Prediction of Human-*Plasmodium vivax* Protein Associations From Heterogeneous Network Structures Based on Machine-Learning Approach

Apichat Suratanee[1], Teerapong Buaboocha[2,3] and Kitiporn Plaimas[3,4]

[1]Department of Mathematics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand. [2]Department of Biochemistry, Faculty of Science, Chulalongkorn University, Bangkok, Thailand. [3]Omics Sciences and Bioinformatics Center, Faculty of Science, Chulalongkorn University, Bangkok, Thailand. [4]Advanced Virtual and Intelligent Computing (AVIC) Center, Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok, Thailand.

ABSTRACT: Malaria caused by *Plasmodium vivax* can lead to severe morbidity and death. In addition, resistance has been reported to existing drugs in treating this malaria. Therefore, the identification of new human proteins associated with malaria is urgently needed for the development of additional drugs. In this study, we established an analysis framework to predict human-*P. vivax* protein associations using network topological profiles from a heterogeneous network structure of human and *P. vivax*, machine-learning techniques and statistical analysis. Novel associations were predicted and ranked to determine the importance of human proteins associated with malaria. With the best-ranking score, 411 human proteins were identified as promising proteins. Their regulations and functions were statistically analyzed, which led to the identification of proteins involved in the regulation of membrane and vesicle formation, and proteasome complexes as potential targets for the treatment of *P. vivax* malaria. In conclusion, by integrating related data, our analysis was efficient in identifying potential targets providing an insight into human-parasite protein associations. Furthermore, generalizing this model could allow researchers to gain further insights into other diseases and enhance the field of biomedical science.

KEYWORDS: Network-based method, *Plasmodium vivax*, human-parasite protein association, host-parasite interaction, machine learning, ranking score, topological profiles

Introduction

Plasmodium is a parasite that has proven to be difficult to eradicate. *Plasmodium vivax* is 1 of the 5 species of the parasite group *Plasmodium* that infects humans.[1] *P. vivax* has the ability to confer virulence to humans and survive in human hosts and has been categorized as a benign infection. At present, *P. vivax* malaria is recognized as a cause of severe morbidity and mortality.[2] Approximately, 14.3 million cases of *P. vivax* infection are recorded annually.[3] Although the global incidence of *P. vivax* malaria infection has decreased by 42% since 2000, the disease burden has increased in the Middle East and South America since 2013.[4] In addition, *P. vivax* is able to evolve its strategy to interact with the host, which has led to the development of drug-resistant parasites. The first-line treatment drug for *P. vivax* is chloroquine to treat blood-stage parasitemia together with primaquine to eradicate persistent liver-stage infection.[3] However, *P. vivax* parasites resistant to their respective first-line therapies have been found in Southeast Asia.[5] Recently, tafenoquine, a promising new drug, has been highlighted as a radical cure for *P. vivax* infection. Results have shown that it resulted in a significantly lower risk of *P. vivax* recurrence than placebo in patients with normal glucose-6-phosphate dehydrogenase (G6PD) activity.[6] However, tafenoquine causes hemolysis in patients with G6PD deficiency. Therefore, there is a need for testing G6PD activity before prescription of tafenoquine.[7-9] The *Plasmodium* parasite has the ability to evade the human immune system, recruit host responses to regulate its life cycle, and adapt to the host environment.[10] Specifically, *P. vivax* invades erythrocytes during blood-stage growth in humans. Duffy antigen receptor for chemokines (DARC), which is a host receptor, is recognized by a critical invasion ligand, *P. vivax* Duffy Binding Proteins (DBP), for the invasion of immature red blood cells.[11] Therefore, DBP has been highlighted as a leading vaccine candidate against *P. vivax* malaria.[12] To control this parasite, we require a better understanding of host-parasite interactions which is crucial in the development and design of therapeutic approaches for this infectious disease.

Although recent technological advances in high-throughput techniques have enabled the characterization of proteins that may be involved in the parasitic invasion of target cells, maintaining a continuous *in vitro* culture for *P. vivax* is still very difficult to standardize.[13] This is the main obstacle to the development of a new effective vaccine. However, computational methods can be employed to solve this problem. One of the most widely used methods is a network-based approach

that focuses on protein-protein interaction (PPI) networks. The analysis of a PPI network has been widely studied in several organisms.[14-17] In *Plasmodium*, several studies have investigated the PPI networks with the aim of revealing many important aspects of protein interactions.[10,18-24] Most studies of PPI networks have applied the calculation of degree and centralities, focusing on a single organism in their analyses. In addition, PPI networks have also been used to study the associations between proteins and diseases[14,25-27] and host-parasite protein associations.[10,18,19,24,28] Saha et al[24] investigated the characteristics of a host-pathogen protein interaction network based on interconnectivity and centrality properties. They analyzed the significance of central, peripheral, hub and non-hub protein nodes in the infection process of malaria. They also found few topologically unimportant but biologically significant proteins between humans and malaria. Notably, most such studies have been performed for *Plasmodium falciparum*. Several studies have used ortholog-based methods to predict the association of proteins across species.[29-33] Specifically, Cuesta-Astroz et al[34] developed a method based on orthologous proteins to identify a transferred interaction between host and parasite proteins. They identified common and specific mechanisms of parasitic infection and survival in 15 human parasites. They also intensively analyzed the human-*Schistosoma mansoni* protein interaction network and revealed biological processes, pathways, and tissue-specific interactions that may be essential in the life cycle of the parasites. Lee et al[29] predicted PPIs between *P. falciparum* calmodulin and *H. sapiens* proteins based on orthologous pairs. From the associations between host and parasite, they found that *P. falciparum* may use calcium-modulating proteins in the host cell to maintain the Ca^{2+} levels. Recently, a heterogeneous network has been developed to propagate interaction information from the human PPI network and the *P. vivax* PPI network to infer new associations between human and *P. vivax* proteins.[19] This method was based on protein interactions that were considered to globally represent of these 2 networks. The study used protein similarities between human and parasite proteins to establish their associations; the idea behind this is that a malaria protein that is homologous to a human protein may interact or work together with human proteins to maintain their lives in the host and be related to the same set of cooperative proteins in humans. Thus, the study of the relationship between similar proteins in humans and malarial parasites is of great interest to investigate their network topology in PPI networks. Similar proteins may also have the same level of importance in the PPI, as the centrality measures reflect the essentiality of a protein in terms of the network topology and connections under a specific aspect of the measure. For example, the betweenness centrality provides an insight into a node that may be involved with the paths of communication of any pairs of nodes in the network.[17,35,36] Therefore, the integration of these network topologies for the recognition of human-parasite protein associations via machine learning has the potential to provide important insights and reveal new associations and protein targets in human hosts.

