


# An Efficient Feature Extraction Technique Based on Local Coding PSSM and Multifeatures Fusion for Predicting Protein-Protein Interactions

Ji-Yong An<sup>1,2</sup> , Yong Zhou<sup>1,2</sup>, Yu-Jun Zhao<sup>1,2</sup> and Zi-Ji Yan<sup>1,2</sup>

<sup>1</sup>School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China. <sup>2</sup>Mine Digitization Engineering Research Center, Ministry of Education, Xuzhou, People's Republic of China.

Evolutionary Bioinformatics  
Volume 15: 1–10  
© The Author(s) 2019  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/1176934319879920



## ABSTRACT

**BACKGROUND:** Increasing evidence has indicated that protein-protein interactions (PPIs) play important roles in various aspects of the structural and functional organization of a cell. Thus, continuing to uncover potential PPIs is an important topic in the biomedical domain. Although various feature extraction methods with machine learning approaches have enhanced the prediction of PPIs. There remains room for improvement by developing novel and effective feature extraction methods and classifier approaches to identify PPIs.

**METHOD:** In this study, we proposed a sequence-based feature extraction method called LCPSSMMF, which combined local coding position-specific scoring matrix (PSSM) with multifeatures fusion. First, we used a novel local coding method based on PSSM to build a new PSSM (CPSSM); the advantage of this method is that it incorporated global and local feature extraction, which can account for the interactions between residues in both continuous and discontinuous regions of amino acid sequences. Second, we adopted 2 different feature extraction methods (Local Average Group [LAG] and Bigram Probability [BP]) to capture multiple key feature information by employing the evolutionary information embedded in the CPSSM matrix. Finally, feature vectors were acquired by using multifeatures fusion method.

**RESULT:** To evaluate the performance of the proposed feature extraction approach, we employed support vector machine (SVM) as a prediction classifier and applied this method to *yeast* and *human* PPI datasets. The prediction accuracies of LCPSSMMF were 93.43% and 90.41% on the *yeast* and *human* datasets, respectively. Moreover, we also compared the proposed method with the previous sequence-based approaches on the *yeast* datasets by using the same SVM classifier. The experimental results indicated that the performance of LCPSSMMF significantly exceeded that of several other state-of-the-art methods. It is proven that the LCPSSMMF approach can capture more local and global discriminatory information than almost all previous methods and can function remarkably well in identifying PPIs. To facilitate extensive research in future proteomics studies, we developed a LCPSSMMFSVM server, which is freely available for academic use at <http://219.219.62.123:8888/LCPSSMMFSVM>.

**KEYWORDS:** Protein-protein interactions, local coding, position-specific scoring matrix, multifeatures fusion, support vector machine

**RECEIVED:** September 4, 2019. **ACCEPTED:** September 11, 2019.

**TYPE:** Machine Learning Models for Multi-omics Data Integration-Original Research

**FUNDING:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported by "The Fundamental Research Funds for the Central Universities (2019XKQYMS88)."

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Yong Zhou, School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 21116, Jiangsu, China. Email: [ajysjm@163.com](mailto:ajysjm@163.com)

## Introduction

Proteins are one of the most fundamental elements in living organisms and make important contributions in nearly all fundamental biological processes in the cell. However, proteins perform their functions by interacting with other proteins. Increasing studies have indicated that protein-protein interactions (PPIs) play a critical role in many important biological processes, including transcription regulation, signal transduction, foreign molecules recognition, translation, and so on. As indicated by increasing evidences, knowledge of PPIs can provide a deeper understanding of the molecular functions of biological processes, suggest novel methods for their practical use in medicine, and yield insight into disease mechanisms. Although many high-throughput methods, including the yeast 2-hybrid system,<sup>1-4</sup> protein chips,<sup>5-7</sup> and immunoprecipitation,<sup>8-10</sup> are typically used to identify PPIs, experimental methods for identifying associations between proteins are expensive and time-consuming and

suffer from high rates of false positives and false negatives.<sup>11-15</sup> Thus, increasing numbers of studies have focused on computational approaches to identify PPIs.<sup>16-18</sup> As a result, a large number of computational approaches using multiple data types, including protein domain, and genomic protein structure information have been developed to predict PPIs. However, most of these methods are limited due to difficulty in computing and dependence on a large number of homologous proteins.<sup>7,13,19-21</sup> Therefore, it is very important to identify the close relationship between proteins by exploiting efficient computational approaches based on protein sequence information.

To date, many computational approaches based on protein information have been developed for predicting PPIs.<sup>19,22-24</sup> An et al<sup>25</sup> proposed a novel feature extraction method based on protein sequence called Local Phase Quantization (LPQ), which was combined with RVM classifier to detect PPIs. Xia et al<sup>26</sup> developed a computational approach, which used rotation forest



as a classifier and autocorrelation to represent protein sequences for identifying PPIs. Guo et al<sup>27</sup> exploited a new computational approach by combining autocovariance (AC) with SVM, where AC could take advantage of the interactions between residues a certain distance apart in the sequence. Yu et al<sup>28</sup> proposed a novel feature extraction method called local descriptors (LD), which accounts for the interactions between sequentially distant but spatially close amino acid residues and can fully capture multiple feature information in region in continuous and discontinuous regions within a protein sequence. Huang et al<sup>29</sup> proposed a novel computational model based on protein sequence by using weighted sparse representation as a prediction classifier (WSRC) and employed global encoding (GE) as feature extraction approach for predicting PPIs, which achieved better prediction results. However, there is still room to improve the prediction accuracy of the existing methods.

