Protein Interaction Prediction Method Based on Feature Engineering and XGBoost

Xiaoman Zhao, Xue Wang*

Institute of Intelligent Machinery, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China

Abstract. Human protein interaction prediction studies occupy an important place in systems biology. The understanding of human protein interaction networks and interactome will provide important insights into the regulation of developmental, physiological and pathological processes. In this study, we propose a method based on feature engineering and integrated learning algorithms to construct protein interaction prediction models. Principal Component Analysis (PCA) and Locally Linear Embedding (LLE) dimensionality reduction methods were used to extract sequence features from the 174-dimensional human protein sequence vector after Normalized Difference Sequence Feature (NDSF) encoding, respectively. The classification performance of three integrated learning methods (AdaBoost, Extratrees, XGBoost) applied to PCA and LLE features was compared, and the best combination of parameters was found using cross-validation and grid search methods. The results show that the classification accuracy is significantly higher when using the linear dimensionality reduction method PCA than the nonlinear dimensionality reduction method LLE. the classification with XGBoost achieves a model accuracy of 99.2%, which is the best performance among all models. This study suggests that NDSF combined with PCA and XGBoost may be an effective strategy for classifying different human protein interactions.

1. Introduction

Proteins are the main performers of cellular activities in living organisms and are involved in various aspects of organism growth and reproduction such as cell signaling, metabolism, apoptosis and necrosis, and regulation of gene expression[1]. Proteins do not exist in isolation in an organism and exert their biological properties alone, but they interact with other proteins in some way to drive or trigger certain biochemical reactions together and synergistically exert their biological properties. Therefore, it is necessary to classify and predict protein interactions. However, there are many studies on protein interaction classification prediction, and previous studies have used traditional low-throughput techniques to detect protein interactions, such as mass spectrometry, nuclear magnetic resonance, chromatographic electrophoresis, and other methods[2]; however, no matter which research method is used, almost all of them discuss and study the macroscopic factors, while the microscopic factors of amino acid sequences are rarely studied. Although the study of macroscopic factors is representative and comprehensive, microscopic factors are also extremely important for the overall assessment of protein interactions. Therefore, this paper suggests that characterization (PCA, LLE dimensionality reduction methods) and integrated learning methods (XGBoost) can be used to study the complex compound properties of protein interactions. The following studies and contributions are made in this paper.

(1) Sample data were obtained and preprocessed. Positive samples were collected from the Human Protein Reference Database[3] (HPRD, version 2007) and negative samples were obtained from Swiss-Prot (http://www.expasy.org/sprot/, version 57.3). After obtaining the data, it was necessary to process the abnormal data to ensure the comparability of the final data.
(2) Protein feature extraction method to select the study features of the samples. Numerical protein sequences contain correlated noise and redundant feature information to some extent. Linear (PCA) and nonlinear dimensionality reduction (LLE) techniques are used to select high-frequency influences, reduce redundancy, shorten training time, and reduce losses, and then various tests are used to determine the existence of significant effects among variables.
(3) XGBoost was built based on python 3.7.3 environment for demonstration and analysis. After the preliminary preparation work is completed, the quantitative indicators and selected data are used as training samples and test samples for the follow-up work.
The research method used in this paper is to select sample data to build an algorithmic model for protein interaction prediction and to use XGBoost to predict the overall evaluation of real estate, which achieves the following three main innovations: first, to study the degree of influence of microscopic factors of amino acid sequences on the overall classification of protein interactions, and to find the best combination of parameters using cross-validation and grid search methods; second, to use

* Corresponding author: 181543681@qq.com

characterization and XGBoost to achieve the overall classification of protein interactions; finally, to fill the gap in the field of overall protein interaction classification prediction beyond macroscopic factors.

