Predicting protein-protein interactions between banana and Fusarium oxysporum race 4 integrating sequence and domain homologous alignment and neural network verification

Hui Fang (fanghui2002@163.com)
Guangxi University https://orcid.org/0000-0002-3992-0440
Cheng Zhong
Guangxi University
Chunyan Tang
Guangxi Univresity

Research

Keywords: Protein-protein interactions, Banana, Fusarium oxysporum race 4, Sequence alignment, Prediction

Posted Date: November 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1033168/v1

Predicting protein-protein interactions between banana and *Fusarium oxysporum* race 4 integrating sequence and domain homologous alignment and neural network verification

Hui Fang[1,2,3], Cheng Zhong[1,2*] Chunyan Tang[2]

*Corresponding author: chzhong@gxu.edu.cn

Abstract

Background: The pathogen of banana *Fusarium oxysporum* race 4 (Foc4) infects almost all banana species, and it is the most destructive. The molecular mechanism of the interactions between *Fusarium oxysporum* and banana still needs to be further investigated.

Methods: We use both the homology-interolog and domain-domain method to predict the protein-protein interactions (PPIs) between banana and Foc4. The predicted protein interaction sequences are encoded by the conjoint triad and autocovariance method respectively to obtain continuous and discontinuous information of protein sequences. This information is used as the input data of the neural network model. The Long Short-Term Memory (LSTM) neural network five-fold cross-validation and independent test methods are used to verify the predicted protein interaction sequences. To further confirm the PPIs between banana and Foc4, the Go functional annotation and interaction network analysis are carried out.

Results: The experimental results show that the PPIs for banana and foc4 predicted by our proposed method may interact with each other in terms of sequence structure and

GO functional annotation, and Foc4 protein plays a more active role in the process of Foc4 infecting banana.

Conclusions: This study obtained the PPIs between banana and Foc4 by using computing means for the first time, which will provide data support for molecular biology experiments.

Keywords: Protein-protein interactions; Banana; *Fusarium oxysporum* race 4; Sequence alignment;Prediction

Introduction

Banana (Musaspp.) is a monocotyledonous perennial plant of the Musa genus in Musaceae. Banana is the largest herbaceous flowering plant in the world, and its fruit is edible. Banana grows in tropical and subtropical regions and is the fourth largest food crop after rice, wheat, and corn in some countries and regions[1]. Banana *Fusarium oxysporum* race 4 (Foc4), also known as yellow leaf disease and Panama disease, is a typical fungal soil-borne disease caused by *Fusarium oxysporum* f.sp.cubense infection, which destroys banana vascular bundles and causes plant death[2]. Foc4, the pathogen of banana *Fusarium oxysporum* race 4, infects almost all banana species, and it is the most destructive[3]. The pathogenic process of *Fusarium oxysporum* needs to go through the identification process between pathogen and banana root. The pathogen reaches and adheres to the surface of the banana root, and *Fusarium oxysporum* produces a series of pathogenic factors, such as secreted effector protein

factors[4],pathogenicity-related enzymes[5], and toxins[6]. The pathogen invades the inside of the host, colonizes in banana, and shows the symptoms on the outside[7]. At present, some progresses have been achieved in the research of banana Foc4. Some pathogenic factors, cell wall degrading enzymes, and toxins of banana Foc4 have been found. Meanwhile, some banana resistance genes, active substances and hormones related to resistance have been discovered through transcriptomics and proteomics. However, up to now, there are no effective measures to control banana Foc4, and its pathogenic mechanism is not completely clear. Therefore, the molecular mechanism of the interactions between *Fusarium oxysporum* and banana still needs to be further investigated.

Pathogenic bacteria affect protein invasive plants can interact with the plants, some of the plants can start the host defense responses against pathogens. Protein-protein interactions(PPIs) between plant protein and pathogenic protein are crucial to studying the molecular basis of pathogenesis[8]. The studying PPIs methods can be divided into biological experiment-based methods and bioinformatics-based methods. The biological experiment-based methods mainly include yeast two-hybrid[9], bimolecular fluorescence complementation[10], and immunoprecipitation[11]. The biological experiment-based methods have some disadvantages, such as time-consuming, high cost, and low coverage. The bioinformatics-based methods have the advantages of high efficiency and low cost, and they have the disadvantage of the existence of false positives. With the rapid development of omics data, the biological experiment-based

methods are difficult to meet the requirement of high-throughput biological data. At present, the public databases DIP[12], HPRD[13], BioGRID[14], IntAct[15], MINT[16], and HPIDB[17]store a large number of experimentally verified PPIs data, which provide data sources for predicting PPIs using bioinformatics methods. The homology-based interolog method and domain-domain method have been used to predict PPIs in some fields. Recently, some researchers used these two methods to predict the intraspecific PPIs among bacterial blight pathogen, rice, corn, and cassava[18-21], and FWHT-RF[22] can be a useful supplementary method to predict potential PPIs in plants.

Interspecies PPI has been reported in the study of human and pathogenic bacteria, which is used to predict the PPIsbetween human and hepatitis C virus[23], between humans and Bacillus anthracis[24], and between humans and Plasmodium falciparum[25]. For the study of PPIs between plant and pathogen, Li et al. predicted 3074 protein interactions between Arabidopsis thaliana and Ralstoniasolanacearum on the database DIP by the homology-based interolog method and domain-domain method. These protein interactions include 119 Ralstoniasolanacearum proteins and 1442 Arabidopsis thaliana proteins. The data set of PPIs was verified by GO functional annotation and network characteristic analysis[26].