In this study, we proposed a sequence-based feature extraction method called LCPSSMMF, which combined local coding position-specific scoring matrix (PSSM) with multifeatures fusion. First, we used a novel local coding method based on PSSM to build a new PSSM (CPSSM); the advantage of this method is that it incorporated global and local feature extraction, which can account for the interactions between residues in both continuous and discontinuous regions of amino acid sequences. Second, we adopted 2 different feature extraction methods (Local Average Group [LAG] and Bigram Probability [BP]) to capture multiple key feature information by employing the evolutionary information embedded in the CPSSM matrix. Finally, feature vectors were acquired by using multifeatures fusion method.

## Result

To evaluate the performance of the proposed feature extraction approach, we employed support vector machine (SVM) as a prediction classifier and applied this method to *yeast* and *human* PPI datasets. The prediction accuracies of LCPSSMMF were 93.43% and 90.41% on the *yeast* and *human* datasets, respectively. Moreover, we also compared the proposed method with the previous sequence-based approaches on the *yeast* datasets by using the same SVM classifier. The experimental results indicated that the performance of LCPSSMMF significantly exceeded that of several other state-of-the-art methods. It is proven that the LCPSSMMF approach can capture more local and global discriminatory information than almost all previous methods and can function remarkably well in identifying PPIs. To facilitate extensive research in future proteomics studies, we developed an LCPSSMMFSVM server, which is freely available for academic use at <http://219.219.62.123:8888/LCPSSMMFSVM>.

## Materials

### Dataset

In this study, *yeast* and *human* datasets that can be obtained from the publicly available Database of Interacting Proteins (DIP)

were used to evaluate the proposed method.<sup>30</sup> To better carry out our method, some protein sequence pairs were removed, if they were fragments with less than 50 residues in length. For eliminating bias of homologous protein sequence pairs, sequence pairs with  $\geq 40\%$  sequence identity were considered to be homologous. Thus, these protein sequence pairs were also removed. As a result, we constructed the *yeast* dataset, which contained 5594 positive protein pairs and 5594 negative protein pairs. In the same way, the *human* dataset was constructed, which contained 3899 positive protein pairs and 4262 negative protein pairs. Consequently, the *yeast* dataset contains 11188 protein pairs and the *human* dataset contains 8161 protein pairs.

## Feature Extraction Method

From a computational perspective, the key to identify PPIs is to develop an effective computational approach based on protein sequences, which is usually divided into the following 2 steps:

Step 1. Analyzing and designing a feature extraction method that not only captures protein-protein interaction information but also extracts more key feature information contained in the protein sequence.

Step 2. Designing and selecting an appropriate prediction classifier.

The above-mentioned 2 steps are closely related and complement each other. A drawback of the design process of either step is that bias will be introduced that may influence the performance of the prediction model. The pattern of feature extraction is generally divided into 2 classes: (1) the original protein sequence is directly represented as a feature vector by mathematical description and (2) a given protein sequence is first transformed a matrix; and second, the feature vector is created by mathematical description based on the sequence matrix.

Increasing studies have demonstrated that prediction accuracy of the second class is better than that of the first class. As a result, we presented a new feature extraction method based in this article on the second class.

### Position-specific scoring matrix

Position-specific scoring matrix (PSSM which was originally used to detect distantly related proteins) is a very helpful tool for representing protein sequences as a matrix. Thus, we also transformed each protein sequence into a PSSM by employing position-specific iterated BLAST (PSI-BLAST)<sup>31</sup>:

$$PSSM = \begin{bmatrix} P_{1,1} & P_{1,2} & P_{1,3} & \cdots & P_{1,20} \\ P_{2,1} & P_{2,2} & P_{2,3} & \cdots & P_{2,20} \\ \vdots & P_{i,j} & \vdots & \vdots & \vdots \\ P_{L,1} & P_{L,2} & P_{L,3} & \cdots & P_{L,20} \end{bmatrix}, \quad (1)$$

