Computational and Structural Biotechnology Journal 20 (2022) 2322–2331

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Proteome-wide prediction and analysis of the *Cryptosporidium parvum* protein–protein interaction network through integrative methods

Panyu Ren, Xiaodi Yang, Tianpeng Wang, Yunpeng Hou *, Ziding Zhang *

State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China

ARTICLE INFO

Article history:
Received 10 January 2022
Received in revised form 8 May 2022
Accepted 9 May 2022
Available online 13 May 2022

Keywords:
Cryptosporidium parvum
Protein–protein interaction
Prediction
Machine learning
Interolog mapping
Domain-domain interaction

ABSTRACT

As one of the most studied Apicomplexan parasite *Cryptosporidium, Cryptosporidium parvum* (*C. parvum*) causes worldwide serious diarrhea disease cryptosporidiosis, which can be deadly to immunodeficiency individuals, newly born children, and animals. Proteome-wide identification of protein–protein interactions (PPIs) has proven valuable in the systematic understanding of the genome-phenome relationship. However, the PPIs of *C. parvum* are largely unknown because of the limited experimental studies carried out. Therefore, we took full advantage of three bioinformatics methods, i.e., interolog mapping (IM), domain-domain interaction (DDI)-based inference, and machine learning (ML) method, to jointly predict PPIs of *C. parvum.* Due to the lack of experimental PPIs of *C. parvum*, we used the PPI data of *Plasmodium falciparum* (*P. falciparum*), which owned the largest number of PPIs in Apicomplexa, to train an ML model to infer *C. parvum* PPIs. We utilized consistent results of these three methods as the predicted high-confidence PPI network, which contains 4,578 PPIs covering 554 proteins. To further explore the biological significance of the constructed PPI network, we also conducted essential network and protein functional analysis, mainly focusing on hub proteins and functional modules. We anticipate the constructed PPI network can become an important data resource to accelerate the functional genomics studies of *C. parvum* as well as offer new hints to the target discovery in developing drugs/vaccines.

1. Introduction

Cryptosporidium spp. are zoonotic Apicomplexan parasites leading to serious diarrheal disease, i.e., cryptosporidiosis, which is the second leading cause of moderate-to-severe diarrheal disease in children. In 2016, cryptosporidiosis caused approximately 4.2 million infections and 57,000 deaths worldwide [1–3]. There are currently more than 40 accepted *Cryptosporidium* species with different host preferences [4,5]. Among them, *Cryptosporidium parvum* (*C. parvum*) is the most studied species reported in both humans and livestock [6]. In humans, malnourished children under the age of five and HIV/AIDS patients are more susceptible to *C. parvum* [7]. *C. parvum* infection is usually accompanied by clinical symptoms such as abdominal pain and moderate to severe diarrhea, which can easily lead to some sequelae, including weight loss, fatigue, and post-infection irritable bowel syndrome [8,9]. Except for the non-fatal diseases, in 2017, a study in Lebanon has also shown a strong association between human colon cancer and *C. parvum* [10]. Livestock, especially neonatal calves and lambs, can be infected by *C. parvum* and get diarrhea and impair gain of body weight, causing losses in meat and milk production [11–13]. Besides, some wild animals and fishes can also be infected [14,15]. Currently, there are only two drugs, nitazoxanide and paromomycin, developed for cryptosporidiosis treatment, but these two drugs are not fully effective in severely immunocompromised individuals and generate toxicity in dehydrated animals [16,17]. Thus, an immense need still exists for the development of more effective drugs/vaccines to treat cryptosporidiosis.

Proteins perform their molecular functions by interacting with each other. For instance, numerous fundamental biological

Abbreviations: C. parvum, Cryptosporidium parvum; *P. falciparum, Plasmodium falciparum*; PPIs, protein-protein interactions; Y2H, yeast two-hybrid; PCA, protein complementation assay; AP-MS, affinity purification coupled with mass spectrometry; SPR, surface plasmon resonance; ITA, isothermal titration calorimetry; IM, interolog mapping; DDI, domain-domain interaction; DMI, domain-motif interaction; AI, artificial intelligence; ML, machine learning; DPC, Di-peptide Composition; AC, Auto Covariance; LD, Local Descriptor; RF, Random Forest; Doc2Vec, Document to vector; DM, distributed-memory; SGD, stochastic gradient descent; ROC, Receiver Operating Characteristic; PR, Precision-Recall; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; HSP70s, 70kDa heat shock proteins; BP, biological process; CC, cellular component; MF, molecular function.

* Corresponding authors.
E-mail addresses: hou@cau.edu.cn (Y. Hou), zidingzhang@cau.edu.cn (Z. Zhang).

https://doi.org/10.1016/j.csbj.2022.05.017

processes such as transcription, translation, and protein trafficking are mediated by protein–protein interactions (PPIs) [18]. Thus, systematic identification of the PPI network of *C. parvum* will be extremely valuable to understand the genome-phenome relationship as well as provide new hints to seek out reliable drug targets and develop effective vaccines rapidly for treating the cryptosporidiosis. Generally, scientific researchers always make full use of experimental techniques to identify PPIs, such as yeast two-hybrid (Y2H) [19,20], protein complementation assay (PCA) [21], affinity purification coupled with mass spectrometry (AP-MS) [22], surface plasmon resonance (SPR) [23], and isothermal titration calorimetry (ITA) [24]. Nonetheless, the wet-lab experimental methods to validate PPIs are costly, time-consuming, and labor-intensive. Moreover, researches on *C. parvum* have been hampered by the long-time unavailable culture for its full lifecycle in vitro as well as the poorly annotated reference genome [25]. Therefore, only a very limited number of *C. parvum* PPIs have been identified [26].

In this context, bioinformatics methods could contribute to the identification of *C. parvum* PPIs without the limitations mentioned above [27]. A plethora of PPI prediction methods have been developed, such as interolog mapping (IM) [28], domain-domain interaction (DDI)-based inference [29], domain-motif interaction (DMI)-based inference [30], and increasingly popular artificial intelligence (AI) technique. Machine learning (ML) [31], as the core of AI technique, has been widely used in the field of PPI identification and has shown its powerful predictive potential [32]. Briefly, ML methods train a binary classification model by using experimentally known PPIs and selecting potential non-PPIs to learn the differences between them. In this way, we can further accurately identify protein interactions from the query protein pairs. To construct an ML model, the key step is to employ effective feature encoding schemes, which convert protein sequences (i.e., the most commonly used input information) to fixed-dimensional feature vectors. Several common sequence-based feature encoding schemes such as Di-peptide Composition (DPC) [33], Auto Covariance (AC) [34], and Local Descriptor (LD) [35] are widely used, in which amino acid composition/physicochemical properties or residue interaction effects in sequences have been taken into account. Recently, an embedding technique (i.e., Doc2Vec) derived from natural language processing has been applied to encode protein sequences to further predict PPIs and has been evaluated to improve the PPI prediction performance significantly [32].

In this work, we combined traditional prediction methods (i.e., IM and DDI inference) and the ML method [i.e., Random Forest (RF)] to predict proteome-scale *C. parvum* PPIs (Fig. 1). Two sequence encoding schemes (i.e., DPC and Doc2Vec) were adopted to capture protein composition and semantic information, respectively. Although few experimental PPI data of *C. parvum* can be used, we developed the ML model based on the largest number of *Plasmodium falciparum* (*P. falciparum*) PPI data in Apicomplexa [36]. Then, the learned knowledge was transferred to predict *C. parvum* PPIs. By means of a 5-fold cross-validation and independent test, we extensively compared the prediction performance of our prediction framework with other popular sequence encodings-based RF methods, suggesting that our pipeline outperforms other approaches. Finally, we predicted proteome-scale *C. parvum* PPIs based on our proposed computational framework and achieved the *C. parvum* PPI network. Unlike the traditional prediction method, we yielded an interaction probability score [37] in each protein pair. According to the network topology of the constructed PPI network, we inferred several hub proteins and subnetworks that can potentially help explore the pathogenesis of cryptosporidiosis and speed up the discovery of effective drugs/vaccines.

