Computational Methods for Protein-Protein Interaction Prediction

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) occur at almost every level of cell functions. The identification of interactions among proteins provides a global picture of cellular functions and biological processes. It is also an essential step in the construction of PPI networks for human and other organisms. PPI prediction has been considered to be a promising alternative to the traditional drug design techniques. 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. This paper investigates most of the relevant existing computational approaches carried out towards this perspective and provides a comparison of these approaches. Technical challenges, unresolved issues, and future research directions in this area are also discussed.*

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

1. Introduction

Proteins are the building blocks of all living organisms. A protein is a polypeptide which is a linear polymer of several amino acids (AAs) connected by peptide bonds. The primary structure of a protein is the linear sequence of its AAs. The secondary structure of a protein is the general three-dimensional form of its local parts without displaying specific atomic positions, which are considered to be a tertiary structure. Domains are the basic functional units of protein tertiary structures. A protein interacts with other proteins in order to perform certain functions and 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 (PPIs), precisely identifying 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 [1]–[3]. Protein interaction prediction is also a fundamental step in the construction of PPI networks for human and other organisms. The identification of specific disease-related protein interactions can lead to the development of a number of medication drugs. Abnormal PPIs have implications in several neurological disorders such as Creutzfeld-Jacob and Alzheimer [4]–[6]. Therefore, the development of accurate and reliable methods for identifying PPIs has very important impacts in several protein research areas.

In this paper we explore the two main categories of computational PPI prediction methods; sequence-based and structure-based methods. We provide a comprehensive comparative study of the existing approaches in PPI prediction and discuss the technical challenges and unresolved issues in this field. The rest of this paper 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 evaluation measures that are typically used in PPI prediction. In Section 4, which is the focus of our paper, comprehensive description and comparison are presented for the most current computational PPI prediction methods. 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.

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 [7]. Some approaches use AA physiochemical properties. Other approaches use pairwise similarity. Some approaches are template-based while others are statistical-based.

There are various challenges that face machine-learning (ML) protein 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 [8].

Furthermore, most PPI prediction approaches have achieved reasonable performance on balanced datasets con-

taining 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 are required to handle such challenging tasks.

3. Evaluation Metrics

There are several assessment measures that are used to evaluate a PPI predictor and compare it with other approaches. The most frequently used evaluation metrics in this field are accuracy, sensitivity, specificity, precision, F1, and MCC. Accuracy ($Ac = \frac{TP+TN}{TP+TN+FN+FP}$, where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively) is defined as the proportion of correctly predicted interacting and non-interacting protein pairs to all of the protein pairs listed in the dataset. Sensitivity, or recall ($R = \frac{TP}{TP+FN}$), is defined as the proportion of correctly predicted interacting protein pairs to all of the interacting protein pairs listed in the dataset. Precision ($P = \frac{TP}{TP+FP}$) is defined as the proportion of correctly predicted interacting protein pairs to all of the predicted interacting protein pairs. Specificity ($Sp = \frac{TN}{TN+FP}$) is defined as the proportion of correctly predicted non-interacting protein pairs to all the non-interacting protein pairs listed in the dataset. The F-measure ($F1 = \frac{2*P*R}{P+R}$) is an evaluation metric that combines precision and recall into a single value and is defined as the harmonic mean of precision and recall [9], [10]. Mathew correlation coefficient ($MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}$) 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.

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 [11], mass spectrometry [12], Tandem Affinity Purification [13], protein chips [14], and hybrid approaches [15]. Although these techniques have succeeded in identifying several important interacting proteins in several species such as Yeast, Drosophila, and Helicobacter-pylori [16], 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 of computational tools has been increased in order to validate physiochemical experimental results and to predict non-discovered PPIs [1], [17].

Several computational methods have been proposed for PPI prediction and can be categorized according to the used protein features into sequence-based and structure-based methods. Sequence-based methods utilize AA features while structure-based methods use three-dimensional structural features [18]. 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 [18]. This section presents and evaluates some of the existing sequence-based approaches.

4.1.1 PIPE

Protein-protein Interaction Prediction Engine (PIPE) was developed by Pitre *et al.* [19] 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 uses protein primary structure to estimate the likelihood of an interaction between a pair of the yeast Saccharomyces cerevisiae proteins 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) [20] and MIPS [21] 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.* [22] 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.

4.1.2 Auto Covariance

Guo *et al.* [23] 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 and its 30 vicinal AAs 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 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 [24] was used to evaluate the model and the average prediction accuracy, sensitivity, and precision achieved are respectively 0.86, 0.85, and 0.87.

AC includes the information of interactions between AAs a longer distance apart in the sequence taking the neighboring effect into account. The long-range interactions are important for representing the PPI information. 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 [25] and has been successfully applied in various classification problems [26]. SVM offers several related computational advantages such as the lack of local minima in the optimization [27].

