J Biosci (2019) 44:104
DOI: 10.1007/s12038-019-9909-z

Review

Machine-learning techniques for the prediction of protein–protein interactions

Debasree Sarkar[1,2] and Sudipto Saha[2]*

[1]*Present Address: SUNY Upstate Medical University, Syracuse, NY, USA*
[2]*Division of Bioinformatics, Bose Institute, Kolkata, India*

**Corresponding author (Email, ssaha4@jcbose.ac.in, ssaha4@gmail.com)*

MS received 7 October 2018; accepted 29 April 2019; published online 16 August 2019

Protein–protein interactions (PPIs) are important for the study of protein functions and pathways involved in different biological processes, as well as for understanding the cause and progression of diseases. Several high-throughput experimental techniques have been employed for the identification of PPIs in a few model organisms, but still, there is a huge gap in identifying all possible binary PPIs in an organism. Therefore, PPI prediction using machine-learning algorithms has been used in conjunction with experimental methods for discovery of novel protein interactions. The two most popular supervised machine-learning techniques used in the prediction of PPIs are support vector machines and random forest classifiers. Bayesian-probabilistic inference has also been used but mainly for the scoring of high-throughput PPI dataset confidence measures. Recently, deep-learning algorithms have been used for sequence-based prediction of PPIs. Several clustering methods such as hierarchical and *k*-means are useful as unsupervised machine-learning algorithms for the prediction of interacting protein pairs without explicit data labelling. In summary, machine-learning techniques have been widely used for the prediction of PPIs thus allowing experimental researchers to study cellular PPI networks.

Keywords. Clustering; deep learning; decision tree; machine-learning techniques; protein–protein interaction; support vector machine

1. Introduction

Protein–protein interaction (PPI) networks play fundamental roles in regulating nearly all biological processes and form the basis for the cellular structure and functions. Deciphering the interaction networks of proteins, therefore, helps in improving our knowledge of functions of proteins, and is also crucial for understanding cellular pathways, and developing effective therapies for the treatment of human diseases. Several high-throughput experimental techniques such as yeast 2-hybrid (Y2H), affinity purification-mass spectrometry (AP-MS) and protein microarrays have been employed for PPI discovery, but their accuracy is questionable due to the occurrence of a large number of false positives (FPs) as well as false negatives (FNs) (Mrowka *et al.* 2001). In this context, it has been noted that fast and scalable machine-learning algorithms have proven extremely useful for the prediction of novel PPIs, and can improve the efficacy of experimental identification of PPIs, when used in conjunction. Furthermore, it has been found that computational predictions of PPIs demonstrated almost similar accuracy levels to those of large-scale experimental PPI datasets (von Mering *et al.* 2002).

So far, it has been a daunting task for the proteomics researchers to generate a comprehensive picture of entire interactomes, especially for complex eukaryotic organisms such as human beings. The entire human interactome has been estimated to comprise ∼130,000 binary PPIs (Venkatesan *et al.* 2009), of which only ∼60,000 PPIs have been currently compiled from various published and unpublished experimental results through the Human Reference Protein Interactome (HuRI) project (Bader *et al.* 2017). Experimental studies have also been conducted to unravel the interactomes of other model organisms such as *Escherichia coli*, *Helicobacter pylori*, *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (roundworm), *Drosophila melanogaster* (fruit fly), *Mus musculus* (mouse) and *Rattus norvegicus* (rat) (Walhout *et al.* 2000; Rain *et al.* 2001; Xenarios *et al.* 2002). However, as in humans, most of these interactomes have also been incomplete to date. For example, the interactome involving only 37.1% of mouse proteins and 20.2% of *C. elegans* proteins is known with a varying degree of certainty (Alonso-López *et al.* 2016).

Nevertheless, several experimental datasets, especially for the yeast interactome (Uetz *et al.* 2000; Gavin *et al.* 2002; Krogan *et al.* 2006), as well as, some PPI databases such as DIP (Salwinski *et al.* 2004), IntAct (Orchard *et al.* 2014) and BioGRID (Chatr-Aryamontri *et al.* 2017), are available in the public domain. In addition, different experimental methods may be able to detect different types of interactions, thereby reporting different subsets of the actual interactome. For example, interaction networks derived from Y2H and AP-MS experiments have different topological and biological properties, because Y2H is better adapted to capture transient interactions between signalling molecules, whereas, AP-MS data are enriched in stable protein complexes (Saha *et al.* 2010). Since PPI networks are inherently dynamic in nature, the lack of tissue-specific or condition-specific (e.g. healthy *vs* disease state) experimental PPI data further complicates the problem.

Therefore, computational methods can be highly beneficial for the study of interactomes, as they can extrapolate the experimental PPI data for elucidating the complete interactome of an organism, and even predict the interactome for homologous organisms. Furthermore, since experimental strategies for PPI identification are expensive in terms of both time and money, *in silico* prediction of PPIs provides a complementary strategy for query proteome/interactome annotation at a low cost and less time. Although protein–protein docking methods are available, their usefulness in PPI prediction is limited, since there is difficulty in docking proteins undergoing large conformational changes, as well as due to the unavailability of experimentally derived 3D structures for a majority of proteins. In this review article, we have provided an overview of different types of machine-learning strategies that have been utilized for predicting PPIs, and discussed their respective advantages and disadvantages. We have also summarized the datasets used and the organisms studied in the prediction methods developed under each of these categories.