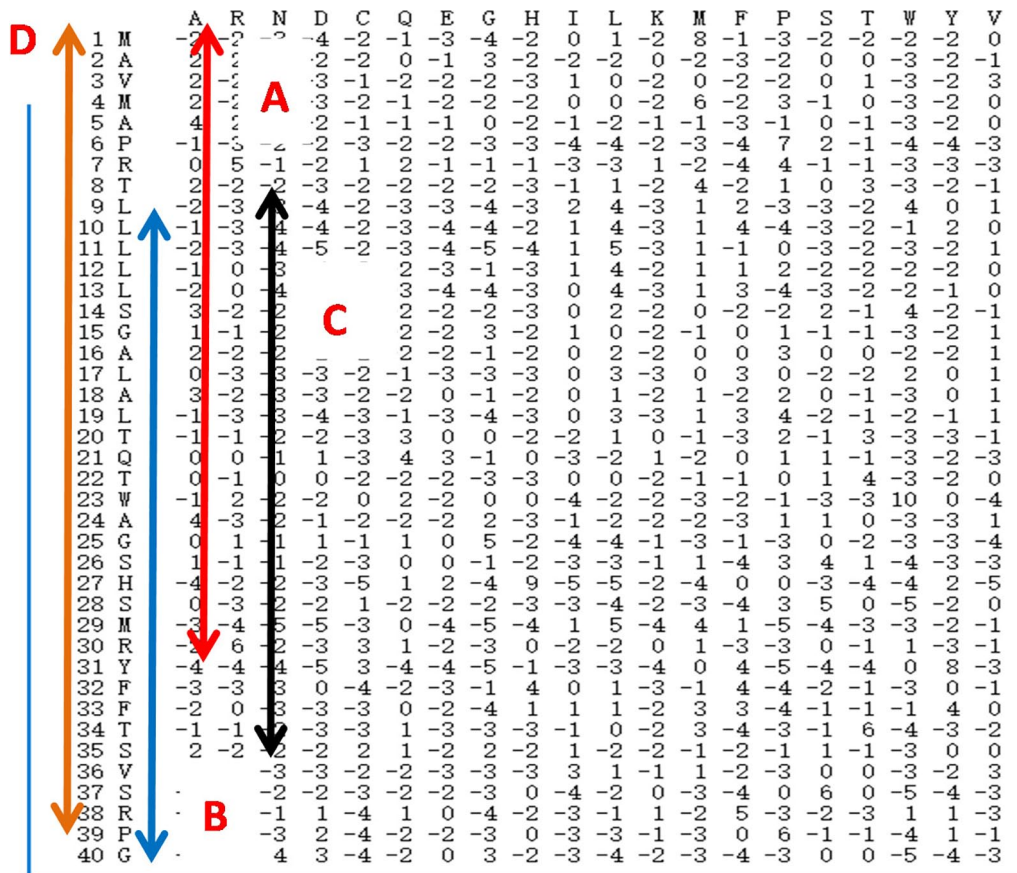


Figure 1. The flow diagram of local coding based on PSSM matrix. PSSM indicates position-specific scoring matrix.

where  $L$  is a given protein sequence length, 20 represent the 20 amino acids, and  $P_{i,j}$  is the probability that the  $i$ th amino acid mutates into the  $j$ th amino acid in the process of biological evolution. Thus, a high score in the PSSM represents a good conservative position, and a low score represents a weak conservative position. The PSSM contains not only protein sequence positional information but also evolutionary information, which can reflect the conservation function of protein sequences. Therefore, this approach is widely applicable transforming a given protein sequence into a PSSM to represent protein sequence characteristics.

In this article, to obtain highly and widely homologous protein sequences, the PSI-BLAST  $e$ -value parameter was set to 0.001, and 3 iterations were selected. Consequently, each protein sequence can be represented as 20-dimensional PSSM, which contain  $L \times 20$  elements, where  $L$  is the length of a given protein sequence and the columns are 20 amino acids.

### Local coding based on PSSM matrix

As is well known, the key characteristics of PPIs are that “PPI usually occur in discontinuous segments in the protein sequence, where distant amino acid residues are brought into spatial proximity by protein folding.” As a result, we developed a novel local coding method based on PSSM; the advantage of the method is that it incorporates both global and local feature

extraction, which can account for the interactions between residues in both continuous and discontinuous regions in amino acid sequences.

The steps for creating the CPSSM matrix are as follows: first, using PSI-BLAST, we constructed each protein sequence PSSM matrix; second, each PSSM was divided into 4 parts: A, B, C, and D. We intercepted the first 75%, the last 75%, and the middle 75% of the PSSM and named these as parts A, B, and C. The whole PSSM matrix is represented in part D. Finally, a new PSSM (CPSSM) is created by merging the A, B, C, and D sub-PSSM in parallel, where the length of the CPSSM is longer than the length of the PSSM. The advantage of the CPSSM matrix is that it can fully account for global and local feature information and the interactions between sequentially distant but spatially close amino acid residues. As a result, the CPSSM matrix can adequately capture multiple overlapping continuous and discontinuous binding patterns within a protein sequence and improve prediction accuracy. Figure 1 shows the procedure of local coding based on PSSM.

### Multi-features fusion

In previous studies, many sequence-based feature extraction methods used a single feature approach, whose drawbacks are that they cannot integrate multiple key feature information-contained protein sequence, and they cannot comprehensively

consider the correlation of the various elements of a protein sequence. Therefore, there is considerable interest in developing a novel feature extraction approach by using multifeatures fusion, which is capable of increasing the quantity of feature vectors information and improving the prediction accuracy. Thus, LAG and BP were used to carry out multifeatures fusion.

*Local Average Group.* The LAG method divided each CPSSM into 20 groups. Each group contained 5% of the length of a CPSSM. Thus, each CPSSM was divided into 20 groups regardless of the length of the protein sequence, where each group consists of 20 features derived from the 20 columns of the CPSSM. The mathematical description is as follows:

$$LAG(k) = \frac{20}{N} \sum_{p=1}^{\frac{N}{20}} Mt \left( p + (i-1) \times \frac{N}{20}, j \right) \quad (2)$$

$$i = 1, \dots, 20; j = 1, \dots, 20; k = j + 20 \times (i-1) \quad (3)$$

where  $N$  represents the length of a CPSSM matrix and  $20/N$  represents 5% of the length of a CPSSM, or specifically, the length of the  $j$ th group.  $Mt(p + (i-1) \times N/20, j)$  represents a  $1 \times 20$  vector, which can be captured from a CPSSM at the  $i$ th position in the  $j$ th group. As a result, we divided each CPSSM into 20 groups and expressed each CPSSM as a 400-dimensional vector. The main advantage of the LAG method is that the residue conservation tendencies in the same domain family are similar and the locations of domains in the same family are closely related to the length of the sequence. Thus, the LAG method transformed each CPSSM into a 400-dimensional feature vector.

