Predicting Protein-Protein Interactions from Protein Sequence Information Using Dual-Tree Complex Wavelet Transform

Jie Pan, Zhu-Hong You(✉), Chang-Qing Yu, Li-Ping Li,
and Xin-ke Zhan

School of Information Engineering, Xijing University, Xi'an 710123, China
zhuhongyou@gmail.com

Abstract. Protein-protein interactions (PPIs) play major roles in most biological processes. Although a number of high-throughput technologies have been established for generating PPIs, it still has unavoidable problems such as time-consuming and labor intensive. In this paper, we develop a novel computational method for predicting PPIs by combining dual-tree complex wavelet transform (DTCWT) on substitution matrix representation (SMR) and weighted sparse representation-based classifier (WSRC). When predicting PPIs of *Yeast* and *Human* datasets, the proposed method obtained remarkable results with average accuracies as high as 97.12% and 97.56%, respectively. The performance of the proposed method is obviously better than the existing methods. Furthermore, we compare it with the superior support vector machine (SVM) classifier for further evaluating the prediction performance of our method. The promising results illustrate that our method is robust and stable for predicting PPIs, and it is anticipated that it would be a useful tool to predict PPIs in a large-scale.

Keywords: Protein-protein interactions · Protein sequence · Dual-tree complex wavelet transform · Weighted sparse representation

1 Introduction

Identification of protein–protein interactions (PPIs) is crucial for studying protein function and deep understanding of biological processes in a cell. In recent years, plenty of high-throughput technologies, such as *Yeast two-hybrid (Y2H) screens* [1, 2], *tandem affinity purification* (TAP) [3] and *mass spectrometric protein complex identification* (MS-PCI) [4], have been developed for the large-scale PPIs detection. However, these previous methods are expensive and require a great deal of human effort. In addition, only a small part of the whole PPIs have been identified. Thus, we can draw a conclusion that only using experimental methods is difficult to identify unknown PPIs.

In recent years, researchers have developed different types of computational methods for the prediction of PPIs [5–16]. For example, Li *et al*. proposed a novel computational model combining Position Weight Matrix (PWM) and Scale-Invariant Feature Transform (SIFT) algorithm [17]. An *et al*. proposed an effective algorithm that using Gray Wolf Optimizer–Based Relevance Vector Machine to identify PPIs [18].

D.-S. Huang and K.-H. Jo (Eds.): ICIC 2020, LNCS 12464, pp. 132–142, 2020.
https://doi.org/10.1007/978-3-030-60802-6_13

Zhu *et al.* proposed a useful tool which used the Position-Specific Scoring Matrices (PSSMs) and ensemble learning algorithm Rotation Forest (RF) [19]. You *et al.* developed a novel method for detecting PPIs by integrating a new protein sequence substitution matrix feature representation and ensemble weighted sparse representation model classifier [20]. Huang *et al.* proposed a sequence-based method based on the combination of weighted sparse representation based classifier (WSRC) and global encoding (GE) of amino acid sequence [21]. Wen *et al.* proposed an effective method based on similar network fusion (SNF) model to integrate the physical and chemical properties of proteins [22]. Leon Wong *et al.* presented a novel computational approach that combining a Rotation Forest Model with a novel PR-LPQ Descriptor [23]. Huang *et al.* developed an effective algorithm, which is based on Extreme Learning Machine (ELM) and combined the concept of Chou's Pseudo-Amino Acid Composition (PseAAC) composition [24].

In this paper, we propose a novel computational method for predicting PPIs, which combines weighted sparse representation based classifier (WSRC) and the dual-tree complex wavelet transform (DTCWT). Specifically, we first select substitute matrix representation (SMR) based on BLOSUM62 to represent protein sequences. Then, we adopt the DTCWT to extract feature vectors from each SMR matrix. Finally, we utilize WSRC to predict PPIs on two different biological datasets: *Yeast* and *Human*. Our model achieves excellent performance results which obtain average accuracies of 97.12% and 97.56%, respectively. In order to further evaluating the proposed method, we compared the WSRC with the state-of-the-art support vector machine (SVM) classifier. The promising results demonstrated that the proposed method is robust and stable for the prediction of PPIs.