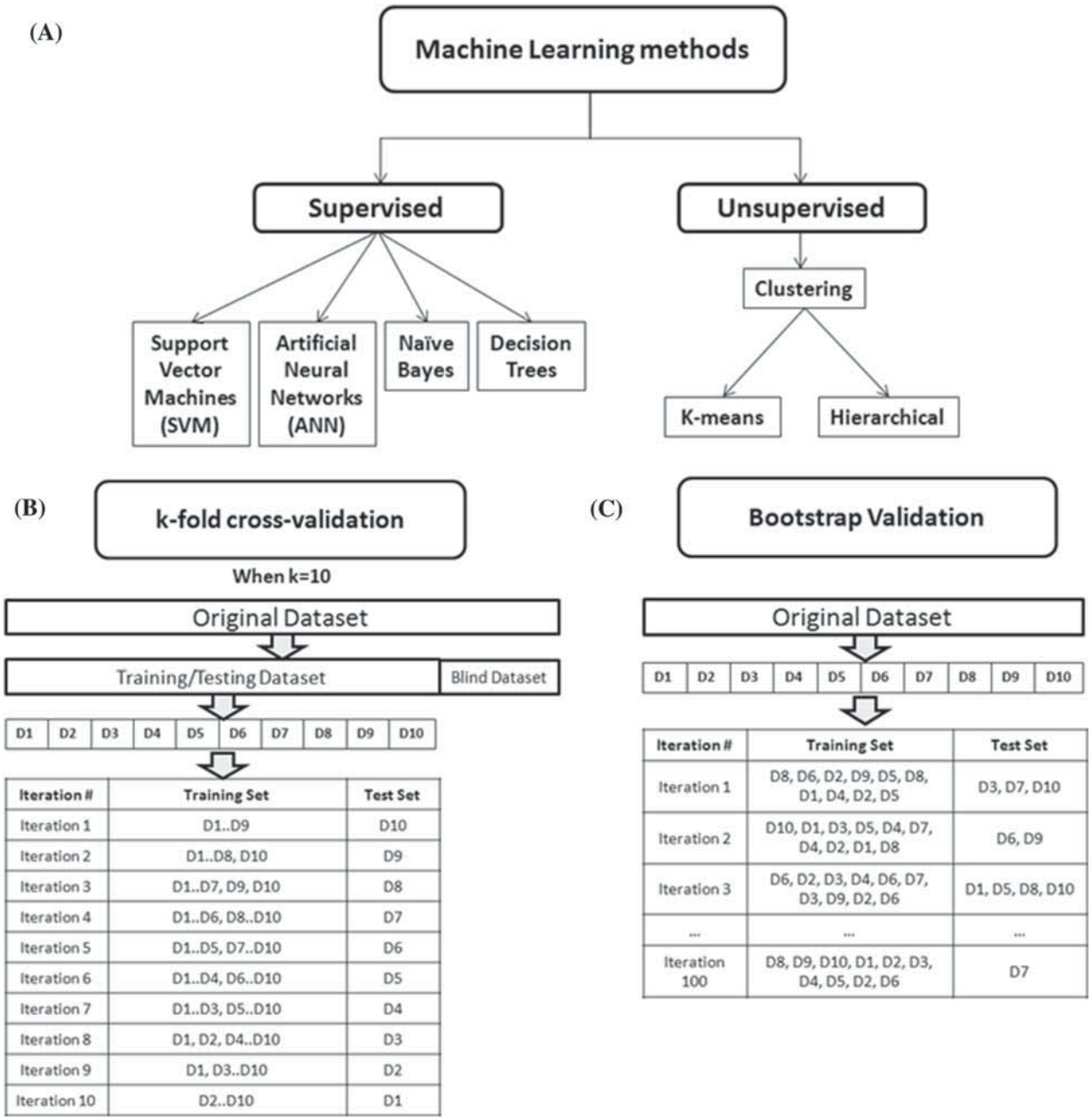
2. Machine-learning techniques

The use of machine-learning techniques in the prediction of PPIs began in 2001 through the independent efforts of a few research groups (Bock and Gough 2001; Sprinzak and Margalit 2001; Zhou and Shan 2001). The prediction of binary PPIs generally involves the sequence or structure of both the interacting proteins as the input and the probability of these two proteins to interact with each other as the output. The observable quantities given as input to a machine-learning algorithm for creating the statistical prediction models are called 'features'. Thus, for PPI-prediction methods, the typically used features include amino acid composition, domain/motif composition or hydrophobicity profiles of input protein sequences, interface properties of protein 3D structures, genomic features such as gene neighbouring or phylogenetic relationship, and network topology-based features such as degree distribution or clustering coefficient.

Machine-learning techniques used for predicting PPIs can be broadly divided into two main categories: supervised and unsupervised, based on whether the input variables need to be labelled according to the expected outcome or not (figure 1A). Supervised learning infers a mapping function from a set example input–output pairs, which could be used for predicting the outcome for other inputs, whereas, unsupervised learning discovers the hidden structure within unlabelled training data for drawing meaningful conclusions. Examples of supervised machine-learning algorithms used for PPI prediction include artificial neural networks (ANNs), Bayesian inference, support vector machines (SVMs) and decision tree-based methods such as random forest (RF). Clustering techniques such as *k*-means, single-linkage and spectral clustering are associated with unsupervised machine-learning methods used for PPI prediction. Supervised machine learning is implemented for classification problems, i.e. mapping of input data into specific classes, where a set of quantitative or categorical features are analysed for identified features that can discriminate the input variables into specified classes. Thus, percentage composition of the 20 standard amino acid residues may be computed using sequences of pairs of interacting proteins. Since defining a high-confidence negative dataset for PPI prediction is difficult, negative datasets have been created by considering randomly selected protein pairs not known to interact with each other (Ben-Hur and Noble 2005), or belonging to different sub-cellular locations (Xia *et al.* 2010).

In binary classification problems such as PPI prediction, there are only two categories: the 'positive' (p) category, containing proteins that interact with each other, and the 'negative' (n) category, containing proteins that do not interact. If the class prediction for each instance is made based on a score that can be represented as a continuous random variable (X), then at a given threshold value (T), the instance is classified as 'positive' if $X > T$, and 'negative' otherwise. There are four possible prediction outcomes from algorithms undertaking such binary classification problems. If two proteins that have been experimentally verified to interact with each other are correctly classified by the algorithm as an interacting protein pair, then it is called a true-positive (TP) prediction; but if wrongly classified as a non-interacting pair, then this is said to be a FN prediction. Conversely, when two proteins not known to interact are classified by the algorithm as non-interactors, a true-negative (TN) prediction occurs, while if they are classified as interactors then it is a FP prediction.

The prediction efficiency of all classification algorithms can be measured using several threshold-dependent parameters such as precision, sensitivity or recall, specificity, accuracy, $F1$-score and Mathew's correlation coefficient (MCC), using the following formulae:



Iteration #	Training Set	Test Set
Iteration 1	D1..D9	D10
Iteration 2	D1..D8, D10	D9
Iteration 3	D1..D7, D9, D10	D8
Iteration 4	D1..D6, D8..D10	D7
Iteration 5	D1..D5, D7..D10	D6
Iteration 6	D1..D4, D6..D10	D5
Iteration 7	D1..D3, D5..D10	D4
Iteration 8	D1, D2, D4..D10	D3
Iteration 9	D1, D3..D10	D2
Iteration 10	D2..D10	D1

Iteration #	Training Set	Test Set
Iteration 1	D8, D6, D2, D9, D5, D8, D1, D4, D2, D5	D3, D7, D10
Iteration 2	D10, D1, D3, D5, D4, D7, D4, D2, D1, D8	D6, D9
Iteration 3	D6, D2, D3, D4, D6, D7, D3, D9, D2, D6	D1, D5, D8, D10
...	...	...
Iteration 100	D8, D9, D10, D1, D2, D3, D4, D5, D2, D6	D7

Figure 1. Diagrammatic representation of machine-learning methods, datasets and methods of validation used in prediction of PPIs. (**A**) Classification of different machine-learning methods into supervised and unsupervised approaches. (**B**) Training, testing and blind datasets for *k*-fold cross-validation. (**C**) Training and testing datasets for bootstrap validation.

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}; \quad \text{Recall/Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}};$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \times 100;$$

$$F1 \text{ score} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}$$

$$\text{MCC} = \frac{(\text{TP})(\text{TN})-(\text{FP})(\text{FN})}{\sqrt{[\text{TP}+\text{FP}][\text{TP}+\text{FN}][\text{TN}+\text{FP}][\text{TN}+\text{FN}]}} \times 100$$

MCC is generally believed to be a more accurate criterion for evaluating the prediction performance of binary classifiers than accuracy and *F*1 score, because it takes into account the balanced ratios of all the four prediction outcomes, viz., TP, FP, TN and FN.

