World Scientific Research Journal Volume 7 Issue 12, 2021
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202112_7(12).0025

Identification of Protein-protein Interactions Based on Weighted Sparse Representation

Fanshu Liu

Institute of statistics, Shandong Institute of Business and Technology, Shandong, 264003, China

Abstract

Protein plays an important role in the cellular process of an organism, and its function is demonstrated by protein interaction. Rich information on protein interactions can facilitate the treatment of diseases and the development of drugs, so accurate prediction of protein interactions is of great significance. High-flux biological experiments can be used to predict new protein pairs, but they are expensive and time-consuming to operate and do not meet the demand for such information. With the rise of machine learning algorithms and the increasingly powerful computing power, the use of scientific computing models to predict each other has become the first choice. This paper mainly studies the application of weighted sparse representation classifiers under protein sequence feature coding. First of all, the composition, transfer and distribution of the physical and chemical properties of amino acids are selected to encode the amino acid sequence. Secondly, according to the characteristic importance of random forest, the feature operator de-dimensionally de-noises. Finally, for the features extracted in this paper, a weighted sparse representation classifier with strong noise resistance is used to classify the feature set. The results of the 50% cross-validation were: accuracy 96.97%, sensitivity 97.51%, accuracy 96.43%, Matthews correlation coefficient 93.91%, Predictive results are better than existing machine learning models.

Keywords

Protein interaction; Weighted sparse expression classifier; Dimension reduction; Feature extraction.

1. INTRODUCTION

In recent years, with the advancement of the Human Genome Project, life sciences have ushered in a new era. With the huge amount of biological data that comes with it, researchers hope to use data mining technology to obtain research-meaningful information and further reveal the nature of life activities. Protein, as an important unit of biological cells, is the carrier and function performer of the main life activities, and participates in the transport and storage of metabolites in the body, the biochemical reaction of the birth, and the regulation of cellular processes and activate immune functions and other life processes. However, protein function is not performed independently by a single protein, but by protein-to-protein interactions (Protein-Protein Interactions, PPI). [1, 2] The interaction between proteins and the protein complex formed by interaction are the main finishers of various basic functions of cells. Almost all important life activities, including DNA replication and transcription, protein synthesis and secretion, signaling and metabolism, and so on, depend on the interaction between proteins [3,4]. Therefore, predicting protein interactions has become a challenging and practical hot issue in the field of biological information.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202112_7(12).0025

Genes are units with genetic functions, and their expression is made up of proteins, which determine the amino acid sequence of proteins. An organism's protein is a specific configuration that folds and curls inthree-dimensional space after 20 amino acids are polymerized by dehydration to form an amino acid sequence (protein sequence). It is also because of the differences between amino acid sequences and the physical interaction between amino acids and spatial mutual support that protein diversity and stability are formed. Protein data is divided into four types: primary structure, secondary structure, third-level structure and fourth-level structure, in which the protein sequence to be explored in this paper is a primary structure, which is easier to represent than other structures in sequence coding. Therefore, it is theoretically feasible to predict protein interaction through protein sequence.

Traditional methods of identifying whether proteins interact are obtained through biological experiments, but this method of identification is time-cost, usually only a few pairs of proteins can be identified in a single experiment, the speed of identification is much lower than the speed of protein sequence discovery. At the same time, due to some artificial factors caused by a number of errors and lack of interaction, to a large extent will interfere with the relevant downstream work. Therefore, it is urgent to develop and design algorithms with high performance to predict protein interaction. Scholars in this field have found that machine learning-based algorithms can be effectively combined with protein sequence information, and the accuracy of qualitative leaps, reducing waste of resources, which will be conducive to the growth of real economic benefits, accelerate scientific research output, and thus promote the rapid development of biology.

Accurate prediction of protein interactions can tap the function of unknown proteins and provide guidance to biological species that have not been experimentally verified. At the same time, the prediction results of computer methods can also be used as verification and supplement to the results of biological experiments. In practical sense, the research of this subject is of indispensable significance for understanding the inner tissue of life, and it is of great application value to the treatment of diseases and the development of drugs.