*Bigram Probabilities.* N-gram models are usually employed to estimate the probability of a random sequence,<sup>32,33</sup> such as the random sequence in the following:

$$S = S_{j-n+1} S_{j-n+2} \dots S_{j-1} S_j \quad (4)$$

$$P(S) = P(S_j | S_{j-n+1} S_{j-n+2} \dots S_{j-1}) \dots P(S_{j-n+2} | S_{j-n+1}) P(S_{j-n+1}) \quad (5)$$

where  $P(S_j | S_{j-n+1} S_{j-n+2} \dots S_{j-1})$  represents Conditional Probability, which reflects the correlation of  $n$  continuous random variables. N-gram models are statistical models, which have been widely used in natural language processing. Common languages such as words and syllables are treated as random variables. The element value of the PSSM indicates the probability of an amino acid mutating into another amino acid in the evolutionary process, and the probability values of the 20 amino acids are correlated. Therefore, we adopted binary bigram mode to extract the key evolutionary feature information contained in the PSSM. The relevant mathematical description is as follows:

$$BP_{mn} = \sum_{i=1}^{L-1} P_{i,m} P_{i+1,n} \quad 1 \leq m \leq 20, 1 \leq n \leq 20 \quad (6)$$

where  $L$  represents the number of the CPSSM row. A CPSSM element  $P_{i,m}$  represents the relative probability that the  $i$ th amino acid mutates into the  $j$ th amino acid. The matrix  $BP_{mn}$  contain 400 elements,

$$BPF = \begin{bmatrix} BP_{1,1}, BP_{1,2} \dots BP_{1,20}, BP_{2,1}, \dots, BP_{2,20}, \dots \\ BP_{20,1}, \dots, BP_{20,20} \end{bmatrix} \quad (7)$$

$$BPF = [\varphi_1, \varphi_2, \varphi_3, \dots, \varphi_u, \dots, \varphi_\theta]$$

where  $\theta = mn = 400$  is the dimensionality of the feature vector BF. The  $\varphi_u$  can be represented as follows:

$$\varphi_u = \begin{cases} BP_{1,u} & (1 \leq u \leq 20) \\ BP_{2,u-20} & (21 \leq u \leq 40) \\ \dots & \\ BP_{20,u-380} & (381 \leq u \leq 400) \end{cases} \quad (8)$$

Finally, each CPSSM of a given protein sequence was converted into a 400-dimensional vector using the BP feature extraction method.

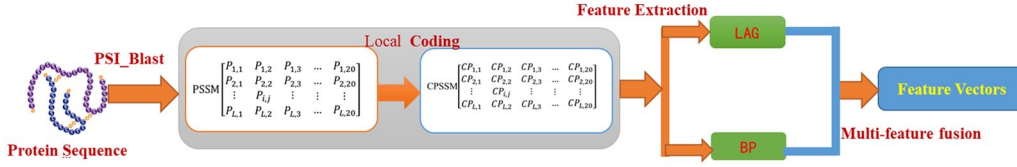
As a result, based on the above analysis, we proposed a new sequence-based feature extraction method by combining local coding based on PSSM and multifeatures fusion in this study. The proposed method fully considers the following key points:

1. Capturing protein interaction information by using local coding based on PSSM;
2. Increasing the amount of feature information through employing LAG and BP to carry out multifeatures fusion.

In this article, we proposed local coding method based on PSSM; the highlight of the method incorporates global and local feature extraction, which could account for the interactions between residues in both continuous and discontinuous regions of amino acid sequences. Using multifeatures fusion integrates multiple key pieces of feature information contained in protein sequences and comprehensively considers the relevant information of each feature element. The flow diagram of our feature extraction algorithm is shown in Figure 2

### Performance evaluation

For evaluating the effectiveness of LCPSSMMF, 4 parameters, including sensitivity, accuracy, Matthews correlation coefficient (MCC), and precision, were calculated. The mathematical descriptions are as follows:



**Figure 2.** The flow diagram of our feature extraction algorithm. BP indicates Bigram Probability; LAG, Local Average Group.

**Table 1.** The abbreviations of different feature extraction methods.

FEATURE EXTRACTION METHOD	ABBREVIATION
Multifeatures fusion based on local coding PSSM matrix	LCPSSMMF
Multifeatures fusion based on original protein sequence PSSM matrix	PSSMMF
Local Average Group based on local coding PSSM matrix	LCPSSMLAG
Bigram Probabilities based on local coding PSSM matrix	LCPSSMBP

Abbreviations: LCPSSMMF, local coding position-specific scoring matrix with multifeatures fusion; PSSM, position-specific scoring matrix; LAG, Local Average Group; BP, Bigram Probabilities.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (9)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (10)$$

$$Precision = \frac{TP}{FP + TP} \quad (11)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \quad (12)$$

where  $FP$  represents false positives,  $TP$  represents true positives,  $FN$  represents false negatives, and  $TN$  represents true negatives. True positives represent the number of true interacting pairs that were correctly predicted. True negatives represent the count of true noninteracting pairs that were predicted correctly. False positives represented the number of noninteracting pairs that were falsely predicted, and false negatives represented true interacting pairs that were falsely predicted to be noninteracting pairs. Moreover, we drew a receiver operating characteristic (ROC) curve to further evaluate the effectiveness of our method.

## Results and Discussion

### Performance of the proposed method

For evaluating the efficiency of our feature extraction approach, we compared it with other feature extraction approaches, as shown in Table 1.