For a threshold-independent overall representation of the prediction performance, a receiver operating characteristics (ROC) curve or a precision-recall (PR) curve is plotted. The area under curve (AUC) statistic for either of these curves, therefore, becomes a threshold-independent measure of

prediction performance. The ROC curve is created by plotting the true-positive rate (TPR) against the false-positive rate (FPR) at various thresholds, where $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$ and $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$. The random variable X that represents the prediction score follows a probability density $f_1(x)$ if the instance actually belongs to the 'positive' category, and $f_0(x)$ if it originates from the 'negative' category. Therefore, $\mathrm{TPR}(T) = \int_T^\infty f_1(x)\mathrm{d}x$ and $\mathrm{FPR}(T) = \int_T^\infty f_0(x)\mathrm{d}x$, and when TPR(T) is plotted parametrically *vs* FPR(T) with threshold T as the varying parameter, then

$$\begin{aligned}\mathrm{AUC} &= \int_{\infty}^{-\infty} \mathrm{TPR}(T)\mathrm{FPR}'(T)\mathrm{d}T \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(T' > T) f_1(T') f_0(T)\mathrm{d}T'\mathrm{d}T \\ &= P(X_1 > X_0)\end{aligned}$$

where X_1 is the score for a positive instance and X_0 is the score for a negative instance.

The aim of PPI predicting machine-learning algorithms is to predict whether a new protein pair will interact with each other or not, after being trained with known PPI pairs. This property is known as generalization and it depends on how well the complexity of the prediction model (hypothesis) generated by the algorithm matches with the complexity of the function underlying the input dataset. If the hypothesis is not complex enough to model the samples, it is called underfitting, whereas, if the hypothesis is too complex and the training samples are not enough to constrain it, this is called overfitting, which leads to problems in generalization. Hence, the generalization ability of a model must be evaluated by dividing the training samples into a larger training dataset and a smaller testing set, for determining the model complexity, as shown in figure 1B. Finally, an independent or blind dataset, containing new input variables not used in training or testing, is also required for evaluating the expected performance of the trained model. For example, Barman *et al.* (2015) have used *E. coli* PPI data downloaded from the Bacterial Protein Interaction Database (*Bacteriome.org*) for training and testing of the prediction models, and *E. coli* PPIs observed in solved X-ray crystal structures obtained from the PDB as a blind set. In ANNs, the training dataset is used to initially fit the parameters during model construction, and the fitted model is successively evaluated using a separate validation dataset to prevent overfitting and minimize errors. Another test dataset is finally used for an unbiased evaluation of whether the model fits the training dataset. For example, Sun *et al.* (2017) have used 33,052 positive and 32,816 negative samples from their initial PPI dataset as a training dataset, 3493 positive and 3507 negative samples as a hold-out validation dataset and several external test sets for final model evaluation.

The datasets for model training, testing and validation can be prepared using techniques such as randomization, cross-validation and bootstrapping (figure 1B and C). If the dataset is large enough, then it can be randomly divided into K parts, and each of these parts can again be randomly divided into training and testing sets, and this process can be repeated K times. This is called randomization, which ensures the random sampling of training and testing sets from the data so that the learning process is independent of the selection of training data. However, since datasets large enough for proper randomization are rarely available, the same dataset is repeatedly split into training and testing sets in different ways by a technique called cross-validation or rotation estimation, which can again be exhaustive or non-exhaustive. Exhaustive cross-validation may involve either leave-p-out cross-validation (LpOCV), where p observations are set aside as the test set and the remaining observations are taken as the training set, or leave-one-out cross-validation (LOOCV), where $p = 1$. The most commonly used form of non-exhaustive cross-validation is k-fold cross-validation, where the original sample is randomly partitioned into k equal-sized subsamples, from which a single subsample is retained as the test set and the remaining subsamples are used as the training set. The process is then repeated k times, with each of the k subsamples used exactly once in the test set, as shown in figure 1B. The most commonly used is the 5-fold cross-validation, where the training dataset is divided into five subsets of which four subsets are used in training the model, and the remaining one is used for testing it, and the process in repeated five times, using a different subset in each iteration.

Furthermore, in the real-world biological classification problems, the number of TN PPI pairs far exceeds the number of TP ones. It is therefore, advisable to use both balanced (equal number of TP and TN) and unbalanced (TN $\gg$ TP) training datasets in such cases (Blagus and Lusa 2010; Yu *et al.* 2010; Barman *et al.* 2015).

2.1 *Support vector machines (SVMs)*

SVM is a statistical learning algorithm developed by Vapnik (1999). SVM-mediated pattern classification is well known for generalization ability, and thus has been widely used for binary classification problems involving biological data, such as classifying interacting and non-interacting protein pairs. The goal of the SVM algorithm is to find an optimal hyperplane that separates the training samples by a maximal margin, with all positive samples lying on one side and all negative samples lying on the other side. Suppose that we are given a training dataset of N instance-labelled pairs $X = [(x_1, y_1), (x_2, y_2)\ldots (x_N, y_N)]$ with input data $x_i \in R^n$ and labelled output data $y_i \in [+1, -1]$, for $i = 1, 2, 3\ldots N$. The SVM algorithm solves the quadratic optimization problem by minimizing the function as below:

$$\min_{w,b,\xi} \frac{1}{2} \parallel w \parallel^2 + \mathrm{C}\sum_{i=1}^{N} \xi_i \tag{1}$$

$$\text{subject to} \quad y_i(\langle w.x_i\rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \tag{2}$$

where w is the weight vector, b is the bias term, C is the penalty factor and ξ_i is the slack variable. Since $\|w\|^2$ is convex, minimizing equation (1) under the linear constraints in (2) can be solved with Lagrange multipliers. The complexity parameter denoted by C is a trade-off between training error and margin, which determines the penalty that should be imposed on training data points that end up on the wrong side of the decision boundary.

Generally, the training examples x_i are mapped into a high-dimensional feature space through some nonlinear function ϕ. This transformation helps us to make the class data distributions linearly separable. The feature vectors in SVM thus appear in the form of dot products of two data points as $K(x_i, x_j) = x_i^{\mathrm{T}}.x_j$, which is called a 'kernel' and represents a measure of similarity between the two data points x_i and x_j. Hence, even though $\phi(x)$ maybe complex and very high dimensional, after the transformation the dot product, $K(x_i, x_j) = \phi(x_i)^{\mathrm{T}}.\phi(x_j)$, also known as the 'kernel function', can have a simple form. Typical kernel functions used in SVM can be linear, polynomial, sigmoid or radial basis function (RBF), as shown below:

- Linear: $K(x_i, x_j) = x_i^{\mathrm{T}} x_j$;
- Polynomial: $K(x_i, x_j) = (\gamma x_i^{\mathrm{T}} x_j + \gamma)^D, \gamma > 0$;
- Sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^{\mathrm{T}} x_j + B)$;
- RBF: $K(x_i, x_j) = \mathrm{e}^{-(\|x_i - x_j\|^2/2\sigma^2)}$, or $K(x_i, x_j) = \mathrm{e}^{-\gamma\|x_i - x_j\|^2}$, $\gamma > 0$.

It has been observed that generally the RBF kernel performs best for complex biological classification problems such as predicting PPI pairs (Barman *et al.* 2014).