At present, the methods of detecting whether proteins interact are mainly divided into experimental methods based on biological information and digital calculation methods. The biological experimental methods for predicting protein interaction include yeast double hybridization system, series affinity and purification, mass spectrometry, immunocoprecipitation technology, etc. But the experimental method is flawed, so scholars turn their attention to the scientific calculation method.

Protein sequence information prediction is mutual. As the basic unit of protein formation, amino acid sequence information is rich in resources and easy to obtain, widely used in the field of predicting protein interaction and has a high recognition rate, which has been favored by scholars. According to the principle of data mining, the recognition rate can be improved from two angles, one is to extract valuable information in the amino acid sequence to build feature vectors, and the other is to optimize the pattern recognition algorithm or machine learning algorithm.

The feature vector is constructed from different angles of amino acid sequence information, and the physical and chemical properties of amino acids are the scientific basis for coding the following methods. For example, as early as 2007, Shen et al. [5] proposed a tripart union feature coding method that takes into account the influence of the left and right neighbors of amino acids on themselves with an accuracy of 83.9%. However, this coding ignores the effects of intermediate and long-range amino acids on itself, so Guo et al. [6] proposed a self-co-anovating coding algorithm, selecting the width of the sliding window to be 30, using support vector machine prediction, with an accuracy of 86.5%. You et al. [7] proposed a multi-scale continuous and discontinuous feature coding algorithm, which recombines the sequence

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202112_7(12).0025

segments, calculates the physical and chemical composition, transformation and distribution of each subsequence, and finally stitches them together. The feature dimension of this construction is too high, so the feature selection method is adopted to remove redundancy, and after this processing, the support vector machine is trained to achieve 91.63% prediction accuracy. In order to give full play to the potential effects of different sequence coding algorithms, Zhang et al. [8] proposed a classification algorithm based on integrated deep learning and integrated protein sequence coding, and configured the corresponding deep learning network for each feature coding method, with a prediction accuracy rate of 95.29%. In order to depict multiple layers of information in the protein sequence, Chen et al.[9] designed the StackPPI Stack Integrated Classifier, which integrates random forests, extreme random trees, and logistic regression with accuracy 96.64%.

While the amino acid sequence feature coding algorithm was developed, the machine learning algorithm for the second classification was widely used to predict each other of proteins. Representative are: support vector machine, Bayesian classifier and neural network, etc, but these methods have some disadvantages. The support vector machine has the absolute defect of hard interval classification, the Bayesian classifier is influenced by the small sample size, which leads to insufficient prior information, and the neural network output method has the problem of network structure determination. Therefore, there is still room for exploration in model design. Weighted sparse mean classifiers overcome the above problems to a certain extent and are worth studying.

2. PROTEINS SEQUENCE INFORMATION EXTRACTION

2.1. Data Preprocessing

In this paper, two highly representative protein interaction data sets are selected: Yeast and Human data in the DIP database. As a positive sample in an experiment, there is a lot of redundancy between them, which can affect the performance of the classifier, so the data set needs to be filtered. First, remove samples with a sequence length of less than 50 in the dataset, and then use cd-hit to cluster the sequence, preserving a protein sequence with homogeneity of less than 40%. After processing, the Yeast dataset contains 5,594 pairs of interacting proteins, and theHuman dataset contains 3,899 pairs of interacting proteins.

Since non-interactive data is not easy to obtain, it is necessary to manually build high-quality negative samples in order to train classifier performance. Research shows that [10], Proteins located in different subcellular locations do not interact. Based on this, the strategy of randomly selecting protein pairings of different subcellular positionings from positive samples to generate negative samples must be met, and the conditions of no repetition of positive and negative samples and equalization of the sample size of opposing sample sets must be met. In this way, the Yeast dataset contains 5594 negative samples, and the Human dataset contains 4262 negative samples. Ultimately, the Yeast dataset contains 11188 pairs of protein sequences, and the Human dataset contains 8,161 pairs of protein sequences.

2.2. Protein Sequence Feature Extraction Method

In order to better adapt to machine learning models, extracting the right characteristics from protein sequence information is a critical step. These characteristics should reflect most of the key information expressed in protein sequences from multiple perspectives.

