Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS

ISSN 2454-6348

Original Research Article **DOI - 10.26479/2017.0303.01**

PREDICTING PROTEIN–PROTEIN INTERACTIONS USING SEQUENCE HOMOLOGY AND MACHINE-LEARNING METHODS

Yifan Tang[1], Cao Wei[1], Kazuya Sumikoshi[1], Shugo Nakamura[1], Tohru Terada[2], Koji Kadota[2], Kentaro Shimizu[1]*

1. Department OF Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan.
2. Agricultural Bioinformatics Research Unit, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan.

ABSTRACT: Protein–protein interactions (PPIs) play an essential role in various biological processes. A range of computational methods have been proposed to predict PPIs from protein sequences. Among these, homology-based methods and machine-learning methods have been widely used. However, to the best of our knowledge, these two methods have not been compared using the same dataset. Thus in this study, we have developed both homology-based and machine-learning methods to predict PPIs from amino-acid sequences and compared the prediction results. In the homology-based method, BLASTP search was used to identify sequence homology. Regarding the machine-learning methods, two popular methods, support vector machine and random forest, as well as six different protein features, were employed to build classifiers. We collected the PPI pairs with high-confidence scores from HitPredict4 to build the positive dataset and we built the negative dataset from the Negatome 2.0 database, in which non-interacting pairs were verified by experiments and 3D structure analysis. Our results show that machine-learning methods achieved better performance than homology-based method but there are many PPIs that are predicted only by the homology-based method. The integration of the two methods is expected to enhance the performance.

KEYWORDS: protein–protein interactions, protein–protein interaction site prediction, support vector machine, machine learning, homology search.

***Corresponding Author: Prof. Kentaro Shimizu** Ph.D.

Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan. * Email Address: shimizu@bi.a.u-tokyo.ac.jp

1.INTRODUCTION

Protein–protein interactions (PPIs) play an essential role in various biological processes and functions in living cells, such as metabolic cycles, gene regulation, and signal transduction [1–3]. Thus, identification of PPIs is critical to understanding the protein functions. Over the past few decades, many experimental techniques, such as yeast two-hybrid systems (Y2H) [4], mass spectrometry (MS) [5], tandem affinity purification (TAP) [6], and protein chip [7], have been developed to detect PPIs. These experimental approaches have provided an enormous amount of PPI data, which have facilitated the development of PPI databases such as IntAct [8], BioGRID [9], and HPRD [10]. However, conducting experiments is labor-intensive and time-consuming [11], and the PPI networks are still incomplete [12]. To overcome these limitations, bioinformatics methods are expected to be useful for identifying PPIs in functional proteomics. Various computational methods have been proposed to predict PPIs, which include genomic context-based methods [13–17], structure-based methods [18–21], and sequence-based methods [22–32]. Genomic context-based methods such as gene-cluster and gene-neighbor methods are based on the search of pairs of genes that show a correlated position or behavior; these genes are assumed to encode proteins that interact with each other [13]. As an example of genomic context-based methods, the phylogenetic profile method constructs a phylogenetic profile of proteins with a binary vector that represents their presence or absence across many organisms. Two proteins shown to have similar profiles by this approach may be functionally related or may interact [14,15]. The gene-fusion method [16] is based on the observation that some single-domain proteins in an organism can fuse to form multidomain proteins in other organisms. This indicates that functionally associated proteins are likely to form a protein complex [17]. Structure-based methods, such as Struct2Net [18], thread protein sequences to all the protein complex structures from the PDB database. Based on the matched structures, logistic regression is used to evaluate the probability of two proteins interacting. PRISM [19] is a template-based method that compares the two sides of the template complex interface with the surfaces of two target monomers in terms of structural alignment. If regions of the target surfaces are similar to the sides of the template interface, these two targets are predicted to interact with each other. Other methods, such as PrePPI [20], predict PPIs by structural alignment combined with secondary structures, while MEGADOCK [21] employs docking simulation to draw inferences on PPIs. The above approaches rely on an abundance of information being available. In the genomic context-based methods, the performance depends on the number and diversity of genomes. It is difficult to predict PPIs of the proteins specific to only one organism. When the structures of the target proteins are known, structure-based methods can achieve high accuracy. However, if modeled structures are used, additional computational time is required and prediction accuracy depends on the model quality. Apart from these methods, sequence-based methods have shown the advantage of generalization because they require information only from amino-acid sequences. Many sequence-based methods

have been proposed and the majority of them use machine-learning techniques for building classifiers. For example, support vector machine (SVM) [33] is an efficient machine-learning algorithm used to identify PPIs. As one example of an SVM approach, Martin et al. [22] developed a novel descriptor called signature product to predict PPIs. The signature product is implemented within SVM as a kernel function and accuracy of 69% could be obtained and applied to a full-yeast dataset from the DIP database. Shen et al. [23] also employed SVM to predict PPIs. In Shen's study, the 20 amino acids were clustered into seven classes according to their dipoles and volumes of their side chains, and then the triads of the classes of adjacent residues were taken as input features for SVM. When applied to predict human PPIs, this method shows good performance with 83.9% accuracy. Guo et al. [24] used an autocovariance model for feature extraction where autocovariance of the indexes representing the physicochemical properties of each amino acid was calculated. This method achieved 87.36% accuracy by SVM under the yeast-core dataset in DIP [25]. Another popular machine-learning method—random forest (RF) [34]—has also been employed in PPI research. For example, Zahiri et al. [26] attempted to combine the position-specific scoring matrix (PSSM) feature with other features, such as post-translational modifications and tissue information, to predict PPIs. Four different classifiers—RF, naïve Bayes (NB), multilayer perceptron, and radial basis function network (RBF)—were applied to prediction. From the prediction results, the RF classifier showed the best performance among these four classifiers. In addition, You et al. [27] proposed a novel multiscale local descriptor (MLD) feature representation. The MLD feature makes it easier to extract multiple continuous binding patterns within a protein sequence. On the basis of the RF classifier with MLD, a high performance of 94.72% accuracy in predicting interactions between proteins was obtained when applied to the yeast-core dataset in DIP. The homology-based method is also applied to predict PPIs from sequences. This method is based on the assumption that two pairs of proteins with high sequence similarity may have similar properties. As the number of reliable PPI data increases, protein interaction mapping becomes useful for the functional annotation of uncharacterized proteins in various species [28]. On the basis of this concept, Yu et al. [29] introduced a new generalized mapping method based on sequence similarity. Specifically, they used a total of 14,911 interactions to investigate the relationships between the sequence similarity and the conservation of interaction. They suggested that sequence similarity above a certain threshold might be a reliable measure for identifying PPIs. Recently, several web servers, such as PPI-Search [30] and BIPS [31], have also been constructed for searching for homologous PPIs based on sequences. These services may assist in the identification of pairs of proteins that potentially interact. Among the sequence-based methods, the homology-based method depends on the conservation between sequences, while the machine-learning method relies on feature extraction and learning algorithms. Both methods were implemented individually to predict PPIs in previous studies, but the evaluation has not been made on the same dataset to compare their performance. Furthermore, the reliability of

the dataset plays a crucial role in the prediction of PPIs. In this study, we developed both homology-based and machine-learning methods to predict PPIs from amino-acid sequences. In the homology-based method, BLASTP search was used to identify sequence homology. BLASTP is a BLAST (Basic Local Alignment Search Tool) program for searching protein databases. In our analysis, NCBI BLAST 2.2.29+ was used. Regarding machine-learning methods, two popular methods, SVM and RF, as well as six different protein features, were employed to build classifiers. These methods give theoretical background and they have achieved good performances in previous studies as described in "Introduction". We collected highly confident positive pairs from HitPredict4 [34] to build the positive dataset. Regarding the collection of the negative dataset, it is still a challenge to ensure the high reliability of results asserting that certain pairs of proteins do not interact. In many previous studies, negative datasets were generated by randomly mapping proteins that appeared in the positive dataset or selecting protein pairs that have different subcellular localizations. In contrast, in this study, we built the negative dataset from the Negatome 2.0 database [36], in which non-interacting pairs were verified by experiments and 3D structure analysis.