2 Materials and Methodology

2.1 Godden Standard Datasets

The first dataset that we chose in this paper is gathered from publicly available database of interacting proteins (DIP). We removed the protein pairs whose length are less than 50 residues because these might be fragments. The pairs with $\geq 40\%$ sequence identity have been deleted too. In this way, the positive dataset is constructed by the remaining 5594 protein pairs. Moreover, we selected 5594 additional protein pairs of different subcellular localizations to build the negative dataset. Consequently, the whole dataset is made up of 11188 protein pairs.

In order to demonstrate the generality of the proposed method, we validated our method on another PPI dataset. We collected the dataset from the Human Protein References Database (HPRD). Those protein pairs which have $\geq 25\%$ sequence identity have been removed. Finally, we used the remaining 3899 protein-protein pairs of experimentally verified PPIs from 2502 different human proteins, so that we can comprise the golden standard positive dataset. Following the previous work [25], we assume that the proteins in different subcellular compartments will not interact with each other and finally obtained 4262 protein pairs from 661 different human proteins as the negative dataset. As a result, the Human dataset is constructed by 8161 protein pairs.

2.2 Substitution Matrix Representation

Substitution matrix representation (SMR) is a modified version of representation method reported by [26]. In this novel matrix representation for proteins, we generated a $N \times 20$ matrix to represent the N-length protein sequence, which are based on a substitution matrix. BLOSUM62 matrix is a powerful substitution matrix and has been utilized in this work for the sequence alignment of proteins. SMR can be defined as follows:

$$SMR(i.j) = B(P(i),j)\, i = 1 \cdots N, j = 1 \cdots 20 \tag{1}$$

In this formula, B means the BLOSUM62 matrix, it is a 20 × 20 substitution matrix and $B(i,j)$ represents the value in row i and column j of BLOSUM62 matrix, this value represents the probability rate of amino acid i converting to amino acid j in the evolution process; $P = (p1, p2 \cdots pN)$ is the given protein sequence constructed by N amino acids.

2.3 The Dual-Tree Complex Wavelet Transform

The dual-tree complex wavelet transform (DTCWT) [27] is a variant of the traditional complex wavelet transform (DWT). It inherited the characteristics of the Multi-scale and Multi-resolution of discrete wavelet transform. At the same time, it makes up for the deficiencies of complex wavelet transform with large amount of calculation and high complexity. Different with the traditional DWT, DTCWT utilized two real DWTs to form a complex transform [28]. The first part symbolizes the real component and the second part represents the imaginary component of this transform.

The DTCWT settled the matters about the "shift-invariant problems" and "directional selectivity in two or more dimensions," which are both weak points of conventional DWT [29]. It acquired directional selectivity by using approximate analytic wavelets. It also has the skills to generate a total of six directionally discriminating subbands oriented in the ±15°, ±45° and ±75° directions, for both the real (R) and imaginary (I) parts. Let $h_i(n)$ and $g_i(n)$ be the filters in the first stage. Let the new stage response of the first filter bank be $H_{new}^{(k)}(e^{jw})$ and second filter bank be $H_{new}^{'(k)}(e^{jw})$; we now have the following result.

Suppose one is provided with CQF pairs $\{h_0(n), h_1(n)\}, \left\{h_0^{'}(n), h_1^{'}(n)\right\}$. For $k > 1$.

$$H_{new}^{(k)}(e^{jw}) = H\left\{H_{new}^{'(k)}(e^{jw})\right\} \tag{2}$$

if and only if

$$h_0^{'(1)}(n) = h_0^{(1)}(n-1) \tag{3}$$

A 2D image $f(x, y)$ can be decomposed by 2D DTCWT over a series of dilations and translations of a complicated scaling function and six complex wavelet function $\varphi^{\theta}_{j,l}$; that is:

$$f(x, y) = \sum_{l \in Z^2} s_{j_0, l\phi j_0} l^{(x,y)} + \sum_{\theta \in \Theta} \sum_{j \geq j_0} \sum_{l \in Z^2} c^{\theta}_{j}, l^{\varphi^{\theta}_{j}}, l^{(x,y)} \tag{4}$$

where $\theta \in \Theta = \{\pm 15^{\circ},\ \pm 45^{\circ},\ \pm 75^{\circ}\}$ gives the directionality of the complex wavelet function.

