Predict protein-protein interactions from protein primary sequences: using wavelet transform combined with stacking algorithm

Pin-San Xu [1] , **Jun Luo** [1] , **Tong-Yi Dou** [Corresp. 1]

[1] Dalian University of Technology, School of Life Science and Medicine, Panjin, China

Corresponding Author: Tong-Yi Dou
Email address: douty@dlut.edu.cn

Most biological processes within a cell are carried out by protein-protein interaction (PPI) networks, or so called interactomics. Therefore, identification of PPIs is crucial to elucidating protein functions and further understanding of various cellular biological processes. Currently, a series of high-throughput experimental technologies for detect PPIs have been presented. However, the time-consuming and labor-driven characteristics of these methods forced people to turn to virtual technology for PPIs prediction. Herein, we developed a new predictor which uses stacking algorithm with information extraction by wavelet transform. When applied on the *Saccharomyces cerevisiae* PPI dataset, the proposed method got a prediction accuracy of 83.35% with sensitivity of 92.95% at the specificity of 65.41%. An independent data set of 2726 *Helicobacter pylori* PPIs was also used to evaluate this prediction model, and the prediction accuracy is 80.39%, which is better than that of most existing methods.

Predict protein-protein interactions from protein primary sequences: using wavelet transform combined with stacking algorithm

Pin-San Xu[1], Jun Luo[1], Tong-Yi Dou[1],

[1] School of Life Science and Medicine, Dalian University of Technology, Panjin, China

Corresponding Author:

Tong-Yi Dou [1]

NO.2 Dagong Road, Panjin City, Liaoning Province, LN 427, China

Email address: douty@dlut.edu.cn

Abstract

Most biological processes within a cell are carried out by protein-protein interaction (PPI) networks, or so called interactomics. Therefore, identification of PPIs is crucial to elucidating protein functions and further understanding of various cellular biological processes. Currently, a series of high-throughput experimental technologies for detect PPIs have been presented. However, the time-consuming and labor-driven characteristics of these methods forced people to turn to virtual technology for PPIs prediction. Herein, we developed a new predictor which uses stacking algorithm with information extraction by wavelet transform. When applied on the *Saccharomyces cerevisiae* PPI dataset, the proposed method got a prediction accuracy of 83.35% with sensitivity of 92.95% at the specificity of 65.41%. An independent data set of 2726 *Helicobacter pylori* PPIs was also used to evaluate this prediction model, and the prediction accuracy is 80.39%, which is better than that of most existing methods.

Introduction

Proteins play critical roles in almost all important biological processes of living cells, for example, metabolic cycles, replication and DNA transcription. However, proteins rarely act alone, but achieve most of their work through PPI networks [1]. Currently, a series of high-throughput experimental technologies for PPI detection are developed, such as yeast two-hybrid screen (Y2H), protein chip technology, and tandem affinity purification tagging (TAP) [1-3]. Although these methods have successfully identified a large number of PPIs [1-3], the relatively time-consuming and labor-driven characteristics lead researchers to looking for more efficient and cost-effective alternative tools [4].

Up to date, a number of bioinformatics method for PPI prediction have been developed, among which algorithms based primarily on sequence conservation, phylogenetic profiles, literature mining, etc. [5-7]. Although these methods give high predictive accuracy, most of the protein information needed by these methods for predict PPIs is normally inaccessible, especially for those less well-characterized proteins. Fundamentally, however, many of the functions and properties of proteins can be informed by the low frequency signals in the amino acid sequence [8]. As reported by recent studies based on protein primary sequence can also achieve satisfactory accuracy for predicting PPIs [4-6]. Meanwhile, wavelet transform [9], an effective feature extraction method, has been widely used in signal extraction of amino acid sequences and achieved good performance. For instance, wavelet transform was utilized for membrane protein prediction [10], protein structural prediction [11], protein classification [12], and PPI prediction [13]. It is well known that wavelet transform takes the advantage over Fourier transform in the extraction of location information, however, none of the above studies had paid attention to simultaneous extraction of both signal strength information and the position information.

On the other side, there were also studies using ensemble classifiers significantly improve the overall performance of the classifier in predicting membrane protein types [14], subcellular localization of protein [15], and of course, in predicting PPIs [8,13].

Inspired by previous researches, here we report a new method that improves the prediction performance in predicting PPIs. The method operates stacking algorithm with information

extracted from protein primary sequences by wavelet transform. First, the physicochemical property of each protein sequence is transformed into series of vectors. Then, stacking algorithm with two layers was adopted to carry out the PPI prediction, first layer of stacking algorithm including four independent classifiers and logistic regression [16] was applied to stacking algorithm as the second layer. Finally, the proposed method was tested on two PPI datasets. The results demonstrated that the proposed approach offers a better performance than any of the current programs under various statistical standards in the two widely-used data-sets by a 5-fold cross validation.