In Table 1, PSSM, LC, and MF represent PSSM, local coding based on PSSM, and multifeatures fusion, respectively.

LAG and BP represent LAG and BP feature extraction method, respectively.

The SVM classifier has been widely used to predict PPIs. As a result, for ensuring fairness, different feature extraction methods used the same SVM classifier on *human* and *yeast* datasets in the experiment. In addition, to avoid over fitting, we divided the datasets into training sets and independent test sets. Specifically, 4out of 5 of the datasets were selected as training sets and the remaining datasets were selected as independent test sets. Simultaneously, 5-fold cross validation was employed to benchmark the effectiveness of our feature extraction method. The LIBSVM tool<sup>34</sup> was used to carry out classification in the experiment. The radial basis function (RBF) kernel parameters of the SVM were optimized by using the grid search method, where  $c$  is 0.001 and  $g$  is 0.3, and other parameters were set to the default value.

It can be seen from Table 2 that the LCPSSMMF obtained an average prediction accuracy, sensitivity, precision, and MCC of 93.14%, 92.50%, 93.90%, and 87.41%, respectively, using the *yeast* dataset. As shown in Tables 3 to 5, the PSSMMF, LCPSSMLAG, and LCPSSMBP achieved an average prediction accuracy, sensitivity, precision, and MCC of 90.48%, 90.26%, 90.58%, and 82.84%; 87.09%, 86.57%, 87.13%, and 77.33%; and 88.83%, 86.22%, 87.67%, and 81.34% using the *yeast* dataset, respectively. Similarly, Table 6 shows that the LCPSSMMF obtained an average accuracy, sensitivity, precision, and MCC of 90.41%, 93.54%, 88.02%, and 82.62% using the *human* dataset. It can be seen from Tables 7 to 9 that the PSSMMF, LCPSSMAB, and LCPSSMBG obtained an average accuracy of 87.58%, 85.86%, and 86.42%; average sensitivity of 87.69%, 84.89%, and 86.03%; average precision of 87.24%, 86.58%, and 87.10%; and average MCC of 77.84%, 75.72%, and 76.85%, respectively, on the *human* dataset. Similarly, as shown in Figures 3 and 4, the ROC curves of LCPSSMMF are also

**Table 2.** The experimental results of the LCPSSMMF method on the yeast dataset.

TESTING TIMES	ACCURACY (%)	SENSITIVITY (%)	PRECISION (%)	MCC (%)
1	94.05	94.14	93.89	88.82
2	93.07	91.93	93.74	87.09
3	93.02	93.02	94.19	87.95
4	92.40	91.26	93.42	85.95
5	93.17	92.19	94.26	87.27
Average	93.14 ± 0.60	92.50 ± 1.1	93.90 ± 0.34	87.41 ± 1.06

Abbreviations: LCPSSMMF, local coding position-specific scoring matrix with multifeatures fusion; MCC, Matthews correlation coefficient.

**Table 3.** The experimental results of the PSSMMF method on the yeast dataset.

TESTING TIMES	ACCURACY (%)	SENSITIVITY (%)	PRECISION (%)	MCC (%)
1	90.75	90.81	90.57	83.20
2	89.85	88.82	90.22	81.74
3	89.58	90.37	89.19	81.33
4	91.46	91.08	91.82	84.38
5	90.76	90.23	91.12	83.55
Average	90.48 ± 0.76	90.26 ± 0.87	90.58 ± 0.98	82.84 ± 1.27

Abbreviation: MCC, Matthews correlation coefficient; PSSMMF, position-specific scoring matrix with multifeatures fusion.

**Table 4.** The experimental results of the LCPSSMAB method on the yeast dataset.

TESTING TIMES	ACCURACY (%)	SENSITIVITY (%)	PRECISION (%)	MCC (%)
1	86.14	86.50	86.35	76.11
2	86.59	85.88	87.57	76.77
3	88.29	87.05	88.98	79.31
4	86.46	86.94	86.09	76.58
5	87.98	86.52	86.67	77.88
Average	87.09 ± 0.97	86.57 ± 0.46	87.13 ± 1.17	77.33 ± 1.28

Abbreviation: MCC, Matthews correlation coefficient.

**Table 5.** The experimental results of the LCPSSMBG method on the yeast dataset.

TESTING TIMES	ACCURACY (%)	SENSITIVITY (%)	PRECISION (%)	MCC (%)
1	88.41	85.32	92.75	80.98
2	87.78	84.78	91.58	79.98
3	89.16	85.78	92.65	80.95
4	88.76	87.15	92.00	81.60
5	90.06	88.07	93.40	83.21
Average	88.83 ± 0.85	86.22 ± 1.36	92.47 ± 0.71	81.34 ± 1.19

Abbreviation: MCC, Matthews correlation coefficient.

**Table 6.** The experimental results of the LCPSSMMF method on the *human* dataset.

TESTING TIMES	ACCURACY (%)	SENSITIVITY (%)	PRECISION (%)	MCC (%)
1	89.42	90.96	88.11	81.07
2	90.89	92.74	89.82	83.41
3	90.25	93.94	88.06	82.29
4	89.87	94.16	86.15	81.75
5	91.61	95.94	88.00	84.60
Average	90.40 ± 0.86	93.54 ± 1.84	88.03 ± 1.30	82.62 ± 1.40

Abbreviation: LCPSSMMF, local coding position-specific scoring matrix with multifeatures fusion; MCC, Matthews correlation coefficient.