2. Materials and methods

2.1. Data collection and dataset construction

2.1.1. C. parvum and P. falciparum proteins

We downloaded the proteomes of the *C. parvum* and *P. falciparum* (isolate 3D7) from the UniProt database (https://www.uniprot.org/) [38], which contained 3,805 and 5,387 proteins, respectively.

2.1.2. The PPIs of P. falciparum (isolate 3D7)

There are limited *C. parvum* PPIs experimentally determined in previous studies. However, the *C. parvum* belongs to the Apicomplexa [36]. By investigating PPI data of Apicomplexa in BioGRID [39], we discovered only *P. falciparum* (isolate 3D7) evolutionarily close to *C. parvum* has a relatively large number of interaction data available to train a PPI prediction model which can be further transferred to predict *C. parvum* PPIs. Therefore, we downloaded *P. falciparum* (isolate 3D7) PPIs from BioGRID [39]. To obtain a reasonable *P. falciparum* (isolate 3D7) protein interaction dataset, we excluded non-physical interactions, redundant interactions, and interactions containing too short/long sequences (i.e., length $\leq$ 30 or $\geq$ 5,000 amino acids). As a result, we obtained 1,968 experimentally verified *P. falciparum* (isolate 3D7) PPIs, which were regarded as the positive samples. Regarding the sam-

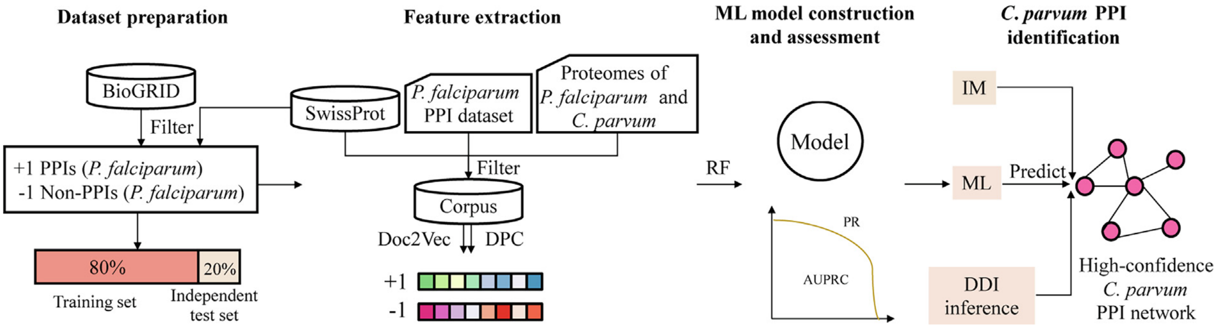
Dataset preparation
BioGRID
Filter
+1 PPIs (P. falciparum)
-1 Non-PPIs (P. falciparum)
80%
20%
Training set
Independent test set
Feature extraction
SwissProt
P. falciparum PPI dataset
Proteomes of P. falciparum and C. parvum
Filter
Corpus
Doc2Vec
DPC
+1
-1
ML model construction and assessment
RF
Model
PR
AUPRC
C. parvum PPI identification
IM
ML
Predict
DDI inference
High-confidence C. parvum PPI network

Fig. 1. Workflow of the proposed computational pipeline to predict *C. parvum* PPIs. In the dataset preparation step, we constructed positive and negative samples based on *P. falciparum* PPI data from BioGRID as well as the Swiss-Prot database and divided the dataset into a training set (80%) and an independent test set (20%). Furthermore, we extracted protein features using the Doc2Vec encoding scheme and DPC encoding scheme. The Doc2Vec model was trained on the compiled protein corpus covering sequences from Swiss-Prot, *P. falciparum* PPI dataset, and proteomes of *P. falciparum* and *C. parvum*. Based on the encoded feature vectors of the *P. falciparum* PPI dataset, we further trained the ML classification model using the RF algorithm and assessed the model's performance. Finally, we transferred the trained ML model to predict *C. parvum* PPIs and combined two other traditional methods (i.e., IM and DDI inference) to obtain the high-confidence *C. parvum* PPI network.

pling of the negative samples, we randomly selected protein pairs from the *P. falciparum* (isolate 3D7) interaction dataset and its proteome, ensuring the selected protein pairs will not occur in the positive samples. Here, an unbalanced ratio of 1:10 positives to negatives was set. In other words, the number of negative samples is 19,680. Moreover, we randomly divided the samples into a training set (80%) and an independent test set (20%) for model training and assessment separately (Supplementary Table S1). *C. parvum* and *P. falciparum* belong to different parasite species. To confirm the transferable generalization ability of the ML model trained on *P. falciparum* dataset to predict *C. parvum* PPIs, we also used another *P. falciparum* dataset partition, in which we divided samples into a training set and an independent test set including 800/8,000 and 200/2,000 positive/negative samples, respectively (Supplementary Table S1). Briefly, in this novel *P. falciparum* dataset partition, proteins in the independent test set will not occur in the training set, i.e., each protein being tested is equivalent to a novel protein unseen in the trained model. Therefore, the performance on the novel dataset partition can indirectly reflect the transferable generalization ability of the model testing on a novel parasite species.

2.2. Document to vector (Doc2Vec) model

Doc2Vec adopts an unsupervised embedding learning framework and trains the model based on the hypothesis that a series of protein sequences constitute a 'document' (also called a corpus). Thus, each sequence represents a sentence in a certain biological language suggesting its biological functions can be semantically interpreted [32]. In our previous study regarding the human-virus PPI prediction issue [32], we employed Doc2Vec to convert protein sequences into fixed-dimensional feature vectors for RF classifier training. The results showed that the Doc2Vec encoding improved model prediction performance and outperformed some traditional sequence-based encoding schemes (e.g., AC and LD) [32]. To implement the Doc2Vec encoding, here we first used the protein sequences from the Swiss-Prot database [40] with a length between 30 and 5,000 amino acids to establish a complete corpus (i.e., training data). Then, we removed the redundancy of the above database by using CD-HIT [41] (sequence identity $\leq$ 0.5). In addition, we also added the protein sequences from our positive/negative samples and the proteomes of *C. parvum* and *P. falciparum* (isolate 3D7). When finishing the above steps, the non-redundant protein sequences were compiled as a corpus for the Doc2Vec model training. Specifically, we split each protein sequence into several small k-mer residue segments regarded as biological words. Next, these completed protein sequences (sentences) and k-mer residue segments (words) were utilized to train the Doc2Vec model. The model made full use of the distributed-memory (DM) model architecture to let us describe each residue segment through the sentence vector and context words. Iteratively, we utilized stochastic gradient descent (SGD) [42] and backpropagation to update model parameters. Finally, we considered the output sentence vectors as the protein sequence features.

We used the Python library Gensim [43] to train the Doc2Vec model. The hyperparameters (e.g., k-mers and the dimensionality of output vectors) were optimized by the 5-fold cross-validation [44]. In particular, we trained multiple RF classifiers based on feature vectors extracted from different Doc2Vec models that were trained on different lengths of k-mers (2 to 7) and different dimensions of output vectors (16, 32, 64, 128, 256) to obtain optimal parameters of the Doc2Vec model. The final dimension of the Doc2Vec encoding for a protein pair is 128 (64 × 2) after parameter optimization.