Protein sequences are made up of 20 amino acids, and the physical and chemical properties of each amino acid closely influence protein interactions. At the same time, 20 categories can lead to feature dimension disasters and redundant information, so amino acids are classified into 3 categories based on information about the 13 physical and chemical properties of amino

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202112_7(12).0025

acids [11], and only a representative classification of 6 physical and chemical properties is listed here. See Table 1 for details.

Table 1. Classification of the physical and chemical properties of amino acids

properties	Category 1	Category 2	Category 3
Hydrophobicity	RKEDQN	GASTPHY	CLVIMFW
Vanderwaard volumn	GASTPDC	NVEQIL	MHKFRYW
Polarity	LIFWCMVY	PATGS	HQRKNE
Polarization rate	GASDT	CPNVEQIL	KMHFRYW
Electric charge	KR	ANCQGHILMFPSTWYV	DE
Secondary structure	EALMQKRH	VIYCWFT	GNPSD

CTD method is a characteristic representation method that describes the composition, order and distribution of amino acid residues in the global sequence. The author gives the following example, in accordance with the above classification guidelines, 'MQRPGPRLWLVLQVMGSCAAISSMDMERP' is expressed as '122333211111211331331331212 23'. Amino acid Composition can be achieved through the following calculation process:

$$Composition(r) = \frac{L(r)}{L}, r \in \{P, N, H\}$$

L(r) in this formula represents the proportion of the class r to which the amino acid belongs in the sequence. The protein sequence can be observed to contain 12 H, 10 N, and 7 P. L represents the length of the protein sequence, so the three amino acid composition characteristics are constructed as:12/(12+10+7)=0.4138, 10/(12+10+7)=0.3448, 7/(12+10+7)=0.2414.

The Transition between aminoacids can be achieved through the following calculation process:

$$Transition(r) = \frac{L(r,s) + L(s,r)}{L-1}, r \in \{(P, N), (N, H), (H, P)\}$$

Where L(r,s) represents the proportion of conversions from r to s and s to r to the categories to which amino acids belong. The H to P or P to H ratio in the sequence is 7/28=0.25, The same conversion N to P or P to N ratio and the conversion N to H or H to N. The percentages are 3/28=0.107,5/28=0.1768.

Feature D represents the distribution of each type of amino acid residue in the sequence distribution position (Distribution), mainly calculating the distribution of each class of amino acids in the total amount of 1%, 25%, 50%, 75%, 100% of the total amino acid distribution throughout the sequence. The sequence contains 10 N-residuals, calculating the first, the second (25% ×10≈2), the third (50% × 10≈5), the fourth (75% ×10≈7), and the fifth The position of the base (10) in the amino acid sequence, and finally in the sequence is 4,5,15,17,21, so the characteristics of D extraction are: (9.09%, 13.64%, 45.45%, 63.64%, 95.45%).

Finally, for each amino acid sequence, three descriptors (C, T, D) are calculated in the same way and stitched together as a 273-dimensional feature vector with a feature dimension of 39 for C, 39 for T, and 195 for D. The two inter-made or non-interoperability protein sequence pairs are then combined to obtain a 546-dimensional feature vector.

2.3. Feature Selection

In general, too high a feature dimension can result in an increase in model calculations, and some features contribute little or no to the model. Therefore, it is necessary to use some

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202112_7(12).0025

methods to filter out redundant features and reduce feature space, which is beneficial to reduce unnecessary waste of resources, reduce time costs, and improve the prediction accuracy of the model. Feature selection is a method to remove irrelevant features, selecting the optimal subset of features from the original feature set, so that the feature selection can achieve similar or better results than before. The feature selection method based on random forest is an efficient dimensional reduction tool.

The Random Forest Algorithm [12] is integrated by multiple decision trees. Its randomness is reflected in the randomness of the training set and the randomness of the selection of candidate separation features, which create the diversity of random forests. First, random forests generate multiple training sets using the bootstrap method. Then, for each training set, construct a decision tree. Finally, when the node selects the feature split, it randomly selects a part of the feature, finds the optimal feature in the pumped feature, applies it to the node, and divides it.