The majority of SVM-based PPI-prediction methods depend on protein-primary sequences as input (Bock and Gough 2001; Ben-Hur and Noble 2005; Martin *et al.* 2005; Shen *et al.* 2007; Guo *et al.* 2008; Barman *et al.* 2015; Srinivasulu *et al.* 2015; Sriwastava *et al.* 2015; You *et al.* 2015b), since sequence information is available for most proteins. SVM-prediction methods have also been trained with input features such as PDB structures (Zhu *et al.* 2006), domain compositions (Chatterjee *et al.* 2011; Hou *et al.* 2012) and gene ontology (GO) annotations (Mei 2013) of interacting proteins. SVM-based methods for the prediction of specific domain-binding linear motif peptides that may mediate PPIs have also been proposed (Kundu *et al.* 2014; Sarkar *et al.* 2018). Several examples of SVM-based PPI-prediction methods are summarized in table 1.

Although SVM is very efficient in classifying datasets with unspecified complexity, one of its major drawbacks is the requirement of a considerable amount of computational resources for training and parameter optimization, since the number of support vectors grows linearly with the scale of the training data. In addition, the outputs are not probabilistic, and the relationship between training data and output may be complex and not clearly comprehensible.

2.2 *Decision tree-based methods*

Decision-tree algorithms (Breiman *et al.* 1984) comprise recursive partitioning of the input space by selecting the best attribute, and expanding the leaf nodes of the tree until a predefined stopping criterion, such as a minimum number of training instances assigned to each leaf node of the tree, is met. The best test condition for splitting is determined by different algorithms using different metrics such as Gini impurity and information gain. Gini impurity is a measure of the misclassification to denote the probability of a randomly chosen element from the set to be incorrectly labelled according to the distribution of labels in the subset. Gini impurity for a set of items with J classes, with $i \in [1, 2, \ldots, J]$, and p_i being the fraction of items labelled with class i in the set, can be computed as

$$I_{\mathrm{G}}(p) = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{J} p_i(1 - p_i) = \sum_{i=1}^{J} (p_i - p_i^2)$$
$$= \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^2 = 1 - \sum_{i=1}^{J} p_i^2$$

Information gain or entropy is another metric used to determine the best feature to be considered for splitting at each step of tree-building. Entropy is defined as $H(T) = I_{\mathrm{E}}(p_1, p_2, \ldots, p_J) = -\sum_{i=1}^{J} p_i \log_2 p_i$, with $p_1, p_2, \ldots, p_J$ being fractions that add up to 1 and represent the percentage of each class present in the child node that results from a split in the tree:

$$I_{\mathrm{G}}(T, \alpha) = H(T) - H(T|\alpha)$$
$$= -\sum_{i=1}^{J} p_i \log_2 p_i - \sum_{a} p(a)$$
$$- \sum_{i=1}^{J} -Pr(i|a)\log_2 Pr(i|a)$$

In an RF classification algorithm, many decision trees are constructed based on random feature vectors sampled independently from a training dataset. For a new input, the feature vector is passed through each of the trees and classification is performed by majority voting among the independent trees. The main advantage of this method is that for a large dataset having a large number of features, no separate feature selection algorithm is needed as the algorithm itself can rank features according to their significance in classification.

Various decision tree-based PPI-prediction methods have been developed using several input features such as protein sequence (Chen and Jeong 2009; Xia *et al.* 2010; Zahiri *et al.* 2013; Wei *et al.* 2015; You *et al.* 2015a; Liu *et al.* 2016; Sze-To *et al.* 2016; Zhou *et al.* 2017) and 3D structure (Li *et al.* 2012; Maheshwari and Brylinski 2017). Some methods also use genomic features (Wang *et al.* 2009) and domain composition (Chen and Liu 2005; Rodgers-Melnick *et al.* 2013) as listed in table 2.

Table 1. SVM-based methods for PPI prediction

Type of feature	Dataset(s) used for model building	Organism(s)	Reference(s)
Sequence-based	Database of interacting proteins (DIPs)	*E. coli*, yeast, *Drosophila*, human, mouse, rat, bovine	Bock and Gough (2001)
		E. coli, *H. pylori*, yeast, human, mouse	Martin *et al.* (2005)
		E. coli, yeast, human	Sriwastava *et al.* (2015)
		Yeast	Guo *et al.* (2008) You *et al.* (2015b) Ben-Hur and Noble (2005)
	Biomolecular Interaction Network Database (BIND)	Yeast	Ben-Hur and Noble (2005) Bandyopadhyay and Mallick (2017)
	Human Protein References Database (HPRD)	Human	Shen *et al.* (2007)
	Su *et al.* (2008)	*E. coli*	Barman *et al.* (2015)
	Kiemer *et al.* (2007)	Yeast	Ruan *et al.* (2018)
	Chen *et al.* (2013), Yugandhar and Gromiha (2014)	–	Srinivasulu *et al.* (2015)
Structure-based	Protein Data Bank (PDB)	–	Bradford and Westhead (2005)
		–	Zhu *et al.* (2006)
GO-based	NCBI HIV-1 Human Interaction Database	Human with HIV-1	Mei (2013)
	Database of interacting proteins (DIPs)	Yeast	Bandyopadhyay and Mallick (2017)
Domain/motif-based	Database of interacting proteins (DIPs)	*E. coli*, yeast, *Drosophila*, human, mouse, rat, bovine	Chatterjee *et al.* (2011)
	VirusMINT	Human with viruses such as HIV-1/SV-40/HBV/HCV, etc.	Barman *et al.* (2014)
	Sparks *et al.* (1996), Cestra *et al.* (1999), Tong *et al.* (2002), Landgraf *et al.* (2004)	Yeast	Hou *et al.* (2012)
	Miller *et al.* (2008), Jones *et al.* (2006), Kaushansky *et al.* (2008), Carducci *et al.* (2012), Stiffler *et al.* (2007), Tonikian *et al.* (2008)	–	Kundu *et al.* (2014)
	Sarkar *et al.* (2015)	–	Sarkar *et al.* (2018)

Decision tree-based classification methods have been found to show superior performance in the prediction of PPIs as compared to other types of machine-learning algorithms (Xia *et al.* 2010; You *et al.* 2015a; Zhou *et al.* 2017). In particular, the RF classifier consistently emerged as one of the top performing algorithms across features and datasets (Qi *et al.* 2006; Zhou *et al.* 2017), as illustrated in table 3. However, decision tree-based classification methods, especially RFs, are sensitive to high noise in the data, are prone to overfitting, and highly correlated features can affect their prediction performance.

2.3 *Probabilistic/Bayesian classification*

Bayesian-probabilistic classifiers are generally preferred by biologists over other machine-learning techniques, because their functionalities are easier to understand as compared to 'black-box' predictors such as SVMs and neural networks, and they can be used with both numerical and categorical data.

Probabilistic classifiers denote algorithms that classify the input data points using conditional probability distributions to model the relationship between the features of the training samples and the class which they belong to. Hence, if the features are denoted by $x_i (i = 1, \ldots, M)$, then the feature vector for each data point can be represented as $x = [x_1, x_2, \ldots, x_M]$, and the probability of the data point belonging to each of the N classes $(c = c_1, c_2, \ldots, c_N)$ as $P(C = c_1|x), P(C = c_2|x), \ldots, P(C = c_N|x)$. The probability for each class can be computed using the Bayes' theorem as

$$P(C|x) = \frac{P(x|C)P(C)}{P(x)} = \frac{P(x|C)P(C)}{\sum_{c_i \in C} P(x|C = c_i)P(C = c_i)}$$

$P(x|C)$ and $P(C)$ can be estimated from the training data for each class. Thus, a new data point is classified as the class with the maximum probability $c^* = \arg\max_{c_j} P\big(C = c_j|x\big)$ and $j = 1, \ldots, N$.