In this study, alternative properties based on local network topology features and machine-learning techniques were used to elucidate new associations between human and *P. vivax* proteins. The associations presented in this study indicate the existence of functional interactions between human and *P. vivax* proteins, implying that these proteins cooperate to perform a task in the underlying mechanisms. A ranking technique was also developed to predict potential protein targets in humans which may be important for the treatment of *P. vivax* malaria. Clustering analysis was performed using information from the heterogeneous network analysis to identify groups of related proteins and functional proteins. Finally, a list of human proteins that are crucial for the cellular mechanisms of *P. vivax* was reported and validated via a literature search. This list may be useful in further studies that wish to develop drugs for the treatment of *P. vivax*.
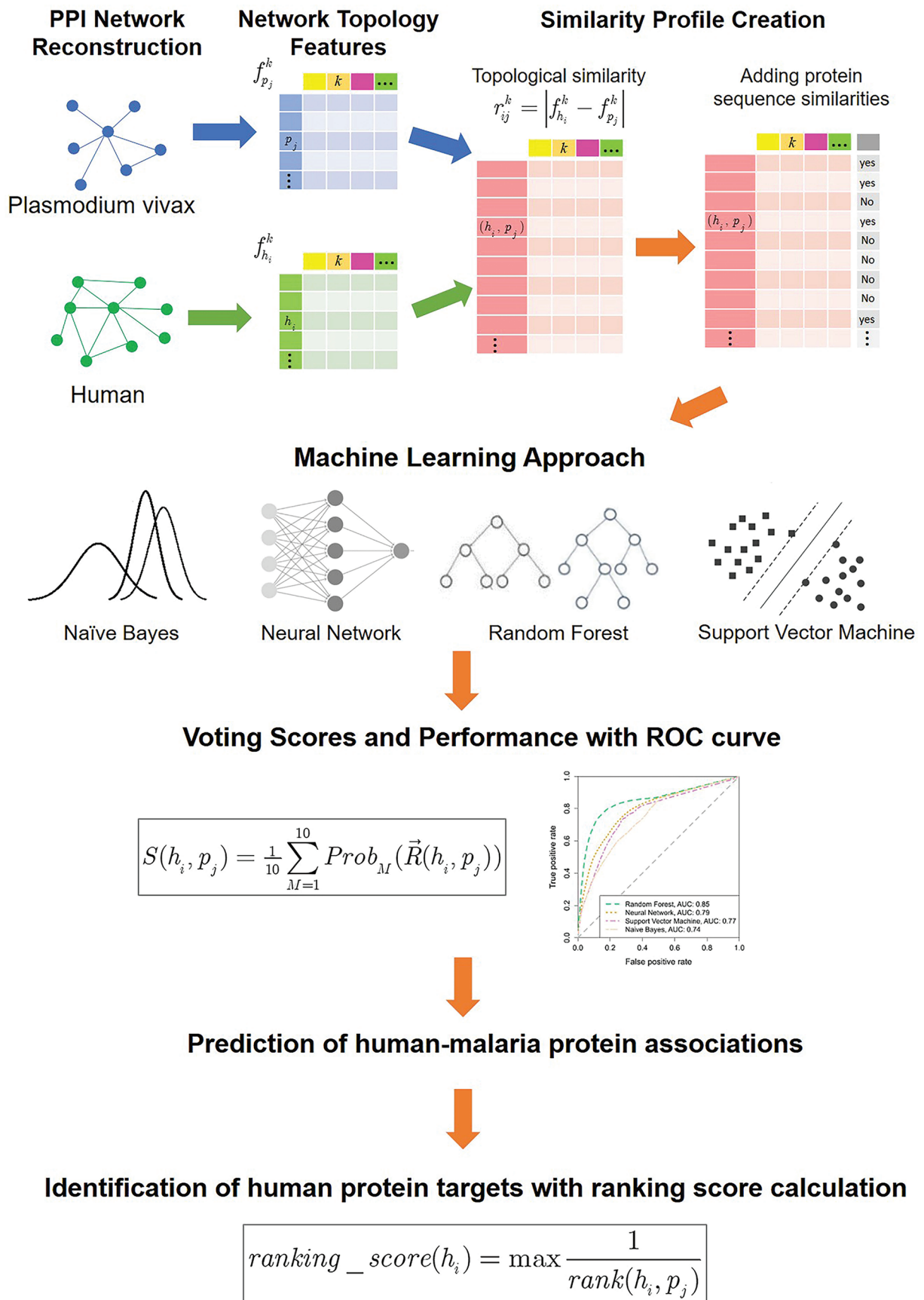
Materials and Methods

Overview of the analysis framework

The analysis framework was initiated with the network reconstruction process as shown in Figure 1. First, PPI networks for humans and malarial parasites were constructed based on the interaction information obtained from the STRING database.[37] Each protein node in each network was then extracted for its network topological features such as the degree and the betweenness centrality. Subsequently, both networks were linked together to form a heterogeneous network based on their protein sequence similarity. Then, the topological features of a pair of human and malaria proteins were compared and evaluated to obtain the strength of the differences and to build a similarity profile of the human-parasite protein pairs. The protein sequence similarities obtained from *BlastP* searches (E-value ⩽ 1e–05) were then used as an initial class label of a pair of human and *P. vivax* proteins. The complete profile was then applied to various machine-learning techniques (naïve Bayes, neural network, random forest, and support vector machine). Cross-validations were performed for each technique, and the performances were measured using the receiver operator characteristic (ROC) curve. The top classifiers from the best technique were selected as models to predict new potential associations. Finally, the human proteins in the list of predicted associations were ranked to identify potential protein targets for malaria invasion in the human host.

Network construction and topology features

Our analysis was performed on PPI networks of human proteins and *P. vivax* proteins. The networks were obtained from the STRING database (version 11.0).[37] To ensure that only reliable interactions were obtained, interactions with a high

PPI Network Reconstruction

Network Topology Features

Similarity Profile Creation

Topological similarity

$$r_{ij}^k = \left| f_{h_i}^k - f_{p_j}^k \right|$$

Adding protein sequence similarities

$f_{p_j}^k$

p_j

Plasmodium vivax

$f_{h_i}^k$

h_i

Human

(h_i, p_j)

(h_i, p_j)

Machine Learning Approach

Naïve Bayes   Neural Network   Random Forest   Support Vector Machine

Voting Scores and Performance with ROC curve

$$S(h_i, p_j) = \frac{1}{10} \sum_{M=1}^{10} Prob_M(\vec{R}(h_i, p_j))$$

Random Forest, AUC: 0.85
Neural Network, AUC: 0.79
Support Vector Machine, AUC: 0.77
Naïve Bayes, AUC: 0.74

True positive rate

False positive rate

Prediction of human-malaria protein associations

Identification of human protein targets with ranking score calculation

$$ranking\_score(h_i) = \max \frac{1}{rank(h_i, p_j)}$$

Figure 1. Analytical framework. An overview of the identification processes to infer human protein targets from human-parasite protein associations obtained using machine-learning methods with network topology features.

confidence score ($>$900) were retained. A total of 12 038 human proteins with 313 359 interactions and 1787 *P. vivax* proteins with 11 477 interactions were obtained. Subsequently, a heterogeneous network was constructed by connecting human-human protein interactions and *P. vivax-P. vivax* protein interactions with the human-*P. vivax* protein associations.

The network topology features of all proteins were extracted based on centrality measurements. Several studies have shown that a relationship exists between gene essentiality and network centrality in PPI networks.[38-40] Thus, we further investigated 5 topological features: betweenness centrality, closeness centrality, degree, eccentricity, and Kleinberg's hub centrality. Each of these features explained different aspects of the measurement. Betweenness centrality reflects an important node in term of overloading paths passing through it in the communication of the network.[35,36] Closeness centrality measures how close a given node is to the other nodes in the network.[35,36] The degree represents the level of the local connections of a given node.[35,36] Eccentricity calculates the local density of the connections among neighboring nodes of a given node. The Kleinberg's hub measures the importance of a given node connecting the other important nodes.[36]

Defining the human–P. vivax protein associations

To define the initial associations between human and *P. vivax* proteins, we used the information obtained from a sequence similarity search. When 2 protein sequences shared significant similarity with the *BlastP* expectation value (E-value) less than 1e−05, they were inferred to be homologous. This means that they did not arise independently, but rather shared a common ancestor.[41] Therefore, we could define an association between 2 sequences when they share more similarity than that would be expected by chance. However, when no statistically significant match was found between the 2 protein sequences, we could not ensure that no homologs were present. Thus, the machine-learning method may be able to reveal hidden homologs. The *P. vivax* protein sequences were retrieved from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database[42,43] using the Rcpi package[44] and then searched against all human protein sequences from the NCBI database. We defined that 2 protein sequences were homologous when *BlastP* (https://blast.ncbi.nlm.nih.gov) gave rise an E-value less than 1e−05. Then the pair of these 2 proteins was labeled to be associated.