2. MATERIALS AND METHODS

2.1 Materials

We used HitPredict 4 (version of September 2015) as the positive data source, which provides a manually curated dataset of 398,696 physical interactions among 70,808 proteins from 105 species. It is reliable resource of experimentally identified, physical protein–protein interactions with confidence scores to indicate their reliability [34]. On the basis of the statistical analysis, a method-based score of >0.485 or an annotation-based score of >0.5 was suggested as thresholds for identifying high-confidence interactions. We built a balanced dataset by collecting a suitable number of PPIs with efficient cut-off points from HitPredict 4 (398,696 pairs) and Negatome 2.0 (6,136 pairs). Since the number of supporting publications can be taken as direct evidence for evaluating the quality of PPIs, this was used as a cut-off in our study. All PPI sequences were derived from UniProtKB [37]. To ensure the reliability of these sequences, we checked the PE entry for each protein in UniProtKB and only retained those that were annotated with "evidence at protein level" or "evidence at transcript level." The two reliable thresholds suggested by HitPredict 4 (method-based score and annotation-based score) were also used to ensure that the selected PPIs had sufficient evidence supported by experiments.We used Negatome 2.0 to build our negative dataset, in which non-interacting protein pairs were derived from manual curation of the literature and by analyzing the 3D structures of protein complexes. An initial stringent dataset of 6,136 pairs was downloaded from the Negatome 2.0 web page and the sequences of these negative pairs were then obtained from the UniProtKB database.In both the positive and the negative datasets, we excluded the following pairs: (1) pairs that existed in both HitPredict 4 and Negatome 2.0; (2) sequences of <50 amino acids; (3) one of a pair of repeat sequences (e.g., A-B and B-A); and (4) sequences containing the amino acid characters B, J, O, U, X,

and Z. This resulted in a total of 9,566 interacting pairs and 4,720 non-interacting pairs being collected to build our reliable dataset. Since a protein pair was taken as a unit in this study, sequence redundancy refers to the redundancy of protein pair sequences. First, we used BLASTClust [38] to cluster all of the protein pair sequences with an identity of 40% and coverage of 70%. On the basis of these clusters, we then defined the similarity of the protein pairs as follows: let A be a pair of protein A1 and protein A2 and B be a pair of B1 and B2; if (A1, B1) and (A2, B2) are both in the same respective sequence clusters or (A1, B2) and (A2, B1) are both in the same respective sequence clusters, then A and B are defined as a similar pair. We then grouped similar protein pairs based on single linkage clustering [39]. As a result, we obtained 8,388 clusters where the members were all positive pairs and 3,867 clusters where the members were all negative pairs. The longest pair (sum of the lengths of proteins in the pair is the longest) was taken to be representative of that cluster. Finally, we ranked the 8,388 positive clusters by the combined interaction score of the representative pairs and selected the top 3,867 positive pairs as the final positive dataset. All 3,867 representative negative pairs in the negative clusters were selected as the final negative dataset. We developed two kinds of methods to predict whether a query protein pair interacts or not, based on the amino-acid sequences (Figure 1).

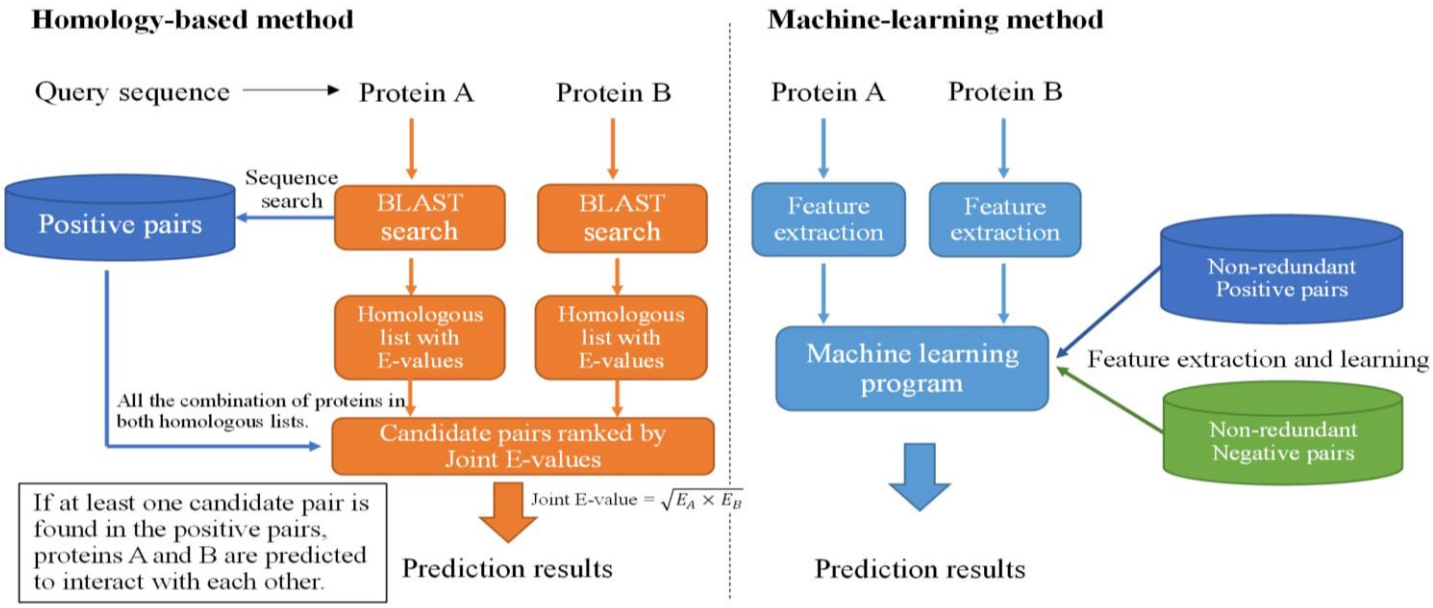

Figure 1 Homology-based method and machine-learning method. This figure provides an outline of the homology-based method and the machine-learning method.

2.2 Homology-based method

Homologs are proteins with significant sequence similarity. Here we used the same definition as provided in [40], whereby two proteins were considered homologous if they had an E-value 10^{-10} in a BLASTP [37] search. This value is small enough to avoid false-positive detection. We also used the joint sequence similarity (JE) [29] to measure the similarity between two protein sequences, expressed by:

$$\text{Joint } E\text{-value} = \sqrt{E_A \times E_B}$$

where E_A and E_B denote the E-values of A and B, respectively. An E-value that is close to 0 indicates higher reliability of the homology between two sequences. However, if either of the two sequences has an E-value of 0, J_E could also be 0, which would overinflate the J_E value when the other E-value is high. To avoid this, the minimum E-value was assigned as 10^{-180} rather than 0. In the homology-based method, we used HINTdb [41] to identify known interactions between PPIs. HINTdb includes 398,696 physical PPIs, all of which are based on empirical evidence. A total of 70,808 sequences registered in HINTdb were employed as a sequence database for determining sequence homology. For each protein pair used as a query, we first obtained two homologous lists using BLASTP to search for the two sequences in the HINTdb sequence database separately. We then calculated J_E for all possible combinations of homologous sequences in the two lists and listed the candidate pairs in ascending order of J_E. A score of "+1" was assigned to candidate pairs that could be found in HINTdb and "0" was assigned to all other cases. Finally, we predicted that a particular query pair interacted with each other when at least one candidate pair had a score of "1." This practical predicting process is shown in Figure 2, which provides an example using a query of the two protein sequences Q12834 and Q13257 (UniProtKB). A BLASTP search was first used to find homologous proteins. We then combined the members within the two homologous lists Q1 and Q2 and calculated J_E for each combination. After ranking these according to J_E, we searched the HINTdb database for all of the homologous pairs, which yielded two pairs with a score of "+1" (Q24044 and Q9VRQ2; and Q9SZA4 and Q9LU93). Therefore, we predicted that proteins Q12834 and Q13257 interact with each other.

