and ML-based methods. This section presents and evaluates some of the state-of-the-art structure-based approaches.

4.2.1 Template Structure-Based Approaches

PRISM

Tuncbag *et al.* [65] developed PRISM as a template-based PPI prediction method based on information regarding the interaction surface of crystalline complex structures. The two sides of a template interface are compared with the surfaces of two target monomers by structural alignment. If regions of the target surfaces are similar to the complementary sides of the template interface, then these two targets are predicted to interact with each other through the template interface architecture. The method can be summarized as follows. First, interacting surface residues of target chains are extracted using Naccess [66]. Second, complementary chains of template interfaces are separated and structurally compared with each of the target surfaces by using MultiProt [67]. Third, the structural alignment results are filtered according to threshold values, and the resulting set of target surfaces is transformed into the corresponding template interfaces to form a complex. Finally, the Fiber-Dock [68] algorithm is used to refine the interactions to introduce flexibility, compute the global energy of the complex, and rank the solutions according to their energies. When the computed energy of a protein pair is less than a threshold of -10 kcal/mol, the pair is determined to interact.

PRISM has been applied for predicting PPIs in a human apoptosis pathway [69] and a p53- protein-related pathway [70], and has contributed to the understanding of the structural mechanisms underlying some types of signal transduction. PRISM obtained a precision of 0.231 when applied to a human apoptosis pathway that consisted of 57 proteins.

PrePPI

Zhang *et al.* [71] proposed PrePPI (Predicting Protein-Protein Interactions) as a structural alignment PPI predictor based on geometric relationships between secondary structure information. Given a pair of query proteins A and B, representative structures for the individual subunits (M_A, M_B) are taken from the PDB (Protein Data Bank) [72] or from the ModBase [73] and SkyBase [74] homology model databases. Close and remote structural neighbors are found for each subunit. A template for the interaction exists if a PDB or PQS [75] structure contains interacting pairs that are structural neighbors of M_A and M_B. A model is constructed by superposing the individual subunits, M_A and M_B, on their corresponding structural neighbors. The likelihood for each model to represent a true interaction is then calculated using a Bayesian Network trained on 11,851 yeast interactions and 7,409 human interactions datasets. Finally the structure-derived score is combined with non-structural information, including co-expression and functional similarity, into a naive Bayes classifier.

Although template-based methods can achieve high prediction accuracy when close templates are retrieved, the accuracy significantly decreases when the sequence identity of target and template is low.

4.2.2 Statistical Structure-Based Approaches

PID Matrix Score

Kim *et al.* [6] presented the Potentially Interacting Domain pair (PID) matrix as a domain-based PPI prediction algorithm. The PID matrix score was constructed as a measure of interactability (interaction probability) between domains. The algorithm analysis was based on the DIP (Database of Interacting Proteins) which contains more than ten thousand of mostly experimentally verified interacting protein pairs. Domain information was extracted from InterPro [76] which is an integrated database of protein families, domains and functional sites. Cross validation was performed with subsets of DIP data (positive datasets) and randomly generated protein pairs from TrEMBL/SwissProt database (negative datasets). The method achieved 0.50 sensitivity and 0.98 specificity. The authors reported that the PID matrix can also be used in the mapping of the genome-wide interaction networks.

PreSPI

Han *et al.* [77], [78] proposed a domain combination-based method which considers all possible domain combinations as the basic units of protein interactions. The domain combination interaction probability is based on the number of interacting protein pairs containing the domain combination pair and the number of domain combinations in each protein. The method considers the possibility of domain combinations appearing in both interacting and non-interacting sets of protein pairs. The ranking of multiple protein pairs were decided by the interacting probabilities computed through the interacting probability equation.

The method was evaluated using an interacting set of protein pairs in yeast acquired from DIP database [29], and a randomly generated non-interacting set of protein pairs. The domain information for the proteins was extracted from the PDB (http://www.ebi.ac.uk/proteome/) [72], [76]. PreSPI achieved a sensitivity of 0.77 and a specificity of 0.95.