2.3. Random Forest algorithm and parameter optimization

Random Forest (RF) is a popular decision tree-based ensemble ML algorithm [45]. In general, the RF model always has a comparatively more robust and perfect performance than other frequently-used ML methods in the issue of PPI prediction [32]. Therefore, based on the *P. falciparum* PPI dataset, we utilized the RF algorithm to train the model. We used a variety of bootstrap samples of the raw data ('bagging') to construct the classification trees. Then, when the classification trees are constructed by isolating each node and using the best among a predictor subnet randomly chosen at that node ('boosting'), RF will change accordingly. We mainly optimized three parameters for RF model training, including 'n_estimators' (the number of trees in the forest), 'max_depth' (the maximum depth of the decision trees), and 'criterion' (feature selection method). The optimal range of these parameters is [100, 500, 1,000, 1,500], [10, 50, 100, 200] and ['entropy', 'gini'], respectively. The above RF algorithm is implemented by a Python library called Scikit-learn [46]. We used 5-fold cross-validation and regarded the 'neg_log_loss' scoring function as an assessment criterion for various sequence encoding schemes-based RF algorithms. Besides, we took advantage of the 'GridSearchCV' function to optimize all parameters [47].

2.4. Other traditional sequence-based encoding schemes

In addition to the Doc2Vec, we also utilized three other sequence-based encoding schemes to train the RF model.

2.4.1. Di-peptide composition (DPC)

DPC represents the ratio of two continuous amino acids composition in the whole protein sequence. The concrete formula is: $S_{DPC}(A_iA_j) = \frac{N_{A_iA_j}}{L-1}, i,j \in (1,2,\ldots,20)$, where A_i and A_j represents 20 standard amino acids separately, $N_{A_iA_j}$ is the number of specific di-peptide in a protein sequence, and L is the length of the corresponding protein sequence. Therefore, the final dimension of DPC encoding for a protein pair is 800 (20 × 20 × 2).

2.4.2. Local Descriptor (LD)

We divided the protein sequence into several subdomains by utilizing the LD encoding scheme and extracted each subdomain's traits, which mainly captures the local feature of the protein [48]. First, we divided the 20 standard amino acids into seven classes (AGV, DE, FILP, HNQW, KR, MSTY, and C) based on the physicochemical properties of the amino acids' side chains. In addition, the protein sequence was separated into ten different regions, and each of them was expressed by three traits [i.e., Composition (C), Transition (T), and Distribution (D)]. Among them, C represents the composition of each class of amino acids, and its dimensionality is 7. T reflects the composition of any two classes of amino acids, and its dimensionality is 21 (6 × 7/2). D represents the distribution (i.e., the first, 25%, 50%, 75%, and 100%) of each class of amino acids, and its dimensionality is 35 (5 × 7). As a consequence, the LD encoding transforms a protein pair into a 1,260-dimensional [(7 + 21 + 35) × 10 × 2] vector.

2.4.3. Auto covariance (AC)

AC encoding takes into account the interaction effect of amino acids spaced at a certain distance. Herein, we employed seven physicochemical properties i.e., hydrophobicity (H1), hydrophilicity (H2), net charge index of side chains (NCI), polarity (P1), polarizability (P2), solvent accessible surface area (SASA), and volume of side chains (V) to represent the protein feature [34]. For each protein sequence, corresponding AC features can be inferred by:

$$S_{AC}(lag,j) = \frac{1}{L-lag}\sum_{i=1}^{L-lag}\left(\mathrm{R}_{i,j} - \frac{1}{L}\sum_{k=1}^{L}\mathrm{R}_{k,j}\right) \times \left(\mathrm{R}_{(i+lag),j} - \frac{1}{L}\sum_{k=1}^{L}\mathrm{R}_{k,j}\right),\ \ j \in (1,2,\ldots,7)$$

In the above formula, i and k represent the i^{th} and k^{th} residue in the protein sequence separately, j stands for one of the seven physicochemical features, $\mathrm{R}_{i,j}$ and $\mathrm{R}_{k,j}$ represent the j^{th} physicochemical feature of the i^{th} and k^{th} residue, and lag is the distance between the i^{th} residue and its adjacent residue, ranging from 1 to 30. Finally, the dimensionality of AC encoding scheme is 420 (30 × 7 × 2).

2.4.4. Combination of encoding schemes

In addition to solely using Doc2Vec, DPC, LD, and AC encoding schemes, we also used the different combinations among them by concatenating their sequence encodings. Firstly, we tried two encoding schemes (i.e., Doc2Vec + LD, Doc2Vec + AC, Doc2Vec + DPC, LD + AC, LD + DPC, AC + DPC). Secondly, we utilized three encoding schemes (i.e., Doc2Vec + LD + AC, Doc2Vec + LD + DPC, Doc2Vec + AC + DPC). Finally, we used all four encoding schemes (i.e., Doc2Vec + LD + AC + DPC).

2.5. Performance evaluation

We conducted a 5-fold cross-validation and an independent test to evaluate the performance of various models. We adopted two metrics to comprehensively assess the models, including the areas under Receiver Operating Characteristic (ROC) curve and Precision-Recall (PR) curve (i.e., AUC and AUPRC, respectively). PR curve and AUPRC have been proved to be more suitable for assessing the prediction model when the ratio of positive-to-negative samples is imbalanced. Generally, the closer the value of AUC/AUPRC is to 1, the better the performance of the PPI prediction model is. We used the R package ROCR to plot ROC and PR curves [49].

2.6. Other traditional prediction methods

2.6.1. The IM method

The IM method predicts interactions based on the homology between protein sequences of known PPIs and unknown protein pairs. Firstly, we downloaded all PPIs of various organisms from five public protein interaction databases including IntAct [50], BioGRID [39], MINT [51], DIP [52] and HPIDB [53] to obtain an interaction template library. Subsequently, we used the scoring strategy of HIPPIE [54] to compute the quality scores of the template interactions according to experimental detection techniques, the number of involved species, and the number of references reporting the template interaction. Furthermore, we used BLAST to align sequences of *C. parvum* proteome against all the sequences in the PPI template library to obtain homologs. Specifically, we used the homology thresholds: sequence identity ≥ 30% and alignment coverage of query protein ≥ 40%. Thirdly, according to our previous study, we calculated the IM probability scores for proteome-scale *C. parvum* protein pairs [37].

2.6.2. The DDI inference method

The DDI inference method predicts PPIs based on the detected interacting domain pairs. First of all, we obtained the Pfam domains for interacting proteins by domain scanning using HMMER [55] (E-value $\leq 10^{-5}$). Next, we identified co-occurrence domain pairs of known protein interactions to construct a DDI library. Domains of *C. parvum* protein pairs were also retrieved by domain scanning based on the Pfam database using the same E-value cut-off. Similar to the IM method, each domain pair in the DDI template library was assigned a confidence score through the expectation maximization (EM) algorithm. Finally, we also retrieved DDI probability scores for proteome-scale *C. parvum* protein pairs [37].

2.7. Network analysis

The enrichment of Gene Ontology (GO) terms [56] and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [57] were performed using DAVID 6.8 (https://david.ncifcrf.gov/). When a PPI network performs its function, some densely-connected subnetworks, also termed as functional modules, play important biological roles [58]. We utilized one plug-in called MCODE [59] in the Cytoscape software [60] to identify the potential functional modules, and the default parameters were set. All of them are clustered in GO terms and KEGG pathways by using ClueGO [61].

3. Results and discussion

3.1. The performance of Doc2Vec-based RF model

According to the previous work [32], the ML algorithm RF has a better performance in PPI prediction, and a new sequence embedding technique Doc2Vec has a robust performance in interspecies PPI prediction. Thus, we employed Doc2Vec to encode protein sequences and further trained the RF classifier based on *P. falciparum* (isolate 3D7) PPI dataset. We utilized a 5-fold cross-validation to optimize the parameters (i.e., k-mers, window size, epoch, and vector size) of the Doc2Vec model by cmparing corresponding RF models for *P. falciparum* (isolate 3D7) PPI prediction. In particular, we set the optimization baseline for k, window size, vector size, and epoch are 5, 3, 32, and 70, respectively, which were used and shown relatively superior performance in our previous work [32]. Subsequently, we optimized one of them while keeping other parameters unchanged and obtained the best one to replace. The order and optimization range of the parameters are shown in Supplementary Table S2. For instance, the performance of RF models under different k-mers can be seen in Supplementary Table S3. As a result, the combination of Doc2Vec with 3-mers, window size 5, vector size 64 and 70 epochs, and RF (Doc2Vec + RF) provided the best performance where the corresponding AUC and AUPRC values were 0.961 and 0.770 in the 5-fold cross-validation. Moreover, we applied an independent test set to assess the RF model in which Doc2Vec + RF achieved an AUC = 0.957 and an AUPRC = 0.762.