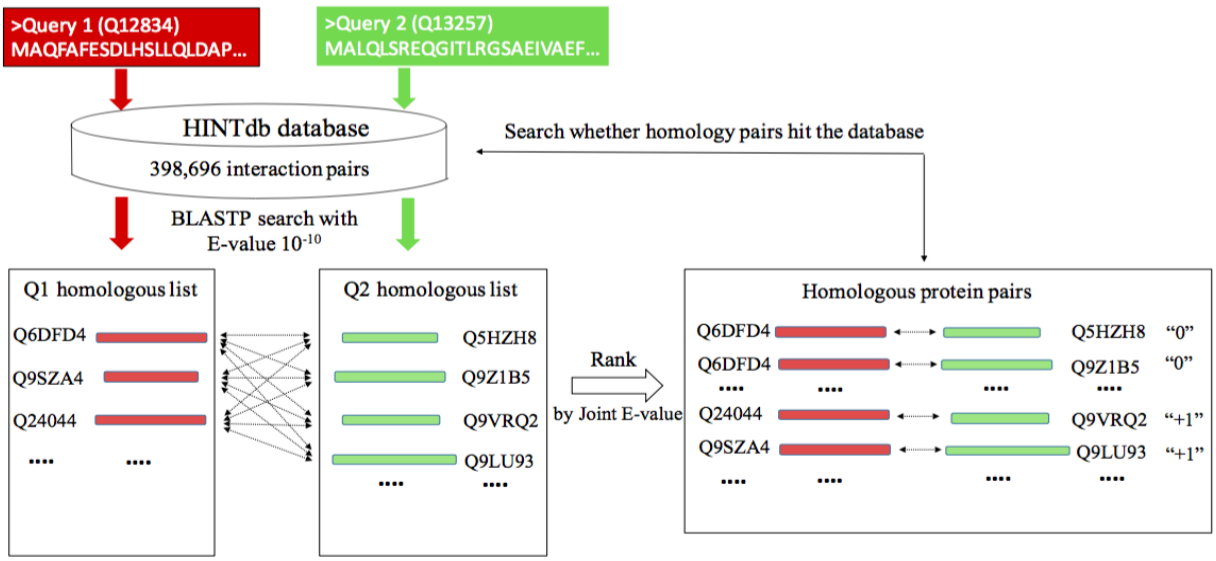

Figure 2 Flow chart of the method used to predict whether a query protein pair interact with each other (UniProtKB: Q12834 and Q13257).

We also considered the impact of J_E on the prediction results by using J_E as a cut-off to select candidate pairs with a similarity that was less than the threshold. A J_E threshold set (10^{-10}, 10^{-30}, 10^{-50}, 10^{-70},

10^{-90}, 10^{-110}, 10^{-130}, and 10^{-150}) was built for setting the acceptance of the similarity for the candidate pairs. Since the E-value cut-off in BLASTP was set to 10^{-10}, a JE value of 10^{-10} was equivalent to the case in which no threshold of E-value was applied, while 10^{-150} reflected the most stringent case.

2.3 Machine-learning methods and feature extraction

We used SVM and RF as the machine-learning algorithms. SVM shows good performance and generalization abilities for classification and regression analysis [23]. We adopted the RBF or Gaussian kernel as a kernel function. The RF algorithm employs a collection of decision trees to improve the stability and accuracy of classification. These algorithms require a fixed length of the feature vector for training and testing. We implemented six different feature extractions that succeeded in representing variable lengths of protein sequences: amino-acid composition (AAC) [42], dipeptide composition (DC) [42], tripeptide composition (TC) [43], pseudo-amino-acid composition (PseAAC) [44], MLD [27], and autocovariance (AC) [24]. The definitions of these features are described below.

2.3.1 Amino-acid composition (AAC)

AAC counts the frequency of each of the 20 amino acids based on the protein sequence. For a given protein sequence P, the feature vector $\Phi(P)$ can be calculated as:

$$\Phi(P) = \left[\ f(x)_{x\in\{A,R,N\ldots V\}}\right]$$

where

$$f(x) = \frac{\text{total number of amino acid } x \text{ in sequence}}{\text{total number of amino acids in sequence}}$$

2.3.2 Dipeptide composition (DC)

DC takes every two consecutive amino acids as a single unit and counts the frequency of all of the dipeptide patterns. It then represents a sequence with a fixed length vector of 400 (= 20 × 20). For a given protein sequence P, the feature vector $\Phi(P)$ is:

$$\Phi(P) = \left[\ f(x)_{x\in\{AA,AR,\ldots VN,\ldots VV\}}\right]$$

where

$$f(x) = \frac{\text{total number of dipeptide } x \text{ in sequence}}{\text{total number of dipeptides in sequence}}$$

2.3.3 Tripeptide composition (TC)

In general, TC calculates the frequency of three consecutive amino acids and results in an 8,000 (= 20 × 20 × 20)-dimensional vector. We reduced the dimensionality of this vector by classifying the 20 amino acids into seven groups based on the dipoles and side-chain volumes (**Table 1**) [23]. Thus, the protein sequence P was transformed according to the names of the seven groups. TC counts the frequency of three consecutive groups and forms a 343 (= 7 × 7 × 7)-dimensional vector according to the following equation:

$$\Phi(P) = \left[\ f(x)_{x\in\{111,112,\ldots,776,777\}}\right]$$

where

$$f(x) = \frac{\text{total number of tripeptide } x \text{ in sequence}}{\text{total number of tripeptides in sequence}}$$

Table 1 The seven groups of amino acids based on their dipoles and side-chain volumes.

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
A, G, V	I, L, F, P	Y, M, T, S	H, N, Q, W	R, K	D, E	C

2.3.4 Pseudo-amino-acid composition (PseAAC)

PseAAC not only incorporates the amino-acid composition, but also considers the sequence order. The sequence order is represented by a series of sequence correlation factors, which are defined by a correlation function that includes hydrophobicity, hydrophilicity, and side-chain volume. The original values of these three physicochemical properties for each amino acid are listed in **Supplementary Table S1**. They were first normalized by:

$$N_{R,j} = \frac{P_{R,j} - P_j}{S_j}$$

where *R* relates to the 20 amino acids and *j* relates to the three properties mentioned above. P_j is the mean of the *j*-th property value across all 20 amino acids; S_j is the standard deviation of the *j*-th property value across all 20 amino acids; and $N_{R,j}$ is the normalized value of the *j*-th property for amino acid *R*. The correlation factors were then calculated according to the following equation:

$$\theta_{lag} = \frac{1}{n - lag} \sum_{i=1}^{n-lag} \sum_{j=1}^{3} \left(N_{X_i,j} - N_{X_{i+lag},j} \right)^2 \times \frac{1}{3}$$

where *lag* is the interval between one residue and its vicinal residue, *n* is the sequence length of protein *X*, and X_i represents the amino acid at the *i*-th position of *X*. *j* relates to the three physicochemical properties. A given protein sequence is presented by a series of sequence correlation factors given below:

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \theta_j}, & (1 \le u \le 20) \\ \dfrac{w\theta_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \theta_j}, & (20+1 \le u \le 20+\lambda) \end{cases}$$

where f_i is the normalized occurrence frequency of the 20 amino acids, θ_j is the correlation factor across distance j, λ controls the range of the sequence order considered, and w is the weight factor for the sequence order effect. Thus, a sequence can be represented by:

$$\Phi(P) = [x_1, x_2, x_u \ldots, x_{20+\lambda}]$$

We calculated PseAAC using the default parameters $\lambda = 10$ and $w = 0.05$, as proposed in [44].