2.4 Weighted Sparse Representation-Based Classification

In the past twenty years, sparse representation based classifier (SRC) [30, 31] has earned considerable attention in the field of signal processing, pattern recognition and computer vision because of the great development of linear representation methods (LRBM) and compressed sensing (CS) theory. Sparse representation attempts to optimize matrix to reveal the relationship between any given test sample and the training set. Therefore, it would be a good trial to use it for building a prediction system for PPIs. In this work, we build a computational model by employing weighted sparse representation-based classifier (WSRC).

Given a training sample matrix $x \in R^{m \times n}$ which is made up of n samples of m dimensions. If there are sufficient training samples belonging to the *kth* class, then the sub-matrix constructed by the samples of the *kth* class can be symbolized as $X_k[l_{k1}, l_{k2} \cdots l_{kn_k}]$, where l_i denotes the class of *ith* sample and n_k is the number of samples belonging to *kth* class. Thus, X can be further rewritten as $X = [X_1, X_2 \cdots X_K]$, where K is the class number of the whole samples. Given a test sample $y \in R^m$ and it can be represented as

$$y = \alpha_{k,1} l_{k,1} + \alpha_{k,2} l_{k,2} + \cdots + \alpha_{k,n_k} l_{k,n_k} \tag{5}$$

when considering the whole training set representation, Eq. (5) can be further symbolized as

$$y = X\alpha_0 \tag{6}$$

where $\alpha_0 = \left[0, \ldots 0, \alpha_{k,2}, \cdots, \alpha_{k,n_k}, 0, \cdots 0\right]^T$. For these reason that the nonzero entries in α_0 are only associated with the *kth* class, so if class number of samples become large, the α_0 would come to be sparse. The key question of SRC algorithm is searching the α vector which can subject to Eq. (6) and minimize the ℓ_0-norm of itself:

$$\begin{aligned} \widehat{\alpha_0} &= \arg\min \|\alpha\|_0 \\ &\text{subject to } y = X\alpha \end{aligned} \tag{7}$$

Problem (7) is an NP-hard problem and it can be achieved but difficultly to be solved precisely. According to the theory of compressive sensing [32, 33] show, if α is sparse enough, we can solve the related problem convex l_1-minimization problem instead of solving the l_0-minimization problem directly.

$$\begin{gathered}\widehat{\alpha_1} = \arg\min\|\alpha\|_1 \\ \text{subject to } y = X\alpha\end{gathered} \tag{8}$$

When dealing with occlusion, we should extend Eq. (8) to the stable ℓ_1-minimization problem

$$\begin{gathered}\widehat{\alpha_1} = \arg\min\|\alpha\|_1 \\ \text{subject to } \|y - X\alpha\| \leq \varepsilon\end{gathered} \tag{9}$$

where $\varepsilon > 0$ represent the tolerance of reconstruction error. Given the solution from Eq. (9), the SRC algorithm assigns the label of test sample y to class c with the reconstruction residual:

$$\min_c r_c(y) = \|y - X\widehat{\alpha_1}^c\|, \, c = 1\cdots K \tag{10}$$

Besides sparse representation, Nearest Neighbor (NN) is another popular classifier which only considering the influence of the Nearest Neighbor in training data to classify the test sample and SRC uses the linearity structure of data and overcomes the drawback of NN. Some researches shows that locality is more essential than sparsity in some cases [34, 35]. Lu *et al.* [36] have proposed a modified version of traditional sparse representation based classifier called weighted sparse representation based classifier (WSRC), it integrates the locality structure of data into basic sparse representation. Specifically, Gaussian distance between single sample and the whole training samples will be first computed and WSRC can use it as the weights of each training sample. The Gaussian distance between two samples, s_1 and s_2 can be described as follow:

$$d_G(s_1, s_2) = e^{-\|s_1 - s_2\|^2/2\alpha^2} \tag{11}$$

where σ is the Gaussian kernel width. In this way, weighted sparse representation based classifier can retain the locality structure of data and then it turned to solve the following problem:

$$\begin{gathered}\widehat{\alpha_1} = \arg\min\|W\alpha\|_1 \\ \text{subject to } y = X\alpha\end{gathered} \tag{12}$$

specifically,

$$diag(W) = \left[d_G(y, x_1^1), \ldots, d_G(y, x_{n_k}^k)\right]^T \tag{13}$$

where W is a block-diagonal matrix of locality adaptor and n_k is the sample number of training set in class k. Dealing with this occlusion, we can eventually solve the following stable ℓ_1-minimization problem:

$$\begin{aligned} &\widehat{\alpha_1} = \arg\min \|W\alpha\|_1 \\ &\text{subject to } \|y - X\alpha\| \leq \varepsilon \end{aligned} \tag{14}$$

where $\varepsilon > 0$ is the tolerance value.