PreSPI suffers from several limitations. First, this method ignores other domain-domain interaction information between the protein pairs. Second, it assumes that one domain combination is independent of another. Third, the method is computationally expensive as all possible domain combinations are considered.

Domain Cohesion and Coupling

Jang *et al.* [79] proposed a domain cohesion and coupling (DCC)-based PPI prediction method using the information of intra-protein domain interactions and inter-protein domain interactions. The method aims to identify which domains are involved in a PPI by determining the probability of the domains causing the proteins to interact irrespective of the number of participating domains. The coupling powers of all domain interaction pairs are stored in an interaction significance (IS) matrix which is used to predict PPIs. The method was evaluated on S. cerevisiae proteins and achieved 0.82 sensitivity and 0.83 specificity. The domain information for the proteins was extracted from Pfam (http://pfam.sanger.ac.uk) [80], which is a protein domain family database that contains multiple sequence alignments of common domain families.

MEGADOCK

Ohue *et al.* [81] developed MEGADOCK as a protein-protein docking software package using the real Pairwise Shape Complementarity (rPSC) score. First, they conducted rigid-body docking calculations based on a simplified energy function considering shape complementarities, electrostatics, and hydrophobic interactions for all possible binary combinations of proteins in the target set. Using this process, a group of high-scoring docking complexes for each pair of proteins were obtained. Then, ZRANK [82] was applied for more advanced binding energy calculation and re-ranked the docking results based on ZRANK energy scores. The deviation of the selected docking scores from the score distribution of high-ranked complexes was determined as a standardized score (Z-score) and was used to assess possible interactions. Potential complexes that had no other high-scoring interactions nearby were rejected using structural differences. Thus binding pairs that had at least one populated area of high-scoring structures were considered. MEGADOCK has been applied for PPI prediction for 13 proteins of a bacterial chemotaxis pathway [83], [84] and obtained a precision of 0.4. MEGADOCK is available at http://www.bi.cs.titech.ac.jp/megadock.

One of the limitations of this approach is the demerit of generating false-positives for the cases in which no similar structures are seen in known complex structure databases.

Meta Approach

Ohue *et al.* [85] proposed a PPI prediction approach based on combining template-based and docking methods. The approach applies PRISM [65] as a template-matching method and MEGADOCK [81] as a docking method. A protein pair is considered to be interacting if both PRISM and MEGADOCK predict that this protein pair interacts. When applied to the human apoptosis signaling pathway, the method obtained a precision of 0.333, which is higher than that achieved using individual methods (0.231 for PRISM and 0.145 for MEGADOCK), while maintaining an F1 of 0.285 comparable to that obtained using individual methods (0.296 for PRISM, and 0.220 for MEGADOCK).

Meta approaches have already been used in the field of protein tertiary structure prediction [86], and critical experiments have demonstrated improved performance of Meta predictors when compared with individual methods. The Meta approach has also provided favorable results in protein domain prediction [87] and the prediction of disordered regions in proteins [88]. Although some true positives may be dropped by this method, the remaining predicted pairs are expected to have higher reliability because of the consensus between two prediction methods that have different characteristics.

4.2.3 Machine Learning Structure-Based Approaches

Random Forest

Chen and Liu [44] introduced a domain-based Random Forest PPI predictor. Protein pairs were characterized by the domains existing in each protein. The protein domain information were collected from Pfam database [89]. Each protein pair was represented by a vector of features where each feature corresponds to a Pfam domain. If a domain exists in both proteins, then the associated feature value is 2. If the domain exists in one of the two proteins, then its associated feature value is 1. If a domain does not exists in both proteins, then the feature value is 0. These domain features were used to train a Random Forest PPI classifier. The random decision forest constructs many decision trees and each is grown from a different subset of training samples and random subset of feature and the final classification of a given protein pair is determined by majority votes among the classes decided by the forest of trees.