By using the homology-based interolog method and domain-domain method, Ma et al. predicted 523 PPIs between rice and Magnaporthegrisea, including 27 rice blast proteins and 236 rice proteins[27]. The obtained PPIs data set was verified by the

machine learning method, and the protein function was analyzed by GO (Gene Ontology) and the database KEGG. Zheng et al.[28] used the structure-based method and generated a global PPI network consisting of 2,018 PPIs involving 1,344 rice and 418 blast fungus proteins. To our knowledge, the research on predicting PPIs between plants and pathogens has only been reported on the model plants Arabidopsis thaliana, rice, and their pathogens. But there are no related reports on predicting PPIs between banana and Foc4 based on the Bioinformatics methods. The study on the interactions between banana and Foc4 has been mainly conducted from the independent perspective of infection of Foc4 pathogenic factors and active substances related to banana resistance. The genes or proteins differentially expressed in bananas could be obtained in previous studies, but the effectors of Foc4 interacting with banana protein could not be accepted.

This paper has the following contributions. We proposed a computing method for predicting PPIs for banana and Foc4 for the first time. We encoded the predicted PPIs sequences for banana and Foc4 by the conjoint triad method and autocovariance method respectively to obtain continuous and discontinuous information of protein sequences, verify the predicted PPIs may interact in sequence structural characteristics by the LSTM neural network five-fold cross-validation and independent test methods, and further functionally verify the PPIs by the Go function annotation and interaction network analysis. The predicted PPIs between banana and Foc4 will provide data support for molecular biology experiments.

Materials and Methods

Datasets

We first downloaded 45856 banana proteins in banana protein sequences from https://banana-genome-hub.southgreen.fr and 14459 Foc4 protein sequences from ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/350/365/GCA_000350365.1_Foc4_1.0, respectively. Secondly, We downloaded all PPIs of six model species, Arabidopsis thaliana, nematode, Drosophila, yeast, Escherichia coli, and human, from the database MINTat https://mint.bio.uniroma2.it/, the database DIP at https://dip.doe-mbi.ucla.edu/dip/main.cgi, the database TAIR at https://www.arabidopsis.org/, the database BioGRID at https://downloads.thebioged.org/biogerid/release-archive/biogerid-3.5.166/, and the database INTACT at https://www.ebi.ac.uk/intact/, respectively. Thirdly, we downloaded 118921 PPIs from the database MINT, 76881 PPIs from the database DIP, 2656 PPIs from the database TAIR, and 183768 PPIs from the database IntAct. Finally, we downloaded 62782 pathogen-host interspecific protein interactions from the databaseHPIDB at http://hpidb.igbb.msstate.edu/.All domain-domain interaction template PPIs were downloaded from the database3DID[29] at https://3did. irbbarcelona.org/. The corresponding protein sequences of the above six species were downloaded from the databaseUniprotat https://www.uniprot.org/. Different databases may use different IDs for the same protein. We used the software tool Biomart[30] to convert the different protein IDs into uniform IDs.

Methods

We first downloaded the experimentally verified intra-species and inter-species PPIs from the database as the interaction template. Next, we applied the interolog method and domain-domain method to predict the data sets of PPIs between banana and Foc4 to find the common PPIs between banana and Foc4. Thirdly, we used the conjoint triad(CT)[31] and auto covariance(AC)[32] to code protein sequence features to obtain the structure information of continuous and discontinuous protein sequences. Fourthly, we verified the predicted PPIs data sets for banana and Foc4 by using LSTM neural network five-fold cross-validation method and independent test method. Finally, we computed the accuracy, sensitivity, specificity, receiver operating characteristic curve (ROC), and area under the curve(AUC) of the predicted results. Figure 1 shows the process of predicting PPIs between banana and Foc4, in which iPPIs indicate interolog PPIs, dPPIs represent domain-domain PPIs, and DDI denotes domain-domain interactions.