The process of feature selection is actually to sort the importance of features. The merits of a feature depend on how relevant it is to the classification label. In this paper, the characteristics are ranked according to the GINI coefficient according to the classification accuracy criterion function of random forests. Set the set T contains a sample of K categories, then the GINI coefficient calculation formula of T is:

$$GINI(T) = 1 - \sum_{j=1}^{K} P_j^2$$

Where P_j is the probability of a Class j sample appearing. After a split, T is divided into m parts, at which point the GINI coefficient is:

$$GINI_{split}(T) = \frac{K_1}{K} GINI(T_1) + \cdots + \frac{K_m}{K} GINI(T_m)$$

The smaller the GINI coefficient, the higher the score on the feature. By setting a threshold, select features that contribute more to the model than the threshold as the final feature set. The reorganization process for the feature vector in this article is shown in Figure 1.

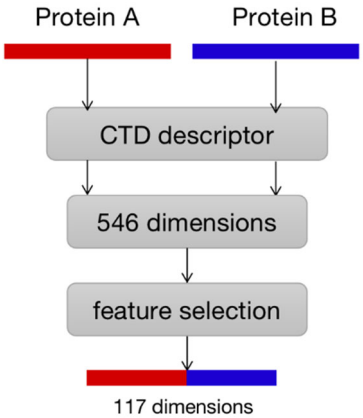

Figure 1. Feature reorganization flowchart

3. PREDICTIVE MODELS AND EVALUATION METHODS

3.1. Weighted Sparse Representation Based Classification

With the advent of the era of big data, analyzing billions of dollars of data is undoubtedly a headache, many researchers note the importance of data reduction methods. At present, sparse representation theory can provide a concise representation of complex redundant information,

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202112_7(12).0025

which is widely used in the fields of pattern recognition and computer vision, and is the focus of research and development direction of compression theory. The Sparse Representation Classifier (SRC)was designed by Wright [13]. And applied to face recognition has achieved more satisfactory results. The basic assumption of the algorithm is that each test sample can be linearly represented by a training sample as a complete base, and the linear combination coefficients shown in the table are sparse and contain the categories to which the sample belongs. SRC based on the above algorithmic ideas has unique noise resistance and good robustness, and performs well in more classification models.

The given dataset X$\in R^{m\times n}$, m is the sample dimension, n is the number of samples.The set sample is divided into k categories, and the sub matrix that makes up the training sample is $X_k = [l_{k,1}, l_{k,2}, \dots, l_{k,n}]$, X_k is the dictionary to be sparse. Sparse indicates that the classifier wants to calculate a sparse matrix A, makes $X_k A$ the optimal approximation representation for the test sample. A test sample y, expressed in the above concept, can be symbolized as:

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

Which $\alpha_{k,i}$ is the coefficient corresponding to the k-class training sample. Because a training sample of a different class than the test sample contributes close to 0 or equal to 0to the testsample, the (1) formula can be represented as:

$$y = X\omega \tag{2}$$

Where $\omega = [0, \dots, 0, \omega_{k,1}, \omega_{k,2}, \dots, \omega_{k,n_k}, 0, \dots, 0]^T$, then the sparse representation coefficient of y in the dictionary. If the ω is ideal, the training sample corresponding to the non-zero coefficient in the ω is more likely to be the same atom of the test sample, while the less the non-zero coefficient, the sparserest. The vector can be solved using the following model:

$$\widehat{\omega}_0 = argmin||\omega||_0 \; subject \; to \; y = X\omega \tag{3}$$

Formula (3) Minimizing l_0 the paradigm is an NP-hard problem. According to the theory of compression perception, l_0 is rare for the vector solved by the paradigm to be strictly equal to zero, which can be transformed into a l_1 convex optimization problem that minimizes the number of models. Therefore, the target function becomes:

$$\widehat{\omega}_0 = argmin||\omega||_1 \; subject \; to \; y = X\omega \tag{4}$$

Because there is more or less noise in the actual situation, the error threshold ε is introduced, ε is greater than zero:

$$\omega_0 = argmin||\omega||_0 \; subject \; to \; \left||y - X\omega|\right| < \varepsilon \tag{5}$$