In addition, the relationship between network topologies and functions has been revealed in several studies with the assumption that for each function, the wiring patterns of the proteins are similar.[45] Different standard network topologies can be used to understand the information contained in the wiring of a protein in the PPI.[45,46] Therefore, we integrated initial associations from the protein sequence similarity search and the similarities from network topological features and fed them into machine-learning algorithms to predict new associations using both types of similarity information. It is worth noting that our method is a homology-based method that relies on sequence similarity, similar to previous studies.[29-34] Protein associations were predicted based on the initial associations from sequence similarity. Moreover, homology-based methods have been used to infer functionally interacting proteins in previous studies.[29-34]

Features of topological differences for machine learning

Based on the 5 network topology features, we established a vector $\vec{R}$, that is a similarity profile, representing a relationship between the topological values of a human protein (h_i) and a *P. vivax* protein (p_j), as follows

$$\vec{R}\left(h_i, p_j\right) = \left(r_{ij}^k\right) \tag{1}$$

where $r_{ij}^k = |\, f_{h_i}^k - f_{p_j}^k\,|$, $i = 1, 2,\ldots, m$ and $j = 1, 2,\ldots, n$. m and n are the number of human and *P. vivax* proteins, respectively. k is the index for each topological feature, ranging from 1 to 5. $f_{h_i}^k$ represents the kth centrality value of a human protein h_i and $f_{p_j}^k$ represents the kth centrality value of a *P. vivax* protein, p_j. Therefore, r_{ij}^k denotes the topological similarity between the kth centrality values of human protein i and *P. vivax* protein j. A low value of r_{ij}^k indicates a high similarity between the topological features k of these 2 different types of proteins.

Training and validating of the association classifiers and calculating association scores

We investigated all possible pairs of proteins to identify human-parasite protein associations. To this end, we employed machine-learning techniques to classify defined and undefined associations. Four classification algorithms, namely naïve Bayes, neural network, random forest, and support vector machine algorithms, were employed. Each of these classifiers is a well-known algorithm for recognizing and creating classifiers in different ways. The naïve Bayes' approach uses the statistics and likelihoods to make a final decision. A neural network calculates a set of optimal weights for a weighted network structure to separate different classes based on the features. Random forest creates complex and hierarchical rules along the features to provide a predicted class. The support vector machine builds a hyperplane to identify an optimal classifier with maximum margin. With the different calculation methods to search for the best solution for the classifier, all 4 classifiers were applied to search for the best classifier. Different parameters of each algorithm were optimized to determine the optimal models of each algorithm.

For the naïve bayes classification, we tuned 3 hyperparameters. The first parameter was to allow to use a kernel density

estimation or a Gaussian density estimation. The second parameter was used to adjust the bandwidth of the kernel density when using kernel density estimation. Using this parameter, we optimized it from 0 to 5. The third parameter was the parameter for the *Laplace* smoother, which we tuned from 0 to 5.

For neural networks, we optimized the number of units in the hidden layers (H) and weight decay to avoid overfitting (d) by employing a grid search with H = 1, 2, 3,. . ., 10 and d = 0.5, 0.1, 1e–2, 1e–3, 1e–4, 1e–5, 1e–6, and 1e–7. The maximum iterations were set to 1000.

For the random forest algorithm, we varied the number of variables randomly sampled at each split time with a value of $2n$ for $n \in \{0, 1, 2, 3, 4, 5\}$.

For the support vector machine, we used a radial basis kernel, and optimized the cost of false classification (C) and kernel width (γ) by employing a grid search with C = {0.75, 1.0, 1.25} and γ = {0.01, 0.015, 0.2}.

Ten 10-fold cross-validations were performed to evaluate the performance of the classifiers. At each time, the undefined association set was randomly selected with an equal size to the defined set. A total of 80% of these data were used to optimize the parameters using the cross-validation technique. At each time of the cross-validation, the defined and undefined associations were randomly split into 10 equal sizes. Nine parts were concatenated and used to train and optimize the parameters. Testing was performed with the remaining part and the performance was measured by comparing the predictions and the true class labels. This experiment was repeated with a randomly undefined set 10 times. Several cutoffs on the probabilities of positive class predictions were calculated, yielding an ROC curve, which is a plot of the true-positive rate (TPR) against the false-positive rate (FPR) at the different cutoffs. Using the ROC curve, a broader view of the performance over various cutoffs could be measured by calculating the area under the curve (AUC). An AUC of 1 indicated the best performance of the classifier in which it can recognize and classify the samples, whereas an AUC of 0.5 indicated that the performace could achieve the same as random prediction by chance.

Subsequently, the AUCs of the aforementioned 4 classification algorithms were compared. The algorithm with the highest AUC was used as the prediction model. Ten classifiers from the final model were employed as the ensemble classifiers. Each classifier provided the probabilities of positive prediction for a human-parasite protein pair. The voting score (S) was calculated from the average probabilities of the 10 classifiers. Therefore, the score was computed as follows

$$S\left(h_i, p_j\right) = \frac{1}{10} \sum_{M=1}^{10} \text{Prob}_M \left(\vec{R}(h_i, p_j)\right) \quad (2)$$

where Prob_M is the probability of a positive prediction derived from the output of the *M*th machine. The score was applied to all defined and undefined associations in this study.

Ranking score calculation for each human protein

Using machine-learning algorithms to perform the classifications, we obtained a promising list of human-parasite protein associations. It would be interesting to use these associations to identify human proteins crucial for the *P. vivax* malaria mechanism. It is worth noting that one human protein could be associated with more than 1 *P. vivax* protein. To identify the impact of a human protein on the list, we applied a ranking method for all human proteins in the list. The probability of a positive prediction for a pair of human and *P. vivax* proteins was used to rank the protein pairs. The pair with the highest probability value was ranked first. Notably, several pairs can have the same probability value. In this case, they were assigned the same rank. The ranking score of a human protein h_i was calculated as follows

$$ranking\_score\left(h_i\right) = \max \frac{1}{rank\left(h_i, p_j\right)} \quad (3)$$

where $rank(h_i, p_j)$ is the rank of a pair of a human protein h_i and *P. vivax* protein p_j, for all possible p_j, according to the prediction probability score of the association.

Gene ontology enrichment analysis

To infer gene functions from the human candidate sets, we employed Gene Ontology (GO) enrichment analysis to determine which GO terms were overrepresented in our candidate proteins. To this end, the Cytoscape 3.7.2[47] plugin ClueGO v2.5.6[48] was used. ClueGO constructed a gene network based on GO terms by employing all differentially expressed genes. A 2-sided hypergeometric test with Benjamin-Hochberg corrections was performed to calculate the significant GO terms. Only GO terms with adjusted *p*-values less than 0.05 were considered.

Results
Network structures and node properties of human and P. vivax networks

In this study, we constructed 2 PPI networks of human and *P. vivax* from the information of the STRING database.[37] The reconstructed human PPI network consisted of 12 038 proteins and 313 359 edges, while the malaria PPI network comprised 1787 proteins and 11 477 edges. The structures of the human PPI network and malaria PPI network followed the power-law distribution (Figure 2A and B, respectively), indicating that there are small numbers of high-degree nodes and large numbers of low-degree nodes in the networks. The topological network features of each protein were calculated based on node properties in the networks, namely betweenness centrality, closeness centrality, degree, eccentricity, and Kleinberg's hub. The deviations of these features are shown as boxplots in Figure 3. Interestingly, both networks had similar average
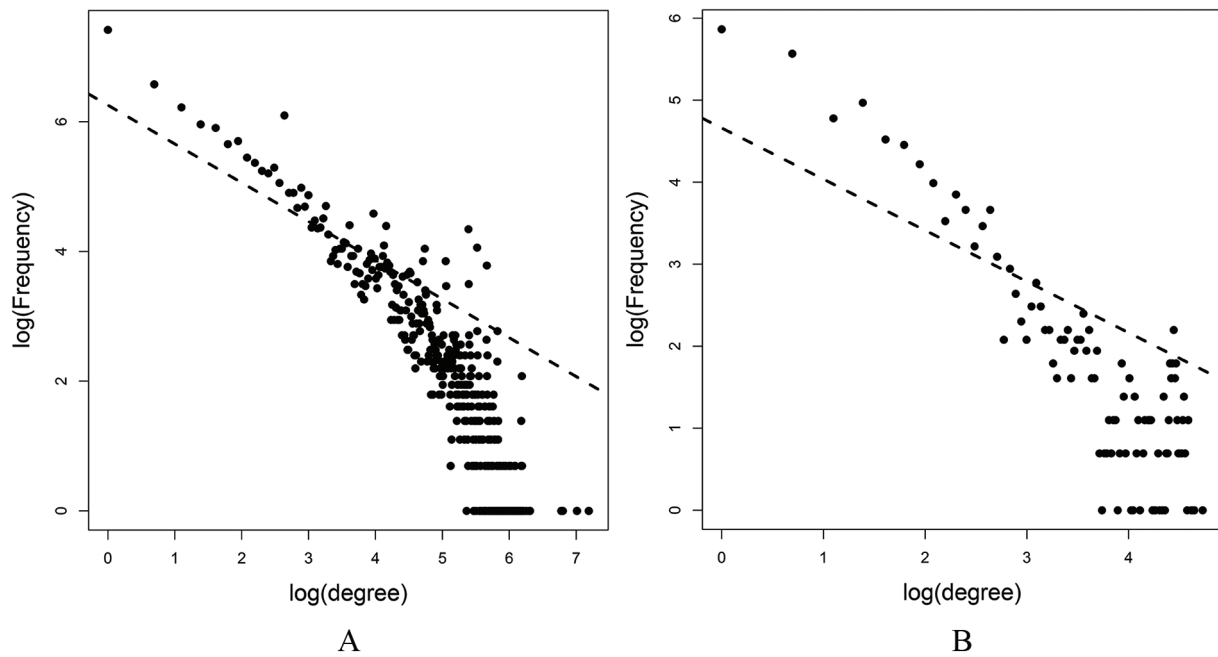
Figure 2. Degree distributions of 2 networks: the degree distributions of (A) human protein-protein interaction network and (B) malaria protein-protein interaction network.