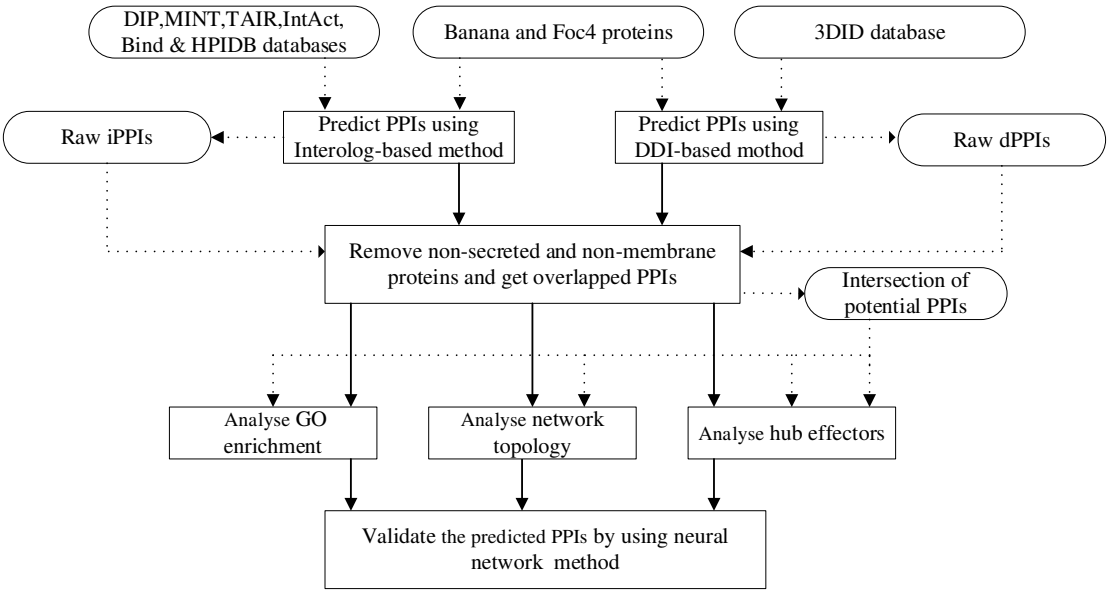

Fig. 1 Process of predicting PPIs between banana and Foc4, where the solid arrow represents 'control flow direction' and the dashed arrow denotes 'data flow direction'.

Predicting PPIs between banana and Foc4

The interolog method is a means for predicting homologous interactions. Its main idea is that homologous proteins may have similar properties. If two proteins A and B interact with each other via verified experiments, and two proteins A' and B' are homologous proteins of A and B respectively, then according to the principle that homologous proteins have similar properties, proteins A' and B' may also interact with each other[23]. The idea of the domain-domain interaction prediction method is that if proteins C and D contain domains C and D which can interact with each other, proteins C and D may interact with each other[24].

Based on the protein sequence data of banana and Foc4, we used the homology-based interolog method and domain-domain method to predict the interactions between banana and Foc4. We selected the transmembrane or secreted proteins in Foc4 as the protein infecting banana[26]and obtained the final PPIs data set between banana and Foc4.

For the interolog method, we used the local sequence alignment tool BLAST to find the homology proteins, where the parameter *E* was set to 0.00001, the sequence identity was set to 30%, and the coverage was set to 80%[26, 27]. Firstly, the protein sequences of six model species are compared with banana and Foc4 to find out the orthologous

proteins between banana and Foc4. Then, the host protein sequences in the database HPIDB are compared with the banana protein sequences and the pathogen protein sequences are compared with the Foc4 protein sequences to obtain interspecific homologous proteins.

We submitted the protein sequences of banana and Foc4 to the database 3DID to find out the domains contained in each protein, where the value of parameter *E* is set to 0.00001 and the sequence identity is set to 90%[26]. If any PPI of banana and Foc4 contains a couple of interactive domains in the database 3DID, it is considered that this pair of proteins for banana and Foc4 may interact with each other[33].

We applied the two software tools signalP[34]and WoLFPSOFT[35] with the default values of their parameters to find secretory proteins. If a protein predicted by signalP contains a signal peptide and it is located as extracellular by WoLFPSOFT, the protein is a secretory protein. In addition, we used the software TMHMM2.0[36] to predict transmembrane proteins in Foc4 proteins. If the number of transmembrane helices predicted by TMHMM is greater than 1, the proteins are considered to be transmembrane proteins[37].

PPIs coding of sequence features

Proteins are biomolecules composed of amino acids, while protein sequences are represented by 20 standard amino acids. Encoding the sequence feature of a protein is to extract the feature vector from the protein sequence. The sequence feature extraction transforms the original sequences into a fixed-length numerical vector. In recent years,

some researchers have proposed some methods to predict PPIs using only protein sequence information, but these methods can not fully capture interaction information from continuous and discontinuous amino acid fragments at the same time.

In order to solve the above problem, the conjoint triad (CT) method and auto covariance(AC) method were used to encode sequence features. By using the CT method, 20 amino acids are divided into seven categories according to the volume of even electrodes and side chain volume. Each three consecutive amino acid are regarded as a basic unit, and the class frequency of all basic units in a protein is counted. The AC method mainly considers the proximity effect and uses both the continuous and discontinuous sequence information in a protein sequence. The number of all possible kinds for each basic unit is 7×7×7=343. Thus, the final feature vector with 686-dimension contains the features of two proteins interacting with each other. Min-max normalization was performed on the feature vectors to map the result of encoding each protein pair into the interval [0,1], so as to remove the influence of protein length on frequency counting. Let f_i represent the *i*-th component of a protein eigenvector, the *i*-th component of a normalized protein feature vector, d_i, is computed as follows[31] :

$$d_i = \frac{f_i - \min\{f_1, f_2, ..., f_{343}\}}{\max\{f_1, f_2, ..., f_{343}\}}, i = 1, 2, 3, \ldots, 343 \qquad (1)$$

The interactions between amino acids are reflected by seven physical and chemical characteristics of amino acids. The seven physical and chemical properties are hydrophobicity, hydrophilicity, net charge index, polarity, polarizability, solvent

accessible surface area, and side chain volume, respectively. Each protein sequence is transformed into a 7-dimensional vector, and each amino acid is represented by a normalized value of seven descriptors. The initial values of seven physical and chemical properties of 20 amino acids can be found in [32]. The variance $AC_{lag,j}$ is computed as follows[32]:

$$AC_{lag,j} = \frac{1}{n-lag}\sum_{i=1}^{n-lag}(X_{i,j} - \frac{1}{n}\sum_{i=1}^{n}X_{i,j})(X_{(i+lag),j} - \frac{1}{n}\sum_{i=1}^{n}X_{i,j}) \quad (2)$$

where *lag* represents the distance between the two amino acid residues, *n* is the length of protein sequence *X*, $X_{i,j}$ represents the *j*-th descriptor in the *i*-th position of a protein sequence. In this paper, seven physical and chemical properties are used and the optimal value of *lag* is set to 30[38]. After AC transformation, each protein sequence has been transformed into a 210-dimensional vector. Combined with using the CT method, each PPI sequence has been transformed into a vector of (343+210)×2=1106 dimensions.