Materials & Methods

Generation of benchmark data sets

Saccharomyces cerevisiae dataset

The PPI data sets employed in this paper are collected from *Saccharomyces cerevisiae* database of interacting proteins (DIP), version 20160731, and it is customized to the standards almost the same way as in Jia et al. [13] The only difference is that in order to get reasonable length of coefficients arrays after the original sequence process from discrete wavelet transform (DWT), proteins in this dataset must contain at least 64 residues. The non-interactive data comprised of two parts: proteins which located at different subcellular localizations and that located at same subcellular localizations but did not appear in the positive dataset. In this case, 17333 positive pairs and additional 32568 negative pairs are generated. The *Saccharomyces cerevisiae* dataset used in this paper can be obtained in https://github.com/deltawing/master_experiment_stacking.

Helicobacter pylori dataset

The *Helicobacter pylori* PPI dataset is also corroborated the effectiveness of the method we proposed. The dataset is prepared just as Martin et al. [17] described, except the series we used must contain at least 64 residues. The final dataset contains 1307 protein pairs that have interactive relationship and 1419 protein pairs without interactive relationship at the same time. This dataset can also be accessed in https://github.com/deltawing/master_experiment_stacking.

Feature vector construction

When identifying protein characteristics using some specific methods, it is valuable to formulate the sequence with an effective mathematical expression, which not only encompasses its sequence order information but also gain the key features [18]. As mostly, the length of protein sequence varies a lot, the formula must transform the original sequence to a vector of features that have unified length which is needed by ordinary machine learning models. The learning models using amino acid sequence to classify the subcellular localization of protein, classify interactive or no-interactive relationship of proteins or identify function of protein, have been developed in recent years [6-8,19-23]. A large part of these studies adopted pseudo amino acid composition [24] method or also known as Chou's PseAAC [25-26].

According to a recent review [27], the general form of Chou's PseAAC for a protein or peptide P can be formulated as:

$$P = [\psi_1 \psi_2 ... \psi_n ... \psi_\Omega]^T \quad (1)$$

where T is the transpose operator, Ω indicate the vector's dimension. The value of Ω together with $\psi_n(n = 1,2,..., \Omega)$ in Eq. (1) are changed with the means of extract methods. In the following, we are about to depict how to analysis principal component in the benchmark dataset. As described in [8`28], a protein's low-frequency spectrum reflects its overall sequence eigenvalues. Therefore, an effective way to extract low-frequency spectrum information may help heighten the success rate in predicting PPIs.

Since introduced by Mallet S. G. in 1989 [9], wavelet transform has been used as an impressive method by scholars in various researches, such as the prediction of promoters [29], predicting protein classify cation [12], protein structural classes [30], G-protein-coupled receptor classes [31], enzyme family classes [32], homo-oligomeric proteins [33], membrane protein classes [34], protein quaternary structural attributes [35], etc. Within this work, we also use the wavelet transform method to extract information from protein sequence.

Physicochemical properties

The physicochemical property of proteins may have a great impact on protein–protein interactions. In this study, seven physicochemical properties of amino acids were selected to reflect the natural features of proteins, which are: hydrophobicity [36], hydrophilicity [37], side-chain volume [38], polarity [39], polarizability [40], solvent-accessible surface area or SASA [41], and side-chain net charge index or NCI [42], respectively. Please note that all these constants are transformed as the following before use:

$$\Phi_{i,j}^{'} = \frac{\Phi_{i,j} - \overline{\Phi_j}}{SD(\Phi_j)} \quad (2)$$

where $\Phi_{i,j}$ represents the j-th physicochemical properties for i-th amino acid, $\overline{\Phi_j}$ is the mean of j-th physicochemical property over the 20 amino acids, and $SD(\Phi_j)$ means the corresponding standard deviation of j-th physicochemical property.

After transformation, normalized values of each kind physicochemical property of a protein sequence are formed into one vector, thus each sequence have seven vectors representing its character.

Discrete wavelet transform

As a multiresolution analysis tool for decompose signal and determining component frequencies, wavelet transform overcomes the resolution shortcoming of Fourier analysis, for it not only analyzing the spectrum of the signal but also taking into account the specific location of the signal in the time domain, especially in a nonstationary process. The nature of DWT analysis make it reflect the sequence-order series more effectively than other techniques. By applying the DWT on any of these seven numerical vectors of a protein, each sequence-order vector is considered as a discrete time series and will put into one half band high-pass filter and one half band low-pass filter. The approximation coefficient series that output from high-pass filter removed all signals which frequency below half of the highest frequency in the sequence represents the high frequency components, while the coefficient series output from low-pass filter removed signals have frequency above half of the highest represents the high-scale components [29]. At every decomposition level, after passed through filters, numerical vector will discard every other sample, in other words subsampling by 2. The length of output from either filter is then half of the length than that of original sequence, and the output signal from the low-pass filter will continue to pass through the same two kinds of filters for some other decomposition until the intended number of iterations is reached, Fig. 1 illustrated a schematic diagram of the procedure of multi-level DWT, and the length of output series from each decomposition level can be described as follows:

$$\ell = floor\left(\frac{\mathcal{L}}{2^{n}}\right) \tag{3}$$

where ℓ represents the length of output series, $\mathcal{L}$ represents the length of input original numerical vectors of the physicochemical property of protein, n means decomposition level, $floor()$ represents the largest integral value that is not greater than the value in parentheses.