Next, the sparse coefficients obtained are classified by the corresponding category, and the test sample can be reconstructed by different types of training samples. Ultimately, the decision rule for sparsely representing classifiers is to compare the residual size of different classes of refactored samples with the test samples, which can be summed up as:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202112_7(12).0025

$$min\gamma_c(y) = \left||y - X\widehat{\omega}_1^c\right||, c = 1 \dots K \quad (6)$$

When refactoring a test sample based on sparse representation classifier, sparseness is paid attention to, but the locality of feature subspace is not taken into account, resulting in poor classification. For this defect, Fan [14] et al. propsed Weighted Sparse Representation Based Classification(WSRC). WSRC focuses on training samples similar to those to be tested, giving such samples more interpretation, and rejecting training samples with large differences in dictionaries, i.e. by weighting local information, can improve the accuracy of classification. WSRC calculates the Gaussian distance between the sample to be measured and all the training samples, using these distances as weight constraint training samples. Mathematically, Gaussian distance of the sample s_1 and s_2 are defined as:

$$d(s_1, s_2) = \exp(-\frac{||s_1 - s_2||^2}{2\sigma^2})$$

Which, σ represents the width of the Gauss core. In this way, the WSRC objective function is further expressed as:

$$\widehat{\omega}_0 = argmin||D\omega||_1 \; subject \; to \left||y - X\omega\right|| < \varepsilon$$

$$D = [d_G(y, l_{1,1}), \dots, d_G(y, l_{k,n_k})]^T$$

After many tuning experiments, it is determined that the experimental parameters used in this paper are: $\varepsilon = 0.001$, $\sigma = 1.5$.

3.2. Model Evaluation Indicators

In order to measure the predictive performance of the two classification models, four more commonly used model evaluation indicators are used in this paper, namely Accuracy (ACC), Sensitivity (SN), Precision (PE) and Matthews correlation coefficient (MCC). Accuracy is the number of sample sets that are correctly identified as a comprehensive measure of model performance. Sensitivity is the ratio of positive samples predicted to be positive, accuracy is the proportion of positive samples in positive samples, and Matthews correlation coefficient is the degree of correlation between proteins that are inter-made before and after prediction. They are mathematically defined as:

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

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

$$PE = \frac{TP}{TP + FP}$$

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

In the upper class, TP (true positive) is the true positive number, i.e. the positive sample is predicted as a positive sample, the FP (false postive) is the false positive number, i.e. the negative sample is misjudged as a positive sample, TN (true negative) is the true negative number, i.e. the negative sample is predicted as negative, and the FN (false negative) is the false negative number, i.e. the positive sample is misjudged as a negative sample.

4. MODEL PREDICTION RESULTS AND ANALYSIS

This chapter will present experimental results for predicting protein interactions using weighted sparse presentation classifiers for Yeast and Human datasets. It will also be compared with some of the methods mentioned in this article. These include: comparative analysis of

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202112_7(12).0025

random forest feature selection, comparative analysis of support vector machine prediction results, and comparative analysis of existing method prediction results.

4.1. The Result of the Model Prediction

After many experiments, it was decided to use the average accuracy, sensitivity, accuracy and Matthews correlation coefficient of 5 fold cross-validation as the predictive performance of the model. The model's predictions on the Yeastdataset are shown in Table 2, where you can see that the model's prediction accuracy on the Yeast dataset is up to 97.78 percent. Average prediction accuracy, average sensitivity, average accuracy, and average Matthews correlation coefficients were 96.97%,97.51%,96.43%, and 93.91%, respectively, And the float between the results of 5 experiments is small, the model has a certain robustness.

Table 2. The results of the Yeast dataset experiment

Testing set	Acc(%)	Sn(%)	Sp(%)	Mcc(%)
Fold-1	96.24	96.98	95.71	92.49
Fold-2	96.72	97.61	95.70	93.27
Fold-3	97.78	97.81	97.62	95.56
Fold-4	96.82	97.67	96.00	93.65
Fold-5	97.30	97.48	97.10	94.60
Average	96.97	97.51	96.43	93.91

The model's predictions on the Human dataset are shown in Table 3, from which you can see that the model's prediction accuracy on the Human dataset is up to 96.27%, the average prediction accuracy, Average sensitivity, average accuracy, and average Matthews correlation coefficients were 95.5%, 96.83%, 93.2%, and 97%, respectively, and the results of 5 experiments fluctuate less, the model has a certain robust type.