betweenness centrality, degree and eccentricity, but large differences in closeness centrality and a small difference in Kleinberg's hub. A node with a high betweenness score was indicative of a node with overloading paths passing through it, that is, the node may act as a bridge between 2 or more communities. The boxplot of betweenness centrality scores showed that both human and parasite networks had a similar mean overload for each node in the entire network. Evidently, there were the similar mean of degrees and eccentricities for both networks.

Closeness centrality provides a good measure of a given node located in the middle location, such that it can reach the other nodes in the shortest way. The human network showed lower values of closeness scores than those of the parasite network. This may be due to the fact that, in the human network, there were several proteins, and several protein interactions caused a protein complex, compared to that in the parasite network. Kleinberg's hub represents the protein nodes that may connect to other important nodes in the network. The boxplot shows that, on average, human proteins are slightly more likely to connect with other important nodes than that are parasite proteins. Although the boxplots show the overall distributions of each node property in the entire network, they do not represent all single differences of each protein in both networks. In addition, these differences may provide a good view of how human and parasite proteins relate to each other in terms of the cooperative community in the network. Thus, the similarity profiles of these topological node properties for each pair of human and *Plasmodium* proteins were determined. This profile was used as a feature to train the machine-learning classifiers.

We calculated the topological similarity of each feature for each pair of human and *Plasmodium* proteins. All possible combinations of these 2 types of proteins resulted in 225 675 478 human-*Plasmodium* protein pairs. Next, the similarity features based on the node properties were calculated (see Materials and Methods) for each pair of human-*Plasmodium* proteins. Initially, we defined 19 939 pairs as positive association pairs based on protein sequence similarities. The remaining pairs, namely 225 655 539 pairs, were defined as an undefined set. These data sets were prepared to be fed into the established classification processes. Before the classification process, it was interesting to analyze the topology features to determine the relationship between proteins in the positive pairs. We then calculated an uncentered correlation of each node property between human and parasite proteins in the positive set, as shown in Table 1. This uncentered correlation provides the value of the relationship, ranging from 0 to 1. As expected, we found a high correlation of closeness centrality between the human and parasite proteins, with a correlation coefficient of 0.9805. In addition, a moderate correlation of eccentricity between the human and parasite proteins with a correlation coefficient of 0.6827 in the positive set was observed. A low correlation of degree and betweenness centrality between human and parasite proteins was observed, with correlation coefficients of 0.3507 and 0.1316, respectively. With Kleinberg's hub, no correlation was observed, with correlation coefficient of 0.0556 between human and parasite proteins. The characterization of the topological features of human and parasite protein interaction networks may help to identify underlying proteins that cooperate with host cell recognition and invasion by parasite proteins.
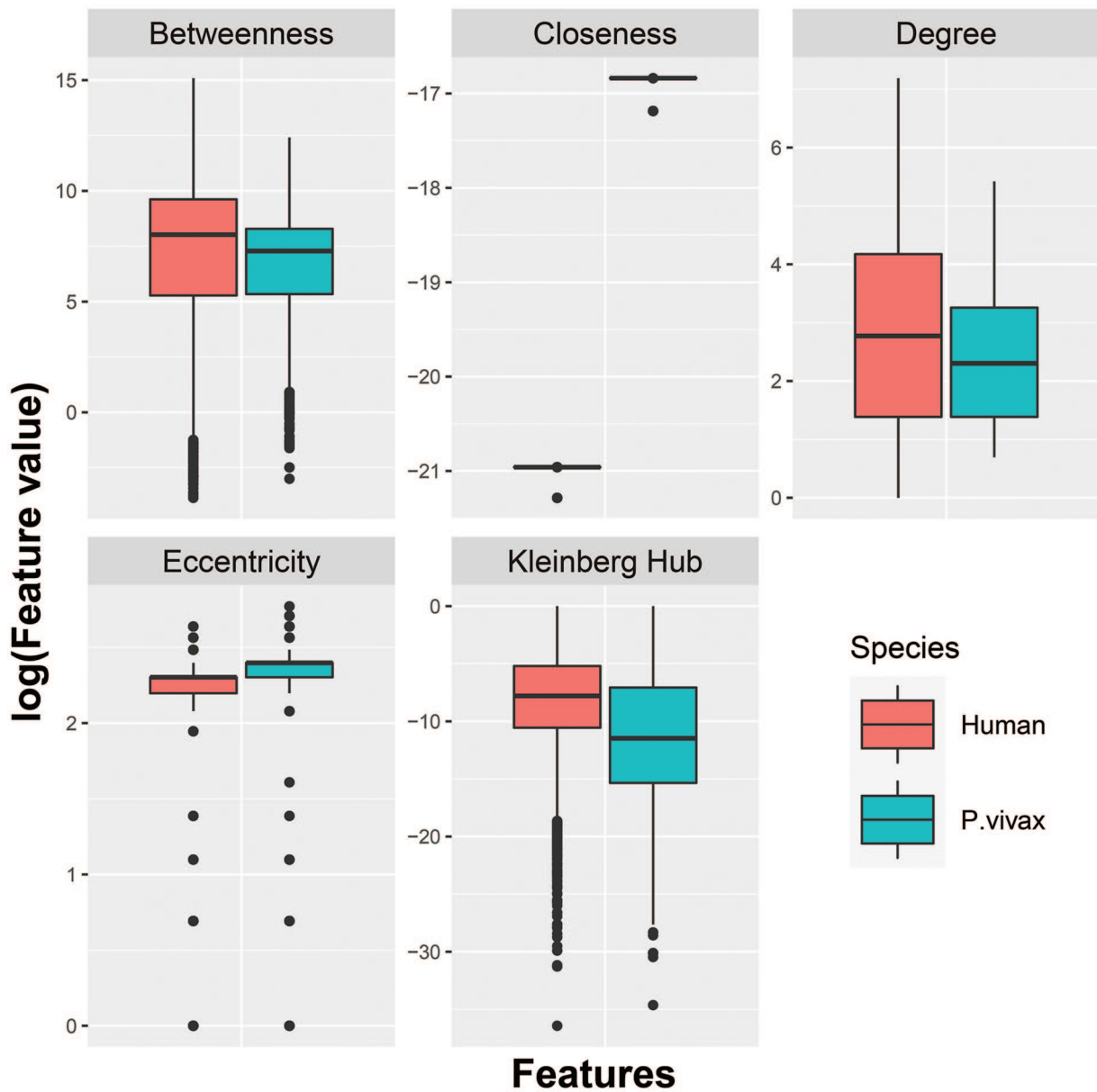
**Figure 3.** Boxplots for the properties of each node.

Table 1. Correlation coefficient values of each topological feature between human and parasite proteins in the positive set.