The frequencies that contain essential information in the original series show high amplitudes in those output series. While those are not protruding in the original series show relatively low values, these values decomposed can be omitted without losing the major part of the information, which allows DWT to lessen the dimensions of the original series effectively. Besides, the locations of these remarkable sample point and the position of these key features in original series have a one-to-one relationship. Given an output vector series with ℓ sample points as expressed by

$$Series = \varphi_1\, \varphi_2\, \varphi_3 \ldots \varphi_m \ldots \varphi_\ell \tag{4}$$

Where φ_1 represents the 1st sample point of output vector series, φ_2 represents the 2nd residue, and so forth. In this study, we use Daubechies db1 wavelet as our wavelet algorithm and use four decomposition level. Consequently, five subsequences can be obtained from the output of the algorithm. In each subsequence, 10 coefficients are extracted to reflect the internal information

of the subsequence, these are (1) mean of the wavelet coefficients in the subsequence, (2) standard deviation of the wavelet coefficients in the subsequence, (3) 4 samples which have the biggest absolute value in the subsequence and their locations. In this paper, we process the original location number to the location value that we use as follows:

$$location\ value = \hbar \times \left(\frac{m}{\ell}\right) \tag{5}$$

where $m = 1,2,\ldots,\ell$ is the sample point of output vector series in Eq. (4), m is the original location number of samples which have the biggest absolute value, ℓ represents the length of vector series just as in Eq. (4), $\hbar$ represents a coefficient, to make sure the $location\ value$, as well as four most remarkable sample point, are in the same order of magnitude. In this study, $\hbar$ is equal to 3. Therefore, the vector's dimension of a protein in Eq. (4) is $\Omega = 7 \times 5 \times 10 = 350$. For two proteins described as P^1 and P^2, the descriptors of the protein pair are formulated by their orthogonal sum [42]; i.e.,

$$P^1 \oplus P^2 = \left[{\psi_1}^1\ {\psi_2}^1 \ldots {\psi_n}^1 \ldots {\psi_{350}}^1\ {\psi_1}^2\ {\psi_2}^2 \ldots {\psi_n}^2 \ldots {\psi_{350}}^2\right]^T \tag{6}$$

thus, a total 700-dimensional vector has been built to represent a pair of proteins.

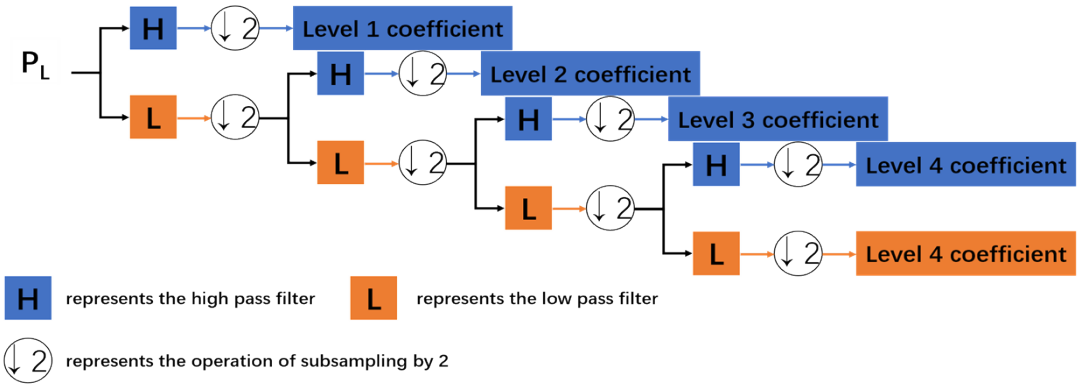

Figure 1. Illustrate of multi-level DWT procedure

Stacking algorithm

The ensemble method used in the present paper is called stacked generalization, or stacking, which is a two-step method. Firstly, subsets of the original data are used to produce a series of ordinary classifiers, the output values of these models are formed as input coefficients of the second step. Then the predictor form the second layer collect coefficients from every former model together and aimed at deciding what models perform well and what badly given these input data [43].