2. Related Work

Currently, more mature and well established studies on microscopic factors of amino acid sequences, such as yeast two-hybrid screening technique[4], fluorescence resonance energy transfer technique[5], phage display technique[6], and tandem affinity purification technique[7] are used to detect protein interactions. However, such conventional methods can only identify a very small number of protein interactions, are not applicable to all proteins of the organism, and the accuracy of the identification results is not high[8]. Therefore, there is a need for a computational prediction method that can support efficient and highly accurate protein interactions. The sequence and idea of the search is shown in Figure 1.

2.1 Genome-based prediction methods

So far, various computational prediction methods for protein interactions such as phylogenetic profiles[9], gene fusion events[10] and gene neighborhoods have been proposed by previous authors. Among them[11], Zhong et al[12] found some interaction relationships between proteins with matching or similar phylogenetic profiles by calculating the phylogenetic profiles of 4290 proteins in E. coli; meanwhile, Souza et al[13] proposed a protein interaction prediction method based on identifying gene fusion events in the complete genome and found that gene fusion could predict protein interactions. Since all the above methods have their own shortcomings, such as the phylogenetic profile method requires the prior construction of the phylogenetic profile of the species, the construction of the phylogenetic profile is tedious, and the frequency of gene fusion events is very low, so the genomic-based prediction methods cannot be widely used.

2.2 Structure-based prediction methods

With the accumulation of protein structure information, some researchers have proposed computational prediction methods for protein interactions based on protein structure information. In 2002, Finn et al[14] proposed a protein interaction prediction method that simulates 3D structure information of known proteins, which can score all possible proteins interacting between two protein families and predicts whether there is an interaction relationship between these proteins; In 2013, LUO et al[15] proposed a protein interaction prediction method based on protein 3D structures and showed that the method obtained good prediction performance and speculated that the reason for the good performance of the method might be the use of homology models and the exploitation of proximity and distance geometric relationships between proteins. In 2022, Liu et al[16] predicted protein interactions based on support vector machines and amino acid index distribution of protein sequences and obtained 94% accuracy. Although the above methods provide a new way to protein interaction prediction, these methods need to ensure that the prior information of the protein is reliable and are limited by the prior information of the protein structure, so these methods cannot be widely used.

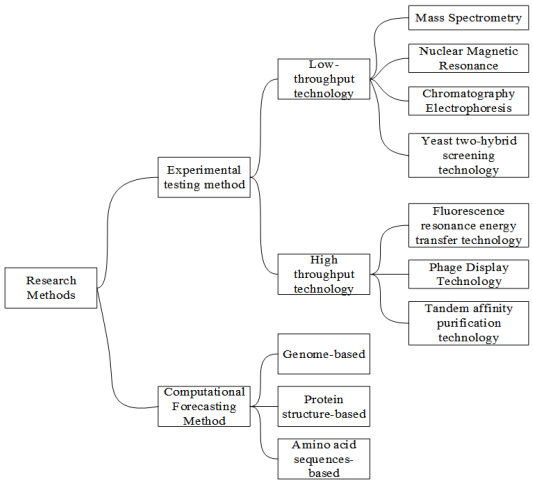

Fig. 1. Protein Interaction Research Methods.

2.3 Sequence-based prediction methods

Compared with the a priori information of proteins, the amino acid sequences of proteins have accumulated more rapidly in recent years. The amino acid sequences of proteins contain rich information and directly determine the secondary and tertiary structures of proteins[17]. Therefore, the study of proteomics based on amino acid sequences has gradually become a widely adopted approach by scholars for subcellular localization, protein structure-function prediction and protein interaction prediction[18].

3. Research Methods

3.1 Data acquisition and preprocessing

In the data set, only human proteins were collected according to the following requirements: (1) only human proteins were collected; (2) sequences annotated with ambiguous or uncertain subcellular location terms (e.g., "potential," "probable," or "by similarity") were excluded; (3) sequences annotated with two or more positions were excluded; (4) sequences with "fragment" annotations were excluded, and sequences with less than 50 amino acid residues in sequence length were removed. "); (3) exclude sequences annotated with two or more positions; (4) exclude sequences annotated with "fragments" and remove sequences with less than 50 amino acid residues in sequence length. After the above process, a total of 2184 human proteins were collected from six different subcellular organelles (cytoplasm, nucleus, endoplasmic reticulum, Golgi apparatus, lysosome and mitochondria). By randomly pairing these proteins with other proteins in different subcellular organelles, a total of 36,480 negative pairs were generated.