**Table 7.** The experimental results of the PSSMMF method on the *human* dataset.

TESTING TIMES	ACCURACY (%)	SENSITIVITY (%)	PRECISION (%)	MCC (%)
1	86.72	84.62	88.35	76.95
2	87.11	88.99	85.04	77.53
3	88.49	87.40	87.29	78.11
4	87.36	88.15	87.61	77.88
5	88.26	89.30	87.91	78.74
Average	87.58 ± 0.75	87.69 ± 1.90	87.24 ± 1.56	77.84 ± 0.67

Abbreviations: PSSMMF, position-specific scoring matrix with multifeatures fusion; MCC, Matthews correlation coefficient.

**Table 8.** The experimental results of the LCPSSMAB method on the *human* dataset.

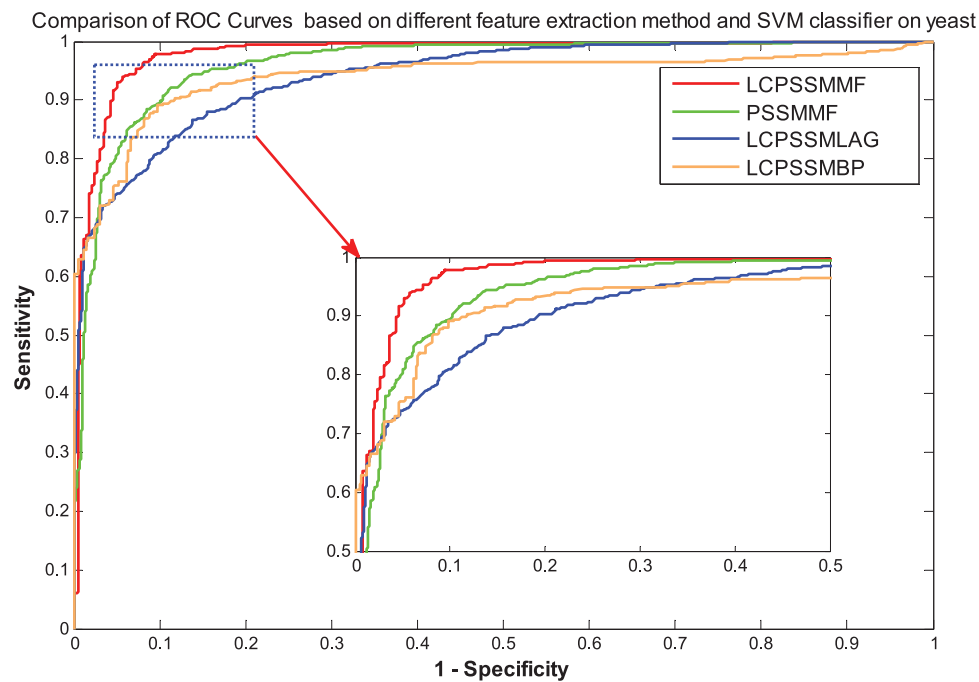
TESTING TIMES	ACCURACY (%)	SENSITIVITY (%)	PRECISION (%)	MCC (%)
1	85.63	84.52	86.30	75.38
2	85.18	83.01	86.69	74.73
3	85.31	85.37	85.48	74.94
4	86.91	85.53	87.37	77.22
5	86.30	86.02	87.10	76.34
Average	85.86 ± 0.72	84.89 ± 1.18	86.59 ± 0.75	75.72 ± 1.04

Abbreviation: MCC, Matthews correlation coefficient.

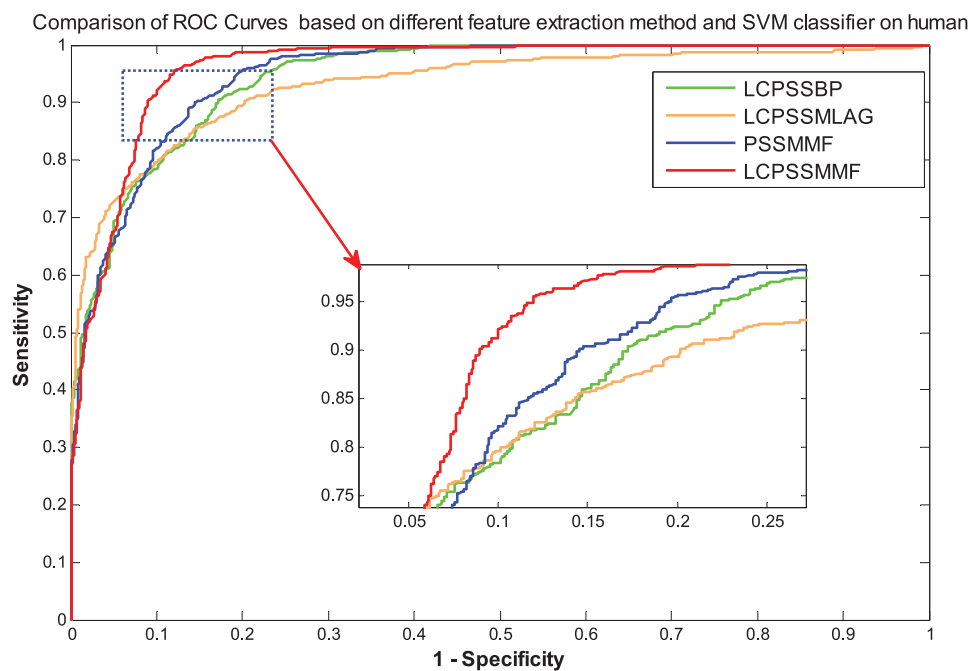
**Table 9.** The experimental results of the LCPSSMBG method on the *human* dataset.