Verification

We used the interolog method and domain-domain method to deal with the proteins of banana and Foc4 to obtain their PPIs, and treated these PPIs as the positive samples with size 739. We randomly selected the interaction pair of proteins from banana and Foc4 proteome as negative samples. We verified the predicted results by the five-fold cross-validation method and independent test method, respectively.

In the five-fold cross-validation test, the Long Short-Term Memory(LSTM) neural network [39] is used to predict PPIs between banana and Foc4. By using the

characteristic coding of the PPIs between banana and Foc4, the original protein sequence is converted into a fixed-length numerical vector which is used as the input of the LSTM neural network. The input layer of LSTM neural network is a feature vector composed of the forward and backward hidden layer output vectors h_f and h_b. The corrected linear unit(relu) is used as the activation function in the hidden layer, and the softmax function is used in the output layer. According to the results of the CT and AC coding schemes, the input sequence is $X = (x_1, x_2, x_3, ..., x_{1106})$ and the prediction model outputs a corresponding result sequence is $Y = \{y_1, y_2, y_3, ..., y_{1106}\}$.

In the actual biological data, the number of positive samples and the number of negative samples are inconsistent, and the positive and negative samples are usually unbalanced data sets. In order to deal with this situation, we randomly constructed the data sets with 1:1 positive and negative samples and 1:10 positive and negative samples respectively to conduct five-fold cross-validation and independent test verification. For the independent test of 1:10 positive and negative samples, if the size of positive samples is *m*, the size of negative samples is10*m*. We selected the samples with a size of 2*m*/3 in the positive samples and the samples with a size of 2*m*/3 in the negative samples to form the training set, and selected the remaining positive samples with a size of *m*/3 and the remaining negative samples with size of 10*m* -2*m*/3=28*m*/3 to form the test set.

In this paper, we used the accuracy *ACC*, sensitivity *Sn*, specificity *Sp*, receiver operating characteristic curve *ROC*, and area under curve *AUC* to evaluate the prediction effect[23]：

$$ACC = \frac{TN + TP}{TN + TP + FN + FP} \tag{3}$$

$$Sn = \frac{TP}{TP + FN} \tag{4}$$

$$Sp = \frac{TN}{TN + FP} \tag{5}$$

where *TN* is the number of true counterexamples, *TP* represents the number of true examples, *FN* denotes the number of false counterexamples, and *FP* is the number of false-positive examples.

Each protein is used as a node and the interaction between each pair of proteins is represented as an edge, a PPIs network is created by all the nodes and edges. We used the software Cytoscape3.7[40] to visualize the PPIs network to conveniently and intuitively observe the characteristics of the network. We used the ClusterViz plug-in in Cytoscape[40]to divide the interaction network into different functional modules. We executed the algorithm ClusterVizuse FAG-EC [41] to partition the network into several subnetworks. The median centrality V_i of node *i* in the network is calculated as follows:

$$V_i = \sum_{s \neq t \neq i} \frac{n_{st}^i}{g_{st}} \tag{6}$$

where g_{st} denotes the number of the shortest paths from node *s* to node *t*, and n_{st}^i represents the number of the shortest paths from node *s* to node *t* via node *i* in the network.

We applied the software TBTools[42] to carry out the GO (Gene Ontology) functional enrichment analysis of PPIs. According to the specification for TBTools, we set the value of parameter $P<0.05$ and used Bonferroni correction[43].

Results

Experimental environment

The computer used is with Intel (R) Xeon (R) W-2133 CPU @ 3.6 GHz processor and memory capacity 8 GB running operating system Windows10. The prediction algorithm was implemented by Python3 programming.

Experimental results

We first predicted 26910 PPIs and 376755 PPIs between banana and Foc4 by using the interolog method and domain-domain method, respectively. Table 1 shows the results of predicted PPIs, where 739 interactions with 515 banana proteins and 81 Foc4 proteins are common overlapping PPIs predicted by the interolog method and domain-domain method, Method1 represents the interolog method, and Method2 denotes the domain-domain method. The detailed data sets of all predicted results are given in Appendix 1.

Table 1 Statistical information of predicted PPIs between banana and Foc4

Prediction method	Number of PPIs	Number of Banana proteins	Number ofFoc4 proteins
Method1	26910	5938	697
Method2	376755	18965	1916
Common parts of predicted results of Method1 and Method2	739	515	81

It can be seen from the results in Table 1 that the number of PPIs predicted by interolog method is less than that of PPIs predicted by the domain-domain method. This

is because the interolog method adopts the homologous sequence-based alignment, which depends on the amount of data in the existing database, while the domain-domain method is based on the interactive domains contained in proteins, and a protein can contain two or more interactive domains[44].