2.3.5 Multiscale local descriptor (MLD)

MLD is a proposed method for transforming the protein sequences into feature vectors by using a binary coding scheme. A protein sequence is transformed into groups based on the dipoles and side-chain volumes. The entire sequence is then divided into multiple sequence segments of varying lengths to describe local regions. In MLD, the protein sequence is divided into four equal-length segments (S1, S2, S3, and S4), following which 16 different combinations are derived using a 4-bit binary coding scheme. For example, 1100 refers to the continuous region constructed by S1 and S2. In MLD, only nine continuous sub-sequences are considered: 0001, 0010, 0011, 0100, 0110, 0111, 1000, 1100, and 1110. For each sub-sequence, the local descriptors Composition, Transition, and Distribution (CTD) [32] are calculated and concatenated. In CTD, the sequence is represented by seven groups of amino acids, which is the same as TC. Composition calculates the frequency of each group, Transition characterizes the frequency with which amino acids in one group are followed by amino acids in another group, and Distribution measures the location of the first, 25%, 50%, 75%, and 100% of the amino acids in the group. For example, the sub-sequence "AGCMTYCCACCCASYAGCCGYG" would be transformed into "11233322122213311221 31" according to the amino-acid classification. The composition is 36.36% (= 8/22) for "1," 36.36% (= 8/22) for "2," and 27.27% (= 6/22) for "3." There are three types of transitions in this transformed sequence, giving a Transition of 28.57% (= 6/21) for "1" to "2" or "2" to "1"; 19% (= 4/21) for "1" to "3" or "3" to "1"; and 9.52% (= 2/21) for "2" to "3" or "3" to "2." In terms of Distribution, eight residues are represented by "1," the rankings of which at the first, 25%, 50%, 75%, and 100% of occurrences are 1st, 2nd (= 8 × 25%), 4th (= 8 × 50%), 6th (= 8 × 75%), and 8th (= 8 × 100%). The locations of "1" at the 1st, 2nd, 4th, 6th, and 8th positions in this sequence are 1, 2, 13, 17, and 22, respectively. Hence, the Distributions for "1" are 4.55% (= 1/22), 9.09% (= 2/22), 59.09% (= 13/22),

77.27% (= 17/22), and 100% (= 22/22). Similarly, the Distributions for "2" and "3" are 13.64%, 31.82%, 45.45%, 54.55%, and 86.36%; and 18.18%, 18.18%, 27.27%, 63.64%, and 95.45%, respectively. For each continuous region, CTD generates a 63-dimensional vector: 7 for composition, 21 (= 7 × [6/2]) for Transition, and 35 (= 7 × 5) for Distribution. Nine sub-sequences are then calculated and concatenated for a 567 (= 63 × 9)-dimensional feature vector.

2.3.6 Autocovariance (AC)

In AC, seven physicochemical properties of amino acids were selected to represent the sequence feature: hydrophobicity, hydrophilicity, amino-acid side-chain volume, polarity, polarizability, solvent-accessible surface area, and net charge index of the amino-acid side chains, respectively. The original values of these seven physicochemical properties for each amino acid are listed in **Supplementary Table S2**. They were first normalized to zero mean and unit standard deviation according to the following equation:

$$N_{R,j} = \frac{P_{R,j} - P_j}{S_j}$$

where R relates to the 20 amino acids and j relates to the seven physicochemical properties. P_j is the mean of the j-th property value across all 20 amino acids, S_j is the standard deviation of the j-th property value across all 20 amino acids, and $N_{R,j}$ represents the normalized value of the j-th property for amino acid R.

AC uses the *lag* (i.e., the interval between one residue and its vicinal residue) to transform the variable length of protein sequences into a fixed length of the feature vector. For a given protein sequence, the AC feature is calculated by the following equation:

$$AC_{lag,j} = \frac{1}{n - lag} \sum_{i=1}^{n-lag} \left(N_{X_i,j} - \frac{1}{n} \sum_{i=1}^{n} N_{X_i,j} \right) \times \left(N_{X_{(i+lag),j}} - \frac{1}{n} \sum_{i=1}^{n} N_{X_i,j} \right)$$

where *lag* represents the distance from its neighbor, j is the j-th property among the seven physicochemical properties, and X_i refers to the amino acid at position i of sequence X. Thus, the total length of AC is *lag* × 7 and is defined by:

$$\Phi(P) = \left[AC_{1,j}, AC_{2,j}, \ldots, AC_{30,j} \right], j \in \{1,2, \ldots,7\}$$

In this study, *lag* was set to 30, which is the optimal value reported by Guo [24].

2.3.7 Representing proteins

All of the above feature extraction methods were used to represent an individual protein sequence, separately. Because our goal was to predict whether pairs of proteins interact, a protein pair was represented by concatenating the feature vectors of two sequences in the protein pair. Two popular machine-learning techniques, SVM and RF, were then applied to build classifiers separately.

2.4 Optimization of parameters

Optimization of the training model is a common step in machine learning, for which grid search is

often the method of choice. Grid search simply involves conducting exhaustive searching over a manually specified subset of candidate parameters. This search is usually carried out using the cross-validation method. The optimal parameters were then selected for the training model based on the performance of all of the parameters employed in the grid search. In this study, grid search was implemented in both SVM and RF to search for the best parameters. In general, an SVM classifier with RBF kernel has at least two parameters that need to be tuned for good performance: the cost parameter, which determines the misclassification penalty; and the gamma parameter, which is used in the RBF kernel function. The values of 1, 10, 100, and 1,000 were used for the cost parameter, while 0.0001, 0.001, 0.01, and 0.1 were used for the gamma parameter in the grid search. Regarding RF, two parameters are tuned: the ensemble size and the maximum number of features in each decision tree. The values of 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 were used for the ensemble size, and the values of 5, 10, 15, 20, 25, and 30 were used for the maximum number of features to search for the best parameter. Two parameters, the maximum number of features and the number of trees, were optimized by grid search through fivefold cross-validation. The jackknife test was also used for comparison of the homology-based method and the machine-learning method.

2.5 Evaluation

The performance of each classifier was measured using the 5-fold cross validation. Several measurements are used to evaluate classifiers and they are sensitivity, specificity, accuracy, the Matthew's correlation coefficient (MCC), and area under the receiver operating characteristic curve (AUC). In the following equations, TP, FN, TN, FP refer to the numbers of true positives, false negatives, false negatives and false positives respectively.

Sensitivity is the percentage of correctly predicted interacting protein pairs and given by:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity is the percentage of correctly predicted non-interacting protein pairs using the following equation:

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

Accuracy is the percentage of correctly identified interacting and non-interacting pairs and given by:

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

MCC is a balanced measurement used to assess the effectiveness of the performance. Its definition is given by:

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

AUC is a measurement independent of the threshold of the decision value based on Receiver Operating Characteristic (ROC) curve. ROC represents the relationship between sensitivity (true positive rate) and 1-specificity (false positive rate). ROC curves help to identify the potential tradeoff between correct predictions and incorrect ones. It can also be summarized as a single value by taking the area under the curve (AUC). In this study, MCC was used in machine learning method for selecting optimal parameters.

3. RESULTS AND DISCUSSION

3.1 Performance of the homology-based method

The performance of the homology-based method using different J_E thresholds is shown in **Table 2**. Several measurements are used to evaluate classifiers, including sensitivity, specificity, accuracy, and the Matthew's correlation coefficient (MCC) described in **section 2.5**. In the case of $J_E = 10^{-10}$, the sensitivity was 74.06%, the specificity was 71.63%, the accuracy was 72.84%, and MCC was 45.70%. As the J_E threshold became more stringent, the overall performance deteriorated, with the exception of specificity. For instance, the most rigorous case of $J_E = 10^{-150}$ had the highest specificity (95.42%), but the lowest values for all other measurements (sensitivity, 27.92%; accuracy, 61.67%; and MCC, 31.64%). Therefore, on the basis of the accuracy and MCC, a threshold of $J_E = 10^{-10}$ was considered optimal for the homology-based method.

Table 2 Performance of the homology-based method using different joint sequence similarity (J_E) thresholds. MCC, Matthew's correlation coefficient

J_E threshold	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC (%)
10^{-10}	74.06	71.63	72.84	45.70
10^{-30}	70.18	74.86	72.52	45.09
10^{-50}	65.63	78.22	71.92	44.21
10^{-70}	61.10	81.17	71.14	43.15
10^{-90}	56.68	84.09	70.39	42.40
10^{-110}	45.15	89.96	67.55	39.28
10^{-130}	35.60	93.74	64.67	36.07
10^{-150}	27.92	95.42	61.67	31.64

3.2 Protein families

A protein family represents a group of proteins that typically have similar structures, functions, and sequence similarity; this information may provide some vital clues in PPI prediction. The Pfam 30.0 [45] database provides an extensive collection of 16,306 protein families and the interactions between protein families based on a 3D structure analysis [46]. A search against the Pfam 30.0 database (*E*-value cut-off = 10^{-8}) allowed us to list the top 10 most frequent Pfam families occurring in 2,864 true positive pairs (**Table 3**).

Table 3 Top 10 most frequent Pfam families in true positive pairs. The relative frequency was calculated as the proportion of occurrences of each family relative to the total number of families.