In this paper, each of the datasets used is divided into two groups, one for whole training process

called “the big training set”, the other for the testing process called “the big testing set”. In the first step, for the different classification model, the training set is divided into five parts, with N times of iteration, whereas N equals to the number of predictors in the first step. In each iteration, four data parts are formed as a training set for each classification model training while one data part is left for classification model prediction. When the iterations complete, a result matrix of $M \times 1$ is obtained, where M represented the number of samples of the training set. After all N classification models have had their prediction result, a $M \times N$ output matrix can be gotten. This output matrix is sent to second step of the algorithm as coefficient. This matrix, together with the real label list are sent to the $(N + 1)$th algorithm for training a model. When it comes to prediction process of the model, “the big testing set” also iterated N times for the same N individual models as in the training process, the output matrix is sent to the same $(N + 1)$th model to predict the final result. To offer an intuitive picture, three overview pictures are given in Fig. 2~4 to illustrate how the training process and testing process works.

For each algorithm of the first layer may shows a better prediction than other algorithms in some specific data, model of the second layer can evaluate the performance of these predictors and find the correspondence between the predictor and the specific data which it has a good performance [43]. Considering this work is easier than the job done by the algorithm of the first layer, logistic regression [16] is chosen as the algorithm for the second layer, for its simplicity and has a fast calculation speed.

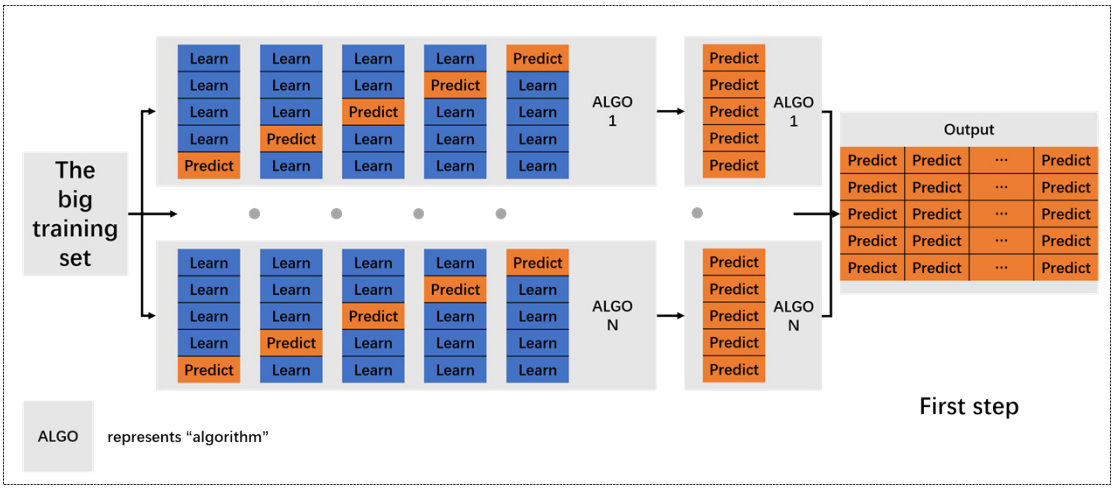

Figure 2. An overview of the first step of the stacking algorithm when learning “the big training set”

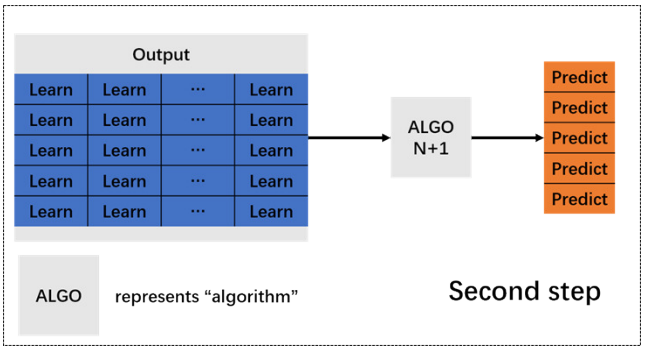

Figure 3. The second step of the stacking algorithm when learning "the big training set"