TESTING TIMES	ACCURACY (%)	SENSITIVITY (%)	PRECISION (%)	MCC (%)
1	85.21	84.92	87.05	76.21
2	84.67	83.46	86.75	74.01
3	86.47	87.58	85.24	76.59
4	86.98	85.46	88.06	77.34
5	88.80	88.76	88.41	80.10
Average	86.42 ± 1.62	86.03 ± 2.21	87.10 ± 1.24	76.85 ± 2.20

Abbreviation: MCC, Matthews correlation coefficient.



**Figure 3.** Comparison of ROC curves based on different feature extraction methods and SVM classifier on the *yeast* dataset. ROC indicates receiver operating characteristic; SVM, support vector machine.



**Figure 4.** Comparison of ROC curves based on different feature extraction method and SVM classifier on the *human* dataset. LCPSSMMF indicates local coding position-specific scoring matrix (PSSM) with multifeatures fusion; ROC, receiver operating characteristic; SVM, support vector machine.

significantly better than those of PSSMMF, LCPSSMLAG, and LCPSSMBP. As a result, these experimental results demonstrated that the predictive capability of LCPSSMMF is superior to other methods. This result clearly demonstrated that LCPSSMMF is an effective feature extraction method for predicting PPIs. The success of LCPSSMMF can be attributed to the following several factories: (1) we exploited local coding based on PSSM matrix, which

incorporates global and local feature extraction, thus accounting for the interactions between residues in both continuous and discontinuous regions of amino acid sequences; (2) serial multifeatures fusion can integrate multiple key feature information contained in the sequence and comprehensively consider the relevant information of each element in the sequence; and (3) the LAG method based on the residue conservation tendencies in the same domain



**Table 10.** The prediction results of different feature extraction methods on the *yeast* dataset.

METHODS	ACCURACY (%)	SENSITIVITY (%)	PRECISION (%)
LCPSSMMF	93.14	92.50	93.90
AC <sup>27</sup>	87.36	87.30	87.82
ACC <sup>27</sup>	89.33	89.93	88.87
GE <sup>29</sup>	91.73	85.05	97.05
LD <sup>28</sup>	88.56	87.37	89.50

Abbreviations: AC, auto covariance; LCPSSMMF, local coding position-specific scoring matrix with multifeatures fusion; LD, local descriptors; ACC, auto cross covariance; GE, global encoding.

family is similar, and the locations of the domains in the same family are closely related to the length of the sequence. The BP approach represented each protein sequence by its PSSM and calculated the Bigram feature using the probability information contained in PSSM. This approach can significantly reduce the sparsity level and helps to improve recognition performance. Thus, information can be effectively captured from the new PSSMs by using the LAG and BP feature extraction methods. Therefore, the efficiency of the LCPSSMMF approach is clearly superior to other feature extraction methods.

#### *Comparison of SVM based on other feature extraction methods*

Meanwhile, to further evaluate the effectiveness of LCPSSMMF, we compared the prediction capability of LCPSSMMF with that of existing methods by using the same SVM classifier on the *yeast* dataset. It is shown in Table 10 that 4 different approaches obtained average prediction accuracy between 87.36% and 91.73%, which is lower than that of the proposed LCPSSMMF method. Similarly, the sensitivity and precision of LCPSSMMF are also superior to other approaches. It is obvious from these experimental results that the proposed LCPSSMMF feature extraction method yielded significantly better prediction results than other existing approaches. All these results indicated that the LCPSSMMF can improve the prediction accuracy relative to current state-of-the-art methods.

## Conclusions

In this study, we proposed a sequence-based feature extraction method called LCPSSMMF, which combined local coding based on PSSM with multifeatures fusion. First, we used a novel local coding method based on PSSM to build a new PSSM (CPSSM), which incorporates global and local feature extraction to account for the interactions between residues in both continuous and discontinuous regions of amino acid sequences. Second, we adopted 2 different feature extraction methods (LAG and BP) to capture multiple key feature information by using the evolutionary information embedded in

CPSSM. Finally, feature vectors were acquired by using the multifeatures fusion method.

The experimental results proved that the predictive capability of the LCPSSMMF is superior to that of other methods. The success of LCPSSMMF can be attributed to the following reasons: (1) we developed local coding based on PSSM, which can fully account for global and local feature information and the interactions between sequentially distant but spatially close amino acid residues. As a result, this method can adequately capture multiple overlapping continuous and discontinuous binding patterns within a protein sequence and improve prediction accuracy. (2) Serial multifeatures fusion can integrate multiple key feature information contained in the sequence and comprehensively consider the relevant information of each element in the sequence. (3) The LAG method based on the residue conservation tendencies in the same domain family is similar and the locations of domains in the same family are closely related to the length of the sequence. The BP approach represented each protein sequence by its PSSM and calculated the Bigram feature using the probability information contained in PSSM, which can significantly reduce the sparsity level and improve the recognition performance. Thus, information can be effectively captured from the new PSSMs by using the LAG and BP feature extraction methods. Therefore, the efficiency of LCPSSMMF approach is obviously superior to other feature extraction methods. This study clearly demonstrated that the LCPSSMMF is an effective feature extraction method for predicting PPIs.

## Acknowledgements

The authors would like to thank all the guest editors and anonymous reviewers for their constructive advice.