No.	Pfam identifier	Family name	Frequency (%)
1	PF00400	WD40	5.7
2	PF00069	Pkinase	2.3
3	PF00076	RRM_1	1.5
4	PF00028	Cadherin	1.2
5	PF05001	AMP-binding	1.3
6	PF00681	Plectin	0.9
7	PF00008	EGF	0.8
8	PF00017	SH2	0.7
9	PF00240	Ubiquitin	0.7
10	PF00630	Filamin	0.7

Most of these families were reasonably common, being close to or above a rate of 1%, and were functionally related to the interactions that were also reported previously [11]. For example, members of the WD40 protein family are functionally related in terms of transduction, transcription, and cell cycle control. WD40 motifs act as a site for PPIs and WD40 proteins contribute to the assembly of protein complexes or act as mediators of transient interplay among other proteins. Another family that frequently occurred among the true positives was the Pkinase (protein kinase) family, the proteins in which are evolutionarily conserved to mediate protein phosphorylation, which plays a key role in cellular processes such as division, proliferation, and differentiation. Phosphorylation usually results in a functional change in the target protein by altering enzyme activity, cellular location, or the association with other proteins, and the catalytic subunits of protein kinases are highly conserved. Other protein families that frequently occurred among the true positives, such as RRM_1 (RNA recognition motif), Cadherin, and AMP-binding, are also related to the act of protein binding.

3.3 Performance of the machine-learning methods

We used the machine-learning methods SVM and RF, in which we implemented fivefold cross-validation and a grid search to choose the optimal parameters based on AUC. All of the measurement scores were calculated as the mean of five subsets through the fivefold cross-validation. The performance of SVM for each feature is shown in **Table 4**. DC outperformed all other features in terms of accuracy (87.32%), MCC (74.64%), and AUC (94.78%), while AC showed the lowest AUC score (92.09%). All features had accuracies >80% and AUC scores >90%. The ROC curves for each of these features are shown in **Figure 3**, with DC showing the best AUC score.

Table 4 Performance of support vector machine (SVM) for each protein feature. MCC, Matthew's correlation coefficient; AUC, area under the receiver operating characteristic curve; AAC, amino-acid composition; DC, dipeptide composition; TC, tripeptide composition; MLD, multiscale local descriptor; PseAAC, pseudo-amino-acid composition; AC, autocovariance

Feature	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC (%)	AUC (%)
AAC	86.35 ± 1.81	84.77 ± 1.90	85.56 ± 1.04	71.13 ± 2.07	93.23 ± 1.03
DC	86.60 ± 2.10	88.03 ± 1.74	87.32 ± 1.63	74.64 ± 3.26	94.78 ± 1.11
TC	85.85 ± 2.96	87.69 ± 1.54	86.77 ± 1.70	73.57 ± 3.39	93.92 ± 1.35
MLD	84.74 ± 2.14	86.58 ± 3.44	85.66 ± 2.21	71.35 ± 4.46	94.05 ± 1.36
PseAAC	75.33 ± 3.52	91.13 ± 1.96	83.23 ± 1.51	67.32 ± 2.77	93.44 ± 1.12
AC	87.02 ± 1.45	82.99 ± 4.40	85.00 ± 2.49	70.07 ± 4.89	92.09 ± 0.93

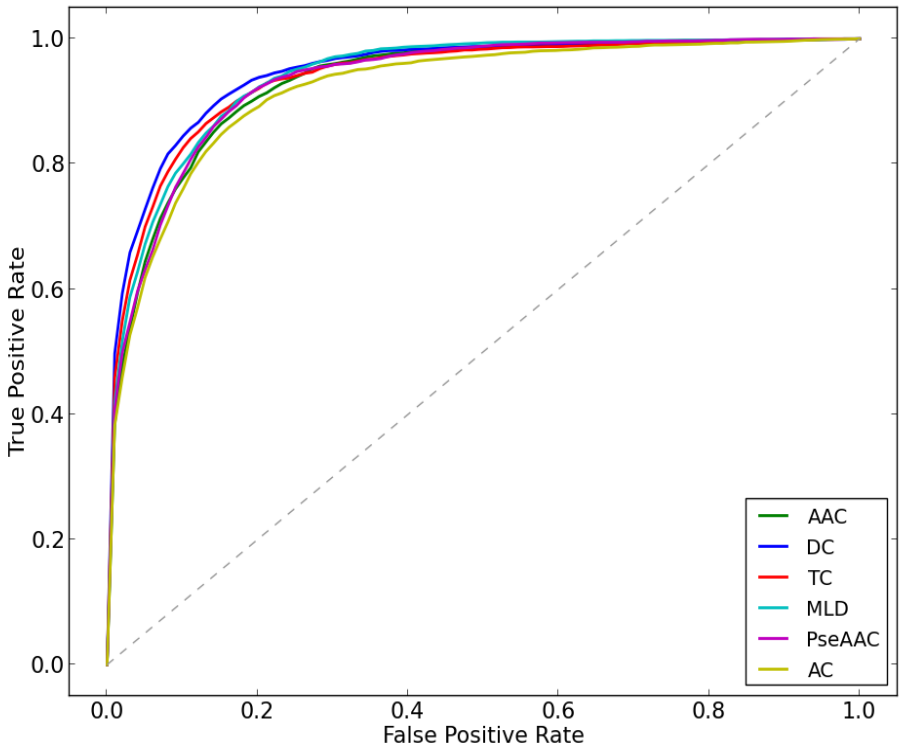

Figure 3 The mean receiver operating characteristic (ROC) curve under fivefold cross-validation with SVM. AAC, amino-acid composition; DC, dipeptide composition; TC, tripeptide composition; MLD, multiscale local descriptor; PseAAC, pseudo-amino-acid composition; AC, auto covariance.

For RF, PseAAC outperformed all other features in terms of accuracy (86.75%), MCC (73.53%), and AUC (95.16%), whereas AC showed the lowest AUC score (94.19%; **Table 5**). All accuracy scores were in the range of 85%–86% and all AUC scores were in the range of 94%–95%. Only slight

differences were observed among the ROC curves of the six features (**Figure 4**).

Table 5 Performance of RF for each protein feature. Other legends are the same as those in **Table 4**.

Feature	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC (%)	AUC (%)
AAC	85.10 ± 1.15	88.26 ± 1.99	86.68 ± 1.14	73.41 ± 2.30	95.14 ± 0.74
DC	84.22 ± 1.63	87.36 ± 2.39	85.79 ± 0.71	71.63 ± 1.46	94.66 ± 1.00
TC	85.31 ± 3.30	86.92 ± 1.82	86.11 ± 2.02	72.25 ± 4.02	94.61 ± 1.16
MLD	85.93 ± 2.14	86.86 ± 1.87	86.40 ± 1.40	72.81 ± 2.79	94.72 ± 1.08
PseAAC	85.52 ± 1.97	87.98 ± 2.79	86.75 ± 1.08	73.53 ± 2.19	95.16 ± 0.70
AC	83.71 ± 1.18	87.51 ± 1.86	85.61 ± 0.62	71.28 ± 1.30	94.19 ± 0.96

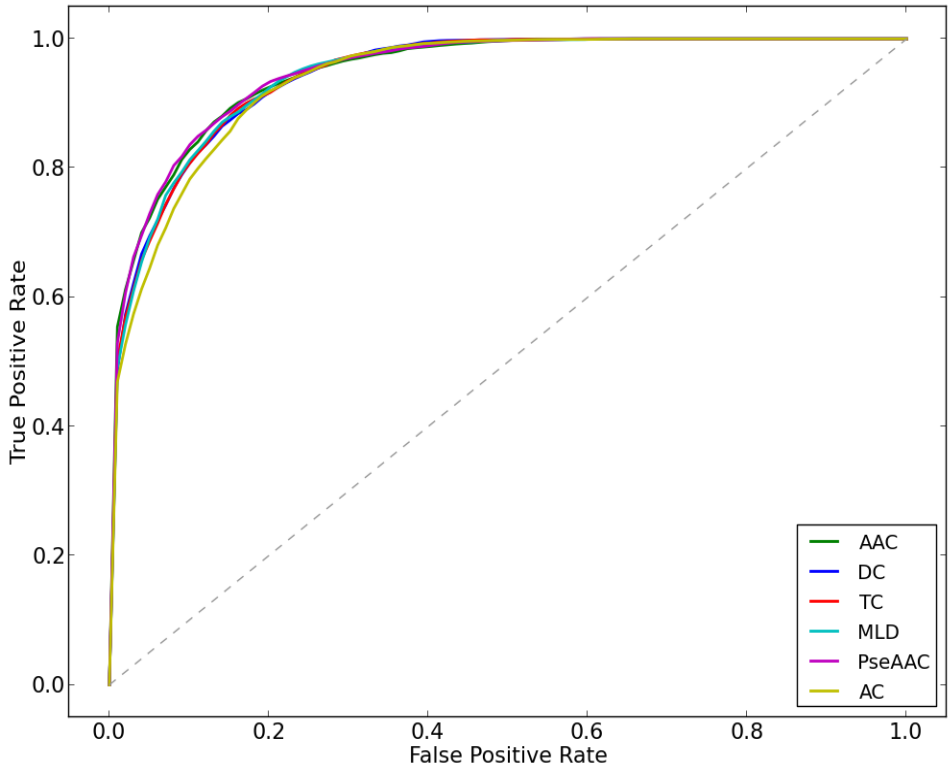

Figure 4 The mean receiver operating characteristic (ROC) curve under fivefold cross-validation with RF. Other legends are the same as those in **Figure 3**.