When evaluated on a dataset containing 9834 yeast protein interaction pairs among 3713 proteins, and 8000 negative randomly generated samples, the method achieved a sensitivity of 0.8 and a specificity of 0.64.. Yeast PPI data was collected from the DIP [29], [37], Deng *et al.* [90], Schwikowski *et al.* [91]. The dataset of Deng *et al.* is a combined interaction data experimentally obtained through two hybrid assays on Saccharomyces cerevisiae by Uetz *et al.* [92] and Ito *et al.* [93]. Schwikowski *et al.* gathered their data from yeast two-hybrid, biochemical and genetic data.

Random Forest classifier has several advantages. It is relatively fast, simple, robust to outliers and noise, easily parallelized, avoids overfitting, and performs well in many classification problems [94], [95]. Random Forest shows a significant performance improvement over the single tree classifiers. It interprets the importance of the features using measures such as decrease mean accuracy or *Gini* importance [96]. RF benefit from the randomization of decision tress as they have low-bias and high variance. Random Forest has few parameters to tune and less dependent on tuning parameters [97], [98]. However, the computational cost of Random Forest increases as the number of generated tress increases. One of the limitations of this approach is that PPI prediction depends on domain knowledge so proteins without domain information cannot provide any useful information for prediction. Therefore, the method excluded the pairs where at least one of the proteins has no domain information.

Struct2Net:

Singh *et al.* [99] introduced Struct2Net as a structure-based PPI predictor. The method predicts interactions by threading each pair of protein sequences into potential structures in the Protein Data Bank (PDB) [72]. Given two protein sequences (or one sequence against all sequences of a species), Struct2Net threads the sequence to all the protein complexes in the PDB and then chooses the best potential match. Based on this match, it uses logistic regression technique to predict whether the two proteins interact.

Later on, Singh *et al.* [100] introduced Struct2Net as a web server with multiple querying options which is available at http://struct2net.csail.mit.edu. Users can retrieve Yeast, fly, and human PPI predictions by gene name or identifier while they can query for proteins of other organisms by AA sequence in FASTA format. Struct2Net returns a list of interacting proteins if one protein sequence is provided and an interaction prediction if two sequences are provided. When evaluated on yeast and fly protein pairs, Struct2Net achieves a recall of 0.80 with a precision of 0.30.

A common limitation of all structure-based PPI prediction approaches is the low coverage as the number of known protein structures is much smaller than the number of known protein sequences, and therefore, such approaches fail when there is no structural template available for the queried protein pair. Table 3 summarizes these structure-based approaches and compared them in terms of features, techniques, tools, and validation datasets.

5. Conclusion

This chapter provides a review of the computational techniques for protein-protein interaction prediction including the open issues and main challenges in this domain. We investigated several relevant existing approaches and provided a categorization and comparison of them. It is clearly noticed that PPI prediction still needs much research effort in order to achieve reasonable prediction accuracy. One of the issues in the PPI prediction methods is that they do not use a uniform dataset and evaluation measure. We recommend creating a freely available standard benchmark dataset taking into consideration the biological properties of proteins and examining the performance of all these methods on this benchmark dataset using a well-defined evaluation measures. This will allow researchers to compare the performance of these prediction methods in a fair and uniform fashion. This work can be extended by investigating more recently published PPI prediction techniques, analyze them in depth, and compare their performance on a uniform dataset according to a uniform evaluation metrics. More focus should be given to the techniques which incorporate biological knowledge into the prediction process.

References

[1] J. M. Berg, J. L. Tymoczko, and L. Stryer, "Protein structure and function," 2002.

Table 3: Structure-based PPI prediction approaches.