In summary, the WSRC algorithm can be summarized by the following steps:

Algorithm1. Weighted Sparse Representation Based Classifier (WSRC)

```
(1). Input：training samples matrix $X \in R^{m \times n}$ and any test sample $y \in R^d$ .
(2). Normalize the columns of $X$ to have unit $\ell_2$ -norm.
(3). Calculate the Gaussian distances between $y$ and each sample in $X$ and make up matrix $W$ .
(4). Solve the stable $\ell_1$ -minimization problem defined in Eq.(12)
(5). Compute each residual of $K$ classes: $r_c(y) = \left\| y - X\overset{c}{\alpha_1} \right\|$ (c=1,2,...,K)
(6). Output: assign to the class $c$ by the rule: $identity(y) = \underset{c}{\arg\min}(r_c(y))$ .
```

3 Results and Discussion

In order to evaluate the performance of the proposed method, the overall prediction accuracy (Acc.), sensitivity (Sen.), precision (PR.), and Matthews's correlation coefficient (MCC.) were calculated. They are defined as follows:

$$Acc. = \frac{TP + TN}{TP + FP + TN + FN} \tag{15}$$

$$Sen. = \frac{TP}{TP + FN} \tag{16}$$

$$PR. = \frac{TP}{TP + FP} \tag{17}$$

$$MCC. = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + TN) \times (TN + FN) \times (TP + FP) \times (TN + FN)}} \tag{18}$$

In these algorithm, true positive (TP) represents the number of true samples which are predicted correctly; false negative (FN) is the number of samples predicted to be non-interacting pairs incorrectly; false positive (FP) is the number of true non-interacting pairs predicted to be PPIs falsely; and true negative (TN) is the number of true noninteracting pairs predicted correctly. What's more, the receiver operating characteristic (ROC) curves are also calculated to evaluate the performance of proposed method. In order to summarize ROC curve in a numerical way, the area under an ROC curve (AUC) was computed.

3.1 Assessment of Prediction Ability

For the sake of fairness, when predicting PPIs of *Yeast* and *Human*, we set the same corresponding parameters of weighted sparse representation-based classifier. We set the $\sigma = 1.5$ and $\varepsilon = 0.00005$. In addition, 5-fold cross-validation [37] was employed in our experiments in order to avoid overfitting and get a stable and reliable model from the proposed method. Specifically, we divided the whole dataset into five subsets. Four of the subsets are used for training and the last part was used for testing. By this way, the results of these experiments in which we used the proposed model to predict PPIs of *Yeast* and *Human* datasets are shown in Tables 1 and 2.

Table 1. 5-fold cross-validation results obtained by using proposed method on *Yeast* dataset.

Testing set	Acc. (%)	PR. (%)	Sen. (%)	MCC (%)	AUC (%)
1	96.65	100	93.24	93.50	96.59
2	97.63	100	95.14	95.36	98.01
3	97.18	100	94.43	94.52	96.92
4	97.00	100	94.02	94.18	97.04
5	97.14	100	94.38	94.44	97.16
Average	**97.12 ± 0.35**	**100 ± 0.00**	**94.24 ± 0.69**	**94.40 ± 0.67**	**97.14 ± 0.53**

Table 2. 5-fold cross-validation results obtained by using proposed method on *Human* dataset.

Testing set	Acc. (%)	PR. (%)	Sen. (%)	MCC (%)	AUC (%)
1	98.28	99.86	96.44	96.60	97.89
2	97.43	99.32	95.19	94.95	96.73
3	96.63	99.74	93.45	93.47	96.20
4	97.49	99.05	95.57	95.07	97.61
5	97.98	99.48	96.35	96.03	98.10
Average	**97.56 ± 0.63**	**99.49 ± 0.32**	**95.40 ± 1.21**	**95.23 ± 1.19**	**97.31 ± 0.81**

When using the proposed method to predict PPIs of the *Yeast* dataset, we obtained the prediction results with average accuracy, precision, sensitivity, and MCC of 97.12%, 100%, 94.24% and 94.40%. The standard deviations of these criteria values are relatively low, which of accuracy, precision, sensitivity and MCC are 0.35%, 0.00%, 0.69%, and 0.67%, respectively. Form Table 2, when exploring the *Human* dataset, the proposed method yielded results of average accuracy, precision, sensitivity and MCC of 97.56%, 99.49%, 95.40%, and 95.23%. The standard deviations are 0.63%, 0.32%, 1.21%, and 1.19%, respectively. The ROC curves performed on these two datasets are shown in Fig. 1. To better evaluate the prediction performance of the proposed method, we computed the AUC values of *Yeast* and *Human* datasets, which are 97.14% and 97.31%, respectively.