3.4 Protein features

Among the six protein features examined, DC and PseAAC showed the highest AUC scores with SVM and RF, respectively, while AC had the lowest AUC scores for both methods. All features except AC incorporated the frequency of amino-acid pattern (single AA or consecutive AAs), indicating that this plays a major role in the classification of PPIs. PseAAC not only includes AAC, but also considers the physicochemical properties of the proteins. PseAAC outperformed AAC in terms of

AUC by 0.21% and 0.02% in SVM and RF, respectively, demonstrating the effectiveness of combining these features. Data on the basic amino acid composition (AAC) are often useful for characterizing a dataset. In total, 3,380 and 2,553 protein sequences occurred in the positive and negative datasets, respectively. To investigate the difference between these, we excluded the 466 sequences that were common to both datasets. A comparison of the frequency of the 20 amino acids showed only a slight difference between the positive and negative datasets (**Figure 5**), indicating that the use of AAC alone may not be the best approach for representing the features of the dataset.

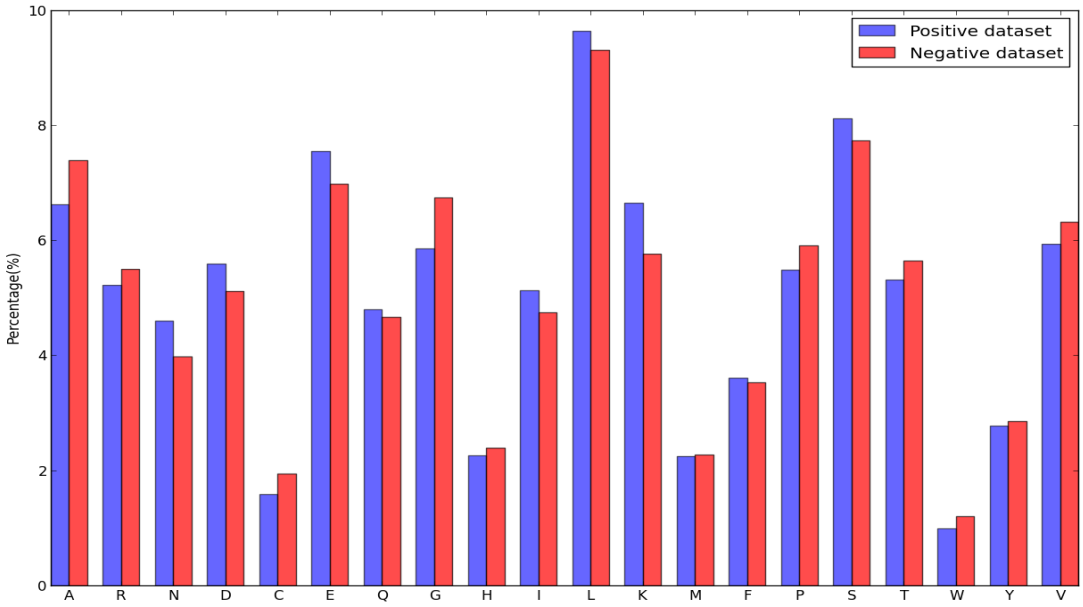

Figure 5 Relative frequencies of the 20 amino acids in the positive and negative datasets.

DC provided the best predictions in the SVM approach because it takes two consecutive amino acids into account and so can extract more patterns from the protein sequences. We listed the top 20 most different occurrences in frequency according to DC (**Figure 6**), which contributed to its good performance. By contrast, PseAAC outperformed all other features for RF because it incorporated both frequency and sequence order information. These sequence order factors use the properties of hydrophobicity, hydrophilicity, and side-chain volume, all of which are related to the physical interactions.

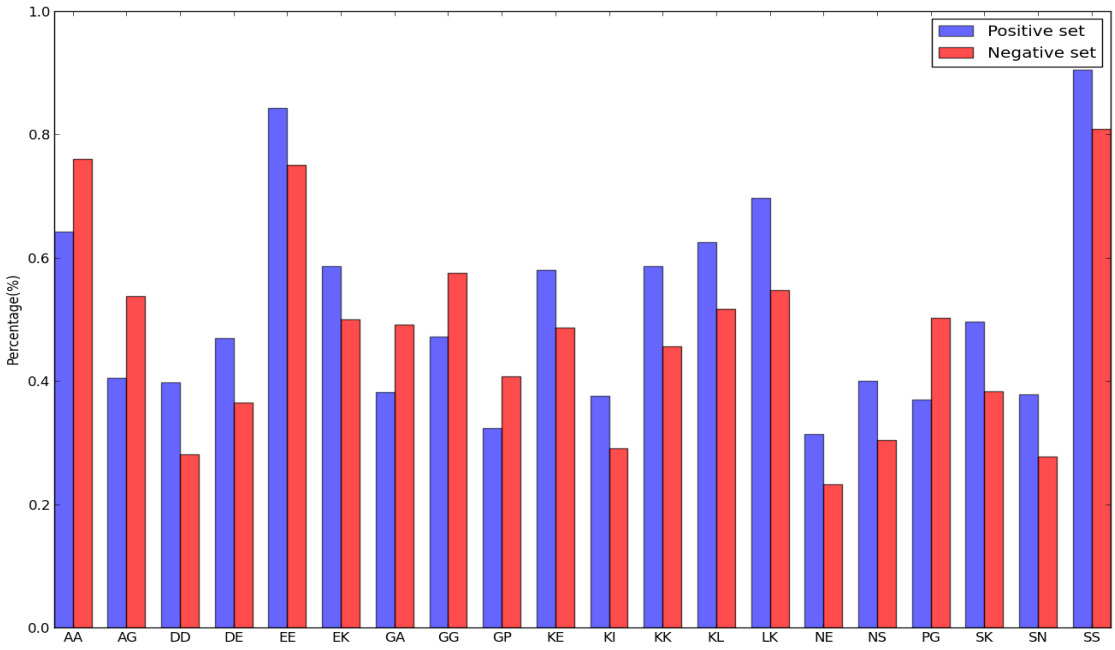

Figure 6 Top 20 most different frequencies of two consecutive amino acids in dipeptide composition.

3.5 Comparison of the homology-based method and the machine-learning methods

One classifier was selected to represent each of the machine-learning methods (DC for SVM and PseAAC for RF) and these were compared directly. PseAAC with RF outperformed DC with SVM in terms of AUC by 0.38%, but had 0.57% lower accuracy. Therefore, since a high accuracy score will provide more instances for comparison with the homology-based method, DC with SVM was selected as the representative classifier for the machine-learning methods, for which the jackknife test was conducted to predict the likelihood of interaction of each pair in turn.The comparison of the two methods is shown in **Table 6**. DC with SVM outperformed the homology-based method ($J_E = 10^{-10}$) in terms of sensitivity (+13.39%), specificity (+16.91%), accuracy (+15.16%), and MCC (+30.31%).

Table 6 Performance of the machine-learning and homology-based methods.