Approach	Extracted Features	Technique/Tool	Datasets
PRISM (Tuncbag *et al.* 2011)	Interaction surface of crystalline complex structures	Naccess, MultiProt, Fiber-Dock	Human Protein (Ozbabacan et al. 2012, Tuncbag *et al.* 2009)
PrePPI (Zhang *et al.* 2012)	Secondary structure	Bayesian networks, Naive Bayes	Yeast protein, Human protein
PID Matrix Score (Kim *et al.* 2002)	Potentially interacting domain pairs	PID matrix	DIP, InterPro, TrEMBL/SwissProt
PreSPI (Han *et al.* 2003, 2004)	Domain combination interaction probability	Interacting probability equation	Yeast protein (DIP), PDB
DCC (Jang *et al.* 2012)	Intra-protein and inter-protein domain interactions	Interaction significance matrix	S. cerevisiae protein, Pfam
MEGADOCK (Ohue *et al.* 2013a)	Shape complementarities, electrostatics, and hydrophobic interactions	rPSC, ZRANK	Bacterial protein (Ohue *et al.* 2012, Matsuzaki *et al.*. 2013)
Meta Approach (Ohue *et al.* 2013b)	Interaction surface of crystalline complex structures, shape complementarities, electrostatics, and hydrophobic interactions	PRISM, MEGADOCK	Human protein
Random Forest (Chen and Liu 2005)	Existence of similar domains	Random Forest	DIP, Deng *et al.*, Schwikowski *et al.*, Pfam
Struct2Net (Singh *et al.* 2006, 2010)	Homology with known protein complexes in PDB	Logistic regression	Yeast, Fly ,and Human protein

[2] C. Chothia, “Proteins. one thousand families for the molecular biologist.” *Nature*, vol. 357, no. 6379, p. 543, 1992.

[3] P. D. Yoo, A. R. Sikder, J. Taheri, B. B. Zhou, and A. Y. Zomaya, “Domnet: protein domain boundary prediction using enhanced general regression network and new profiles,” *NanoBioscience, IEEE Transactions on*, vol. 7, no. 2, pp. 172–181, 2008.

[4] N. Zaki, S. Lazarova-Molnar, W. El-Hajj, and P. Campbell, “Protein-protein interaction based on pairwise similarity,” *BMC bioinformatics*, vol. 10, no. 1, p. 150, 2009.

[5] I. Xenarios and D. Eisenberg, “Protein interaction databases,” *Current Opinion in Biotechnology*, vol. 12, no. 4, pp. 334–339, 2001.

[6] W. K. Kim, J. Park, J. K. Suh, *et al.*, “Large scale statistical prediction of protein-protein interaction by potentially interacting domain (pid) pair,” *Genome Informatics Series*, pp. 42–50, 2002.

[7] G. D. Bader and C. W. Hogue, “Analyzing yeast protein–protein interaction data obtained from different sources,” *Nature biotechnology*, vol. 20, no. 10, pp. 991–997, 2002.

[8] C. Von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, “Comparative assessment of large-scale data sets of protein–protein interactions,” *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[9] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman, “Evaluation of different biological data and computational classification methods for use in protein interaction prediction,” *Proteins: Structure, Function, and Bioinformatics*, vol. 63, no. 3, pp. 490–500, 2006.

[10] M. Shatnawi, “Computational methods for protein-protein interaction prediction,” in *BIOCOMP’14*, 2014.

[11] K.-C. Chou, “Some remarks on protein attribute prediction and pseudo amino acid composition,” *Journal of theoretical biology*, vol. 273, no. 1, pp. 236–247, 2011.

[12] J. C. Melo, G. Cavalcanti, and K. Guimaraes, “Pca feature extraction for protein structure prediction,” in *Neural Networks, 2003. Proceedings of the International Joint Conference on*, vol. 4. IEEE, 2003, pp. 2952–2957.

[13] Y. Sasaki, “The truth of the f-measure,” *Teach Tutor mater*, pp. 1–5, 2007.

[14] D. Powers, “Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[15] T. Fawcett, “An introduction to roc analysis,” *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[16] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (roc) curve.” *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[17] C. E. Metz, “Basic principles of roc analysis,” in *Seminars in nuclear medicine*, vol. 8, no. 4. Elsevier, 1978, pp. 283–298.

[18] P. L. Bartel and S. Fields, *The yeast two-hybrid system*. Oxford University Press, 1997.

[19] A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, *et al.*, “Functional organization of the yeast proteome by systematic analysis of protein complexes,” *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[20] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Séraphin, “A generic protein purification method for protein complex characterization and proteome exploration,” *Nature biotechnology*, vol. 17, no. 10, pp. 1030–1032, 1999.