3.2. Comparison with other popular sequence encoding schemes

We also compared Doc2Vec with three other traditional sequence-based encoding schemes (i.e., DPC, LD, and AC) under the computational framework of RF. To ensure a fair comparison, all the RF classifiers based on different encoding schemes were trained on the same datasets and evaluated on the same independent test sets. In this study, we mainly assessed the performance of various models depending on the AUPRC values since the ratio of positive-to-negative samples of training sets is highly unbalanced (i.e., 1:10). Firstly, we randomly chose a training set and an independent test set, which means these two sets may have the same proteins (see Materials and methods for details). Among the individual sequence encodings, we found Doc2Vec (AUPRC = 0.762) outperformed DPC (AUPRC = 0.745), LD (AUPRC = 0.710) and AC (AUPRC = 0.687) (Fig. 2**A and** Table 1). To seek the best combination of different encoding schemes, moreover, we applied various encoding combinations, and we observed that Doc2Vec + DPC showed the best performance (AUPRC = 0.768) which outperformed the corresponding performance of both individual encod-

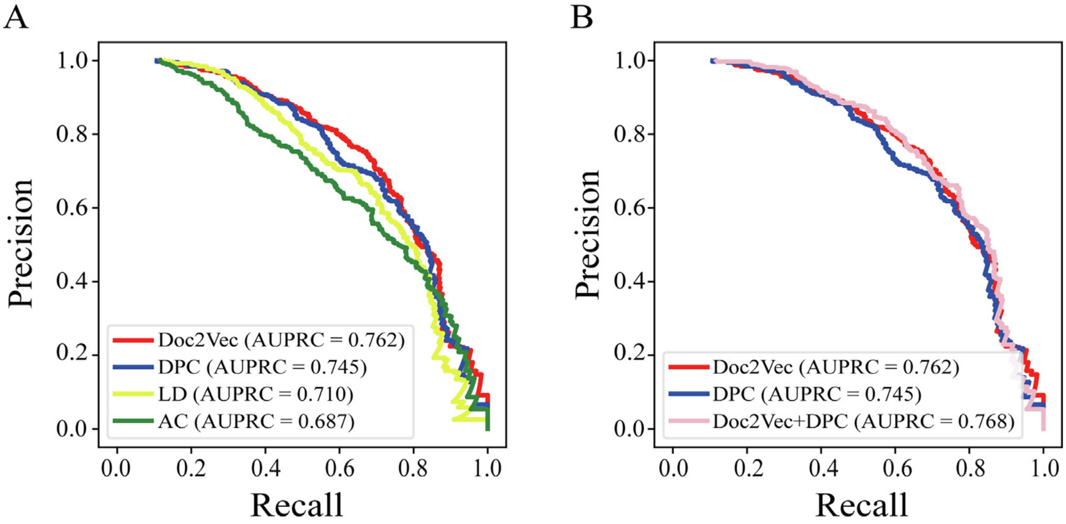

Fig. 2. (**A**) Performance of different individual sequence encoding schemes-based Random Forest (RF) classifiers in predicting *P. falciparum* PPIs. Areas under the Precision-Recall curves (AUPRC) indicate that Document to Vector (Doc2Vec) outperformed Di-peptide composition (DPC), Local Descriptor (LD), and Auto Covariance (AC) applying an independent test set. **(B)** Performance of two best individual sequence encoding schemes (i.e., Doc2Vec and DPC)-based RF classifiers and their sequence encoding combination (Doc2Vec + DPC)-based RF classifier in predicting *P. falciparum* PPIs. AUPRC indicates that the combination sequence encoding scheme slightly outperformed each individual sequence encoding scheme.

Table 1
Performance of individual or combined encoding schemes-based RF classifiers.

	Method	AUC	AUPRC
Individual encoding scheme	Doc2Vec	0.957	0.762
	DPC	0.955	0.745
	LD	0.950	0.710
	AC	0.928	0.687
Combined encoding schemes	Doc2Vec + DPC	0.963	0.768
	Doc2Vec + LD	0.958	0.742
	Doc2Vec + AC	0.957	0.752
	DPC + LD	0.954	0.732
	DPC + AC	0.953	0.741
	LD + AC	0.951	0.712
	Doc2Vec + DPC + LD	0.959	0.752
	Doc2Vec + DPC + AC	0.956	0.755
	Doc2Vec + LD + AC	0.958	0.742
	DPC + LD + AC	0.954	0.731
	Doc2Vec + DPC + LD + AC	0.959	0.749

ing schemes and other combined encoding schemes (Fig. 2**B and** Table 1). To investigate the robustness of the models, we further added another two repeats by random sampling. Corresponding average values and standard deviations of performance of models are listed in Supplementary Table S4, which suggests a robust performance of Doc2Vec + DPC in comparison to others.

While both *P. falciparum* (isolate 3D7) and *C. parvum* belong to Apicomplexa, they are not from the same species. Therefore, to assess the transfer-ability of the trained model based on the *P. falciparum* PPI dataset to predict *C. parvum* PPIs, we adopted a more rigorous method to divide another training set and independent test set. In this new dataset partition, each protein in the test set was unseen in the training set. As expected, the performance of RF models based on both individual and combined sequence encodings decreased during this assessment. Specifically, among the individual encodings, AUPRC values of Doc2Vec + RF and DPC + RF decreased around 15 and 12 percentiles but were still higher than that of RF models based on other individual encoding schemes (Fig. 3**A and** Table 2). In particular, the best RF model based on combined encodings, i.e., RF-based on Doc2Vec + DPC (AUPRC = 0.742), only slightly decreased compared to the previous model (AUPRC = 0.768), and it maintained the best combination of different encoding schemes (Fig. 3**B and** Table 2). The results suggest the two sequence encodings are highly complementary in which Doc2Vec and DPC can effectively capture context semantic information and amino acid composition information of protein sequences, respectively. We also added another two repeats of this sample partition approach, and the corresponding performance of different models is available in Supplementary Table S5.

The best combination of the encoding schemes is Doc2Vec + DPC. However, the performance dropped a little when utilizing the stricter dataset, implying that Doc2Vec + DPC-based RF model has a good generalization ability. Thus, we employed the RF model based on the combined sequence encoding, i.e., Doc2Vec + DPC, to predict proteome-scale *C. parvum* PPIs in the end. The training set and test set of the model are displayed in Supplementary Table S6. In addition, due to the lack of experimental PPIs of *C. parvum*, we used PPIs of *P. falciparum* as substitutes for training the prediction model since they both belong to Apicomplexan parasites. But it should be noted that this strategy of model training may inevitably introduce some biases since it heavily learned protein features of *P. falciparum*.