Method	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC (%)
DC with SVM	87.45	88.54	88.00	76.01
Homology-based	74.06	71.63	72.84	45.70

In addition, the two methods were assessed by comparing the number of TP and TN pairs. More than 65.1% of the 7,734 pairs (3,867 positive pairs plus 3,867 negative pairs) were predicted correctly using both the homology-based method and the SVM model (**Figure 7**). However, nearly 23.5% of the 3,867 positive pairs and 22.2% of the 3,867 negative pairs were beyond the ability of the homology-based method, but were predicted well using SVM. By contrast, 10.2% of the 3,867 positive pairs and 5.3% of the 3,867 negative pairs failed to be identified in SVM, but were successfully predicted using the homology-based method.

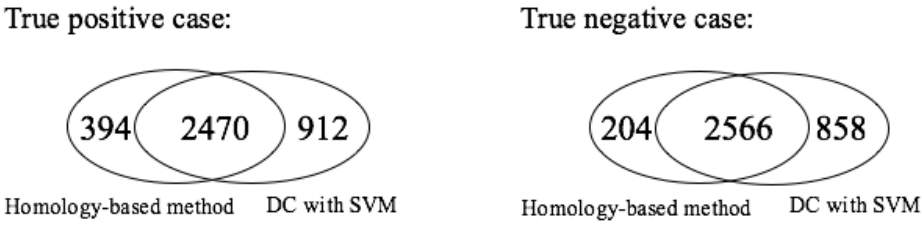

Figure 7 Comparison of the machine-learning and homology-based methods.

The homology-based method depends on the local sequence alignment, meaning that the measure of homology, acceptance of the *E*-value, or pairwise similarity had a large impact on the prediction results in our experiments, with an inability to predict interacting pairs without homologs correctly. However, these pairs could be identified using SVM. Regarding DC with SVM, several vectors were selected as support vectors to determine the hyperplane of classification in the collected dataset. However, those vectors that are in close proximity to the hyperplane may be misclassified. Therefore, it is expected that an approach integrating these two methods would improve the performance for predicting PPIs. Sequences that are annotated as belonging to the same family typically contain conserved regions and are functionally related. Across the entire dataset, 40 protein families were selected according to the ranking order, with the number of occurrences ranging from 1,466 to 59. **Table 7** illustrates the accuracy for each protein family using each of the two methods. This shows, for instance, that the accuracy of the protein pairs containing PF00168 (No. 23; C2) was 83.5% in the SVM model but 43.8% in the homology-based model. This family is a C2 domain involved in targeting proteins to cell membranes. It shows wide range of lipid selectivity for the major components of cell membranes, while protein pairs containing PF00134 (No. 40; Cyclin_N) achieved accuracy of 89.2% in the homology-based method and 85.7% in the SVM model. These different results could be used to select the optimal method for obtaining the best result. It is expected that the consideration of multiple domain effects may provide a more precise result in specific cases and will contribute to the further study of PPI prediction.

Table 7 Performance of SVM and homology-based methods for each protein family

No.	Protein family	Average length of the domain	Average identity of full alignment (%)	Average coverage of the sequence by the domain (%)	No. of occurrences	Accuracy with homology (%)	Accuracy with SVM (%)
1	PF00240	70.30	37 %	25.31	1466	65.8	98.9
2	PF00069	238.10	21	38.74	363	73.5	82.0
3	PF00400	39.40	25	19.92	304	68.4	88.1
4	PF07714	230.00	24	35.37	204	68.1	78.4
5	PF00018	47.20	29	7.11	179	70.9	86.5
6	PF00017	77.60	28	14.81	173	77.4	85.5
7	PF07654	82.10	24	33.37	172	89.5	97.0
8	PF00227	172.20	21	72.55	130	47.6	87.6
9	PF00076	67.40	22	22.89	129	62.0	89.9
10	PF00271	116.90	20	13.28	122	72.1	92.6
11	PF00071	152.20	29	59.73	112	45.5	78.5
12	PF01248	92.10	23	48.69	103	94.1	98.0
13	PF00104	179.20	19	38.75	102	68.6	79.4
14	PF00105	67.20	46	14.63	98	69.3	79.5
15	PF00170	61.00	28	16.59	96	79.1	90.6
16	PF00010	53.90	29	13.98	95	85.2	84.2
17	PF00179	132.60	27	45.04	94	89.3	87.2
18	PF00439	83.90	26	10.61	94	74.4	77.6
19	PF12796	87.80	23	30.24	87	66.6	80.4
20	PF07686	103.10	17	36.94	79	92.4	91.1
21	PF00096	23.20	40	20.25	75	57.3	90.6
22	PF13912	26.00	34	7.24	75	57.3	90.6
23	PF00168	105.90	18	20.29	73	43.8	83.5
24	PF00467	34.00	31	17.01	71	70.4	100.0
25	PF00569	44.40	30	4.59	66	75.7	74.2
26	PF10584	22.90	62	8.51	65	24.6	92.3
27	PF00270	169.90	22	22.25	64	64.0	93.7
28	PF01423	69.40	25	50.36	63	55.5	79.3
29	PF00169	104.40	17	13.31	63	73.0	77.7
30	PF00249	47.50	26	13.98	63	71.4	79.3
31	PF00888	460.40	22	65.28	62	93.5	93.5
32	PF13920	48.80	30	9.09	62	80.6	85.4
33	PF00531	81.70	20	9.09	62	70.9	79.0
34	PF00046	56.00	32	14.83	60	41.6	85.0
35	PF13181	32.30	19	6.93	60	71.6	93.3
36	PF00397	29.90	36	6.04	59	83.0	89.8
37	PF00628	50.00	29	5.85	59	71.1	89.8
38	PF00782	123.80	20	29.62	58	77.5	96.5
39	PF01479	46.80	26	17.51	57	63.1	100.0
40	PF00134	124.00	19	30.75	56	89.2	85.7

In this study, we demonstrated that an appropriate protein feature combined with the learning algorithm is useful for classifying protein pairs into those that do and do not interact with each other. We also found that 10.2% (394) of interacting pairs and 5.3% (204) of non-interacting pairs failed to be identified by the SVM model, but were successfully predicted using the homology-based method.

This suggests that integrating these two methods may enhance the performance in further study. Finally, across the entire dataset, we listed the top 40 protein families that frequently appeared by searching Pfam database. We showed the accuracy of the predictions of interaction for each of these families using each of the two methods. The difference in performance among these families may provide a way of predicting when given a protein pair belongs to a particular protein family.

4. CONCLUSION

In this study, we demonstrated that an appropriate protein feature combined with the learning algorithm is useful for classifying positive and negative protein pairs. We also found that 10.2% (394) of positive pairs and 5.3% (204) of negative pairs failed to be identified by the SVM model, but were successfully predicted using the homology-based method. This suggests that integrating these two methods may enhance the performance in further study. Finally, across the entire dataset, we listed the top 40 protein families with the most occurrences by the searching Pfam database. We showed the accuracy of the predictions of interaction for each of these families using each of the two methods. The difference in performance among these families may provide a way of predicting when given a protein pair belongs to a particular protein family.

CONFLICT OF INTEREST

The authors have no conflict of interest.

REFERENCES

1. Roslan R, Othman RM, Shah ZA, Kasim S, Asmuni H, Taliba J, et al. Utilizing shared interacting domain patterns and Gene Ontology information to improve protein-protein interaction prediction. Comput. Biol. Med. Elsevier; 2010;40:555–64.

2. Papin J, Subramaniam S. Bioinformatics and cellular signaling. Curr. Opin. Biotechnol. 2004;15:78–81.

3. Tucker CL, Gera JF, Uetz P. Towards an understanding of complex protein networks. Trends Cell Biol. 2001;11:102–6.

4. Braun P, Gingras AC. History of protein-protein interactions: From egg-white to complex networks. Proteomics. 2012;12:1478–98.

5. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature.2002;415:180–3.

6. Gavin A-C, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 2002;415:141–7.

7. Zhu H, Snyder M. Protein chip technology. Curr. Opin. Chem. Biol. 2003;7:55–63.

8. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014;42:358–63.

9. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34:D535-9.

10. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database - 2009 update. Nucleic Acids Res. 2009;37:767–72.

11. Bock JR, Gough DA. Predicting protein – protein interactions from primary structure. 2001;17:455–60.

12. Hart GT, Ramani AK, Marcotte EM. How complete are current yeast and human protein-interaction networks? Genome Biol. 2006;7:120.