[21] H. Zhu, M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek, *et al.*, “Global analysis of protein activities using proteome chips,” *science*, vol. 293, no. 5537, pp. 2101–2105, 2001.

[22] A. H. Y. Tong, B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, *et al.*, “A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules,” *Science*, vol. 295, no. 5553, pp. 321–324, 2002.

[23] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, “Predicting protein–protein interactions based only on sequences information,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 11, pp. 4337–4341, 2007.

[24] A. Szilágyi, V. Grimm, A. K. Arakaki, and J. Skolnick, “Prediction of physical protein–protein interactions,” *Physical biology*, vol. 2, no. 2, p. S1, 2005.

[25] A. Porollo and J. Meller, “Computational methods for prediction of protein-protein interaction sites,” *Protein-Protein Interactions-Computational and Experimental Tools; W. Cai and H. Hong, Eds. InTech*, vol. 472, pp. 3–26, 2012.

[26] F. Pazos and A. Valencia, “Similarity of phylogenetic trees as in-

dicator of protein–protein interaction," *Protein engineering*, vol. 14, no. 9, pp. 609–614, 2001.

[27] A. D. McLachlan, "Tests for comparing related amino-acid sequences. cytochrome c and cytochrome c_{551}," *Journal of molecular biology*, vol. 61, no. 2, pp. 409–424, 1971.

[28] S. Pitre, F. Dehne, A. Chan, J. Cheetham, A. Duong, A. Emili, M. Gebbia, J. Greenblatt, M. Jessulat, N. Krogan, *et al.*, "Pipe: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs," *BMC bioinformatics*, vol. 7, no. 1, p. 365, 2006.

[29] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The database of interacting proteins: 2004 update," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D449–D451, 2004.

[30] H.-W. Mewes, D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Münsterkötter, S. Rudd, and B. Weil, "Mips: a database for genomes and protein sequences," *Nucleic acids research*, vol. 30, no. 1, pp. 31–34, 2002.

[31] S. Pitre, C. North, M. Alamgir, M. Jessulat, A. Chan, X. Luo, J. Green, M. Dumontier, F. Dehne, and A. Golshani, "Global investigation of protein–protein interactions in yeast saccharomyces cerevisiae using re-occurring short polypeptide sequences," *Nucleic acids research*, vol. 36, no. 13, pp. 4286–4294, 2008.

[32] C. H. Liu, K.-C. Li, and S. Yuan, "Human protein–protein interaction prediction by a novel sequence-based co-evolution method: co-evolutionary divergence," *Bioinformatics*, vol. 29, no. 1, pp. 92–98, 2013.

[33] A. Matsuya, R. Sakate, Y. Kawahara, K. O. Koyanagi, Y. Sato, Y. Fujii, C. Yamasaki, T. Habara, H. Nakaoka, F. Todokoro, *et al.*, "Evola: Ortholog database of all human genes in h-invdb with manual curation of phylogenetic trees," *Nucleic acids research*, vol. 36, no. suppl 1, pp. D787–D792, 2008.

[34] T. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, *et al.*, "Human protein reference database-2009 update," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D767–D772, 2009.

[35] T. Dandekar, B. Snel, M. Huynen, and P. Bork, "Conservation of gene order: a fingerprint of proteins that physically interact," *Trends in biochemical sciences*, vol. 23, no. 9, pp. 324–328, 1998.

[36] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences," *Nucleic acids research*, vol. 36, no. 9, pp. 3025–3030, 2008.

[37] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions," *Nucleic acids research*, vol. 30, no. 1, pp. 303–305, 2002.

[38] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[39] N. Zaki, S. Wolfsheimer, G. Nuel, and S. Khuri, "Conotoxin protein classification using free scores of words and support vector machines," *BMC bioinformatics*, vol. 12, no. 1, p. 217, 2011.

[40] V. N. Vapnik, "Statistical learning theory (adaptive and learning systems for signal processing, communications and control series)," 1998.