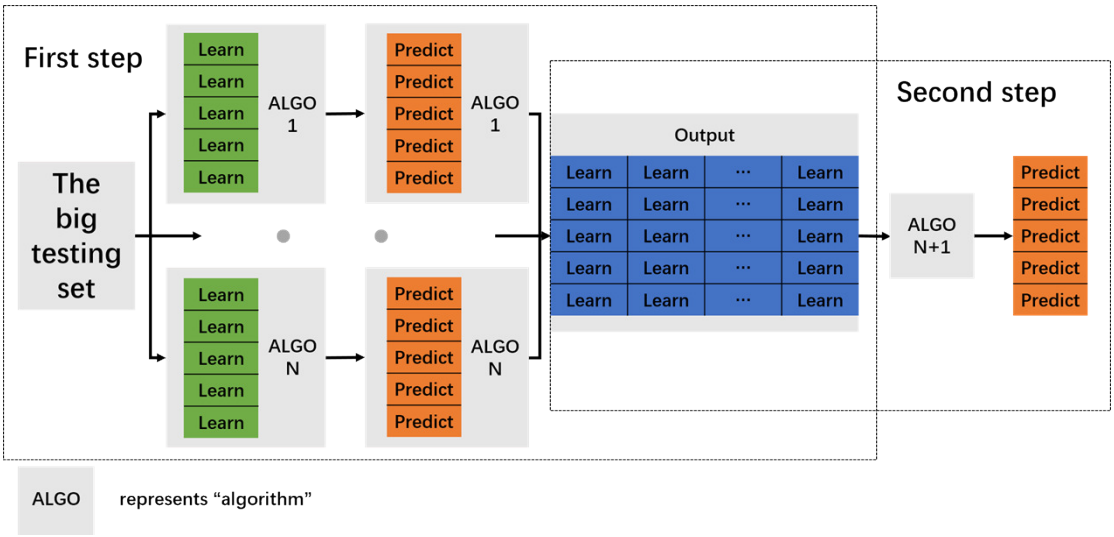

Figure 4. A flowchart to show how the stacking algorithm works when predicting "the big testing set"

Evaluation of the predictive performance

To evaluation the proposed method, 5-fold cross validation as well as several metrics which are widely used are adopted in this paper, which are (1) sensitivity, (2) specificity, (3) overall accuracy, (4) F-score, (5) Mathew's correlation coefficient, and (6) the area under ROC curve or AUC. Some of these measures are calculated by:

$$\begin{cases} \mathrm{Sn} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ \mathrm{Sp} = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\ \mathrm{Acc} = \dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \\ \mathrm{F}_1 = \dfrac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \\ \mathrm{Mcc} = \dfrac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}} \end{cases} \quad (7)$$

where TP represents the true positive; TN, the true negative; FP, the false positive; FN, the false negative; Sn, the sensitivity; Sp, the specificity; Acc, the overall accuracy; and Mcc, the Mathew's correlation coefficient.

Results and discussion

Predictors used for the first layer of the stacking algorithm

The predictor for the first layer of stacking algorithm can be any of the widely-used machine learning algorithms, including simple classifiers likewise some noted ensembled classifications, or the assembly of these algorithms. To avoid overfitting, herein 5-fold cross-validation is used to assess the performances of eight widely-used algorithms. The data used by algorithms is obtained from the *Saccharomyces cerevisiae* dataset by the method mentioned above. Algorithms include random forest classifier [44], gradient boosting classifier [45], extra-trees algorithm [46], adaboost classifier [47], k-nearest neighbors [48], linear discriminant analysis, quadratic discriminant analysis [49], and support vector machine [50].

Table 1

Comparison of the performance by some widely used method of the yeast dataset. The definition of Acc, Mcc, F_1, Sn and Sp, please refer to Eq (7)

Method	Acc (%)	Mcc (%)	Sn (%)	Sp (%)	F_1 (%)	AUC (%)	Time consumption(s)
Random forest classifier	81.75	58.59	94.39	58.02	68.85	62.32	159.83
Gradient boosting classifier	82.85	61.23	84.03	80.02	73.23	69.01	2506.97
Extra trees classifier	82.89	61.23	82.38	84.38	71.67	70.28	91.20
Adaboost classifier	73.29	37.80	75.46	66.51	54.75	68.77	330.66
K-neighbors classifier	79.10	52.71	81.73	72.98	67.76	51.30	513.83
Linear discriminant analysis	70.39	29.98	72.82	61.69	47.71	50.68	43.75
Quadratic discriminant analysis	81.23	72.88	85.46	73.18	58.54	61.51	54.10
Support vector machine	83.75	50.13	81.74	79.86	60.02	73.18	13564.35

As we can see from Table 1, six of the eight algorithms have a predict accuracy above or around 80%. Considering that the algorithm should be different from one another, and the final method should have a reasonable time consumption, finally four algorithms are chosen: gradient boosting classifier, extra trees classifier, k-neighbors classifier, and quadratic discriminant analysis. The essential parameters of these estimators are set as follows: number of boosting stages in gradient boosting classifier is set to 150, contribution of each tree is set to 0.3, maximum depth of the individual regression estimators is set to 7; number of trees in the extra trees classifier is set to 200, use Gini impurity to measure the quality of a split; in k-neighbors classifier, points in each neighborhood has a weighting decided by the inverse of their distance; while quadratic discriminant analysis does not have any particular parameters that need to set.