13. Shoemaker BA, Panchenko AR. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comput. Biol. 2007. p. 595–

601.

14. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. U. S. A. 1999;96:4285–8.

15. Zahiri J, Bozorgmehr JH, Masoudi-nejad A. Computational Prediction of Protein – Protein Interaction Networks : Algo- rithms and Resources. 2013;397–414.

16. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature. 1999;402:86–90.

17. Rao VS, Srinivas K, Sujini GN, Sunand Kumar GN. Protein-Protein Interaction Detection: Methods and Analysis. 2014;2014.

18. Singh R, Park D, Xu J, Hosur R, Berger B. Struct2Net: A web service to predict protein-protein interactions using a structure-based approach. Nucleic Acids Res. 2010;38.

19. Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A. PRISM: Protein interactions by structural matching. Nucleic Acids Res. 2005;33:331–6.

20. Zhang QC, Petrey D, Garzón JI, Deng L, Honig B. PrePPI: A structure-informed database of protein-protein interactions. Nucleic Acids Res. 2013;41:828–33.

21. Ohue M, Matsuzaki Y, Uchikoga N, Ishida T, Akiyama Y. MEGADOCK: An All-to-All Protein-Protein Interaction Prediction System Using Tertiary Structure Data. Protein Pept. Lett. 2014;21:766–78.

22. Martin S, Roe D, Faulon JL. Predicting protein-protein interactions using signature products. Bioinformatics. 2005;21:218–26.

23. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, et al. Predicting protein-protein interactions based only on sequences information. Proc. Natl. Acad. Sci. U. S. A. 2007;104:4337–41.

24. Guo Y, Yu L, Wen Z, Li M. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res. 2008;36:3025–30.

25. Xenarios I, Salwínski L, Duan XJ, Higney P, Kim S-M, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. 2002;30:303–5.

26. Zahiri J, Mohammad-Noori M, Ebrahimpour R, Saadat S, Bozorgmehr JH, Goldberg T, et al. LocFuse: Human protein–protein interaction prediction via classifier fusion using protein localization information. Genomics. 2014;104:496– 503.

27. You Z-H, Chan KCC, Hu P. Predicting Protein-Protein Interactions from Primary Protein Sequences Using a Novel Multi-Scale Local Feature Representation Scheme and the Random Forest. PLoS One. 2015;10:e0125811.

28. Walhout AJM, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, et al. Protein interaction mapping in C. elegans using proteins involved in vulval development. Science. American Association for the Advancement of Science; 2000;287:116–22.

29. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JDJ, et al. Annotation transfer between genomes: Protein-protein interrologs and protein-DNA regulogs. Genome Res. 2004;14:1107–18.

30. Chen CC, Lin CY, Lo YS, Yang JM. PPISearch: A web server for searching homologous protein-protein interactions across multiple species. Nucleic Acids Res. 2009;37:369–75.

31. Garcia-Garcia J, Schleker S, Klein-Seetharaman J, Oliva B. BIPS: BIANA Interolog Prediction Server. A tool for protein--protein interaction inference. Nucleic Acids Res. 2012;40:147–51.

32. 51. Zhou Y, Gao Y, Zheng Y. Prediction of Protein-Protein Interactions Using Local Description of Amino Acid Sequence. Adv. Comput. Sci.2011;254–62.

33. Vapnik VN, Vapnik V. Statistical learning theory. Wiley New York; 1998.

34. Breiman L. Random forests. Mach. Learn. Springer; 2001;45:5–32.

35. López Y, Nakai K, Patil A. HitPredict version 4: comprehensive reliability scoring of physical protein-protein interactions from more than 100 species. Database (Oxford). 2015;2015:1–10.

36. Blohm P, Frishman G, Smialowski P, Goebels F, Wachinger B, Ruepp A, et al. Negatome 2.0: A database of noninteracting proteins derived by literature mining, manual annotation and protein structure analysis. Nucleic Acids Res. 2014;42:1–5.

37. Consortium U, others. UniProt: a hub for protein information. Nucleic Acids Res. Oxford Univ Press; 2014;gku989.

38. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. Oxford

Univ Press; 1997;25:3389–402.

39. Dondoshansky I, Wolf Y. BLASTCLUST-BLAST score-based singlelinkage clustering. 2000.

40. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, et al. Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Interactions or Interologs Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Inte. 2001;2120–6.

41. Patil A, Nakamura H. HINT: a database of annotated protein-protein interactions and their homologs. Biophysics (Oxf). 2005;1:21–4.

42. Bhasin M, Raghava GPS. Classification of nuclear receptors based on amino acid composition and dipeptide composition. J. Biol. Chem. ASBMB; 2004;279:23262–6.

43. Cao D-S, Xu Q-S, Liang Y-Z. propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics. Oxford Univ Press; 2013;29:960–2.

44. Liu L, Cai Y, Lu W, Feng K, Peng C, Niu B. Prediction of protein-protein interactions based on PseAA composition and hybrid feature selection. Biochem. Biophys. Res. Commun. Elsevier Inc.; 2009;380:318–22.

45. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. Oxford Univ Press; 2016;44:D279--D285.

46. Finn RD, Miller BL, Clements J, Bateman A. iPfam: a database of protein family and domain interactions found in the Protein Data Bank. Nucleic Acids Res. Oxford Univ Press; 2014;42:D364-D373.

SUPPLEMENTARY FILES

Table S1 The original values of the three physicochemical properties for each amino acid. H1: hydrophobicity; H2: hydrophilicity; M: the mass of side chain.

AA	H_1	H_2	M
A	0.62	-0.5	15.0
R	-2.53	3.0	101.0
N	-0.78	0.2	58.0
D	-0.90	3.0	59.0
C	0.29	-1.0	47.0
Q	-0.85	0.2	72.0
E	-0.74	3.0	73.0
G	0.48	0.0	1.0
H	-0.40	-0.5	82.0
I	1.38	-1.8	57.0
L	1.06	-1.8	57.0
K	-1.50	3.0	73.0
M	0.64	-1.3	75.0
F	1.19	-2.5	91.0
P	0.12	0.0	42.0
S	-0.18	0.3	31.0
T	-0.05	-0.4	45.0
W	0.81	-3.4	130.0
Y	0.26	-2.3	107.0
V	1.08	-1.5	43.0

Table S2 The original values of the seven physicochemical properties for each amino acid. H1: hydrophobicity; H2: hydrophilicity; V: volume of side chains; P1: polarity; P2: polarizability; SASA: solvent accessible surface area; NCI: net charge index of side chains.

AA	H_1	H_2	V	P_1	P_2	SASA	NCI
A	0.62	-0.5	27.5	8.1	0.046	1.181	0.007187
R	-2.53	3.0	105	10.5	0.291	2.56	0.043587
N	-0.78	0.2	58.7	11.6	0.134	1.655	0.005392
D	-0.90	3.0	40	13	0.105	1.587	-0.02382
C	0.29	-1.0	44.6	5.5	0.128	1.461	-0.03661
Q	-0.85	0.2	80.7	10.5	0.18	1.932	0.049211
E	-0.74	3.0	62	12.3	0.151	1.862	0.006802
G	0.48	0.0	0	9.0	0.0	0.881	0.179052
H	-0.40	-0.5	79	10.4	0.23	2.025	-0.01069
I	1.38	-1.8	93.5	5.2	0.186	1.81	0.021631
L	1.06	-1.8	93.5	4.9	0.186	1.931	0.051672
K	-1.50	3.0	100	11.3	0.219	2.258	0.017708
M	0.64	-1.3	94.1	5.7	0.221	2.034	0.002683
F	1.19	-2.5	115.5	5.2	0.29	2.228	0.037552
P	0.12	0.0	41.9	8.0	0.131	1.468	0.239531
S	-0.18	0.3	29.3	9.2	0.062	1.298	0.004627
T	-0.05	-0.4	51.3	8.6	0.108	1.525	0.003352
W	0.81	-3.4	145.5	5.4	0.409	2.663	0.037977
Y	0.26	-2.3	117.3	6.2	0.298	2.368	0.023599
V	1.08	-1.5	71.5	5.9	0.14	1.645	0.057004