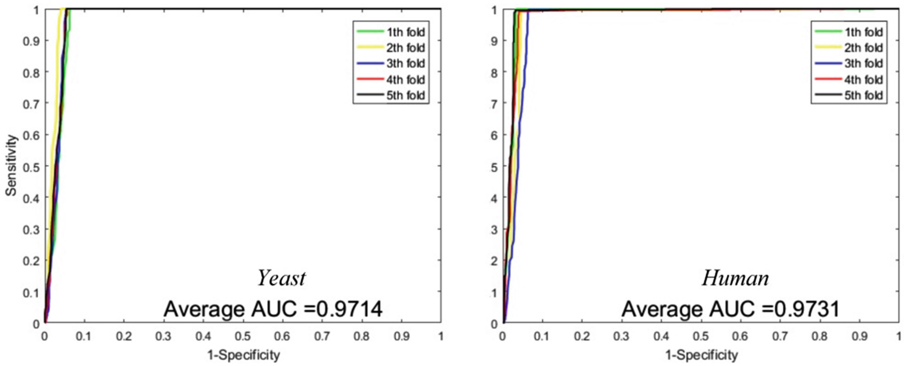

Fig. 1. ROC from proposed method result for *Yeast* and *Human* PPIs dataset.

The high accuracies show that WSRC based model combining the SMR-DTCWT descriptors is feasible and effective for predicting PPIs. All experimental results demonstrate the feasibility, effectiveness and robustness of the proposed method.

3.2 Comparison with SVM-Based Method

Table 3. Comparison with support vector machine on *Yeast* and *Human* datasets

Dataset	Classifier	Acc. (%)	PR. (%)	Sen. (%)	MCC (%)	AUC (%)
Yeast	SVM	84.74 ± 0.78	83.44 ± 0.96	86.71 ± 1.61	74.11 ± 1.08	91.99 ± 0.74
	WSRC	**97.12 ± 0.35**	**100.00 ± 0.00**	**94.24 ± 0.69**	**94.40 ± 0.67**	**97.14 ± 0.53**
Human	SVM	83.14 ± 0.83	81.92 ± 1.19	83.08 ± 2.17	71.92 ± 1.09	90.57 ± 0.41
	WSRC	**97.56 ± 0.63**	**99.49 ± 0.32**	**95.40 ± 1.21**	**95.23 ± 1.19**	**97.31 ± 0.81**

In order to further evaluate the performance of the proposed method, we compared the prediction performance of the proposed method with the state-of-the-art SVM classifier on the dataset of *Human* and *Yeast*. We utilized the same feature extraction method and a grid search method to optimize two corresponding parameters of SVM c and g. Here, we set the $c = 0.3$ and $g = 0.3$.

Table 3 shows that when using SVM to predict PPIs of *Yeast* dataset, we obtained relatively poor results with the average accuracy, precision, sensitivity, MCC, and AUC of 84.74%, 83.44%, 86.71%, 74.11%, and 91.99%, respectively. When exploring the *Human* dataset with the SVM-based method yielded relatively low results with the average accuracy, precision, sensitivity, MCC, and AUC of 83.14%, 81.92%, 83.08%, 71.92%, and 90.57%, respectively. Considering the comparison result and higher values for criteria and lower standard deviations, the prediction performance of SVM-based method is lower than that of WSRC. The ROC curves performed by SVM classifier on the two datasets are shown in Fig. 2.