DEGREE	CLOSENESS	BETWEENNESS	ECCENTRICITY	KLEINBERG'S HUB
0.3507	0.9805	0.1316	0.6827	0.0556

Performance of the classifications used to recognize human–parasite protein associations

Four classification algorithms, naïve Bayes, neural network, random forest, and support vector machine, were used to recognize human-parasite protein associations. Their performances were compared to select the best classifier for the recognition of human-parasite protein similarities, based on topological features. Ten 10-fold cross-validations were applied for each algorithm, which yielded the performance in terms of

an ROC curve with an AUC, as shown in Figure 4. The random forest algorithm provided the best classifier, with an AUC of 0.85. The neural network algorithm yielded a slightly lower performance, with an AUC of 0.79. Similarly, the support vector machine achieved an AUC of 0.77. The naïve Bayes classifier yielded a slightly lower performance compared with that of the neural network and support vector machine with an AUC of 0.74. Notably, the random forest algorithm provided the best performance, with an AUC that was relatively far from that of
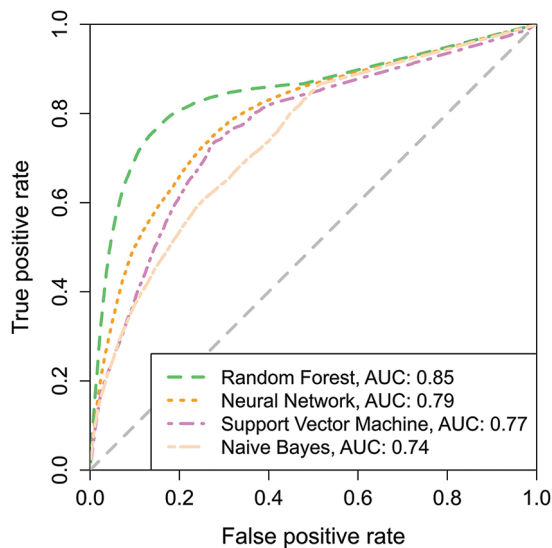
Figure 4. Receiver operating characteristic (ROC) curves for the predictions of human-parasite protein associations of each machine-learning algorithm.

AUC indicates area under the curve; ROC, receiver operating characteristic.

Figure 5. Histogram showing the frequency of ranking scores in logarithm scale for human proteins in the predicted human-parasite associations.

the other algorithms. This is of great interest because the results obtained for this algorithm indicate its potential in identifying new human-parasite protein associations and, furthermore, in selection of key human proteins for the parasite.

The classifier showed a better performance than that did random selection, which may result in 50% correct predictions. Moreover, we attempted to demonstrate the reliability of the relationship between sequence similarity and network topologies by performing several random experiments. These experiments could be performed by randomly shuffling class labels and retraining the random forest classifiers. Ten 10-fold cross-validations were performed in the same procedures. An AUC of 0.5 was obtained for these random experiments. This was also a good indication that the network topologies of protein nodes in the PPI networks could be used to infer the relationship between human and parasite proteins in terms of sequence similarity, reflecting the homologs and similar cooperation in the network community.

Based on the best performance and the results of the random forest classifiers, we defined a voting score for a pair of human and parasite proteins. Ten probability values of the positive prediction for a pair of human and parasite proteins were obtained. The average of these probability values was calculated and defined as a voting score for a pair of human and parasite proteins (see Materials and Methods). This score was used to define the stringency of predicting human-parasite protein associations. Initially, we identified 12 038 human proteins in the human PPI network and 1787 parasite proteins in the parasite PPI network. This resulted in a total of 225 675 478 human-parasite protein pairs. A total of 19 939 pairs were initially defined as positive association pairs based on protein sequence similarities. After performing the random forest classification, the average voting score was calculated for each pair.
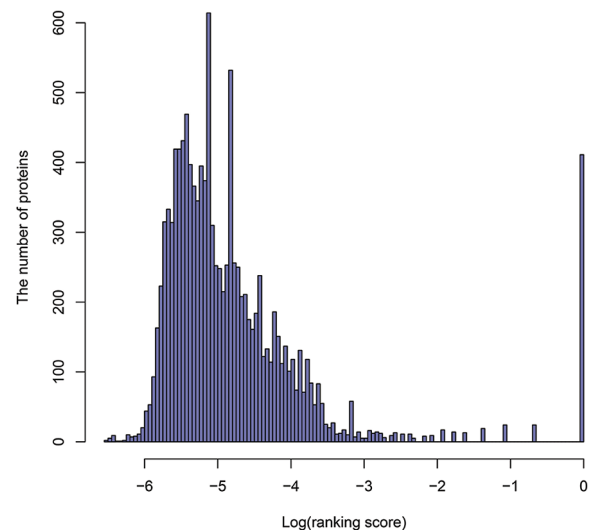
It is worth noting that these scores indicated associations based on the network topological profiles of the human-parasite protein pairs using machine learning. It was also interesting to combine these scores with the other association scores from other aspects such as the heterogeneous network study.[19] With the heterogeneous network model, the network propagation algorithm with a decay factor of 0.1 was performed on the network to prioritize human-parasite protein associations.[19] A total of 21 511 906 overlap pairs from both machine-learning and network propagation techniques with scores greater than 0 were obtained and used for the further analysis and selection of key human proteins. Of these pairs, 831 had the highest voting scores of the predictions according to our machine-learning analysis (Supplementary Table S1).

Identifying promising key human proteins from predicted associations

All human proteins among the 21 511 906 pairs were ranked to calculate their ranking scores under the assumption that human proteins in association with high ranking scores may be important for parasite mechanisms. The final ranking score for each human protein was obtained by the production of the ranking score (see section "Ranking score calculation for each human protein") calculated from the ranked pairs obtained using the machine-learning method and the ranking score calculated from the ranked pairs using the network propagation methods. The histogram of the logarithmic transformation of the final ranking scores of all 12 038 human proteins is shown in Figure 5. Notably, most of the ranking scores were less than 0.0001, while the top best-ranking score was 1 (the logarithm of 1 is 0). Using this top-ranking score, we obtained 411 human proteins. These human proteins were defined as the first list of promising target proteins in human hosts. A complete list of these 411
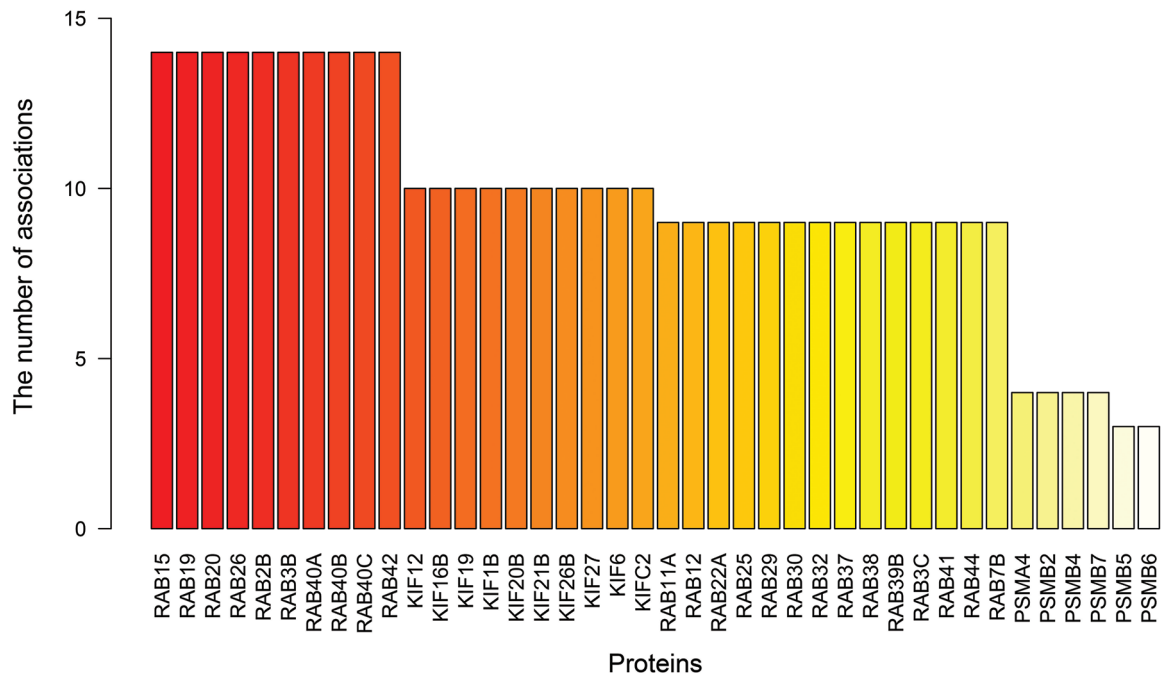
Figure 6. Bar plot illustrating the number of the highest-score associations of each human protein. Only proteins associated with more than 2 pairs were presented.

human proteins is provided in Supplementary Table S2. The bar plot representing the number of highest-score associations for these 411 proteins is shown in Figure 6. Note that only proteins found in more than 2 association pairs are presented in the figure. Overall, we identified Ras-related proteins, kinesin family members, and proteasome 20 S subunit alpha and beta in the list.