[41] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.

[42] N. Zaki, S. Deris, and H. Alashwal, "Protein-protein interaction detection based on substring sensitivity measure," *International journal of biomedical sciences*, vol. 2, no. 1, pp. 148–154, 2006.

[43] N. Zaki, "Protein-protein interaction prediction using homology and inter-domain linker region information," *Advances in Electrical Engineering and Computational Science*, vol. 67, no. 4, pp. 635–645, 2007.

[44] X.-W. Chen and M. Liu, "Prediction of protein–protein interactions using random decision forest framework," *Bioinformatics*, vol. 21, no. 24, pp. 4394–4400, 2005.

[45] L. Liao and W. S. Noble, "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships," *Journal of computational biology*, vol. 10, no. 6, pp. 857–868, 2003.

[46] H. Saigo, J.-P. Vert, N. Ueda, and T. Akutsu, "Protein homology detection using string alignment kernels," *Bioinformatics*, vol. 20, no. 11, pp. 1682–1689, 2004.

[47] S. Roy, D. Martinez, H. Platero, T. Lane, and M. Werner-Washburne, "Exploiting amino acid composition for predicting protein-protein interactions," *PloS one*, vol. 4, no. 11, p. e7813, 2009.

[48] S. M. Gomez, W. S. Noble, and A. Rzhetsky, "Learning to predict protein–protein interactions from protein sequences," *Bioinformatics*, vol. 19, no. 15, pp. 1875–1881, 2003.

[49] S. Martin, D. Roe, and J.-L. Faulon, "Predicting protein–protein interactions using signature products," *Bioinformatics*, vol. 21, no. 2, pp. 218–226, 2005.

[50] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "Biogrid: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D535–D539, 2006.

[51] S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.-O. Vidalain, J.-D. J. Han, A. Chesneau, T. Hao, *et al.*, "A map of the interactome network of the metazoan c. elegans," *Science*, vol. 303, no. 5657, pp. 540–543, 2004.

[52] C.-Y. Yu, L.-C. Chou, and D. T. Chang, "Predicting protein-protein interactions in unbalanced data using the primary structure of proteins," *BMC bioinformatics*, vol. 11, no. 1, p. 167, 2010.

[53] Y.-J. Oyang, S.-C. Hwang, Y.-Y. Ou, C.-Y. Chen, and Z.-W. Chen, "Data classification with radial basis function networks based on a novel kernel density estimation algorithm," *Neural Networks, IEEE Transactions on*, vol. 16, no. 1, pp. 225–236, 2005.

[54] S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. Gandhi, M. Gronborg, *et al.*, "Development of human protein reference database as an initial platform for approaching systems biology in humans," *Genome research*, vol. 13, no. 10, pp. 2363–2371, 2003.

[55] G. R. Mishra, M. Suresh, K. Kumaran, N. Kannabiran, S. Suresh, P. Bala, K. Shivakumar, N. Anuradha, R. Reddy, T. M. Raghavan, *et al.*, "Human protein reference databaseâĂŤ2006 update," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D411–D414, 2006.

[56] G. T. Valente, M. L. Acencio, C. Martins, and N. Lemke, "The development of a universal in silico predictor of protein-protein interactions," *PloS one*, vol. 8, no. 5, p. e65587, 2013.

[57] C. Kern, A. J. Gonzalez, L. Liao, and K. Vijay-Shanker, "Predicting interacting residues using long-distance information and novel decoding in hidden markov models," *Nanoscience, IEEE Transactions on*, vol. 12, no. 13, pp. 158–164, 2013.

[58] T. Friedrich, B. Pils, T. Dandekar, J. Schultz, and T. Müller, "Modelling interaction sites in protein domains with interaction profile hidden markov models," *Bioinformatics*, vol. 22, no. 23, pp. 2851–2857, 2006.

[59] A. Stein, R. B. Russell, and P. Aloy, "3did: interacting protein domains of known three-dimensional structure," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D413–D417, 2005.