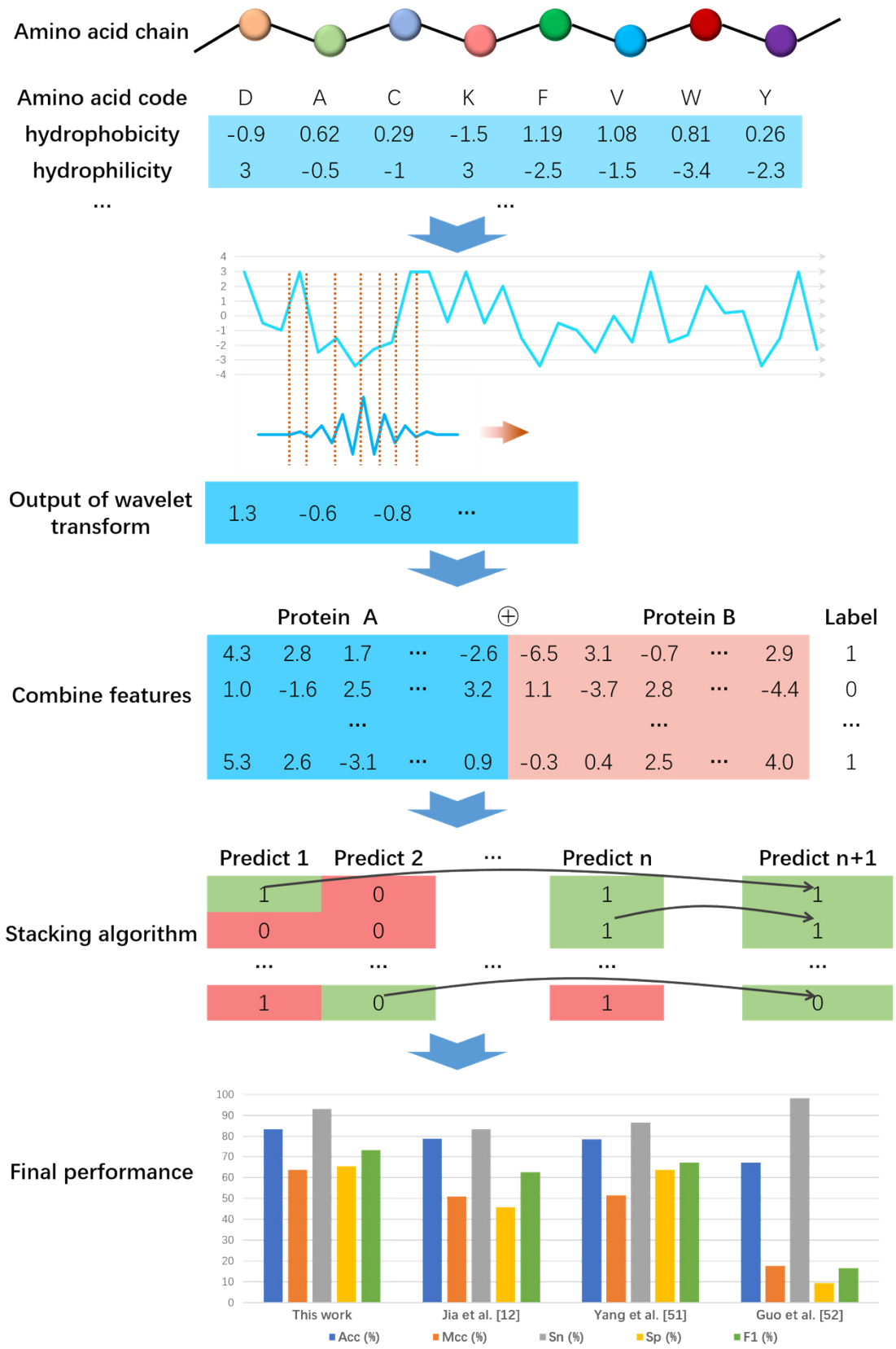

Figure 5. Schematic diagram of the overview algorithm process used in this work

Prediction performance of proposed model

The proposed method was firstly tested on the *Saccharomyces cerevisiae* dataset. The methods from other published papers are used as contrasts. Under 5-fold cross-validation, the method proposed in this paper achieved higher scores in evaluation criterions like Acc, Mcc, F_1, Sn and Sp, than some state-of-the-art methods. Fig. 5 illustrated a schematic diagram of the overall prediction process performed in this work and the results are given in Table 2. Additionally, as recommended by majority literatures, the proposed method was also tested using the *H. pylori* dataset (the results are given in Table 3). As shown Table 3, the algorithm presented herein achieved outstanding performance on the *H. pylori* data set, and the test scores are significantly higher than the other methods available at this stage. The above test results demonstrate that the proposed novel approach can effectively improve the predictive performance of protein interaction and has good robustness as well.

Table 2

Comparison of the performance by the proposed method and other available methods on the yeast dataset*. The definition of Acc, MCC, Sn, Sp and F_1, refer to Eq (7)

Method	Test set	Acc (%)	Mcc (%)	Sn (%)	Sp (%)	F_1 (%)
This work		**83.35**	**63.77**	**92.95**	**65.41**	**73.24**
Jia et al. [12]	iPPI-Esml	78.74	50.98	83.33	45.77	62.65
Yang et al. [51]	LD (Cod4)	78.54	51.52	86.38	63.69	67.23
Guo et al. [52]	AC+ SVM	67.37	17.74	98.33	9.22	16.42

* 5-fold cross-validation was used

Table 3

Comparison with other available methods on the *H. pylori* dataset**.