Clusters of human protein candidates associated to malaria

As mentioned in section "Identifying promising key human proteins from predicted associations," we integrated the association scores from our machine-learning techniques and the heterogeneous network model. First, the association scores of candidate human-parasite protein pairs from the heterogeneous network method were ranked to calculate their ranking scores for each protein in the same manner as in our study (see Materials and Methods). Next, we combined the ranking scores of these 2 methods as the attributes to cluster the human proteins using hierarchical clustering. The aim was to group human proteins with similar levels of importance in both aspects. Figure 7 shows the hierarchical clustering of these proteins. By selecting the cut height of the dendrogram tree as 8, we obtained 7 groups of proteins consisting of 2 groups of Ras-related proteins, a single group of histone H2B proteins, kinesin family members, ubiquitin specific peptidase 17 like family members, zinc finger proteins, and a remaining group of mixed types of proteins. Figure S1 shows the high-resolution circular dendrogram of the clustering analysis. The complete list of these proteins in each cluster is provided in Supplementary
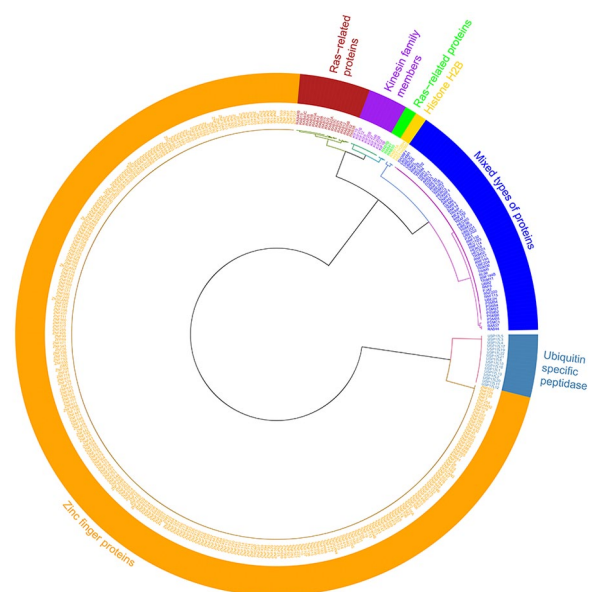
Figure 7. Circular dendrogram of the hierarchical clustering analysis.

Table S3. Ras proteins are members of a superfamily of small GTPases that are involved in many processes of cell growth control. Ubiquitin-specific peptidase 17 like family members regulate different cellular processes, such as cell proliferation, cell migration, progression through the cell cycle, apoptosis, and cellular response to viral infection.[49-51]

*Functional characteristics of annotated human proteins*

Interpreting the functions of these 411 annotated human proteins may reveal the related mechanisms of the human

Table 2. Nine functional groups based on principal gene ontology (GO) terms.

CLUSTER NUMBER	GO ID	PRINCIPLE GO TERM	ADJUSTED P VALUE*	PERCENTAGE OF ASSOCIATED PROTEINS
1	GO:0006355	Regulation of transcription, DNA-templated	8.52E–112	7.29
2	GO:0003700	DNA-binding transcription factor activity	6.50E–27	6.80
3	GO:0032482	Rab protein signal transduction	9.87E–20	30.26
4	GO:0070647	Protein modification by small protein conjugation or removal	9.98E–20	6.85
5	GO:0006511	Ubiquitin-dependent protein catabolic process	8.17E–12	7.23
6	GO:0090382	Phagosome maturation	5.41E–03	11.11
7	GO:0097494	Regulation of vesicle size	5.67E–03	21.43
8	GO:0001217	DNA-binding transcription repressor activity	1.43E–02	4.53
9	GO:0006904	Vesicle docking involved in exocytosis	2.67E–02	8.33

Abbreviation: GO, gene ontology.
*P values were adjusted according to Benjamini-Hochberg correction method.
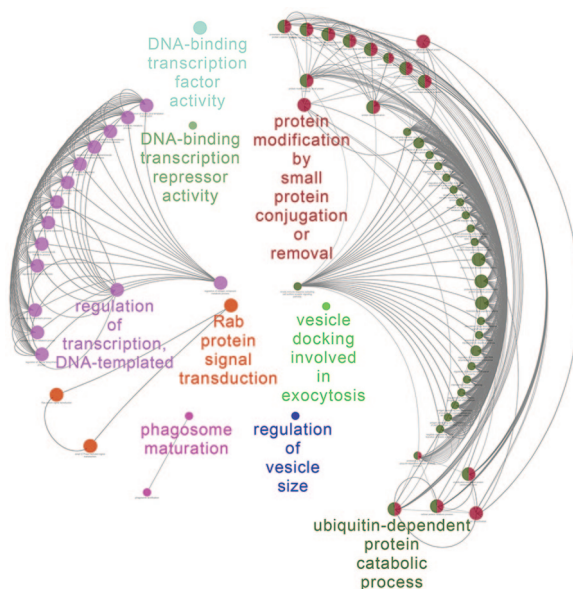
Figure 8. Representative network of gene ontology (GO) terms of our candidate human proteins using ClueGO.
GO indicates gene ontology.

host and parasite. We investigated these human proteins using functional enrichment analyses. Gene ontology annotations were performed to obtain an overview of the biological processes. The analysis was performed using Cytoscape plugins, ClueGO. Gene ontology associations based on biological processes were selected using intermediate detail in the panel setting of ClueGO. This covered 3 to 8 levels of GO terminology. Based on the PPI of STRING, a second enrichment analysis was performed with a group of genes that were connected in the GO network using CluePedia (version 1.5.6). This analysis revealed 9 functional groups of GO terms, as shown in Table 2 and Figure 8. The complete

list of these overrepresented GO terms in the biological process category is provided in Supplementary Table S4. Interestingly, we found the term of regulation of transcription, DNA-templated (GO:0006355), with the most significant term. In addition, Rab protein signal transduction (GO:0032482) and regulation of vesicle size (GO:0097494) were found in a high proportion of our candidate proteins. Rab proteins are a subfamily of the Ras protein family[52] and commonly possess a GTPase fold. These Rab GTPases regulate the processes of membrane trafficking, vesicle formation, and membrane fusion.[52-54] Most of our candidate proteins are involved in the regulation of membrane and vesicle formation. These proteins may assist parasite transports in the host and could be potential targets for the treatment of malaria. Figure 8 presents the network of the main enriched GO terms of the 9 clusters, denoted as 9 different colors. Each cluster contained associated GO terms and was named with its principal GO term.

Protein complexes to potential protein targets

To identify sets of these 411 proteins that interact with each other and play essential roles in regulatory processes, cellular functions, and signaling cascades, we performed enrichment analysis in protein complexes. Enrichment analysis of these proteins was performed on the CORUM protein complex database (version 3.0).[55] Four protein complexes were found using Bonferroni-adjusted P values for the enrichment tests <0.05. These 4 protein complexes consisted of the 20S proteasome, 26S proteasome, PA28gamma-20S proteasome, and PA28-20S proteasome. Most of the proteins overrepresented in these protein complexes were PSMA4, PSMB2, PSMB4, PSMB5, PSMB6, and PSMB7. Only the 26S proteasome

Table 3. The list of protein complexes enriched in 411 promising candidate proteins.