Table 3. The results of the Human dataset experiment

Testing set	Acc(%)	Sn(%)	Sp(%)	Mcc(%)
Fold-1	95.44	96.22	93.70	90.82
Fold-2	95.07	96.33	93.15	90.14
Fold-3	96.27	97.80	93.95	92.53
Fold-4	95.19	97.22	92.13	90.38
Fold-5	95.55	96.57	93.11	90.97
Average	95.50	96.83	93.20	90.97

4.2. Model Performance Comparison

4.2.1 Evaluation of random forest feature selection

In order to improve the prediction accuracy of the model, filter the effective features and filter the noise, the dimension of the feature vector is proposed above. In this regard, this section will devote some length to the validity of the feature selection method. Specifically, you compare a model that uses feature selection with a model that does not. With the other model parameters unchanged, the 546-dimensional feature is reduced, and the first 117-dimensional feature is trained and predicted according to the feature importance sorting mechanism, in which the accuracy of the prediction reaches a peak. The results of the experiment are shown in Table 4:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202112_7(12).0025

Table 4. Yeast Dataset Feature Selection Comparison Experimental Results

model	dimension	Acc(%)	Sn(%)	Sp(%)	Mcc(%)
original model	546	96.97	97.51	96.43	93.91
dimension-lowering model	117	94.41	96.65	91.67	88.91

It can be seen from the table that the model effect has improved significantly after feature selection, which shows that the method of selecting by use of feature is effective.

4.2.2 Comparative analysis of the experimental results of the support vector machine

Support vector machines (SVMs) are widely used in machine learning models that predict proteininteractions. In order to verify the predictive effect of the weighted sparse representation classifier, the classic SVM classifier is used as the control model. Because the dataset is non-linearly smearable, SVM uses radial base core functions for operations. Among them, the model parameters are calculated by grid search method, and their values are: $C = 1.5$, $\gamma = 3.56$. This article compares the experimental results of WSRC and SVM on the Yeast dataset, as shown in Table 5:

Table 5. Yeast dataset WSRC and SVM comparison experiments result

model	Acc(%)	Sn(%)	Sp(%)	Mcc(%)
WSRC	96.97	97.51	96.43	93.91
SVM	92.44	94.79	89.79	85.01

As can be seen from the table, WSRC is more suitable for the characteristics of this article than SVM, and it is proved that WSRC is robust to noise.

4.2.3 A comparative analysis of the experimental results of existing methods

Many researchers have designed methods to predict protein interactions on the Yeast dataset and have made good predictions. To reflect the benefits of WSRC, we compared it with different approaches. The results are shown in Table 6.

Table 6. Comparison of method results on the Yeast dataset

model	Acc(%)	Sn(%)	Sp(%)	Mcc(%)
WSRC+CTD	96.97	97.51	96.43	93.91
SVM+LD	88.56	87.37	89.50	77.15
HOG+SVD+RF [15]	94.83	92.40	97.10	89.77
StackPPI [9]	94.64	92.81	96.46	89.34
WSRC+PseAA [16]	92.50	95.87	88.82	86.09

YiJie Ding et al. in 2016 proposed HOG+SVD+RF, is a new matrix-based coding method that uses integrated learning classification. The protein sequence is represented first by an alternative matrix (SMR), then by using a directional gradient histogram (HOG)and singular value decomposition (SVD)to extract features from the matrix, and finally by entering the feature vector into a random forest for classification.

StackPPI was proposed by Cheng Chen et al. in 2020. The author constructs the feature operator using 7 feature coding methods, such as pseudoamino acid composition, autocorrelated descriptor, location-specific scoring matrix, and then uses XGBoost to reduce

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202112_7(12).0025

dimension, and finally uses a stacked integrated classifier consisting of random forest, extreme random tree and logistic regression algorithm to predict.