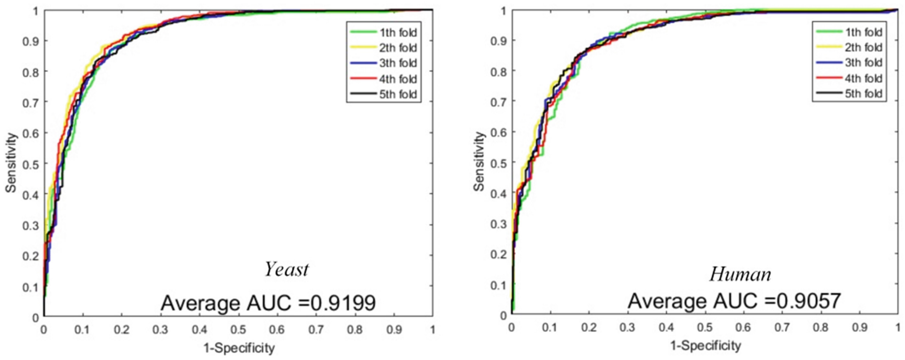

Fig. 2. ROC from SVM-based method result for *Yeast* and *Human* PPIs dataset.

4 Conclusions and Discussion

It is becoming more and more important to develop an effective and accurate method for predicting PPIs. In this work, we explore a novel computation model for predicting PPIs by combing weighted sparse representation-based classifier and the dual-tree complex wavelet transform. In the step of feature extraction, it has been proven that it is effective to combine the SMR matrix and dual-tree complex wavelet transform. Compared with the previous methods, the main improvement of the proposed method is to adopt a novel protein feature representation and utilizing a powerful classifier. In addition, good experiment results indicate that the proposed method performs well in PPIs prediction and has great generalization ability.

Competing Interests. The authors declare that they have no competing interests.

Acknowledgements. This work is supported by the National Science Foundation of China under Grants 61873212.

References

1. Nishihara, T., Nishikawa, J., Kanayama, T., Dakeyama, F., Saito, K., Imagawa, M., et al.: Estrogenic activities of 517 chemicals by yeast two-hybrid assay. J. Health Sci. **46**, 282–298 (2000)
2. Sato, T., Hanada, M., Bodrug, S., et al.: Interactions among members of the Bcl-2 protein family analyzed with a yeast two-hybrid system. Proc. Natl. Acad. Sci. **91**, 9238–9242 (1994)
3. Puig, O., Caspary, F., Rigaut, G., et al.: The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods **24**, 218–229 (2001)
4. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.-L., et al.: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature **415**, 180–183 (2002)

5. You, Z.H., Zhou, M.C., Luo, X., Li, S.: Highly efficient framework for predicting interactions between protein. IEEE Trans. Cybernet. **47**, 731–743 (2016)
6. Wang, L., et al.: Using two-dimensional principal component analysis and rotation forest for prediction of protein-protein interactions. Scientific Reports **8**, 1–10 (2018)
7. Wang, L., You, Z.H., Xia, S.X., Liu, F., Chen, X., Yan, X., Zhou, Y.: Advancing the prediction accuracy of protein-protein interactions by utilizing evolutionary information from position-specific scoring matrix and ensemble classifier. J. Theor. Biol. **418**, 105–110 (2017)
8. Li, Z.W., et al.: Accurate prediction of protein-protein interactions by integrating potential evolutionary information embedded in PSSM profile and discriminative vector machine classifier. Oncotarget **8**(14), 23638 (2017)
9. 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. Neurocomputing **228**, 277–282 (2017)
10. Wang, L., et al.: An ensemble approach for large-scale identification of protein-protein interactions using the alignments of multiple sequences. Oncotarget **8**, 5149 (2017)
11. Wang, Y.B., et al.: Predicting protein-protein interactions from protein sequences by stacked sparse auto-encoder deep neural network. Molecular BioSyst. **13**, 1336–1344 (2017)
12. An, J.Y., You, Z.H., Chen, X., Huang, D.S., Yan, G.Y.: Robust and accurate prediction of protein self-interactions from amino acids sequence using evolutionary information. Molecular BioSyst. **12**, 3702–3710 (2016)
13. Huang, Y.A., et al.: Construction of reliable protein-protein interaction networks using weighted sparse representation based classifier with pseudo substitution matrix representation. Neurocomputing **218**, 131–138 (2016)
14. An, J.Y., Meng, F.R., You, Z.H., Chen, X.: Improving protein-protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model. Protein Sci. **25**, 1825–1833 (2016)
15. You, Z.H., Chan, K.C.C.: Prediction of protein-protein interactions from primary protein sequence using random forest model with a novel multi-scale local feature representation. PLoS ONE **10**, e0131091 (2015)
16. You, Z.H., Zhu, L., Zheng, C.H., Yu, H.J., Deng, S.P., Ji, Z.: Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. BMC Bioinform. **15**(15), s9 (2014)
17. Li, J., Shi X., You, Z.H., Yi, H.C., Chen, Z., Lin, Q., et al.: Using weighted extreme learning machine combined with scale-invariant feature transform to predict protein-protein interactions from protein evolutionary information. IEEE/ACM Trans. Comput. Biol. Bioinform. (2020)
18. An, J.-Y., You, Z.H., Zhou, Y., Wang, D.F.: Sequence-based prediction of protein-protein interactions using gray wolf optimizer-based relevance vector machine. Evol. Bioinform. **15**, 1176934319844522 (2019)
19. Zhu, H.J., You, Z.H., Shi, W.L., Xu, S.K., Jiang, T.H., Zhuang, L.H.: Improved prediction of protein-protein interactions using descriptors derived from PSSM via Gray Level Co-Occurrence Matrix. IEEE Access **7**, 49456–49465 (2019)
20. You, Z.H., Huang, W.Z., Zhang, S., Huang, Y.A., Yu, C.Q., Li, L.P.: An efficient ensemble learning approach for predicting protein-protein interactions by integrating protein primary sequence and evolutionary information. IEEE/ACM Trans. Comput. Biol. Bioinform. **16**, 809–817 (2018)
21. Huang, Y.A., You, Z.H., Chen, X., Chan, K., Luo, X.: Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding. BMC Bioinform. **17**, 184 (2016)