We extracted the feature vector of proteins in banana-Foc4 PPIs, and analyzed the reliability of banana-Foc4 PPIs predicted by the LSTM neural network-based five-fold cross-validation method and independent test method. Table 2 shows the results of sensitivity *Sn*, specificity *Sp*, accuracy *ACC*, and receiver operating characteristic curve *ROC* of the predicted banana-Foc4 PPIs.

Table 2 Values of *Sn*, *Sp*, *ACC* and *ROC* of predicted banana-Foc4 PPIs

Method	Ratio of positive and negative samples	*Sn*	*Sp*	*ACC*	*ROC*
LSTM five-fold cross validation	1:1	0.9075	0.9852	0.9445	0.94
LSTM five-fold cross validation	1:10	0.8581	0.9285	0.8978	0.87
LSTM independent test	1:1	0.9163	0.9485	0.9366	0.94
LSTM independent test	1:10	0.8753	0.9466	0.9047	0.89

We can see from Table 2 that when the ratio of positive and negative samples was 1:1, the results predicted by LSTM neural network-based five-fold cross-validation method were better than that when the ratio of positive and negative samples was 1:10. Similarly, when the ratio of positive and negative samples was 1:1, the results predicted

by LSTM independent test verification method were better than that when the ratio of positive and negative samples was 1:10. The experimental results show that the PPIs between banana and Foc4 predicted by LSTM neural network-based five-fold cross-validation and LSTM neural network-based independent test methods have high structural similarity. The PPIs between banana and Foc4 may interact in sequence structure characteristics.

The following is to analyze the network structure characteristics of the PPIs between banana and Foc4 predicted by the experiment. By using Cytoscape, each protein in the interactions between banana andFoc4is treated as a node, and each interaction between banana and Foc4 is treated as an edge. The result of the PPIs network between banana and Foc4 is shown in Figure 2.

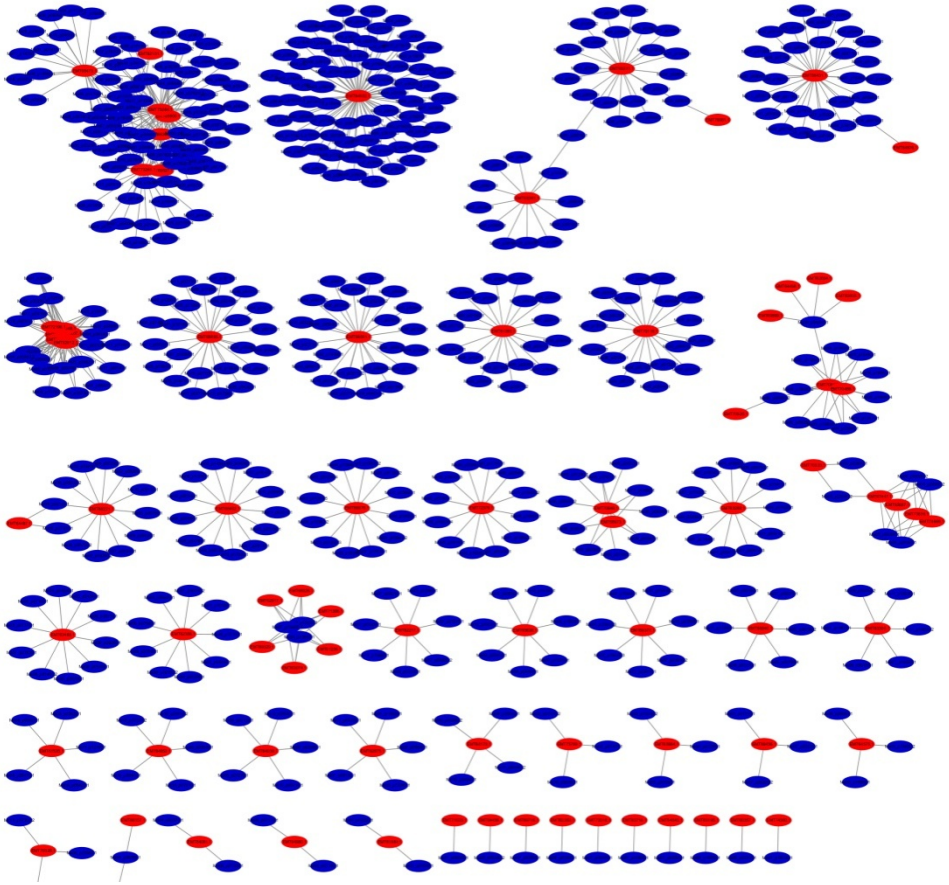

Fig. 2 PPIs network between banana and Foc4, where the red node represents Foc4 protein, and the blue node denotes banana protein

In the PPI network, the connectivity of a protein is defined as the number of all other proteins linking to this protein. The connectivity is an index of evaluating the importance of a protein in the network. From Figure 2 we can see that the average connectivity of Foc4 protein was 9.12 and the average connectivity of banana protein was 1.43. This indicates that the connectivity of Foc4 protein was higher than that of banana protein in the PPI network for banana and Foc4, and Foc4 protein played a more active role, which affected a series of biological processes of banana infected by Foc4. It can also be seen from Figure 2 that the PPI network for banana and Foc4 was divided into 51 sub-networks, in which the largest sub-network contains 86 nodes, the smallest

sub-network has only two nodes, and there are 30 sub-networks with more than or equal to 6 nodes. Some complex sub-networks with more nodes contain multiple Foc4 proteins, some sub-networks only contain one Foc4 protein, and the smallest sub-network only has one banana interacting with Foc4 protein. In addition, we found that three proteins of Foc4, namely EMT64532.1, EMT73264.1 and EMT73245.1, interacted with 72, 58, and 29 proteins of banana, respectively. This illustrates that these three proteins of Foc4 play an important role in the interactions, and these results will provide a basis for future biological experiments.