Method	Test set	Acc (%)	Mcc (%)	Sn (%)	Sp (%)	F_1 (%)
This paper		**80.39**	**61.15**	**76.57**	**84.54**	**80.50**
Jia et al. [12]	iPPI-Esml	78.62	57.13	81.22	75.79	77.25
Yang et al. [51]	LD (Cod4)	70.21	42.51	89.26	49.47	61.39
Guo et al. [52]	AC+ SVM	63.31	32.14	95.01	28.81	42.92

** 5-fold cross-validation was used

Conclusion

Prediction of the protein-protein interactions (PPIs) is nowadays a critical research issue, as it can facilitate revealing the biological processes within living cells. In this work, a novel classifier is developed for predicting PPIs based on the stacking algorithm and information extraction by wavelet transform. Our results on the PPI data of *Saccharomyces cerevisiae* showed that the proposed method with the assistance of wavelet transform is capable of extracting maximum

information from primary protein sequence. Meanwhile, the combination of stacking algorithm can significantly improve on the performance of single classifier in distinguishing interacting and non-interacting protein pairs. In addition, the results on the independent data set of the *H. pylori* PPIs further demonstrated the stable performance of our classifier. In conclusion, this new classifier model might be another effective tool for the prediction of PPIs.

References

[1] Gavin A C, Bösche M, Krause R, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes[J]. Nature, 2002, 415(6868): 141-147.

[2] Krogan N J, Cagney G, Yu H, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae[J]. Nature, 2006, 440(7084): 637-643.

[3] Ito T, Chiba T, Ozawa R, et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome[J]. Proceedings of the National Academy of Sciences, 2001, 98(8): 4569-4574.

[4] Shen J, Zhang J, Luo X, et al. Predicting protein–protein interactions based only on sequences information[J]. Proceedings of the National Academy of Sciences, 2007, 104(11): 4337-4341.

[5] You Z H, Lei Y K, Zhu L, et al. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis[J]. BMC bioinformatics, 2013, 14(8): S10.

[6] You Z H, Li X, Chan K C C. An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers[J]. Neurocomputing, 2017, 228: 277-282.

[7] Zubek J, Tatjewski M, Boniecki A, et al. Multi-level machine learning prediction of protein–protein interactions in Saccharomyces cerevisiae[J]. PeerJ, 2015, 3: e1041.

[8] Chou K C, Mao B. Collective motion in DNA and its role in drug intercalation[J]. Biopolymers, 1988, 27(11): 1795-1815.

[9] Mallat S G. A theory for multiresolution signal decomposition: the wavelet representation[J]. IEEE transactions on pattern analysis and machine intelligence, 1989, 11(7): 674-693.

[10] Liu H, Wang M, Chou K C. Low-frequency Fourier spectrum for predicting membrane protein types[J]. Biochemical and biophysical research communications, 2005, 336(3): 737-739.

[11] Li Z C, Zhou X B, Dai Z, et al. Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis[J]. Amino acids, 2009, 37(2): 415.

[12] Nanni L, Brahnam S, Lumini A. Wavelet images and Chou's pseudo amino acid composition for protein classification[J]. Amino Acids, 2012, 43(2): 657-665.

[13] Jia J, Liu Z, Xiao X, et al. iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC[J]. Journal of theoretical biology, 2015, 377: 47-56.

[14] Wang S Q, Yang J, Chou K C. Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition[J]. Journal of Theoretical Biology, 2006, 242(4):