4.1.3 Pairwise Similarity

Zaki *et al.* [1] proposed a PPI predictor based on pairwise similarity and protein primary structure. Each protein sequence is 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 [28]. These vectors are then used to compute the kernel matrix which is exploited in conjunction with a RBF-kernel SVM. Two proteins may interact by the means of the scores similarities they produce [29], [30]. Each sequence in the testing set is aligned against each sequence in the training set, count the number of positions that have identical AAs, and then divide by the total length of the alignment.

The method was evaluated on 100 interacting yeast *Saccharomyces cerevisiae* protein pairs within 157 proteins and 100 non-interacting protein pairs within 77 proteins from the DIP [20] and MIPS [21] databases. The method achieved a sensitivity of 0.81 and a specificity of 0.744.

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 [31]. 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 [32]. 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.

4.1.4 AA Composition

Roy *et al.* [33] examine the role of amino acid composition (AAC) in PPI prediction and it performance against well-known features such as domains, tuple feature, and signature product feature. Every protein pair is represented by AAC and domain features. AAC is 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 is created. The real-valued composition is 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 are represented as binary features with each feature identified by a domain name. To compare AAC against other non-domain sequence-based features, tuple features [34] and signature products [35] 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 ML 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 [36], 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) [37] and fly, *D. melanogaster* [36] 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.

Most computational PPI prediction methods use domain knowledge as they capture conserved information of interaction surfaces. However, approaches relying on domains as features cannot be applied to proteins without any domain information. On contrast, AAC is simple feature, computationally cheep, applicable to any protein sequence, and can be used when there is lack of domain information. AAC can be used alone or combined with other features to predict new and validate existing interactions.

4.1.5 AA Triad

Yu *et al.* [38] 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) [39] is employed to predict PPIs based on AA triad information. The method is summarized as follows. Each protein sequence is 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 [16]. The method then scans triads one by one along the sequence, and each scanned triad is counted in an occurrence vector, O. Subsequently, a significance vector, S, is 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 is then encoded as a feature vector by concatenating the two significance vectors of the two individual proteins. Finally, the feature vector is 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) [40], [41]. 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 [39].

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

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 [18]. This section presents and evaluates some of the state-of-the-art structure-based approaches.

4.2.1 Struct2Net

Singh *et al.* [42] 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) [43]. 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.* [44] 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.

4.2.2 PRISM

Tuncbag *et al.* [45] 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 [46]. Second, complementary chains of template interfaces are separated and structurally compared with each of the target surfaces by using MultiProt [47]. 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 [48] algorithm is used to refine the interactions to introduce

Table 1: Sequence-based PPI prediction approaches.

Approach	Extracted Features	Technique/Tool	Datasets
PIPE (Pitre *et al.* 2006), PIPE2 (Pitre *et al.* 2008)	Short AA polypeptides	Similarity measure	Yeast protein (DIP and MIPS)
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)

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 [49] and a p53- protein-related pathway [50], 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.

Because PRISM is a template-based method, its prediction accuracy depends on the template dataset prepared. Only PPIs whose interaction surface structures are conserved are expected to be predicted.

4.2.3 MEGADOCK

Ohue *et al.* [51] 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 [52] 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 [53], [54] 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.

4.2.4 Meta Approach

Ohue *et al.* [55] proposed a PPI prediction approach based on combining template-based and docking methods. The approach applies PRISM [45] as a template-matching method and MEGADOCK [51] 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 [56], 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 [57] and the prediction of disordered regions in proteins [58]. 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.

4.2.5 PrePPI

Zhang *et al.* [59] 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) [43] or from the ModBase [60] and SkyBase [61] homology model databases. Close and remote structural neighbors are found for each subunit. A template for the interaction exists if a PDB or PQS [62] 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.

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 2 summarizes these structure-based approaches including the features that are used, the technique and/or the tools applied, and the validation datasets used.

5. Conclusion

This survey is focused on protein-protein interaction prediction. The main challenges that face PPI prediction were presented. We investigated several relevant existing approaches and provided a 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 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 us to compare the performance of these prediction methods in a fair and uniform fashion. This work can also 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 such as structural and functional information into the prediction process.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

Table 2: Structure-based PPI prediction approaches.

Approach	Extracted Features	Technique/Tool	Datasets
Struct2Net (Singh *et al.* 2006, 2010)	Homology with known protein complexes in PDB	Logistic regression	Yeast, Fly ,and Human protein
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)
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
PrePPI (Zhang *et al.* 2012)	Secondary structure	Bayesian networks, Naive Bayes	Yeast protein, Human protein

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] S. J. Hubbard and J. M. Thornton, "Naccess," *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, vol. 2, no. 1, 1993.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

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

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.