After modelling the class conditional probabilities, the probabilistic approach seeks to classify the input data points to the class with the maximum probability. In case of a binary classification problem this amounts to computing the ratio $\Upsilon = \frac{P(x|C=c_1)}{P(x|C=c_2)}$, and then choosing c_1 if $\Upsilon > 1$, and c_2 otherwise, because the decision boundary is formed by the region of the feature space where $\Upsilon = 1$.

Table 2. Decision tree-based methods for PPI prediction

Type of feature	Dataset(s) used for model building	Organism(s)	Reference(s)
Sequence-based	Database of interacting proteins (DIPs)	Yeast	You *et al.* (2015a)
		Yeast and *H. pylori*	Xia *et al.* (2010)
	Human Protein References Database (HPRD)	Human	Zahiri *et al.* (2013)
	Protein Data Bank (PDB)	–	Chen and Jeong (2009)
	Murakami and Mizuguchi (2010)	–	Wei *et al.* (2015)
	Murakami and Mizuguchi (2010)	–	Liu *et al.* (2016)
	Wu *et al.* (2009)	Yeast	Sze-To *et al.* (2016)
	Martin *et al.* (2005), You *et al.* (2014), Huang *et al.* (2015)	*H. pylori*, yeast and human	Zhou *et al.* (2017)
Structure-based	Protein Data Bank (PDB)	–	Maheshwari and Brylinski (2017)
	Database of Three-dimensional Interacting Domains (3did)	–	Li *et al.* (2012)
Genomic feature-based	Munich Information Center for Protein Sequences (MIPS)	Yeast	Wang *et al.* (2009)
Domain/motif-based	Database of interacting proteins (DIPs)	Yeast	Chen and Liu (2005)
		Yeast, *Arabidopsis*, *Drosophila*, human, mouse	Rodgers-Melnick *et al.* (2013)

Table 3. Comparison of SVM- and RF-based PPI-prediction methods on PPI datasets from different organisms (Zhou *et al.* 2017)

Organism	Classifier	Accuracy (%)	*F*-score (%)	MCC (%)
Yeast	SVM	93.25	92.97	86.51
	RF	94.61	94.44	89.37
H. pylori	SVM	85.94	85.90	71.91
	RF	86.28	86.27	72.58
Human	SVM	96.45	96.20	92.96
	RF	97.57	97.44	95.15

Probabilistic methods based on log-odds scoring schemes have been widely used in PPI prediction, as well as for filtering high-throughput experimental datasets that can potentially include several FPs (Gavin *et al.* 2006; Krogan *et al.* 2006; Saha *et al.* 2010; Choi *et al.* 2011). Bayesian-probabilistic frameworks have also been proposed for PPI prediction using genomic features such as co-expression values, essentiality and co-localization (Jansen *et al.* 2003), structural features (Murakami and Mizuguchi 2010; Zhang *et al.* 2012) and sequence signatures (Sprinzak and Margalit 2001). Some prominent examples of PPI-prediction methods based on Bayesian conditional probability relationships are summarized in table 4.

The major disadvantage of Bayesian classification is the forced assumption of independence among the features, which is difficult to ensure in real-world problems. In addition, specifying a prior probability and computing a posterior probability for each data point can be extremely difficult and computationally infeasible in some cases.

2.4 *Artificial neural networks*

Artificial neural network (ANN) is one of the oldest machine-learning algorithms that can be used to perform nonlinear statistical modelling and develop binary classification models, and has now evolved into state-of-the-art deep-learning algorithms such as stacked autoencoders. ANNs require less formal statistical training, and are able to implicitly detect complex nonlinear relationships between dependent and independent variables, as well as, detect all possible interactions between predictor variables.

ANNs consist of a network of connections, each of which transfers the output of a neuron i to the input of a neuron j. Thus, i is the predecessor of j and j is the successor of i. Each connection is assigned a weight w_{ij}. The propagation function $p_j(t) = \sum_i o_i(t) w_{ij}$, computes the input $p_j(t)$ to the neuron j from the outputs $o_i(t)$ of predecessor neurons. Thus, a neuron with label j receives an input $p_j(t)$ from predecessor neurons, and consists of the following components:

- an activation $a_j(t)$, depending on a discrete time parameter,
- a threshold θ_j, which stays fixed unless changed by a learning function,
- an activation function f that computes the new activation at a given time $t+1$from $a_j(t)$, θ_j and the net input $p_j(t)$ giving $a_j(t+1) = f\left(a_j(t), p_j(t), \theta_j\right)$ and
- an output function f_{out} for computing the output from the activation as

$$o_j(t) = f_{\text{out}}\left(a_j(t)\right)$$

PPI-prediction methods based on ANNs and deep learning have been developed using several sequence (You *et al.* 2013; Yousef and Moghadam Charkari 2013; Sun *et al.* 2017; Wang *et al.* 2017b; Huang *et al.* 2018) and structural (Zhou and Shan 2001; Fariselli *et al.* 2002; Ofran and Rost 2007; Wang *et al.* 2010; Du *et al.* 2016) features as listed in table 5.

Table 4. Probabilistic/Bayesian methods for PPI prediction

Type of feature	Dataset(s) used for model building	Organism(s)	Reference(s)
PPI data analysis	Gavin *et al.* (2006)	Yeast	Gavin *et al.* (2006)
	Krogan *et al.* (2006)		Krogan *et al.* (2006)
	Sardiu *et al.* (2008), Sowa *et al.* (2009)		Choi *et al.* (2011)
	Krogan *et al.* (2006), Yu *et al.* (2008)		Saha *et al.* (2010)
	Collins *et al.* (2013), Teo *et al.* (2015)	–	Teo *et al.* (2016)
Sequence-based	Database of interacting proteins (DIPs)	Yeast	Sprinzak and Margalit (2001)
		Yeast and *H. pylori*	An *et al.* (2016)
			Li *et al.* (2017)
			Wang *et al.* (2017a)
Structure-based	Protein Data Bank (PDB)	Yeast and human	Zhang *et al.* (2012)
		–	Murakami and Mizuguchi (2010)
Genomic feature-based	Munich Information Center for Protein Sequences (MIPS)	–	Jansen *et al.* (2003)
	Human Protein References Database (HPRD)	Human	Scott and Barton (2007)
Domain/motif-based	Munich Information Center for Protein Sequences (MIPS)	Yeast	Sprinzak and Margalit (2001)
	Human Protein References Database (HPRD)	Human	Scott and Barton (2007)
	Landgraf *et al.* (2004), Tonikian *et al.* (2009)	Yeast	Jain and Bader (2016)

Table 5. ANN-based methods for PPI prediction

Type of feature	Dataset(s) used for model building	Organism(s)	Reference(s)
Sequence-based	Database of interacting proteins (DIPs)	Yeast	You *et al.* (2013)
			Huang *et al.* (2018)
		Yeast and *H. pylori*	Yousef and Moghadam Charkari (2013)
			Wang *et al.* (2017b)
	Human Protein References Database (HPRD)	Human	Sun *et al.* (2017)
			Huang *et al.* (2018)
Structure-based	Protein Data Bank (PDB)	–	Zhou and Shan (2001)
		–	Fariselli *et al.* (2002)
		–	Ofran and Rost (2007)
		–	Wang *et al.* (2010)
	Database of Three-dimensional Interacting Domains (3did)	–	Du *et al.* (2016)

The disadvantages of using ANN methods include their 'black box' nature, greater computational burden, proneness to overfitting and the empirical nature of model development.