3.2 Feature selection

Dimensionality reduction in machine learning is achieved by mapping the high-dimensional space data to a low-dimensional space representation, which is divided into linear and nonlinear mappings, and PCA[19] is commonly used in linear mappings and LLE[20] in nonlinear mappings. Therefore, we use PCA by characterizing the covariance matrix with the aim of reducing the dimensionality of the data while maintaining the maximum contribution of the data set to the variance. Using the idea of data dimensionality reduction, a multivariate analysis method is used to transform multiple indicators into a few less comprehensive indicators with the loss of less data information. Each principal component is a linear combination of the original variables, which are not correlated with each other, and the principal component analysis takes the variance as the measure of information and takes the components with large cumulative contribution as the principal components. LLE is a data dimensionality reduction method based on stream shape learning[21], and stream shape can be understood as embedding a subspace in a high-dimensional Euclidean space. We used LLE to perform protein sequence feature extraction on the encoded results to ensure that the topology of the original data is maintained after dimensionality reduction. The features are extracted again from the encoded amino acid sequences of NDSF to reduce the computational complexity. Since the encoded sequence vector has 174 dimensions, we roughly chose the range of vector dimensionality scaling based on the interpretable variance plotted as dimensionality. The optimal data dimension was found precisely by repeating the experiment.

3.3 Overview of XGBoost algorithm

XGBoost is a gradient boosting algorithm (based on the integrated tree model, the boosting algorithm generates the weak learning model at each step based on the gradient direction of the loss function, which is called gradient boosting), XGBoost uses a stepwise forward addition model, except that after each iteration the weak learner is generated without computing a coefficient. The XGBoost algorithm reduces the risk of overfitting by adding a loss function with a regular term penalty to achieve weak learner generation, and instead of using the usual search method, the XGBoost algorithm directly uses the first-order derivative and second-order derivative of the loss function by Taylor expansion, and improves the performance of the algorithm by pre-ranking and weighted quantile techniques. The performance of the algorithm is greatly improved.

The difference of the XGBoost model is that it further increases the generalization ability of the model by customizing a set of loss functions with the help of Taylor expansions. Its gradient boosting tree based algorithm adds a regularization term to the objective function, which can reduce the complexity of the model and avoid overfitting, and its objective function are as follows.

$$Obj(\Phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (1)$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\omega^2 \quad (2)$$

where yi is the predicted value, Ω(fk) is the regular term, fk is the decision tree, T represents the number of leaf nodes, ω represents the proportion of leaf nodes, γcontrols the number of leaf nodes, and λcontrols the proportion of leaf nodes.

The XGBoost algorithm performs an iterative operation as well as a second-order Taylor expansion during the solution of the objective function, as shown in (3)

$$Obj(\Phi) = \sum_{i=1}^{n}\left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) \quad (3)$$

where equations (4) and (5) are the first-order and second-order derivatives of the loss function, respectively.

$$g_i = \alpha_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \quad (4)$$

$$h_i = \alpha^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \quad (5)$$

4. Experimentation and Analysis

4.1 Experimentation Data and Environment

To achieve numerical representation of amino acids, 20 common amino acids were classified into seven categories according to B3LYP/6-31G in density generalization theory [] and molecular modeling methods, as shown in Table 1.

Table 1. Amino acids are grouped based on dipole and side chain volume.