941-946.
[15] Chou K C, Shen H B. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization[J]. Biochemical and biophysical research communications, 2006, 347(1): 150-157.
[16] Christopher M. Bishop: Pattern Recognition and Machine Learning, Chapter 4.3.4
[17] Martin S, Roe D, Faulon J L. Predicting protein–protein interactions using signature products[J]. Bioinformatics, 2005, 21(2): 218-226.
[18] Chou K C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology[J]. Current Proteomics, 2009, 6(4): 262-274.
[19] Gacesa R, Barlow D J, Long P F. Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions[J]. PeerJ Computer Science, 2016, 2: e90.
[20] Xu Y, Shao X J, Wu L Y, et al. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins[J]. PeerJ, 2013, 1: e171.
[21] Chen W, Lin H, Feng P M, et al. iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties[J]. PloS one, 2012, 7(10): e47843.
[22] Chen W, Feng P M, Lin H, et al. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition[J]. Nucleic acids research, 2013: gks1450.
[23] You Z H, Li S, Gao X, et al. Large-scale protein-protein interactions detection by integrating big biosensing data with computational model[J]. BioMed research international, 2014, 2014.
[24] Chou K C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes[J]. Bioinformatics, 2005, 21(1): 10-19.
[25] Du P, Wang X, Xu C, et al. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions[J]. Analytical biochemistry, 2012, 425(2): 117-119.
[26] Du P, Gu S, Jiao Y. PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets[J]. International Journal of Molecular Sciences, 2014, 15(3): 3495-3506.
[27] Chou K C. Some remarks on protein attribute prediction and pseudo amino acid composition[J]. Journal of theoretical biology, 2011, 273(1): 236-247.
[28] Chou K C. Low-frequency resonance and cooperativity of hemoglobin[J]. Trends in biochemical sciences, 1989, 14(6): 212.
[29] Zhou X, Li Z, Dai Z, et al. Predicting promoters by pseudo-trinucleotide compositions based on discrete wavelets transform[J]. Journal of theoretical biology, 2013, 319: 1-7.
[30] Chen C, Shen Z B, Zou X Y. Dual-layer wavelet SVM for predicting protein structural class via the general form of Chou's pseudo amino acid composition[J]. Protein and peptide letters, 2012, 19(4): 422-429.
[31] Qiu J D, Huang J H, Liang R P, et al. Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform[J]. Analytical biochemistry, 2009, 390(1): 68-73.

[32] Qiu J D, Huang J H, Shi S P, et al. Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform[J]. Protein and peptide letters, 2010, 17(6): 715-722.
[33] Qiu J D, Suo S B, Sun X Y, et al. OligoPred: A web-server for predicting homo-oligomeric proteins by incorporating discrete wavelet transform into Chou's pseudo amino acid composition[J]. Journal of Molecular Graphics and Modelling, 2011, 30: 129-134.
[34] Rezaei M A, Abdolmaleki P, Karami Z, et al. Prediction of membrane protein types by means of wavelet analysis and cascaded neural networks[J]. Journal of theoretical biology, 2008, 254(4): 817-820.
[35] Sun X Y, Shi S P, Qiu J D, et al. Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform[J]. Molecular BioSystems, 2012, 8(12): 3178-3184.
[36] Tanford C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins[J]. Journal of the American Chemical Society, 1962, 84(22): 4240-4247.
[37] Hopp T P, Woods K R. Prediction of protein antigenic determinants from amino acid sequences[J]. Proceedings of the National Academy of Sciences, 1981, 78(6): 3824-3828.
[38] Krigbaum W R, Komoriya A. Local interactions as a structure determinant for protein molecules: II[J]. Biochimica et Biophysica Acta (BBA)-Protein Structure, 1979, 576(1): 204-228.
[39] Grantham R. Amino acid difference formula to help explain protein evolution[J]. Science, 1974, 185(4154): 862-864.
[40] Charton M, Charton B I. The structural dependence of amino acid hydrophobicity parameters[J]. Journal of theoretical biology, 1982, 99(4): 629-644.
[41] Rose G D, Geselowitz A R, Lesser G J, et al. Hydrophobicity of amino acid residues in globular proteins[J]. Science, 1985, 229: 834-839.
[42] Zhou P, Tian F, Li B, et al. Genetic algorithm-based virtual screening of combinative mode for peptide/protein[J]. ACTA CHIMICA SINICA-CHINESE EDITION-, 2006, 64(7): 691.
[43] Geller J. Data mining: practical machine learning tools and techniques with java implementations[J]. SIGMOD Record, 2002, 31(1): 77.
[44] Svetnik V, Liaw A, Tong C, et al. Random forest: a classification and regression tool for compound classification and QSAR modeling[J]. Journal of chemical information and computer sciences, 2003, 43(6): 1947-1958.
[45] Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001: 1189-1232.
[46] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees[J]. Machine learning, 2006, 63(1): 3-42.
[47] Freund Y, Schapire R E. A desicion-theoretic generalization of on-line learning and an application to boosting[C]//European conference on computational learning theory. Springer Berlin Heidelberg, 1995: 23-37.
[48] Altman N S. An introduction to kernel and nearest-neighbor nonparametric regression[J].

418 The American Statistician, 1992, 46(3): 175-185.
419 [49] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2001[J]. NY
420 Springer, 2001.
421 [50] Chang C C, Lin C J. LIBSVM: a library for support vector machines[J]. ACM Transactions
422 on Intelligent Systems and Technology (TIST), 2011, 2(3): 27.
423 [51] Yang L, Xia J F, Gui J. Prediction of protein-protein interactions from protein sequence
424 using local descriptors[J]. Protein and Peptide Letters, 2010, 17(9): 1085-1090.
425 [52] Guo Y, Yu L, Wen Z, et al. Using support vector machine combined with auto covariance to
426 predict protein–protein interactions from protein sequences[J]. Nucleic acids research, 2008,
427 36(9): 3025-3030.