2.5 *Clustering*

Clustering is the major form of unsupervised machine-learning technique applied to classification problems, which tries to segregate data points into groups such that data points placed in the same group are more similar to each other than to those in other groups. Clustering is useful in exploratory pattern analysis, pattern classification, decision making and also for outlier detection. The main advantage of clustering is its ability to determine the intrinsic classification within a set of unlabelled data, hence not requiring a separate training stage. Hence, clustering is generally used in cases where the class labels are not known in advance.

The different distance metrics used by clustering algorithms include: (a) Euclidean distance metric, (b) Euclidean squared distance metric, (c) Manhattan (city-block) distance, (d) Chebyshev distance, (e) Pearson's correlation coefficient,

Table 6. Clustering-based methods for PPI prediction

Type of feature	Dataset(s) used for model building	Organism(s)	Reference(s)
Network topology-based	Database of interacting proteins (DIPs)	Yeast	Liu *et al.* (2015)
	Munich Information Center for Protein Sequences (MIPS)	Yeast	Bader and Hogue (2003)
			Spirin and Mirny (2003)
	Others	Yeast	Xu and Guan (2014)
Structure-based	Protein Data Bank (PDB)	–	Fukuhara and Kawabata (2008)

(f) squared Pearson's correlation coefficient and (g) Spearman's rank correlation coefficient.

Clustering techniques have been used for PPI prediction using mainly the overall network topology of known interaction networks (Bader and Hogue 2003; Spirin and Mirny 2003; Xu and Guan 2014; Liu *et al.* 2015), except for one method that uses structural features (Fukuhara and Kawabata 2008), as shown in table 6.

The disadvantages of clustering include the inability to process high-dimensional datasets, getting stuck into local optima, ambiguity regarding cluster descriptors and the inherent randomness of the method, which causes problems with reproducibility. In addition, the final output shows the similarity of the object to a single cluster only, whereas, in reality, the object might have similarities to other clusters also, and this information is lost.

3. Discussion

The use of machine-learning methods for prediction of PPIs can help in discovery of novel PPIs by filtering out the large proportion of FPs and FNs reported by high-throughput experimental procedures and improving interactome coverage. Predicted PPIs from machine-learning methods can also be used to prioritize pairs of proteins for experimental assays, and may provide further insight into the specific context for the PPI such as tissue or phenotype. PPI networks discovered through computational prediction and subsequent experimental validation can be used for prediction of gene function (Mostafavi and Morris 2012), identification of disease genes (Navlakha and Kingsford 2010) and the discovery of novel therapeutics (Barabasi *et al.* 2011). Computational analyses of various physico-chemical, structural and functional attributes of interacting protein pairs also provide a better understanding of the molecular basis of PPIs. Nevertheless, the sets of features that have been used to predict PPIs are still not able to fully capture the dynamic and intricate specifications that can identify the true PPIs unambiguously.

The prediction of interacting protein pairs is a complex problem, since it is dependent not only on the sequence or structure, but also on other parameters such as cellular concentration and localization. Initially, various sequence- and structure-based features were used for training of SVMs, ANNs and Bayesian models for prediction of novel PPIs. It was observed that the prediction performance of RF-based methods is better than that of the more popular SVM-based methods. Several studies described in tables 1, 2, 4–6 have used a combination of more than one feature in a single method for prediction of PPIs with higher accuracy. However, the evaluation and incorporation of an optimum set of features encompassing all biochemical, contextual and structural information associated with proper identification of PPIs still remains elusive. Hence, the field of prediction of PPIs using machine-learning techniques is still open in terms of selecting new features, developing new algorithms and parameter optimization.

Acknowledgements

DS acknowledges the DBT-sponsored project titled, 'Centre of Excellence (CoE) in Bioinformatics Centre at Bose Institute' for financial support. This work is dedicated to the Centenary of Bose Institute.

References

Alonso-López D, Gutiérrez MA, Lopes KP, Prieto C, Santamaría R and De Las Rivas J 2016 APID interactomes: Providing proteome-based interactomes with controlled quality for multiple species and derived networks. *Nucleic Acids Res.* **44** W529–W535

An JY, You ZH, Meng FR, Xu SJ and Wang Y 2016 RVMAB: Using the relevance vector machine model combined with average blocks to predict the interactions of proteins from protein sequences. *Int. J. Mol. Sci.* **17** 757

Bader GD and Hogue CW 2003 An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinf.* **4** 2

Bader GR, Roth FP, Tavernier J and Vidal M 2017 *HuRI: The human reference protein interactome mapping project* (Canada: Bader Lab, The Donnelly Centre, The University of Toronto)

Bandyopadhyay S and Mallick K 2017 A new feature vector based on gene ontology terms for protein–protein interaction prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf./IEEE, ACM* **14** 762–770

Barabasi AL, Gulbahce N and Loscalzo J 2011 Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **12** 56–68

Barman RK, Saha S and Das S 2014 Prediction of interactions between viral and host proteins using supervised machine learning methods. *PLoS ONE* **9** e112034

Barman RK, Jana T, Das S and Saha S 2015 Prediction of intra-species protein–protein interactions in enteropathogens facilitating systems biology study. *PLoS ONE* **10** e0145648

Ben-Hur A and Noble WS 2005 Kernel methods for predicting protein–protein interactions. *Bioinformatics* **21** (Suppl 1) i38–i46

Blagus R and Lusa L 2010 Class prediction for high-dimensional class-imbalanced data. *BMC Bioinf.* **11** 523

Bock JR and Gough DA 2001 Predicting protein–protein interactions from primary structure. *Bioinformatics* **17** 455–460

Bradford JR and Westhead DR 2005 Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics* **21** 1487–1494

Breiman L, Friedman J, Stone CJ and Olshen RA 1984 *Classification and regression trees. Wadsworth statistics/probability* (Belmont, California: Chapman & Hall/CRC)

Carducci M, Perfetto L, Briganti L, Paoluzi S, Costa S, Zerweck J, Schutkowski M, Castagnoli L and Cesareni G 2012 The protein interaction network mediated by human SH3 domains. *Biotechnol. Adv.* **30** 4–15

Cestra G, Castagnoli L, Dente L, Minenkova O, Petrelli A, Migone N, Hoffmüller U, Schneider-Mergener J and Cesareni G 1999 The SH3 domains of endophilin and amphiphysin bind to the proline-rich region of synaptojanin 1 at distinct sites that display

an unconventional binding specificity. *J. Biol. Chem.* **274** 32001–32007

Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, Stark C, Breitkreutz BJ, Dolinski K and Tyers M 2017 The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45** D369–D379

Chatterjee P, Basu S, Kundu M, Nasipuri M and Plewczynski D 2011 PPI_SVM: Prediction of protein–protein interactions using machine learning, domain–domain affinities and frequency tables. *Cell. Mol. Biol. Lett.* **16** 264–278