Amino acid type	Grouping
Ala,Gly,Val	***1***
IIe,Leu,Phe,Pro	***2***
Tyr,Met,Thr,Ser	***3***
His,Asn,Gln,Trp	***4***
Arg,Lys	***5***
Asp,Glu	***6***
Cys	7

Based on Python 3.7.3 environment, the quantified eigenvalues are used as input factors of NDSF to obtain 174-dimensional vector output in this paper. All the XGBoosts selected in this paper use cross-validation and grid search methods to find the best combination of parameters (learning rating and n_estimator) so as to achieve parameter optimization. The models were evaluated by comparing their accuracy, loss rate, and AUC. From 36545 pairs of positive samples and 36323 pairs of negative samples, 1~29236 pairs of positive samples and 1~29058 pairs of negative samples were selected as training samples, and 1~3709 pairs of positive samples and 7265 pairs of negative samples were selected as test samples to test the feasibility of the model.

In order to evaluate the performance of the protein interaction prediction model based on amino acid sequences proposed in this paper, three widely used evaluation criteria, including Accuracy, Recall, and Loss, were used in this experiment, which were calculated by the following equations (6), (7), and (8).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Re\,c\,all = \frac{TP}{TP+FN} \quad (7)$$

$$LOSS = -(y\,log(p) + (1-y)\,log(1-p)) \quad (8)$$

Where, TP(True Positives) denotes the number of times the initial positive samples are correctly predicted as positive by the model, TN(True Negatives) denotes the number of times the initial negative samples are correctly predicted as negative by the model, FP(False Positives) denotes the number of times the initial positive samples are incorrectly predicted as negative by the model, and FN(False Negatives) denotes the number of times the initial negative sample was incorrectly predicted as a positive sample by the model. y denotes the true label of the sample (1 or 0), and p denotes the probability that the model predicts a positive sample.

4.2 Analysis of Effect and Efficiency

In order to retain most of the information of the original features as much as possible, while avoiding the influence of correlation between serial features on the classification results. Before using PCA for dimensionality reduction, it is necessary to select the appropriate dimensionality and plot the interpretable variance as a function of dimensionality. There is usually an inflection point on the curve where the interpretable variance stops increasing rapidly, so 45 dimensions is chosen as the termination point, as in Figure 2.

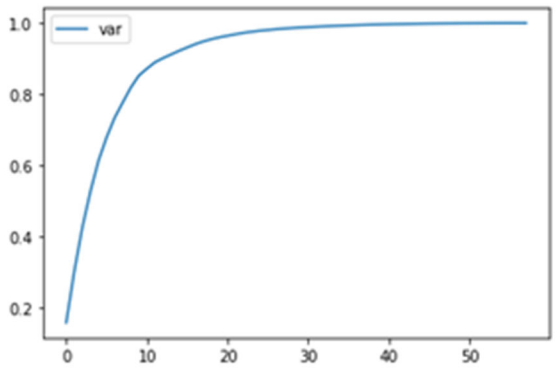

Fig. 2. Feature selection for PCA_NDSF.

To maintain consistency, LLE also chooses 45 dimensions as the termination point.

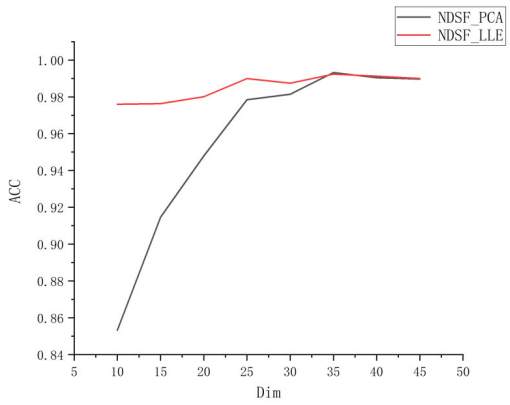

Fig. 3. Comparison of the accuracy of models with different feature dimensions.

Table 2. Comparison of test set and training set of integrated learning algorithms.