To annotate the GO function of PPIs for banana and Foc4, we first aligned the banana protein with SwissProt protein by the software BLAST. Then, we compared the obtained Foc4 protein with SwissProt protein. Finally, we used the TBTools to annotate the GO function PPIs for banana and Foc4. The top 20 annotated results of proteins for Foc4 are shown in Table 3, and the annotated results of proteins for banana are shown in Table 4.

Table 3 Top 20 annotated results of proteins for Foc4

GO Name in Biological Process	GO ID	*P*_value	Hit Counts
membrane fusion	GO:0061025	6.87E-08	8
export from cell	GO:0140352	1.49E-07	17
transport	GO:0006810	2.63E-07	47
establishment of localization	GO:0051234	4.96E-07	47
vesicle fusion	GO:0006906	8.42E-07	6
export across plasma membrane	GO:0140115	8.91E-07	8
localization	GO:0051179	1.56E-06	49
organelle membrane fusion	GO:0090174	2.50E-06	6

vesicle organization	GO:0016050	3.15E-06	9
organelle fusion	GO:0048284	3.18E-06	8
transmembrane transport	GO:0055085	4.18E-06	28
membrane organization	GO:0061024	4.82E-06	15
xenobiotic detoxification by transmembrane export across the plasma membrane	GO:1990961	7.81E-06	6
xenobiotic transport	GO:0042908	9.77E-06	6
intracellular transport	GO:0046907	2.64E-05	21
organophosphate ester transport	GO:0015748	2.67E-05	7
cellular localization	GO:0051641	3.48E-05	27
organic substance transport	GO:0071702	4.36E-05	33
mitochondrial transport	GO:0006839	8.27E-05	8
establishment of localization in cell	GO:0051649	1.00E-04	22

Table 4 Annotated results of proteins for banana

GO Name in Biological Process	GO ID	*P*_value	Hit Counts
transport	GO:0006810	3.33E-16	262
translation	GO:0006412	3.53E-10	63
catabolic process	GO:0009056	1.54E-05	129
protein metabolic process	GO:0019538	2.48E-05	194
tropism	GO:0009606	1.45E-04	20
cellular homeostasis	GO:0019725	5.13E-04	34
embryo development	GO:0009790	1.15E-02	61
cellular component organization	GO:0016043	1.16E-02	206
cell-cell signaling	GO:0007267	2.25E-02	21

It can be seen from Table 3 that in the annotated GO function results of Foc4 protein, the top three ones are membrane fusion, export from cell, and transport respectively. In addition, we can also see that Foc4 protein annotates vesicle fusion, export across membrane, transmembrane transport, and membrane organization, which are all related

to cell membrane function. Foc4 protein must cross the cell membrane if it wants to enter banana and interact with banana protein.

Table 4 shows that in the annotated GO function results of banana proteins, the top three ones are transport, translation, and catabolic process respectively. Some banana R-proteins are annotated with tropism, cellular homeostasis, cell-cell signaling, and other functions, all of which are related to the response of cells to external stress. Foc4 protein enters the banana, and the banana uses the specificity of intracellular resistance proteins to recognize the effector and trigger immune response[45].

The annotated results of predicted PPIs between Foc4 and banana show that Foc4 protein annotated the functions related to cell membrane such as vesicle fusion, transmembrane export, transmembrane transport and membrane tissue, and banana protein annotated the functions related to external stress response such as transport, tropism, cell automatic regulation and cell signal transduction. This illustrates that the PPIs between banana and Foc4 predicted by our method are possible from the perspective of GO functional annotation.

Discussion

One of the characteristics of this study is that the intra-species and inter-species PPIs downloaded from the database were used as interaction templates, the PPIs between banana and Foc4 are predicted by the homology-based interolog method and the domain-domain method respectively, and the intersection of PPIs predicted by these

two methods is taken as the final predicted result which is more accurate. In addition, the problem studied here is inter-species protein interaction, which uses not only intra-species protein interaction of model species as prediction template but also uses inter-species protein interaction of multiple species as prediction template. The template of interspecific interaction prediction comes from the database HPIDB, which contains PPIs of 66 species of animals, plants and many pathogens, including interspecific protein interactions between animals and microorganisms and the ones between plants and microorganisms.

In this paper, we coded the sequence of PPIs by the combined use of CT method and AC method. The CT method regards every three consecutive amino acids as a basic unit and counts the class frequency of all basic units in protein, while the AC method mainly pays close attention to proximity effect. In this way, the continuous and discontinuous sequence information of proteins can be used at the same time, which makes the prediction result more accurate. We verified the PPIs dataset between banana and Foc4 by LSTM neural network-based five-fold cross-validation method and independent test method.