Chen XW and Liu M 2005 Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics* **21** 4394–4400

Chen XW and Jeong JC 2009 Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics* **25** 585–591

Chen J, Sawyer N and Regan L 2013 Protein–protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci.: A Publ. Protein Soc.* **22** 510–515

Choi H, Larsen B, Lin ZY, Breitkreutz A, Mellacheruvu D, Fermin D, Qin ZS, Tyers M, Gingras AC and Nesvizhskii AI 2011 SAINT: Probabilistic scoring of affinity purification-mass spectrometry data. *Nat. Methods* **8** 70–73

Collins BC, Gillet LC, Rosenberger G, Röst HL, Vichalkovski A, Gstaiger M and Aebersold R 2013 Quantifying protein interaction dynamics by SWATH mass spectrometry: Application to the 14-3-3 system. *Nat. Methods* **10** 1246–1253

Du T, Liao L, Wu CH and Sun B 2016 Prediction of residue-residue contact matrix for protein–protein interaction with Fisher score features and deep learning. *Methods* **110** 97–105

Fariselli P, Pazos F, Valencia A and Casadio R 2002 Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.* **269** 1356–1361

Fukuhara N and Kawabata T 2008 HOMCOS: A server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. *Nucleic Acids Res.* **36** W185–W189

Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, *et al*. 2002 Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415** 141–147

Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, *et al*. 2006 Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440** 631–636

Guo Y, Yu L, Wen Z and Li M 2008 Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Res.* **36** 3025–3030

Hou T, Li N, Li Y and Wang W 2012 Characterization of domain–peptide interaction interface: Prediction of SH3 domain-mediated protein–protein interaction network in yeast by generic structure-based models. *J. Proteome Res.* **11** 2982–2995

Huang YA, You ZH, Gao X, Wong L and Wang L 2015 Using weighted sparse representation model combined with discrete cosine transformation to predict protein–protein interactions from protein sequence. *BioMed Res. Int.* **2015** 902198

Huang L, Liao L and Wu CH 2018 Completing sparse and disconnected protein–protein network by deep learning. *BMC Bioinf.* **19** 103

Jain S and Bader GD 2016 Predicting physiologically relevant SH3 domain mediated protein–protein interactions in yeast. *Bioinformatics* **32** 1865–1872

Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF and Gerstein M 2003 A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* **302** 449–453

Jones RB, Gordus A, Krall JA and MacBeath G 2006 A quantitative protein interaction network for the ErbB receptors using protein microarrays. *Nature* **439** 168–174

Kaushansky A, Gordus A, Chang B, Rush J and MacBeath G 2008 A quantitative study of the recruitment potential of all intracellular tyrosine residues on EGFR, FGFR1 and IGF1R. *Mol. BioSyst.* **4** 643–653

Kiemer L, Costa S, Ueffing M and Cesareni G 2007 WI-PHI: A weighted yeast interactome enriched for direct physical interactions. *Proteomics* **7** 932–943

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, *et al*. 2006 Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440** 637–643

Kundu K, Mann M, Costa F and Backofen R 2014 MoDPepInt: An interactive web server for prediction of modular domain–peptide interactions. *Bioinformatics* **30** 2668–2669

Landgraf C, Panni S, Montecchi-Palazzi L, Castagnoli L, Schneider-Mergener J, Volkmer-Engert R and Cesareni G 2004 Protein interaction networks by proteome peptide scanning. *PLoS Biol.* **2** E14

Li BQ, Feng KY, Chen L, Huang T and Cai YD 2012 Prediction of protein–protein interaction sites by random forest algorithm with mRMR and IFS. *PLoS ONE* **7** e43927

Li ZW, You ZH, Chen X, Li LP, Huang DS, Yan GY, Nie R and Huang YA 2017 Accurate prediction of protein–protein interactions by integrating potential evolutionary information embedded in PSSM profile and discriminative vector machine classifier. *Oncotarget* **8** 23638–23649

Liu GH, Shen HB and Yu DJ 2016 Prediction of protein–protein interaction sites with machine-learning-based data-cleaning and post-filtering procedures. *J. Membr. Biol.* **249** 141–153

Liu P, Yang L, Shi D and Tang X 2015 Prediction of protein–protein interactions related to protein complexes based on protein interaction networks. *BioMed Res. Int.* **2015** 259157

Maheshwari S and Brylinski M 2017 Across-proteome modeling of dimer structures for the bottom-up assembly of protein–protein interaction networks. *BMC Bioinf.* **18** 257

Martin S, Roe D and Faulon JL 2005 Predicting protein–protein interactions using signature products. *Bioinformatics* **21** 218–226

Mei S 2013 Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins. *PLoS ONE* **8** e79606

Miller ML, Jensen LJ, Diella F, Jørgensen C, Tinti M, Li L, Hsiung M, Parker SA, *et al*. 2008 Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal.* **1** ra2

Mostafavi S and Morris Q 2012 Combining many interaction networks to predict gene function and analyze gene lists. *Proteomics* **12** 1687–1696

Mrowka R, Patzak A and Herzel H 2001 Is there a bias in proteome research? *Genome Res.* **11** 1971–1973

Murakami Y and Mizuguchi K 2010 Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics* **26** 1841–1848

Navlakha S and Kingsford C 2010 The power of protein interaction networks for associating genes with diseases. *Bioinformatics* **26** 1057–1063

Ofran Y and Rost B 2007 ISIS: Interaction sites identified from sequence. *Bioinformatics* **23** e13–e16

Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, *et al*. 2014 The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42** D358–D363

Qi Y, Bar-Joseph Z and Klein-Seetharaman J 2006 Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins* **63** 490–500

Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schächter V, Chemama Y, Labigne A and Legrain P 2001 The protein–protein interaction map of *Helicobacter pylori*. *Nature* **409** 211–215

Rodgers-Melnick E, Culp M and DiFazio SP 2013 Predicting whole genome protein interaction networks from primary sequence data in model and non-model organisms using ENTS. *BMC Genomics* **14** 608

Ruan P, Hayashida M, Akutsu T and Vert JP 2018 Improving prediction of heterodimeric protein complexes using combination with pairwise kernel. *BMC Bioinf.* **19** 39

Saha S, Kaur P and Ewing RM 2010 The bait compatibility index: Computational bait selection for interaction proteomics experiments. *J. Proteome Res.* **9** 4972–4981

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU and Eisenberg D 2004 The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **32** D449–D451

Sardiu ME, Cai Y, Jin J, Swanson SK, Conaway RC, Conaway JW, Florens L and Washburn MP 2008 Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proc. Natl. Acad. Sci. USA* **105** 1454–1459

Sarkar D, Jana T and Saha S 2015 LMPID: A manually curated database of linear motifs mediating protein–protein interactions. *Database: J. Biol. Databases Curation* **2015** bav014

Sarkar D, Jana T and Saha S 2018 LMDIPred: A web-server for prediction of linear peptide sequences binding to SH3, WW and PDZ domains. *PLoS One* **13** e0200430

Scott MS and Barton GJ 2007 Probabilistic prediction and ranking of human protein–protein interactions. *BMC Bioinf.* **8** 239

Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y and Jiang H 2007 Predicting protein–protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. USA* **104** 4337–4341