		AdaBoost		Etratress		XGBoost	
		Train	Test	Train	Test	Train	Test
NDSF_LLE	*Loss*	*15.14*	*15.26*	*11.45*	*1.30*	*0.22*	*4.58*
	Acc	*87.86*	*87.74*	*95.43*	*98.71*	*98.74*	*98.70*
	Recall	*86.93*	*86.81*	*99.92*	*98.53*	*98.81*	*98.90*
NDSF_PCA	*Loss*	*14.01*	*14.78*	*11.45*	*1.32*	*0.91*	*1.19*
	Acc	*87.00*	*86.00*	*99.01*	*98.98*	*99.32*	*99.29*
	Recall	*87.13*	*87.15*	*98.77*	*98.73*	*99.10*	*98.54*

In the NDFS coding method, the integrated learning XGBoost algorithm combined with PCA and LLE shows a trend of increasing and then decreasing accuracy of the model in the range of 0 to 45 dimensions. The best performance is achieved when it is reduced to 35 dimensions. The comparison of model accuracy in different feature dimensions is shown in Fig. 3. the difference between LLE and PCA is only within 2%. Feature extraction of protein sequence data can effectively retain sufficient information, remove redundant data, and reduce training time, while obtaining data accuracy of up to 99.2%, which has practical application value.

This study uses random search to achieve the best parameter settings, then follows the principle of taking smaller combinations of parameters at a time, and finally sets a reasonable range of parameter values to achieve the training of the model. After selecting the optimal dimensionality reduction method, the results of the three integrated learning methods are compared, as shown in Table 2 for the comparison of the test set and training set of the integrated learning algorithm. Among the integrated algorithm models in this study, by comparing the prediction models constructed by Bagging (AdaBoost, XGBoost) and Boosting approaches (Extratrees), we further find that XGBoost has better prediction results, and the accuracy of the training and test sets of the NDSF_PCA model based on the XGBoost classifier highest reached 99.32% and 99.29%, respectively, with loss rates of 0.91% and 1.19%, and recall rates of 99.10% and 98.54%, respectively.

In the NDSF_PCA_XGBoost model, the accuracy reached 99.2%, the best performance among all models, and the best results were achieved on loss, recall, and acc. The results illustrate that using the gradient boosting algorithm based on learning classification and regression tree (CART) to calculate the complexity of each tree leaf node and to minimize the loss of finding the best prediction score, thus avoiding overfitting the learning model, effectively controlling the complexity of the model and improving the model accuracy.

4.3 Summary and Conclusions

Through the above study, this paper demonstrates that the prediction model constructed using the integrated learning method XGBoost has better classification results and can effectively identify positive and negative protein interaction effects, proving that our model approach is efficient and usable.

Meanwhile, the extraction of features should not be neglected. The scientific selection of microscopic factors affecting amino acid sequences is a prerequisite for

constructing an ideal model for the overall assessment of protein interactions by extracting protein sequence features from the coded results by two dimensionality reduction methods, PCA and LLE. In this study, we also found that the integrated learning method achieves better prediction results for highly unbalanced data, and the complexity of the model is effectively controlled because the algorithm can avoid over-fitting of the learned model. With the continuous improvement of various algorithms, the research work based on feature engineering and XGBoost has laid a good foundation for protein interaction prediction studies.

Acknowledgement

Supported by the Dean's Fund of the Hefei Research Institute of the Chinese Academy of Sciences for "Research on Prediction of Protein Interaction Based on Deep Learning"(YZJJ2021QN26).