[60] L. Rabiner and B.-H. Juang, "An introduction to hidden markov models," *ASSP Magazine, IEEE*, vol. 3, no. 1, pp. 4–16, 1986.

[61] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler, "Hidden markov models in computational biology: Applications to protein modeling," *Journal of molecular biology*, vol. 235, no. 5, pp. 1501–1531, 1994.

[62] S. R. Eddy, "Profile hidden markov models." *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.

[63] A. Krogh *et al.*, "An introduction to hidden markov models for biological sequences," *New Comprehensive Biochemistry*, vol. 32, pp. 45–63, 1998.

[64] B.-J. Yoon, "Hidden markov models and their applications in biological sequence analysis," *Current genomics*, vol. 10, no. 6, p. 402, 2009.

[65] N. Tuncbag, A. Gursoy, R. Nussinov, and O. Keskin, "Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using prism," *Nature protocols*, vol. 6, no. 9, pp. 1341–1354, 2011.

[66] S. J. Hubbard and J. M. Thornton, "Naccess," *Computer Program,*

Department of Biochemistry and Molecular Biology, University College London, vol. 2, no. 1, 1993.

[67] M. Shatsky, R. Nussinov, and H. J. Wolfson, "A method for simultaneous alignment of multiple protein structures," *Proteins: Structure, Function, and Bioinformatics*, vol. 56, no. 1, pp. 143–156, 2004.

[68] E. Mashiach, R. Nussinov, and H. J. Wolfson, "Fiberdock: flexible induced-fit backbone refinement in molecular docking," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 6, pp. 1503–1519, 2010.

[69] S. E. Acuner Ozbabacan, O. Keskin, R. Nussinov, and A. Gursoy, "Enriching the human apoptosis pathway by predicting the structures of protein–protein complexes," *Journal of structural biology*, vol. 179, no. 3, pp. 338–346, 2012.

[70] N. Tuncbag, G. Kar, A. Gursoy, O. Keskin, and R. Nussinov, "Towards inferring time dimensionality in protein–protein interaction networks by integrating structures: the p53 example," *Molecular BioSystems*, vol. 5, no. 12, pp. 1770–1778, 2009.

[71] Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, *et al.*, "Structure-based prediction of protein-protein interactions on a genome-wide scale," *Nature*, vol. 490, no. 7421, pp. 556–560, 2012.

[72] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[73] U. Pieper, N. Eswar, F. P. Davis, H. Braberg, M. S. Madhusudhan, A. Rossi, M. Marti-Renom, R. Karchin, B. M. Webb, D. Eramian, *et al.*, "Modbase: a database of annotated comparative protein structure models and associated resources," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D291–D295, 2006.

[74] N. Mirkovic, Z. Li, A. Parnassa, and D. Murray, "Strategies for high-throughput comparative modeling: Applications to leverage analysis in structural genomics and protein family organization," *Proteins: Structure, Function, and Bioinformatics*, vol. 66, no. 4, pp. 766–777, 2007.

[75] K. Henrick and J. M. Thornton, "Pqs: a protein quaternary structure file server," *Trends in biochemical sciences*, vol. 23, no. 9, pp. 358–361, 1998.

[76] R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. R. Croning, *et al.*, "The interpro database, an integrated documentation resource for protein families, domains and functional sites," *Nucleic acids research*, vol. 29, no. 1, pp. 37–40, 2001.

[77] D. Han, H.-S. Kim, J. Seo, and W. Jang, "A domain combination based probabilistic framework for protein-protein interaction prediction," *GENOME INFORMATICS SERIES*, pp. 250–260, 2003.

[78] D.-S. Han, H.-S. Kim, W.-H. Jang, S.-D. Lee, and J.-K. Suh, "Prespi: a domain combination based prediction system for protein–protein interaction," *Nucleic acids research*, vol. 32, no. 21, pp. 6312–6320, 2004.

[79] W.-H. Jang, S.-H. Jung, and D.-S. Han, "A computational model for predicting protein interactions based on multidomain collaboration," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 9, no. 4, pp. 1081–1090, 2012.