3.3. A high-confidence proteome-scale C. Parvum PPI network

To obtain highly reliable prediction data, we also employed another two traditional prediction methods (i.e., IM and DDI-inference), which can reduce the bias of the ML method introduced by *P. falciparum* training data. Therefore, a comprehensive computational framework was established by three computational biology methods (IM, DDI-inference, and ML). After utilizing these three methods, we obtained a high-confidence *C. parvum* PPI network under a false positive rate control of 5% (Fig. 4). The number of PPIs from these three methods is 119,803, 148,143, and 487,048, separately. A high-confidence *C. parvum* PPI network consisted of 554 proteins with 4,578 PPIs jointly existing in these three methods. In this constructed PPI network (see Supplementary Table S7 for the full list), the average network degree for each protein is 16.5. Generally, the proteins ranked as top high-degree in the network are defined as hub proteins, which may perform important cellular functions involved in different biological processes. Therefore, we focus on hub proteins with a high-degree in our predicted

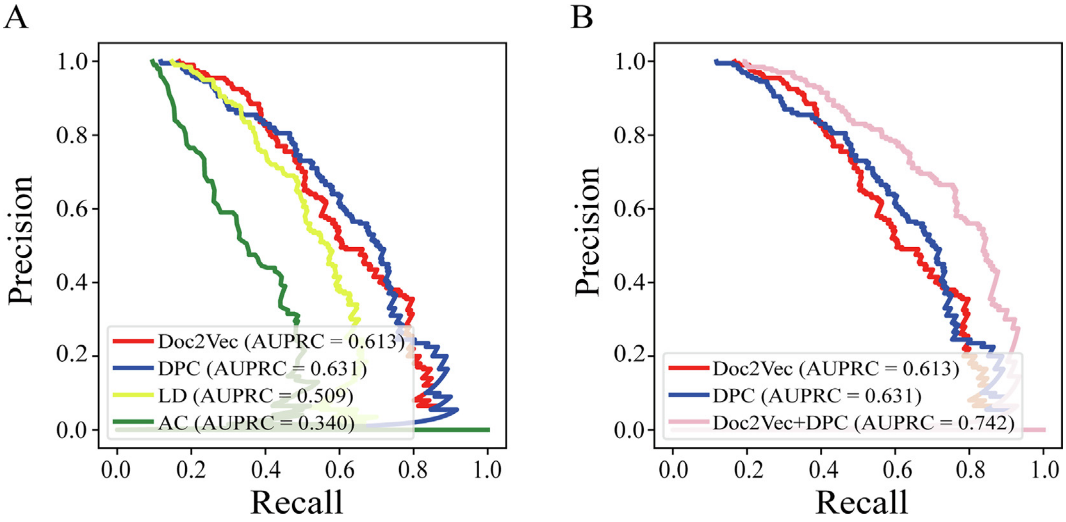

Fig. 3. **(A)** Performance of different individual sequence encoding schemes-based Random Forest (RF) classifiers in predicting *P. falciparum* PPIs based on the novel partition of the dataset (i.e., non-overlapped proteins between the training set and test set). Areas under the Precision-Recall curves (AUPRC) indicate that Di-peptide composition (DPC) outperformed Document to Vector (Doc2Vec), Local Descriptor (LD), and Auto Covariance (AC) applying an independent test set. **(B)** Performance of two best individual sequence encoding schemes (i.e., DPC and Doc2Vec)-based RF classifiers and their combination (Doc2Vec + DPC)-based RF classifier in predicting *P. falciparum* PPIs. AUPRC indicates that the sequence encoding scheme combination significantly improves the performance compared to each sequence encoding scheme.

Table 2
Performance of each individual or combined encoding schemes-based RF classifiers by using the novel dataset partition (i.e., non-overlapped proteins between training set and test set).

	Method	AUC	AUPRC
Individual encoding scheme	Doc2Vec	0.936	0.613
	DPC	0.930	0.631
	LD	0.919	0.509
	AC	0.813	0.340
Combined encoding schemes	Doc2Vec + DPC	0.960	0.742
	Doc2Vec + LD	0.942	0.635
	Doc2Vec + AC	0.949	0.700
	DPC + LD	0.928	0.558
	DPC + AC	0.929	0.615
	LD + AC	0.918	0.527
	Doc2Vec + DPC + LD	0.944	0.650
	Doc2Vec + DPC + AC	0.958	0.732
	Doc2Vec + LD + AC	0.942	0.644
	DPC + LD + AC	0.927	0.572
	Doc2Vec + DPC + LD + AC	0.945	0.656

high-confidence *C. parvum* PPI network. In particular, all of the top three high-degree hub proteins are the heat shock proteins, including 70 kDa heat shock proteins (HSP70s, encoded by cgd7_360 and cgd2_20) and 105 kDa heat shock protein (encoded by cgd4_3270), suggesting potential important functional roles in the lifecycle of the *C. parvum*. Specifically, HSP70s play important roles in various cellular protein folding and remodeling processes, and it has been widely used as the molecular marker in cryptosporidiosis epidemiology [4,62]. Simultaneously, HSPs in the *P. falciparum* are potentially functionally associated with human proteins to facilitate parasite survival and pathogenicity [63].

A total of 383 interactions with cgd7_360, 360 interactions with cgd2_20, 292 interactions with cgd4_3270, were predicted, and GO and KEGG enrichments for these protein sets indicated the potential function in the development, reproduction, and alimentation of *C. parvum* (Fig. 5). In biological process (BP) terms, protein folding, DNA replication initiation, and protein catabolic process were all significantly enriched in three protein sets. In cellular component (CC) terms, cytoplasm, proteasome complex, MCM complex were significantly enriched. In molecular function (MF) terms, ATP binding, helicase activity, GTP binding were significantly enriched. In KEGG pathways, Aminoacyl-tRNA biosynthesis is significantly enriched. In addition, these three genes are highly expressed at the intracellular stage (24 h post-infection) according to CryptoDB [64,65]. Combined with the enrichment analysis, these three hub genes in our predicted interaction network probably play crucial roles related to *C. parvum* development and infection.

3.4. Functional module analysis of C. parvum PPIs

In total, six modules were identified through MCODE [59], and the full list of PPIs involved in these modules is available in Supplementary Table S8. To explore the biological significance of the identified modules, we conducted further analysis on one module with the highest score, which consists of 26 proteins and 254 interactions (Fig. 6). Functional enrichment analysis indicated this module shares possible functions related to *C. parvum* proliferation (Fig. 6). In this module, three genes (cgd7_2920, cgd4_970, and cgd2_1600) encoding MCM proteins were enriched not only for the biological process DNA replication initiation (GO:0006261) (Fig. 6**A**) but also for the DNA replication KEGG pathway (KEGG:03030) (Fig. 6**D**). We also found evidence at the transcription level that those three genes were highly expressed in 24 h cultures [66]. *C. parvum* completes its lifecycle in a single host, and the observation of trophozoites after 24 h means the ongoing process of mitosis. Besides, six genes encoding proteasome subunits (cgd2_1350, cgd4_1170, cgd4_2540, cgd4_3950, cgd6_920 and cgd8_840) were enriched at the proteasome pathway (KEGG:03050) (Fig. 6**D**). The proteasome was reported to be responsible for the regulated degradation of intracellular proteins [67]. Several related GO terms were also enriched in part of these proteins, such as positive regulation of protein catabolic process (GO:0045732) (Fig. 6**A**) in biological process, peptidase complex (GO:1905368) (Fig. 6**B**) in cellular component and ATPase activity (GO:0016887) (Fig. 6**C**) in molecular function. Moreover, several proteins, including two encoding HSP90 proteins (cgd3_3770 and cgd7_3670), were annotated with unfolded protein binding (GO:0051082) (Fig. 6**C**). Collectively, we inferred this module might play important roles in *C. parvum* development, especially in the DNA replication stage. Similarly, the enrichment analysis