WSRC and PseAA were proposed by Yuan Huang in 2016. This method combines continuous wavelet transformation with pseudo-amino acid composition to construct features, and uses weighted sparse to represent classifier prediction interoperability.

5. CONCLUSIONS AND PROSPECTS

This paper first introduces the background knowledge and theory related to this article. These include the properties of protein physical and chemical, feature coding methods, feature selection methods, and the basic theory of weighted sparse representation classifiers. Then a weighted sparse representation model based on protein sequence feature extraction is proposed, and the process of modeling is: first obtain the protein sequence and construct the positive and negative samples, then carry out feature extraction and feature selection of the protein sequence, and finally use the weighted sparse to represent the classifier classification. Finally, the experimental results of 50% cross-validation are given, and the characteristic selection method, the classical support vector machine method and the existing method are compared and evaluated. It is proved that the feature coding algorithm and the prediction model are effectively combined, which provides a new way of thinking for the model selection.

Although the accuracy of predicting protein interactions is high, there are some aspects that can be optimized. First, a more efficient feature coding algorithm can capture most of the information in the protein sequence while making it less relevant to existing coding features, such as the secondary structure of the protein. In this way, feature coding with large fusion differences may improve prediction accuracy. Second, the design of a reasonable integration strategy, the advantages of different classifiers, weaken the single classifier bias on data characteristics. Effective combination of different models is a problem that can be further explored.

REFERENCES

[1] Palopoli N, Edwards R. Large-scale prediction of short linear motifs using structural information from protein-protein interactions[J]. Bmc Bioinformatics, 2015, 11(6):1-9.

[2] Archakov AI, Govorun VM, Dubanov AV, Ivanov YD, Veselovsky AV, Lewi P, et al. Protein-protein interactions as a target for drugs in proteomics[J]. Proteomics, 2003,3:380-391.

[3] Foltman M, Sanchez-Diaz A. Studying Protein–Protein Interactions in Budding Yeast Using Co-immunoprecipitation[J]. Methods in Molecular Biology, 2016, 1369: 239-256.

[4] Kawahashi Y, Doi N, Takashima H, Tsuda C, Oishi Y, Oyama R, et al. In vitro protein microarrays for detecting protein-protein interactions: application of a new method for fluorescence labeling of proteins[J]. Proteomics, 2003; 3:1236-1243.

[5] 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.

[6] Guo Y, Yu L, Wen Z, et al. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences[J]. Nucleic Acids Research, 2008, 36(9): 3025-3030.

[7] You Z H, Zhu L, Zheng C H, et al. Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set[C]. BMC Bioinformatics, 2014, 15(15): S9.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202112_7(12).0025

[8] L. Zhang, G. Yu, D. Xia, J. Wang, Protein-protein interactions prediction based on ensemble deep neural networks, Neurocomputing 324 (2018) 10-19.

[9] Chen Cheng et al. Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier[J]. Computers in Biology and Medicine, 2020, 123.

[10] Guo Y, Yu L, Wen Z, et al. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences[J]. Nucleic Acids Research, 2008, 36(9):3025-3030.

[11] L Rao H B, Zhu F, Yang G B, et al. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence[J]. Nucleic Acids Research, 2011, 39(Web Server issue): W385

[12] BreimanL. Random forests[J]. Machinelearing, 2001, 45(1):5-32.

[13] Wright J., Yang A. Y., Ganesh A. etal. Robust face recognition via sparse representation[J]. Ieee T Pattern Anal, 2009, 31(2):210-227.

[14] Fan Z. Z., Ni M., Zhu Q. etal. Weighted sparse representation for face recognition[J]. Neurocomputing, 2015, 151:304-309.

[15] Ding Y.J, Tang J.J, Guo F. Identification of Protein–Protein Interactions via aNovel Matrix-Based Sequence Representation Modelwith Amino Acid Contact Information[J]. Int J Mol Sci, 2016, 17(10):163.

[16] Y.A. Huang, Z.H. You, X. Chen, G.Y. Yan, Improved protein-protein interactions prediction via weighted sparse representation model combining continuous wavelet descriptor and PseAA composition, BMC Syst. Biol. 10 (2016) 120.