## Author Contributions

Ji-Yong An and Zhou You conceived the algorithm, carried out analyses, prepared the data sets, carried out experiments, and wrote the manuscript; Yu-Jun Zhao and Zi-Ji Yan designed, performed, and analyzed experiments and wrote the manuscript; and all authors read and approved the final manuscript.

## Availability of data and material

In this study, our experimental datasets contain *yeast* and *human* data, which can be obtained from the publicly available DIP.<sup>30</sup>

## ORCID iD

Ji-Yong An  <https://orcid.org/0000-0001-9546-3654>

## REFERENCES

- Gavin AC, Bösch M, Krause R, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *J Econ Sur.* 2002;415:141-147.
- Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A.* 2001;98:4569-4574.
- Wong JH, Alfatah M, Sin MF, et al. A yeast two-hybrid system for the screening and characterization of small-molecule inhibitors of protein-protein interactions identifies a novel putative Mdm2-binding site in p53. *BMC Biol.* 2017;15:108.
- Zhang XF, Ou-Yang L, Hu X, Dai DQ. Identifying binary protein-protein interactions from affinity purification mass spectrometry data. *BMC Genomics.* 2015;16:745-714.
- Ho Y, Gruhler A, Heilbut A, et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature.* 2002;415:180-183.
- Palopoli N, Edwards R. Large-scale prediction of short linear motifs using structural information from protein-protein interactions. *BMC Bioinformatics.* 2015;11:1-9.
- Foltman M, Sanchez-Diaz A. Studying protein-protein interactions in budding yeast using co-immunoprecipitation. *Methods Mol Biol.* 2016;1369:239-256.
- Zhu H, Bilgin M, Bangham R, et al. Global analysis of protein activities using proteome chips. *Science.* 2001;293:2101-2105.
- Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. *Nature Biotech.* 2000;18:1257-1261.
- Rain JC, Selig L, De Reuse H, et al. The protein-protein interaction map of *Helicobacter pylori*. *Nature.* 2001;409:211-215.
- Dalvit C, Caronni D, Mongelli N, Veronesi M, Vulpetti A. NMR-based quality control approach for the identification of false positives and false negatives in high throughput screening. *Curr Drug Discov Technol.* 2006;3:115-124.
- An JY, Zhou Y, Zhang L, Niu Q, Wang DF. Improving self-interacting proteins prediction accuracy using protein evolutionary information and weighed-extreme learning machine. *Current Bioinform.* 2019;14:115-122.
- Jia J, Xiao X, Liu B. Prediction of protein-protein interactions with physico-chemical descriptors and wavelet transform via random forests. *J Lab Autom.* 2015;21:368-377.
- Smits AH, Vermeulen M. Characterizing protein-protein interactions using mass spectrometry: challenges and opportunities. *Trends Biotechnol.* 2016;34:825-834.
- Deng Y, Gao L, Wang B. ppiPre: predicting protein-protein interactions by combining heterogeneous features. *BMC Syst Biol.* 2013;7:S8.
- Hue M, Riffle M, Vert JP, Noble WS. Large-scale prediction of protein-protein interactions from structures. *BMC Bioinform.* 2010;11:144-149.
- Lu L, Lu HJ. Multiprospector: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Prot Struct Func Bioinform.* 2010;49:350-364.
- Chen Y, Xu D. Computational analyses of high-throughput protein-protein interaction data. *Curr Protein Pept Sci.* 2003;4:159-181.
- An JY, Zhang L, Zhou Y, Zhao YJ, Wang DF. Computational methods using weighed-extreme learning machine to predict protein self-interactions with protein evolutionary information. *J Cheminform.* 2017;9:47.
- Keskin O, Tuncbag N, Gursoy A. Predicting protein-protein interactions from the molecular to the proteome level. *Chem Rev.* 2016;116:4884-4909.
- You ZH, Li X, Chan KC. An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers. *Neurocomputing.* 2016;228:277-282.
- Zahiri J, Yaghoubi O, Mohammad-Noori M, Ebrahimpour R, Masoudi-Nejad A. PPLevo: protein-protein interaction prediction from PSSM based evolutionary information. *Genomics.* 2013;102:237-242.
- Zhang S. Accurate prediction of protein structural classes by incorporating PSSS and PSSM into Chou's general PseAAC. *Chem Intell Lab Syst* 2015;142:28-35.
- Zahiri J, Mohammad-Noori M, Ebrahimpour R, et al. LocFuse: human protein-protein interaction prediction via classifier fusion using protein localization information. *Genomics.* 2014;104:496-503.
- An JY, Meng FR, You ZH, Fang YH, Zhao YJ, Zhang M. Using the relevance vector machine model combined with local phase quantization to predict protein-protein interactions from protein sequences. *Biomed Res Int.* 2016;2016:4783801.
- Xia JF, Han K, Huang DS. Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor. *Protein Pept Lett.* 2010;17:137-145.
- 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;36:3025-3030.
- Yu ZZ, Yun G, Ying YZ. Prediction of protein-protein interactions using local description of amino acid sequence. *Commun Comput Inform Sci.* 2011;202:254-262.
- Huang YA, You ZH, Chen X, Chan K, Luo X. Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding. *BMC Bioinform.* 2016;17:184.
- Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 2002;30:303-305.
- Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem Sci.* 1998;23:444-447.
- Goldberg Y. A primer on neural network models for natural language processing. *Comput Sci.* 2016, <https://arxiv.org/abs/1510.00726>.
- Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform.* 2003;36:462-477.
- Chang CC, Lin CJ. LIBSVM—a library for support vector machines, 2011, <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>.