Sowa ME, Bennett EJ, Gygi SP and Harper JW 2009 Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138** 389–403

Sparks AB, Rider JE, Hoffman NG, Fowlkes DM, Quillam LA and Kay BK 1996 Distinct ligand preferences of Src homology 3 domains from Src, Yes, Abl, Cortactin, p53bp2, PLCgamma, Crk, and Grb2. *Proc. Natl. Acad. Sci. USA* **93** 1540–1544

Spirin V and Mirny LA 2003 Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA* **100** 12123–12128

Sprinzak E and Margalit H 2001 Correlated sequence-signatures as markers of protein–protein interaction. *J. Mol. Biol.* **311** 681–692

Srinivasulu YS, Wang JR, Hsu KT, Tsai MJ, Charoenkwan P, Huang WL, Huang HL and Ho SY 2015 Characterizing informative sequence descriptors and predicting binding affinities of heterodimeric protein complexes. *BMC Bioinf.* **16** (Suppl 18) S14

Sriwastava BK, Basu S and Maulik U 2015 Predicting protein–protein interaction sites with a novel membership based fuzzy SVM classifier. *IEEE/ACM Trans. Comput. Biol. Bioinf./IEEE, ACM* **12** 1394–1404

Stiffler MA, Chen JR, Grantcharova VP, Lei Y, Fuchs D, Allen JE, Zaslavskaia LA and MacBeath G 2007 PDZ domain binding selectivity is optimized across the mouse proteome. *Science* **317** 364–369

Su C, Peregrin-Alvarez JM, Butland G, Phanse S, Fong V, Emili A and Parkinson J 2008 *Bacteriome.org*: an integrated protein interaction database for *E. coli*. *Nucleic Acids Res.* **36** D632–D636

Sun T, Zhou B, Lai L and Pei J 2017 Sequence-based prediction of protein–protein interaction using a deep-learning algorithm. *BMC Bioinf.* **18** 277

Sze-To A, Fung S, Lee EA and Wong AKC 2016 Prediction of protein–protein interaction via co-occurring aligned pattern clusters. *Methods* **110** 26–34

Teo G, Kim S, Tsou CC, Collins B, Gingras AC, Nesvizhskii AI and Choi H 2015 mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. *J. Proteomics* **129** 108–120

Teo G, Koh H, Fermin D, Lambert JP, Knight JD, Gingras AC and Choi H 2016 SAINTq: Scoring protein–protein interactions in affinity purification: mass spectrometry experiments with fragment or peptide intensity data. *Proteomics* **16** 2238–2245

Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, *et al*. 2002 A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295** 321–324

Tonikian R, Zhang Y, Sazinsky SL, Currell B, Yeh JH, Reva B, Held HA, Appleton BA, *et al*. 2008 A specificity map for the PDZ domain family. *PLoS Biol.* **6** e239

Tonikian R, Xin X, Toret CP, Gfeller D, Landgraf C, Panni S, Paoluzi S, Castagnoli L, *et al*. 2009 Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. *PLoS Biol.* **7** e1000218

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, *et al*. 2000 A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403** 623–627

Vapnik VN 1999 An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **10** 988–999

Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, *et al*. 2009 An empirical framework for binary interactome mapping. *Nat. Methods* **6** 83–90

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S and Bork P 2002 Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417** 399–403

Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N and Vidal M 2000 Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287** 116–122

Wang J, Li C, Wang E and Wang X 2009 Uncovering the rules for protein–protein interactions from yeast genomic data. *Proc. Natl. Acad. Sci. USA* **106** 3752–3757

Wang B, Chen P, Wang P, Zhao G and Zhang X 2010 Radial basis function neural network ensemble for predicting protein–protein interaction sites in heterocomplexes. *Protein Pept. Lett.* **17** 1111–1116

Wang Y, You Z, Li X, Chen X, Jiang T and Zhang J 2017a PCVMZM: Using the probabilistic classification vector machines model combined with a zernike moments descriptor to predict protein–protein interactions from protein sequences. *Int. J. Mol. Sci.* **18**(5) E1029

Wang YB, You ZH, Li X, Jiang TH, Chen X, Zhou X and Wang L 2017b Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. *Mol. BioSyst.* **13** 1336–1344

Wei ZS, Yang JY, Shen HB and Yu DJ 2015 A cascade random forests algorithm for predicting protein–protein interaction sites. *IEEE Trans. Nanobiosci.* **14** 746–760

Wu J, Vallenius T, Ovaska K, Westermarck J, Mäkelä TP and Hautaniemi S 2009 Integrated network analysis platform for protein–protein interactions. *Nat. Methods* **6** 75–77

Xenarios I, Salwínski L, Duan XJ, Higney P, Kim SM and Eisenberg D 2002 DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30** 303–305

Xia JF, Han K and Huang DS 2010 Sequence-based prediction of protein–protein interactions by means of rotation forest and autocorrelation descriptor. *Protein Pept. Lett.* **17** 137–145

Xu B and Guan J 2014 From function to interaction: A new paradigm for accurately predicting protein complexes based on protein-to-protein interaction networks. *IEEE/ACM Trans. Comput. Biol. Bioinf./IEEE, ACM* **11** 616–627

You ZH, Lei YK, Zhu L, Xia J and Wang B 2013 Prediction of protein–protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinf.* **14** (Suppl 8) S10

You ZH, Zhu L, Zheng CH, Yu HJ, Deng SP and Ji Z 2014 Prediction of protein–protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. *BMC Bioinf.* **15** (Suppl 15) S9

You ZH, Chan KC and Hu P 2015a Predicting protein–protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. *PLoS One* **10** e0125811

You ZH, Li J, Gao X, He Z, Zhu L, Lei YK and Ji Z 2015b Detecting protein–protein interactions with a novel matrix-based protein sequence representation and support vector machines. *BioMed Res. Int.* **2015** 867516

Yousef A and Moghadam Charkari N 2013 A novel method based on new adaptive LVQ neural network for predicting protein–protein interactions from protein sequences. *J. Theor. Biol.* **336** 231–239

Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, *et al*. 2008 High-quality binary protein interaction map of the yeast interactome network. *Science* **322** 104–110

Yu CY, Chou LC and Chang DT 2010 Predicting protein–protein interactions in unbalanced data using the primary structure of proteins. *BMC Bioinf.* **11** 167

Yugandhar K and Gromiha MM 2014 Feature selection and classification of protein–protein complexes based on their binding affinities using machine learning approaches. *Proteins* **82** 2088–2096

Zahiri J, Yaghoubi O, Mohammad-Noori M, Ebrahimpour R and Masoudi-Nejad A 2013 PPIevo: Protein–protein interaction prediction from PSSM based evolutionary information. *Genomics* **102** 237–242

Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A and Honig B 2012 Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature* **490** 556–560

Zhou HX and Shan Y 2001 Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins* **44** 336–343

Zhou C, Yu H, Ding Y, Guo F and Gong XJ 2017 Multi-scale encoding of amino acid sequences for predicting protein interactions using gradient boosting decision tree. *PLoS ONE* **12** e0181426

Zhu H, Domingues FS, Sommer I and Lengauer T 2006 NOXclass: Prediction of protein–protein interaction types. *BMC Bioinf.* **7** 27

Corresponding editor: BJ Rao