22. Wen, Y.T., Lei, H.J., You, Z.H., Lei, B.Y., Chen, X., Li, L.P.: Prediction of protein-protein interactions by label propagation with protein evolutionary and chemical information derived from heterogeneous network. J. Theor. Biol. **430**, 9–20 (2017)
23. Wong, L., You, Z.-H., Li, S., Huang, Y.-A., Liu, G.: Detection of protein-protein interactions from amino acid sequences using a rotation forest model with a novel PR-LPQ descriptor. In: Huang, D.-S., Han, K. (eds.) ICIC 2015. LNCS (LNAI), vol. 9227, pp. 713–720. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22053-6_75
24. Huang, Q.Y., You, Z.H., Li, S., Zhu, Z.X.: Using Chou's amphiphilic pseudo-amino acid composition and extreme learning machine for prediction of protein-protein interactions. In: IEEE, pp. 2952–2956 (2014)
25. You, Z.H., Yu, J.Z., Zhu, L., Li, S., Wen, Z.K.: A MapReduce based parallel SVM for large-scale predicting protein–protein interactions. Neurocomputing **145**, 37–43 (2014)
26. Yu, X., Zheng, X., Liu, T., Dou, Y., Wang, J.: Predicting subcellular location of apoptosis proteins with pseudo amino acid composition: approach from amino acid substitution matrix and auto covariance transformation. Amino Acids **42**, 1619–1625 (2012)
27. Selesnick, I.W., Baraniuk, R.G., Kingsbury, N.G.: The dual-tree complex wavelet transform. IEEE Signal Process. Mag. **22**, 123–151 (2005)
28. Kingsbury, N.: Complex wavelets for shift invariant analysis and filtering of signals. Appl. Comput. Harmon. Anal. **10**, 234–253 (2001)
29. Barri, A., Dooms, A., Schelkens, P.: The near shift-invariance of the dual-tree complex wavelet transform revisited. J. Math. Anal. Appl. **389**, 1303–1314 (2012)
30. Yin, J., Liu, Z., Jin, Z., Yang, W.: Kernel sparse representation based classification. Neurocomputing **77**, 120–128 (2012)
31. Gao, Y., Ma, J., Yuille, A.L.: Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples. IEEE Trans. Image Process. **26**, 2545–2560 (2017)
32. Candes, E.J., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory **52**, 5406–5425 (2006)
33. Chen, S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**, 33–61 (1998)
34. Liu, C.H.W.: Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans. Image Process. **11**, 467–476 (2002)
35. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: IEEE, pp. 3360–3367 (2010)
36. Lu, C.Y., Min, H., Gui, J., Zhu, L., Lei, Y.K.: Face recognition via weighted sparse representation. J. Vis. Commun. Image Rep. **24**, 111–116 (2013)
37. Zhang, P.: Model selection via multifold cross validation. Ann. Stat. **21**, 299–313 (1993)