PROTEIN COMPLEX	ADJUSTED *P* VALUE	ASSOCIATED PROTEINS
20S proteasome	8.34E−03	PSMA4, PSMB2, PSMB4, PSMB5, PSMB6, PSMB7
26S proteasome	1.29E−02	PSMA4, PSMB2, PSMB4, PSMB5, PSMB6, PSMB7, PSMC1
PA28gamma-20S proteasome	1.35E−02	PSMA4, PSMB2, PSMB4, PSMB5, PSMB6, PSMB7
PA28-20S proteasome	2.10E−02	PSMA4, PSMB2, PSMB4, PSMB5, PSMB6, PSMB7

contained 1 more protein (PSMC1) in the list. Thus, these proteins may be interesting targets in future studies. Table 3 presents a list of the overrepresented protein complexes.

Furthermore, to examine the importance of the proposed human proteins, these proteins were searched for in the Drugbank database.[56] Interestingly, Proteasome 20S Subunit Beta 2 (PSMB2) and Proteasome 20S Subunit Beta 5 (PSMB5) were identified, which are known to be drug targets, in the Drugbank database. PSMB2 and PSMB5 play several roles. They were found to be enriched in the principal GO terms of regulation of transcription, DNA-templated, protein modification by small protein conjugation or removal, and ubiquitin-dependent protein catabolic process. Interestingly, PSMB2 was found to be a drug target of carfilzomib (DB08889), while PSMB5 is a drug target of carfilzomib and bortezomib (DB00188). Carfilzomib is a synthetic proteasome inhibitor. It is an analogue of the natural product epoxomicin, which effectively kills parasites. Bortezomib is the first therapeutic proteasome inhibitor to be tested in humans, which induces cell cycle arrest and apoptosis. Bortezomib interrupts the degradation of proapoptotic proteins in cancerous cells. It is currently used for the treatment of relapsed multiple myeloma and mantle cell lymphoma. Both carfilzomib and bortezomib have been reported to be related to malaria treatment.[57] Carfilzomib has been reported to potently block *P. falciparum* replication at effective concentrations as well as killing asexual blood-stage *P. falciparum*.[58] Bortezomib exhibits antiplasmodial activities and has been examined for efficacy against *P. falciparum*.[59] PSMB2 and PSMB5 were found in all our resulting protein complexes (Table 3). Thus, these complexes may be a valuable starting point for further studies aiming to design and develop drugs against malaria. In addition, PSMB2 and PSMB5 were observed in mixed types of protein group of 62 proteins in our clustering results (see section "Clusters of human protein candidates associated to malaria" and Supplementary Table S3). Therefore, the remaining 60 proteins in the same cluster of these proteins may be promising therapeutic targets for *P. vivax* malaria. A list of these proteins is provided in Supplementary Table S5. In addition, the relationship of these 411 human proteins and *P. vivax* malaria was evaluated to determine orthologous proteins of *P. vivax* and the 411 human proteins from EggNOG database (version 5.0).[60] The results are presented in Supplementary Table S6.

Discussion

Our understanding of the invasion mechanism of *P. vivax* remains deficient due to the lack of a robust *in vitro* culture system for this parasite. In an attempt to resolve this, the host-parasite interactions were studied, including direct interactions at the protein level inside the cell. In this study, we initially reconstructed the human and parasite PPI networks, and compared their network structures. In principle, both networks follow the power distribution, and the analysis of network topologies between these 2 networks revealed a correlation of the connections within their own network between human and parasite proteins in the positive set. The high correlation of closeness centrality between these proteins indicated that most of the similar proteins between human and parasite responded to minimum paths that connect the other proteins. These proteins also formed a similar local community around them, as the high correlation was observed in terms of eccentricity. Although the degree, betweenness centrality, and Kleinberg's hub did not show significant correlations among these proteins, the machine-learning approaches applied here may help reveal several more human and parasite protein associations in future studies.

A ranking score calculation for the human proteins was developed based on the rank of the associations according to their voting scores. A total of 411 human proteins with the best-ranking score were selected as promising target candidates. Based on the histogram shown in Figure 5, the second-best score had a gap jumping from the top best, while the rest of the scores were far away from the best one. The majority of these proteins had a ranking score of approximately 0.00001, which was very low in terms of the probability of being a reliable association. Thus, these 411 proteins were selected for further analysis together with heterogeneous network prioritization and qualified in terms of clusters, functions, and protein complexes.

The results showed that Ras-related proteins, a single group of histone H2B proteins, kinesin family members, ubiquitin-specific peptidase 17 like family members, and zinc finger proteins were the most prominent in our candidate list. These proteins are involved in several processes of cell growth control and regulation of membrane and vesicle formation. Several proteins related to proteasome 20S subunits have been previously reported as promising multistage targets for malaria therapy.[59]

These proteins may be used for the invasion of parasites to the host cell and have been identified as potential drug targets in the human host.

Conclusion

In this study, we established an analysis framework that uses machine-learning approach based on a heterogeneous network structure. We used the network topology features of proteins in the human PPI network and the *P. vivax* PPI network and integrated protein sequence similarities to the framework to predict human-parasite protein associations. We also developed a ranking score calculation to identify promising protein targets in humans for the treatment of malaria infections. The candidate human proteins that were selected as promising targets were then qualified by clustering analysis together with the information on the existing targets from the heterogeneous network prioritization, as well as by functional and protein complex enrichment analyses. We found that proteins in the cluster of PSMB2 and PSMB5 (known drug targets), human proteins involved in the regulation of membrane and vesicle formation, and complexes such as the 20S proteasome, 26S proteasome, and PA28gamma/-20S proteasomes are potential targets for the design and development of drugs for the treatment of malaria.

In conclusion, the integration of data related to network topologies and sequence similarity provides us with an opportunity to define associations between human and *P. vivax* proteins. Human protein candidates extracted from these associations were used to compile a list of promising targets in humans for further validation in wet-laboratory experiments in future studies. An enhanced understanding of potential host proteins at the molecular level will provide insights to support malaria control efforts and the production of novel antimalarial drugs.

Acknowledgements

Author Contributions

Conceptualization and writing—review and editing were performed by A.S, T.B. and K.P.; A.S. contributed in data curation, funding acquisition, and writing the original draft; formal analysis, methodology and validation was by A.S. and K.P. All authors have read and agreed to the published version of the manuscript.