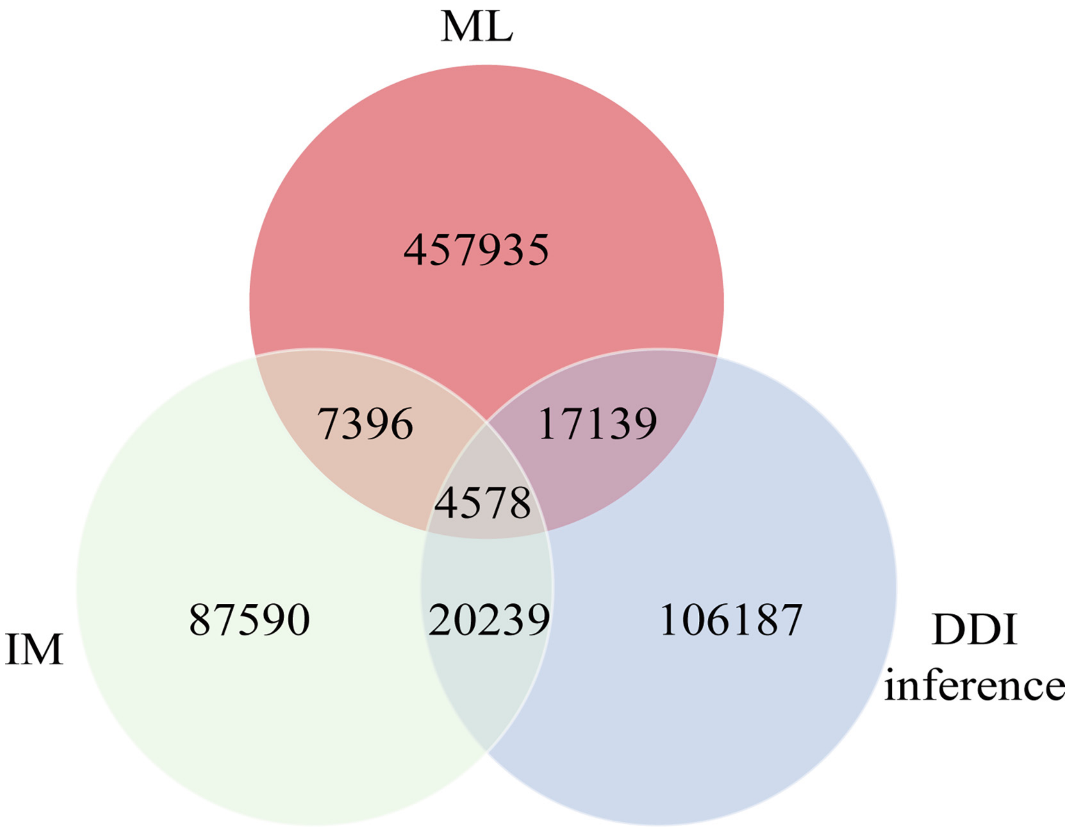

Fig. 4. Overlaps of predicted *C. parvum* PPIs among three computational prediction methods.

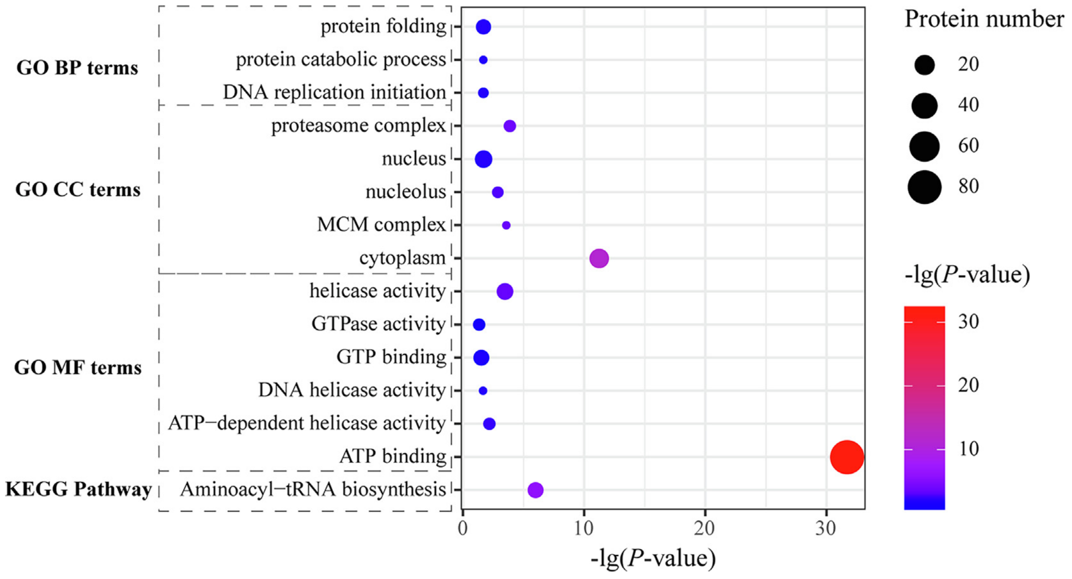

Fig. 5. Enriched GO terms and KEGG pathways of the predicted interactors of the hub proteins.

of the other five modules also indicates module functions related to *C. parvum* development (Supplementary Table S9).

Due to limited experimental progress, researches on drug targets need clues from diverse development-related pathways [68]. Several attempts have been focused on the minimalistic metabolic capacities of *Cryptosporidium*, aiming at suppressing parasite development. For instance, DNA replication is indispensable when the parasite resides in the small intestinal epithelium of different hosts. Inosine 5′-monophosphate dehydrogenase (IMPDH) is required for the biosynthesis of guanine nucleotides in *Cryptosporidium* as the inability to de novo synthesize nucleotides, making it as a promising drug target [69,70]. However, Pawlowic et al. proposed the existence of possible alternative pathway(s) to salvage nucleotides, leading to more challenges for drug development [71]. By testing the influence of *C. parvum* on host cellular metabolic signatures, Velez et al. have reported that glycolysis can

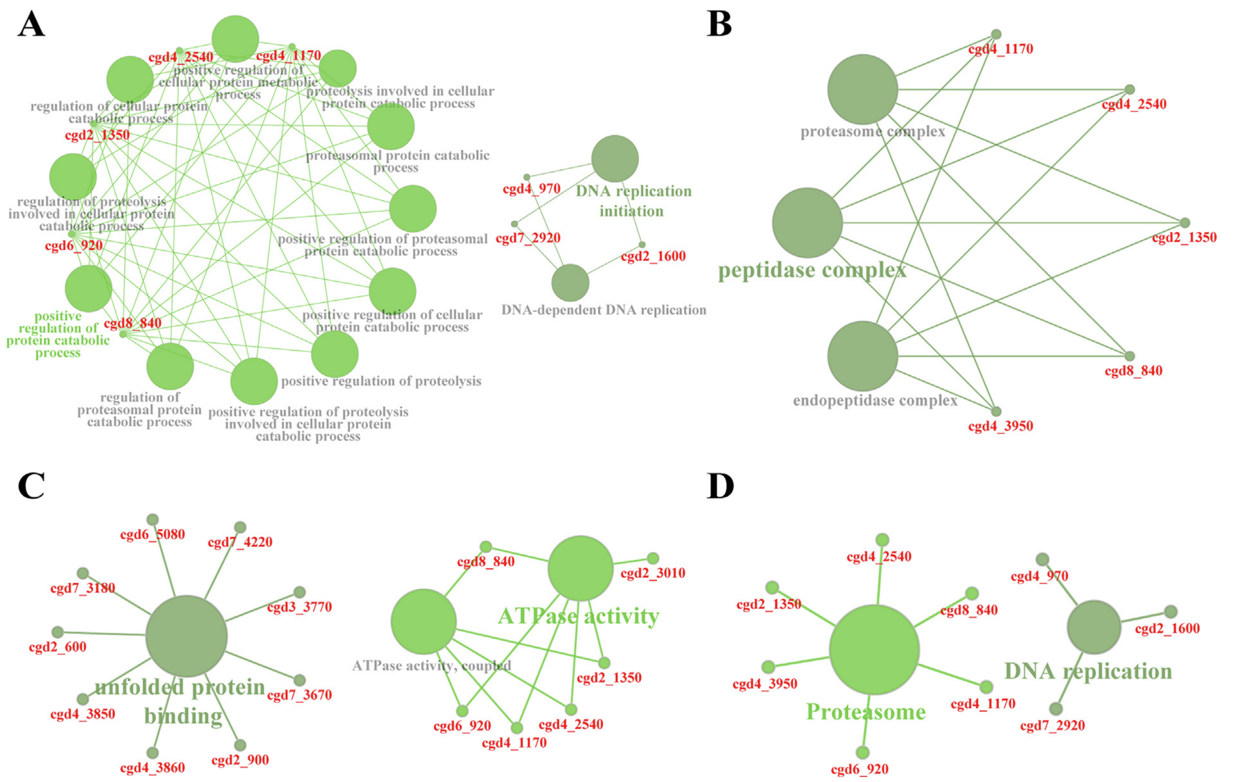
A
cgd4_2540
cgd4_1170
positive regulation of cellular protein metabolic process
proteolysis involved in cellular protein catabolic process
regulation of cellular protein catabolic process
cgd2_1350
proteasomal protein catabolic process
regulation of proteolysis involved in cellular protein catabolic process
cgd6_920
positive regulation of proteasomal protein catabolic process
positive regulation of protein catabolic process
cgd8_840
positive regulation of cellular protein catabolic process
positive regulation of proteolysis
regulation of proteasomal protein catabolic process
positive regulation of proteolysis involved in cellular protein catabolic process
cgd4_970
DNA replication initiation
cgd7_2920
cgd2_1600
DNA-dependent DNA replication
B
cgd4_1170
cgd4_2540
proteasome complex
cgd2_1350
peptidase complex
cgd8_840
endopeptidase complex
cgd4_3950
C
cgd6_5080
cgd7_4220
cgd7_3180
cgd3_3770
cgd2_600
unfolded protein binding
cgd7_3670
cgd4_3850
cgd4_3860
cgd2_900
cgd8_840
cgd2_3010
ATPase activity
ATPase activity, coupled
cgd2_1350
cgd6_920
cgd4_2540
cgd4_1170
D
cgd4_2540
cgd2_1350
cgd8_840
cgd4_3950
Proteasome
cgd4_1170
cgd6_920
cgd4_970
cgd2_1600
DNA replication
cgd7_2920