References

1. DU B-X, QIN Y, JIANG Y-F, et al. Compound-protein interaction prediction by deep learning: Databases, descriptors and models [J]. DRUG DISCOVERY TODAY, 2022, 27(5): 1350-66.
2. Y. HAN, L. CHENG, W. SUN. Analysis of Protein-Protein Interaction Networks through Computational Approaches [J]. PROTEIN AND PEPTIDE LETTERS, 2020, 27(4): 265-78.
3. Z. DU, H. SU, W. WANG, et al. The trRosetta server for fast and accurate protein structure prediction [J]. NATURE PROTOCOLS, 2021, 16(12): 5634-51.
4. L. HU, X. WANG, Y-A.HUANG, et al. A survey on computational models for predicting protein-protein interactions [J]. BRIEFINGS IN BIOINFORMATICS, 2021, 22(5):
5. M S. KHATUN, W.SHOOMBUATONG, M M.HASAN, et al. Evolution of Sequence-based Bioinformatics Tools for Protein-protein Interaction Prediction [J]. CURRENT GENOMICS, 2020, 21(6): 454-63.
6. MEI L-C, HAO G-F, YANG G-F. Computational methods for predicting hotspots at protein-RNA interfaces [J]. WILEY INTERDISCIPLINARY REVIEWS-RNA, 2022, 13(2):
7. D. SARKAR, S. SAHA. Machine-learning techniques for the prediction of protein-protein interactions [J]. JOURNAL OF BIOSCIENCES, 2019, 44(4):
8. C. SHI, J. CHEN, X. KANG, et al. Deep Learning in the Study of Protein-Related Interactions [J]. PROTEIN AND PEPTIDE LETTERS, 2020, 27(5): 359-69.
9. O. SLATER, B. MILLER, M. KONTOYIANNI. Decoding Protein-protein Interactions: An Overview [J]. CURRENT TOPICS IN MEDICINAL CHEMISTRY, 2020, 20(10): 855-82.
10. D. SUN, S. LIU, X. GONG. Review of multimer protein-protein interaction complex topology and structure prediction* [J]. CHINESE PHYSICS B, 2020, 29(10):
11. Y. TABEI. Scalable Prediction of Compound-protein Interaction on Compressed Molecular Fingerprints [J]. MOLECULAR INFORMATICS, 2020, 39(1-2):
12. L. ZHONG, Z. MING, G. XIE, et al. Recent Advances on the Semi-Supervised Learning for Long Non-Coding RNA-Protein Interactions Prediction: A Review [J]. PROTEIN AND PEPTIDE LETTERS, 2020, 27(5): 385-91.
13. P C T. SOUZA, R. ALESSANDRI, J. BARNOUD, et al. Martini 3: a general purpose force field for coarse-grained molecular dynamics [J]. NATURE METHODS, 2021, 18(4): 382-+.
14. RD. FINN, A. BATEMAN, J. CLEMENTS, et al. Pfam: the protein families database [J]. NUCLEIC ACIDS RESEARCH, 2014, 42(D1): D222-D30.
15. B. LEWANDOWSKI, JW. WARD, et al. Sequence-Specific Peptide Synthesis by an Artificial Small-Molecule Machine [J]. SCIENCE, 2013, 339(6116): 189-93.
16. L. Liu, PY. Gong, M. Tang, HK. Hu, YB. Zhang, CH. Liu. Advances in the study of novel coronavirus pneumonia based on proteomics technology[J]. Advances in Biochemistry and Biophysics, 2022, 49(10): 1848-1865. DOI:10.16476/ j.pibb.2022.0029.
17. H. OULDALI, K. SARTHAK, T. ENSSLEN, et al. Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore [J]. NATURE BIOTECHNOLOGY, 2020, 38(2): 176-+.
18. S. WANG, Z-Y TSUN, R L WOLFSON, et al. Lysosomal amino acid transporter SLC38A9 signals arginine sufficiency to mTORC1 [J]. SCIENCE, 2015, 347(6218): 188-94.
19. J. MCCARTY, K. DELANEY, S. DANIELSEN, et al. Complete Phase Diagram for Liquid-Liquid Phase Separation of Intrinsically Disordered Proteins [J]. JOURNAL OF PHYSICAL CHEMISTRY LETTERS, 2019, 10(8): 1644-52.
20. M. BAEK, F. DIMAIO, I. ANISHCHENKO, et al. Accurate prediction of protein structures and interactions using a three-track neural network [J]. SCIENCE, 2021, 373(6557): 871-+.
21. L. Jia, and Y. Luan, Multi-feature Fusion Method Based on Linear Neighborhood Propagation Predict Plant LncRNA-Protein Interactions. Interdisciplinary sciences, computational life sciences 14 (2022) 545-554.