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

REFERENCES

1. Sultan AA. Molecular mechanisms of malaria sporozoite motility and invasion of host cells. *Int Microbiol*. 1999;2:155-160.
2. Baird JK. Evidence and implications of mortality associated with acute *Plasmodium vivax* malaria. *Clin Microbiol Rev*. 2013;26:36-57.
3. Sa JM, Kaslow SR, Moraes Barros RR, et al. *Plasmodium vivax* chloroquine resistance links to pvcrt transcription in a genetic cross. *Nat Commun*. 2019;10:4300.
4. Battle KE, Lucas TCD, Nguyen M, et al. Mapping the global endemicity and clinical burden of *Plasmodium vivax*, 2000-17: a spatial and temporal modelling study. *Lancet*. 2019;394:332-343.
5. Noisang C, Prosser C, Meyer W, et al. Molecular detection of drug resistant malaria in Southern Thailand. *Malar J*. 2019;18:275.
6. Lacerda MVG, Llanos-Cuentas A, Krudsood S, et al. Single-dose tafenoquine to prevent relapse of *Plasmodium vivax* malaria. *N Engl J Med*. 2019;380:215-228.
7. Chu CS, Freedman DO. Tafenoquine and G6PD: a primer for clinicians. *J Travel Med*. 2019;26:taz023.
8. Rueangweerayut R, Bancone G, Harrell EJ, et al. Hemolytic potential of tafenoquine in female volunteers heterozygous for glucose-6-phosphate dehydrogenase (G6PD) deficiency (G6PD Mahidol variant) versus G6PD-normal volunteers. *Am J Trop Med Hyg*. 2017;97:702-711.
9. Commons RJ, McCarthy JS, Price RN. Tafenoquine for the radical cure and prevention of malaria: the importance of testing for G6PD deficiency. *Med J Aust*. 2020;212:152-153.e151.
10. Acharya P, Garg M, Kumar P, Munjal A, Raja KD. Host-parasite interactions in human malaria: clinical implications of basic research. *Front Microbiol*. 2017;8:889.
11. Batchelor JD, Malpede BM, Omattage NS, DeKoster GT, Henzler-Wildman KA, Tolia NH. Red blood cell invasion by *Plasmodium vivax*: structural basis for DBP engagement of DARC. *PLoS Pathog*. 2014;10:e1003869.
12. Grimberg BT, Udomsangpetch R, Xainli J, et al. *Plasmodium vivax* invasion of human erythrocytes inhibited by antibodies directed against the Duffy binding protein. *PLoS Med*. 2007;4:e337.
13. Bermudez M, Moreno-Perez DA, Arevalo-Pinzon G, Curtidor H, Patarroyo MA. *Plasmodium vivax* in vitro continuous culture: the spoke in the wheel. *Malar J*. 2018;17:301.
14. Caufield JH, Wimble C, Shary S, Wuchty S, Uetz P. Bacterial protein meta-interactomes predict cross-species interactions and protein function. *BMC Bioinformatics*. 2017;18:171.
15. Wuchty S, Uetz P. Protein-protein Interaction Networks of E. coli and S. cerevisiae are similar. *Sci Rep*. 2014;4:7187.
16. Rajagopala SV, Sikorski P, Kumar A, et al. The binary protein-protein interaction landscape of Escherichia coli. *Nat Biotechnol*. 2014;32:285-290.
17. Zhang X, Xiao W, Hu X. Predicting essential proteins by integrating orthology, gene expressions, and PPI networks. *PLoS ONE*. 2018;13:e0195410.
18. Paul G, Deshmukh A, Kumar Chourasia B, et al. Protein-protein interaction studies reveal the Plasmodium falciparum merozoite surface protein-1 region involved in a complex formation that binds to human erythrocytes. *Biochem J*. 2018;475:1197-1209.
19. Suratanee A, Plaimas K. Heterogeneous network model to identify potential associations between *Plasmodium vivax* and human proteins. *Int J Mol Sci*. 2020;21:1310.
20. Hillier C, Pardo M, Yu L, et al. Landscape of the Plasmodium interactome reveals both conserved and species-specific functionality. *Cell Rep*. 2019;28:1635-1647.e1635.
21. LaCount DJ, Vignali M, Chettier R, et al. A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature*. 2005;438:103-107.
22. Liu X, Huang Y, Liang J, et al. Computational prediction of protein interactions related to the invasion of erythrocytes by malarial parasites. *BMC Bioinformatics*. 2014;15:393.
23. Pierrot C, Freville A, Olivier C, Souplet V, Khalife J. Inhibition of protein-protein interactions in Plasmodium falciparum: future drug targets. *Curr Pharm Des*. 2012;18:3522-3530.
24. Saha S, Sengupta K, Chatterjee P, Basu S, Nasipuri M. Analysis of protein targets in pathogen-host interaction in infectious diseases: a case study on Plasmodium falciparum and Homo sapiens interaction network. *Brief Funct Genomics*. 2018;17:441-450.
25. Suratanee A, Plaimas K. Identification of inflammatory bowel disease-related proteins using a reverse k-nearest neighbor search. *J Bioinform Comput Biol*. 2014;12:1450017.
26. Suratanee A, Plaimas K. DDA: a novel network-based scoring method to identify disease-disease associations. *Bioinform Biol Insights*. 2015;9:175-186.
27. Suratanee A, Plaimas K. Network-based association analysis to infer new disease-gene relationships using large-scale protein interactions. *PLoS ONE*. 2018;13:e0199435.

28. Cook HV, Doncheva NT, Szklarczyk D, von Mering C, Jensen LJ. Viruses. STRING: a virus-host protein-protein interaction database. *Viruses*. 2018;10:519.

29. Lee SA, Chan CH, Tsai CH, et al. Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics*. 2008;9:S11.

30. Hu Y, Vinayagam A, Nand A, et al. Molecular Interaction Search Tool (MIST): an integrated resource for mining gene and protein interaction data. *Nucleic Acids Res*. 2018;46:D567-D574.

31. Matthews LR, Vaglio P, Reboul J, et al. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs." *Genome Res*. 2001;11:2120-2126.

32. Yu H, Luscombe NM, Lu HX, et al. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res*. 2004;14:1107-1118.

33. von Mering C, Jensen LJ, Snel B, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*. 2005;33:D433-D437.

34. Cuesta-Astroz Y, Santos A, Oliveira G, Jensen LJ. Analysis of predicted host-parasite interactomes reveals commonalities and specificities related to parasitic lifestyle and tissues tropism. *Front Immunol*. 2019;10:212.

35. Plaimas K, Eils R, Konig R. Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Syst Biol*. 2010;4:56.

36. Sciarra C, Chiarotti G, Laio F, Ridolfi L. A change of perspective in network centrality. *Sci Rep*. 2018;8:15269.

37. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019;47:D607-D613.

38. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature*. 2001;411:41-42.

39. Hahn MW, Kern AD. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol*. 2005;22:803-806.

40. Maslov S, Sneppen K. Specificity and stability in topology of protein networks. *Science*. 2002;296:910-913.

41. Pearson WR. An introduction to sequence similarity ("homology") searching. *Curr Protoc Bioinformatics*. 2013; Chapter 3:Unit3.1.

42. Kanehisa M. The KEGG database. *Novartis Found Symp*. 2002;247:91-101; discussion 101-103, 119-128, 244-152.

43. Tanabe M, Kanehisa M. Using the KEGG database resource. *Curr Protoc Bioinformatics*. 2012; Chapter 1:Unit1.12.

44. Cao DS, Xiao N, Xu QS, Chen AF. Rcpi: R/bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics*. 2015;31:279-281.

45. Davis D, Yaveroglu ON, Malod-Dognin N, Stojmirovic A, Przulj N. Topology-function conservation in protein-protein interaction networks. *Bioinformatics*. 2015;31:1632-1639.

46. Newman M. *Networks: An Introduction*. New York, NY: Oxford University Press, Inc.; 2010.

47. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13:2498-2504.

48. Bindea G, Mlecnik B, Hackl H, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*. 2009;25:1091-1093.

49. Burrows JF, McGrattan MJ, Johnston JA. The DUB/USP17 deubiquitinating enzymes, a multigene family within a tandemly repeated sequence. *Genomics*. 2005;85:524-529.

50. McFarlane C, Kelvin AA, de la Vega M, et al. The deubiquitinating enzyme USP17 is highly expressed in tumor biopsies, is cell cycle regulated, and is required for G1-S progression. *Cancer Res*. 2010;70:3329-3339.

51. Fukuura K, Inoue Y, Miyajima C, et al. The ubiquitin-specific protease USP17 prevents cellular senescence by stabilizing the methyltransferase SET8 and transcriptionally repressing p21. *J Biol Chem*. 2019;294:16429-16439.

52. Colicelli J. Human RAS superfamily proteins and related GTPases. *Sci STKE*. 2004;2004:RE13.

53. Bhuin T, Roy JK. Rab proteins: the key regulators of intracellular vesicle transport. *Exp Cell Res*. 2014;328:1-19.

54. Kiral FR, Kohrs FE, Jin EJ, Hiesinger PR. Rab GTPases and membrane trafficking in neurodegeneration. *Curr Biol*. 2018;28:R471-R486.

55. Giurgiu M, Reinhard J, Brauner B, et al. CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res*. 2019;47:D559-D563.

56. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*. 2018;46:D1074-D1082.

57. Kirkman LA, Zhan W, Visone J, et al. Antimalarial proteasome inhibitor reveals collateral sensitivity from intersubunit interactions and fitness cost of resistance. *Proc Natl Acad Sci USA*. 2018;115:E6863-E6870.

58. Li H, Ponder EL, Verdoes M, et al. Validation of the proteasome as a therapeutic target in Plasmodium using an epoxyketone inhibitor with parasite-specific toxicity. *Chem Biol*. 2012;19:1535-1545.

59. Aminake MN, Arndt HD, Pradel G. The proteasome of malaria parasites: a multi-stage drug target for chemotherapeutic intervention. *Int J Parasitol Drugs Drug Resist*. 2012;2:1-10.

60. Huerta-Cepas J, Szklarczyk D, Heller D, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*. 2019;47:D309-D314.