Fig. 6. GO term/KEGG pathway enrichment analysis of the DNA replication module in the categories of biological process **(A)**, cellular component **(B)**, molecular function **(C)**, and KEGG pathway **(D)**.

be taken as the anti-cryptosporidial target, and also proved glutaminolysis and lactate release as necessities of the parasite replication [72]. Moreover, gene silencing of the nucleoside-diphosphate kinase (NDK) markedly inhibited parasite development [73]. Although our network cannot directly map modules to those pathways, the modules we identified could be a good supplement as they show a high relevance to the parasite development.

4. Conclusions

C. parvum is the leading cause of waterborne and foodborne diarrhea with limited vaccine or medicine. In this study, we predicted proteome-wide *C. parvum* PPIs for the first time to fill the gap in a few experimentally validated PPIs in *Cryptosporidium*. We utilized traditional PPI prediction methods (IM and DDI inference) and the ML-based method to predict PPIs and employed the overlapping interactions as the final high-confidence PPIs (4,578 PPIs covering 554 proteins). It is worth mentioning that the ML-based method (i.e., the usage of RF and Doc2Vec encoding scheme) plays a key role in our final prediction model. To explore the value of the constructed PPI network, some essential network and functional analyses were carried out. We discovered three important hub genes (cgd2_20, cgd7_360, and cgd4_3270) encoding HSPs, which may play important roles in the cell cycle or life cycle of *C. parvum*. Thus, we suppose that an in-depth investigation on hub genes may provide new hints for the prevention of cryptosporidiosis [74]. We also found one functional module that could potentially contribute to the development of *C. parvum*. Given the lack of effective drugs in treating the cryptosporidiosis, our identified functional modules can provide clues for experimental scientists to further analyze the pathogenesis of cryptosporidiosis so as to discover drug targets. Regarding future development, we will pay more attention to another important research direction, i.e., the identification of PPIs between *C. parvum* and host cells, which can directly help to accelerate the discovery of therapeutic targets. Taken together, we hope the constructed high-confidence PPI network will become an important data resource to understand the genome-phenome relationship of *C. parvum* as well as speed up target discovery for the exploration of effective drugs/vaccines.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2021YFD1300903, 2017YFD0500404). We are grateful to Prof. Lihua Xiao and Prof. Yaoyu Feng at the College of Veterinary Medicine, South China Agricultural University for the stimulating discussion on this topic. We also thank the support of the high-performance computing platform of the State Key Laboratory of Agrobiotechnology.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.05.017.

References

[1] Zahedi A, Ryan U. *Cryptosporidium* - an update with an emphasis on foodborne and waterborne transmission. Res Vet Sci 2020;132:500–12.

[2] Naghavi M, Abajobir AA, Abbafati C, Abbas KM, Abd-Allah F, et al. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the global burden of disease study 2016. Lancet 2017;390:1151–210.

[3] Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, et al. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the global enteric multicenter study, GEMS): a prospective, case-control study. Lancet 2013;382:209–22.

[4] Feng Y, Ryan UM, Xiao L. Genetic diversity and population structure of *Cryptosporidium*. Trends Parasitol 2018;34:997–1011.

[5] Zahedi A, Durmic Z, Gofton AW, Kueh S, Austen J, et al. *Cryptosporidium* homai n. Sp. (Apicomplexa: Cryptosporidiiae) from the guinea pig (Cavia porcellus). Vet Parasitol 2017;245:92–101.

[6] Innes EA, Chalmers RM, Wells B, Pawlowic MC. A one health approach to tackle cryptosporidiosis. Trends Parasitol 2020;36:290–303.

[7] Wang RJ, Li JQ, Chen YC, Zhang LX, Xiao LH. Widespread occurrence of *Cryptosporidium* infections in patients with HIV/AIDS: epidemiology, clinical feature, diagnosis, and therapy. Acta Trop 2018;187:257–63.

[8] Insulander M, Silverlas C, Lebbad M, Karlsson L, Mattsson JG, et al. Molecular epidemiology and clinical manifestations of human cryptosporidiosis in Sweden. Epidemiol Infect 2013;141:1009–20.

[9] Stiff RE, Davies AP, Mason BW, Hutchings HA, Chalmers RM. Long-term health effects after resolution of acute *Cryptosporidium* parvum infection: a 1-year follow-up of outbreak-associated cases. J Med Microbiol 2017;66:1607–11.

[10] Osman M, Benamrouz S, Guyot K, Baydoun M, Frealle E, et al. High association of *Cryptosporidium* spp. infection with colon adenocarcinoma in Lebanese patients. PLoS One 2017;12. e0189422.

[11] Olson ME, O'Handley RM, Ralston BJ, McAllister TA, Thompson RC. Update on *Cryptosporidium* and giardia infections in cattle. Trends Parasitol 2004;20:185–91.

[12] Abreu BS, Pires LC, Santos KR, Luz CSM, Oliveira MRA, et al. Occurrence of *Cryptosporidium* spp. and its association with ponderal development and diarrhea episodes in nellore mixed breed cattle. Acta Veterinaria Brasilica 2019;13:24–9.

[13] Jacobson C, Williams A, Yang R, Ryan U, Carmichael I, et al. Greater intensity and frequency of *Cryptosporidium* and giardia oocyst shedding beyond the neonatal period is associated with reductions in growth, carcase weight and dressing efficiency in sheep. Vet Parasitol 2016;228:42–51.

[14] Zahedi A, Monis P, Gofton AW, Oskam CL, Ball A, et al. *Cryptosporidium* species and subtypes in animals inhabiting drinking water catchments in three states across Australia. Water Res 2018;134:327–40.

[15] Certad G, Follet J, Gantois N, Hammouma-Ghelboun O, Guyot K, et al. Prevalence, molecular identification, and risk factors for *Cryptosporidium* infection in edible marine fish: a survey across sea areas surrounding France. Front Microbiol 2019;10:1037.

[16] Grinberg A, Markovics A, Galindez J, Lopez-Villalobos N, Kosak A, et al. Controlling the onset of natural cryptosporidiosis in calves with paromomycin sulphate. VET REC 2002;151:606–8.

[17] Sparks H, Nair G, Castellanos-Gonzalez A, White Jr AC. Treatment of *Cryptosporidium*: what we know, gaps, and the way forward. Curr Trop Med Rep 2015;2:181–7.