By observing the results of GO function annotation and PPIs network analysis, we found that there were many Foc4 interacting with host protein in PPIs between banana and Foc4. In addition, we also discovered that many Foc4 protein GO annotations were related to vesicle fusion, export across membrane, transmembrane transport, and membrane organization. This indicates that Foc4 protein needs to be secreted outside

the cell and must cross the cell membrane in order to infect bananas. At the same time, we can see that in the predicted PPIs between banana and Foc4, the functions of proteins related to external stress, cellular homeostasis, and cell-cell signaling are enriched, and the pathogenic molecules in vitro are recognized by proteins in banana and a series of immune responses downstream are stimulated. Therefore, these enriched proteins may be involved in the identification of pathogenic proteins of Foc4. This illustrates that the PPIs between banana and Foc4 proteins predicted by our method are possible from the perspective of GO functional annotation.

Conclusion

The innovation and characteristic of this paper is that both the homology-interolog method and domain-domain method are applied to predict the PPIs between banana and Foc4, and the dataset of PPIs between banana and Foc4 is obtained by computing means for the first time. The combination of the CT and AC methods is used to encode protein characteristics to obtain the continuous and discontinuous sequence information of proteins. The predicted banana-Foc4 PPIs dataset was verified by LSTM neural network-based five-fold cross-validation method and independent test method. The GO annotation and interaction network analysis of banana and Foc4 protein interactions shows that there are indeed PPIs between banana and Foc4, and several Foc4 proteins interacting with host protein together. The dataset of PPIs between banana and Foc4 predicted by computing method will provide a base for the study of banana Fusarium wilt, and also offer a new means for analyzing the molecular mechanism of interactions between banana and Foc4. In the future, we will investigate the biological experiment

method to verify whether there may be some false positives in the protein mutual network between banana and Foc4 constructed by the computation method.

Abbreviations

GO: Gene Ontology; Foc4: Fusarium oxysporum race 4; PPIs: Protein-protein interactions; CT: Conjoint triad; AC: Auto covariance; LSTM: Long Short-Term Memory; ROC: Receiver operating characteristic curve; AUC: Area under the curve; iPPIs: interolog Protein-protein interactions; dPPIs: domain Protein-protein interactions

Acknowledgments

Not applicable.

Authors' contributions

CZ conceived the study. HF and CT designed the project, performed experiments, carried out data analysis, and wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China under Grant No.61962004, and the Natural Science Foundation of Guangxi under Grant No. 2020GXNSFAA259004. Funding support was also provided by Guangxi Academy of Agricultural Sciences (GuiNongKe 2020YM106).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Medical College, State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, Guangxi University, Nanning, Guangxi, 530004, China. [2] School of Computer, Electronics and Information, Guangxi University, Nanning, Guangxi, 530004, China. [3] Guangxi Crop Genetic Improvement and Biotechnology Laboratory, Guangxi Academy of Agricultural Sciences, Nanning, Guangxi, 530007, China.

References

1. Saravanan, T., M. Muthusamy, and T. Marimuthu, *Development of integrated approach to manage the fusarial wilt of banana.* Crop Protection, 2003. **22**(9): p. 1117-1123.
2. Xingshen, L., et al., *Proteomic analysis of Fusarium oxysporum f. sp. cubense tropical race 4-inoculated response to Fusarium wilts in the banana root cells.* Proteome Science, 2013. **11**(1): p. 11-41.
3. Dita, M.A., et al., *A molecular diagnostic for tropical race 4 of the banana fusarium wilt pathogen.* Plant Pathology, 2010. **59**(2): p. 348-357.
4. Ploetz, R.C., *Fusarium Wilt of Banana Is Caused by Several Pathogens Referred to as Fusarium oxysporum f. sp. cubense.* Phytopathology, 2006. **96**(6): p. 648.
5. Kubicek, C.P., T.L. Starr, and N.L. Glass, *Plant Cell Wall–Degrading Enzymes and Their Secretion in Plant-Pathogenic Fungi.* Annual Review of Phytopathology, 2014. **52**(1): p. 427.
6. Singh, V.K., H.B. Singh, and R.S. Upadhyay, *Role of fusaric acid in the development of Fusarium wilt symptoms in tomato: Physiological, biochemical and proteomic perspectives.* Plant Physiology & Biochemistry Ppb, 2017. **118**: p. 320.
7. Michielse, C.B. and M. Rep, *Pathogen profile update: Fusarium oxysporum.* Mol Plant Pathol, 2009. **10**(3): p. 311-24.
8. Andrés, P., et al., *Targeted metabolic reconstruction: a novel approach for the characterization of plant-pathogen interactions.* Briefings in Bioinformatics, 2010. **12**(2): p. 151-162.
9. Ito, T., et al., *A comprehensive two-hybrid analysis to explore the yeast protein interactome.* Proceedings of the National Academy of Sciences, 2001. **98**(8): p. 4569-4574.
10. Hu, C.D., Y. Chinenov, and T.K. Kerppola, *Visualization of interactions among bZIP and Rel family proteins in living cells using bimolecular fluorescence complementation.* Molecular Cell, 2002. **9**(4): p. 789-798.
11. Miklos, G., et al., *Interactions of the NPXY microdomains of the low density lipoprotein receptor-related protein 1.* Proteomics, 2010. **9**(22): p. 5016-5028.
12. Xenarios, I., et al., *DIP: the database of interacting proteins.* Nucleic Acids Res, 2001. **29**(1): p. 239-241.