[80] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, *et al.*, "The pfam protein families database," *Nucleic acids research*, p. gkr1065, 2011.

[81] M. Ohue, Y. Matsuzaki, N. Uchikoga, T. Ishida, and Y. Akiyama, "Megadock: An all-to-all protein-protein interaction prediction system using tertiary structure data." *Protein and peptide letters*, 2013.

[82] B. Pierce and Z. Weng, "Zrank: reranking protein docking predictions with an optimized energy function," *Proteins: Structure, Function, and Bioinformatics*, vol. 67, no. 4, pp. 1078–1086, 2007.

[83] M. Ohue, Y. Matsuzaki, T. Ishida, and Y. Akiyama, "Improvement of the protein–protein docking prediction by introducing a simple hydrophobic interaction model: An application to interaction pathway analysis," in *Pattern Recognition in Bioinformatics*. Springer, 2012, pp. 178–187.

[84] Y. Matsuzaki, M. Ohue, N. Uchikoga, and Y. Akiyama, "Protein-protein interaction network prediction by using rigid-body docking tools: Application to bacterial chemotaxis." *Protein and peptide letters*, 2013.

[85] M. Ohue, Y. Matsuzaki, T. Shimoda, T. Ishida, and Y. Akiyama, "Highly precise protein-protein interaction prediction based on consensus between template-based and de novo docking methods," in *BMC Proceedings*, vol. 7, no. Suppl 7. BioMed Central Ltd, 2013, p. S6.

[86] H. Zhou, S. B. Pandit, and J. Skolnick, "Performance of the pro-sp3-tasser server in casp8," *Proteins: Structure, Function, and Bioinformatics*, vol. 77, no. S9, pp. 123–127, 2009.

[87] H. K. Saini and D. Fischer, "Meta-dp: domain prediction meta-server," *Bioinformatics*, vol. 21, no. 12, pp. 2917–2920, 2005.

[88] T. Ishida and K. Kinoshita, "Prediction of disordered regions in proteins based on the meta approach," *Bioinformatics*, vol. 24, no. 11, pp. 1344–1348, 2008.

[89] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, *et al.*, "The pfam protein families database," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D138–D141, 2004.

[90] M. Deng, S. Mehta, F. Sun, and T. Chen, "Inferring domain–domain interactions from protein–protein interactions," *Genome research*, vol. 12, no. 10, pp. 1540–1548, 2002.

[91] B. Schwikowski, P. Uetz, and S. Fields, "A network of protein–protein interactions in yeast," *Nature biotechnology*, vol. 18, no. 12, pp. 1257–1261, 2000.

[92] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, *et al.*, "A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae," *Nature*, vol. 403, no. 6770, pp. 623–627, 2000.

[93] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki, "Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins," *Proceedings of the National Academy of Sciences*, vol. 97, no. 3, pp. 1143–1147, 2000.

[94] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[95] R. Caruana, N. Karampatziakis, and A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 96–103.

[96] K. Y. Chang and J.-R. Yang, "Analysis and prediction of highly effective antiviral peptides based on random forests," *PloS one*, vol. 8, no. 8, p. e70166, 2013.

[97] G. Izmirlian, "Application of the random forest classification algorithm to a seldi-tof proteomics study in the setting of a cancer prevention trial," *Annals of the New York Academy of Sciences*, vol. 1020, no. 1, pp. 154–174, 2004.

[98] Y. Qi, "Random forest for bioinformatics," in *Ensemble Machine Learning*. Springer, 2012, pp. 307–323.

[99] R. Singh, J. Xu, and B. Berger, "Struct2net: Integrating structure into protein-protein interaction prediction." in *Pacific Symposium on Biocomputing*, vol. 11. Citeseer, 2006, pp. 403–414.

[100] R. Singh, D. Park, J. Xu, R. Hosur, and B. Berger, "Struct2net: a web service to predict protein–protein interactions using a structure-based approach," *Nucleic acids research*, vol. 38, no. suppl 2, pp. W508–W515, 2010.