[18] Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 2020;583:459–68.

[19] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 2000;403:623–7.

[20] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 2001;98:4569–74.

[21] Ding Z, Liang J, Lu Y, Yu Q, Songyang Z, et al. A retrovirus-based protein complementation assay screen reveals functional akt1-binding partners. Proc Natl Acad Sci U S A 2006;103:15014–9.

[22] Mellacheruvu D, Wright Z, Couzens AL, Lambert JP, St-Denis NA, et al. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. Nat Methods 2013;10:730–6.

[23] Nguyen HH, Park J, Kang S, Kim M. Surface plasmon resonance: a versatile technique for biosensor applications. Sensors (Basel) 2015;15:10481–510.

[24] Keller S, Vargas C, Zhao H, Piszczek G, Brautigam CA, et al. High-precision isothermal titration calorimetry with automated peak-shape analysis. Anal Chem 2012;84:5066–73.

[25] Relat RMB, O'Connor RM. *Cryptosporidium*: host and parasite transcriptome in infection. Curr Opin Microbiol 2020;58:138–45.

[26] Li M, Zhang X, Gong P, Li J. *Cryptosporidium* parvum rhomboid1 has an activity in microneme protein CpGP900 cleavage. Parasit Vectors 2016;9:438.

[27] Markowetz F. All biology is computational biology. PLoS Biol 2017;15: e2002050.

[28] Pavithra SR, Kumar R, Tatu U. Systems analysis of chaperone networks in the malarial parasite *Plasmodium falciparum*. PLoS Comput Biol 2007;3:1701–15.

[29] Lee H, Deng M, Sun F, Chen T. An integrated approach to the prediction of domain-domain interactions. BMC Bioinf 2006;7:269.

[30] Akiva E, Friedlander G, Itzhaki Z, Margalit H. A dynamic view of domain-motif interactions. PLoS Comput Biol 2012;8:e1002341.

[31] Xu C, Jackson SA. Machine learning and complex biological data. Genome Biol 2019;20:76.

[32] Yang X, Yang S, Li Q, Wuchty S, Zhang Z. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. Comput Struct Biotechnol J 2020;18:153–61.

[33] Manavalan B, Shin TH, Kim MO, Lee G. AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest. Front Pharmacol 2018;9:276.

[34] Guo Y, Yu L, Wen Z, Li M. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res 2008;36:3025–30.

[35] Chen J, Shan S, He C, Zhao G, Pietikainen M, et al. WLD: a robust local image descriptor. IEEE Trans Pattern Anal Mach Intell 2010;32:1705–20.

[36] Templeton TJ, Iyer LM, Anantharaman V, Enomoto S, Abrahante JE, et al. Comparative analysis of apicomplexa and genomic diversity in eukaryotes. Genome Res 2004;14:1686–95.

[37] Lian X, Yang X, Shao J, Hou F, Yang S, et al. Prediction and analysis of human-herpes simplex virus type 1 protein-protein interactions by integrating multiple methods. Quant Biol 2020;8:312–24.

[38] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2004;32:D115–9.

[39] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 2006;34:D535–9.

[40] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365–70.

[41] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28:3150–2.

[42] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:2481–95.

[43] Silverberg MS, Daly MJ, Moskovitz DN, Rioux JD, McLeod RS, et al. Diagnostic misclassification reduces the ability to detect linkage in inflammatory bowel disease genetic studies. Gut 2001;49:773–6.

[44] Yu NY, Wagner JR, Laird MR, Melli G, Rey S, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 2010;26:1608–15.

[45] Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics 2012;99:323–9.

[46] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[47] Dong W, Huang Y, Lehane B, Ma G. XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring. Autom Constr 2020;114:1–11.

[48] Davies MN, Secker A, Freitas AA, Clark E, Timmis J, et al. Optimizing amino acid groupings for gpcr classification. Bioinformatics 2008;24:1980–6.

[49] Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics 2005;21:3940–1.

[50] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, et al. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 2014;42:D358–63.

[51] Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 2012;40: D857–61.

[52] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, et al. The database of interacting proteins: 2004 update. Nucleic Acids Res 2004;32:D449–51.

[53] Ammari MG, Gresham CR, McCarthy FM, Nanduri B. HPIDB 2.0: a curated database for host-pathogen interactions. Database 2016;2016:1–9.

[54] Schaefer MH, Fontaine JF, Vinayagam A, Porras P, Wanker EE, et al. HIPPIE: Integrating protein interaction networks with experiment based quality scores. PLoS ONE 2012;7:e31826.

[55] Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, et al. HMMER web server: 2018 update. Nucleic Acids Res 2018;46:W200–4.

[56] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.

[57] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30.

[58] Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics 2008;24:i223–31.

[59] Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinf 2003;4.

[60] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–504.

[61] Mlecnik B, Galon J, Bindea G. Automated exploration of gene ontology term and pathway networks with ClueGO-REST. Bioinformatics 2019;35:3864–6.

[62] Rosenzweig R, Nillegoda NB, Mayer MP, Bukau B. The Hsp70 chaperone network. Nat Rev Mol Cell Biol 2019;20:665–80.

[63] Daniyan MO, Przyborski JM, Shonhai A. Partners in mischief: functional networks of heat shock proteins of *Plasmodium falciparum* and their influence on parasite virulence. Biomolecules 2019;9:295.

[64] Amos B, Aurrecoechea C, Barba M, Barreto A, Basenko EY, et al. VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. Nucleic Acids Res 2021;50:D898–911.

[65] Mirhashemi ME, Noubary F, Chapman-Bonofiglio S, Tzipori S, Huggins GS, et al. Transcriptome analysis of pig intestinal cell monolayers infected with *Cryptosporidium parvum* asexual stages. Parasit Vectors 2018;11:176.

[66] Tandel J, English ED, Sateriale A, Gullicksrud JA, Beiting DP, et al. Life cycle progression and sexual development of the apicomplexan parasite *Cryptosporidium parvum*. Nat Microbiol 2019;4:2226–36.

[67] Budenholzer L, Cheng CL, Li Y, Hochstrasser M. Proteasome structure and assembly. J Mol Biol 2017;429:3500–24.

[68] Wang B, Castellanos-Gonzalez A, White Jr AC. Novel drug targets for treatment of cryptosporidiosis. Expert Opin Ther Targets 2020;24:915–22.

[69] Jefferies R, Yang R, Woh CK, Weldt T, Milech N, et al. Target validation of the inosine monophosphate dehydrogenase (impdh) gene in *Cryptosporidium* using phylomer((r)) peptides. Exp Parasitol 2015;148:40–8.

[70] Maurya SK, Gollapalli DR, Kirubakaran S, Zhang M, Johnson CR, et al. Triazole inhibitors of *Cryptosporidium parvum* inosine 5'-monophosphate dehydrogenase. J Med Chem 2009;52:4623–30.

[71] Pawlowic MC, Somepalli M, Sateriale A, Herbert GT, Gibson AR, et al. Genetic ablation of purine salvage in *Cryptosporidium parvum* reveals nucleotide uptake from the host cell. Proc Natl Acad Sci U S A 2019;116:21160–5.

[72] Velez J, Velasquez Z, Silva LMR, Gartner U, Failing K, et al. Metabolic Signatures of *Cryptosporidium parvum*-infected hct-8 cells and impact of selected metabolic inhibitors on *C. parvum* Infection under physioxia and hyperoxia. Biology (Basel) 2021;10.

[73] Castellanos-Gonzalez A, Martinez-Traverso G, Fishbeck K, Nava S, White Jr AC. Systematic gene silencing identified *Cryptosporidium* nucleoside diphosphate kinase and other molecules as targets for suppression of parasite proliferation in human intestinal cells. Sci Rep 2019;9:12153.

[74] Kissinger JC. Evolution of *Cryptosporidium*. Nat Microbiol 2019;4:730–1.