13. Joseph, J.A., et al., *Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans.* Genome Research, 2003. **13**(10): p. 2363-2371.
14. Chris, S., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Research, 2006. **34**(suppl_1): p. 535-939.
15. Sandra, O., et al., *The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases.* Nucleic Acids Research, 2014. **42**(D1): p. 358-363.
16. Zanzoni, A., et al., *MINT: a Molecular INTeraction database.* Febs Letters, 2002. **513**(1): p. 135-140.
17. Ammari, M.G., et al., *HPIDB 2.0: a curated database for host-pathogen interactions.* Database, 2016. **2016**: p. baw103.
18. Guo, J., et al., *Prediction and characterization of protein–protein interaction network in Xanthomonas oryzae pv. oryzae PXO99A.* Research in microbiology, 2013. **164**(10): p. 1035-1044.
19. Gu, H., et al., *PRIN: a predicted rice interactome network.* BMC bioinformatics, 2011. **12**(1): p. 1-13.
20. Zhu, G., et al., *PPIM: a protein-protein interaction database for maize.* Plant physiology, 2016. **170**(2): p. 618-626.
21. Thanasomboon, R., et al., *Prediction of cassava protein interactome based on interolog method.* Scientific reports, 2017. **7**(1): p. 1-15.
22. Pan, J., et al., *FWHT-RF: A Novel Computational Approach to Predict Plant Protein-Protein Interactions via an Ensemble Learning Method.* Scientific Programming, 2021. **2021**(9): p. 1-11.
23. Cui, G., C. Fang, and K. Han, *Prediction of protein-protein interactions between viruses and human by an SVM model.* BMC Bioinformatics, 2012. **13(Suppl 7)**: p. S5.
24. Ahmed, I., P. Witbooi, and A. Christoffels, *Prediction of human-Bacillus anthracis protein-protein interactions using multi-layer neural network.* Bioinformatics, 2018. **34**(24): p. 4159-4164.
25. Ramakrishnan, G., et al., *Homology-Based Prediction of Potential Protein–Protein Interactions between Human Erythrocytes andPlasmodium falciparum.* Bioinformatics & Biology Insights, 2015. **9**(9): p. 195-206.
26. Li, Z.G., F. He, and Z. Zhang, *Prediction of protein–protein interactions between Ralstonia solanacearum and Arabidopsis thaliana.* Amino Acids, 2012. **42**(6): p. 2363-2371.
27. Ma, S., et al., *Prediction of protein-protein interactions between fungus (Magnaporthe grisea) and rice (Oryza sativa L.).* Briefings in Bioinformatics, 2017. **20**(2): p. 448-456.
28. Zheng, C., et al., *Predicting Protein–Protein Interactions Between Rice and Blast Fungus Using Structure-Based Approaches.* Frontiers in Plant Science, 2021. **12**: p. 690124.
29. Mosca, R., et al., *3did: a catalog of domain-based interactions of known three-dimensional structure.* Nucleic Acids Research, 2014. **42**(D1): p. D374-D379.

30. Syed, H., et al., *BioMart Central Portal—unified access to biological data.* Nucleic Acids Research, 2009. **37**(suppl_2): p. W23-W27.
31. Shen, J., et al., *Predicting protein-protein interactions based only on sequences information.* Proc Natl Acad Sci U S A, 2007. **104**(11): p. 4337-4341.
32. Guo, Y., et al., *Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences.* Nucleic acids research, 2008. **36**(9): p. 3025-3030.
33. Li, B.Q., et al., *Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS.* PLoS One, 2012. **7**(8): p. e43927.
34. Nielsen, H., et al., *A Brief History of Protein Sorting Prediction.* The Protein Journal, 2019. **38**(3): p. 200-216.
35. Paul, H., et al., *WoLF PSORT: protein localization predictor.* Nucleic acids research, 2007. **35**(suppl_2): p. W585-W587.
36. Krogh, A., et al., *Predicting transmembrane protein topology with a hidden markov model: application to complete genomes.* Journal of Molecular Biology, 2001. **305**(3): p. 567-580.
37. Tanaka, M., et al., *Whole genome sequencing of Entamoeba nuttalli reveals mammalian host-related molecular signatures and a novel octapeptide-repeat surface protein.* PLoS Negl Trop Dis, 2019. **13**(12): p. e0007923.
38. Wang, X., et al., *A novel conjoint triad auto covariance (CTAC) coding method for predicting protein-protein interaction based on amino acid sequence.* Mathematical biosciences, 2019. **313**: p. 41-47.
39. Sønderby, S.K., et al., *Convolutional LSTM Networks for Subcellular Localization of Proteins*, in *Proceedings of the 2015 International Conference on Algorithms for Computational Biology*. 2015, Springer: Cham. p. 68-80.
40. Lopes, C.T., et al., *Cytoscape Web: an interactive web-based network browser.* Bioinformatics, 2010. **26**(18): p. 2347-2348.
41. Wang, J., et al., *ClusterViz: A Cytoscape APP for Cluster Analysis of Biological Network.* IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2015. **12**(4): p. 815-822.
42. Chen, C., et al., *TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data.* Molecular Plant, 2020. **13**(8): p. 1194-1202.
43. Schulz, C.P., *Multiple correlations and bonferroni's correction.* Biological Psychiatry, 1998. **44**(5): p. 775-777.
44. Han, Y.C., et al., *Prediction and characterization of protein-protein interaction network in Bacillus licheniformis WX-02.* Sci Rep, 2016. **6**(1): p. 19486.
45. Dodds, P.N. and J.P. Rathjen, *Plant immunity: towards an integrated view of plant–pathogen interactions.* Nature Reviews Genetics, 2010. **11**(8): p. 539-548.