A Novel Collaborative Filtering Model-Based Method for Identifying Essential Proteins

Xianyou Zhu¹,²*, Xin He³*, Linai Kuang⁹, Zhiping Chen⁴ and Camara Lancine⁵

¹College of Computer Science and Technology, Hengyang Normal University, Hengyang, China, ²Hunan Provincial Key Laboratory of Intelligent Information Processing and Application, Hengyang, China, ³College of Computer, Xiangtan University, Xiangtan, China, ⁴College of Computer Engineering and Applied Mathematics, Changsha University, Changsha, China, ⁵The Social Sciences and Management University of Bamako, Bamako, Mali

Considering that traditional biological experiments are expensive and time consuming, it is important to develop effective computational models to infer potential essential proteins. In this manuscript, a novel collaborative filtering model-based method called CFMM was proposed, in which, an updated protein–domain interaction (PDI) network was constructed first by applying collaborative filtering algorithm on the original PDI network, and then, through integrating topological features of PDI networks with biological features of proteins, a calculative method was designed to infer potential essential proteins based on an improved PageRank algorithm. The novelties of CFMM lie in construction of an updated PDI network, application of the commodity-customer-based collaborative filtering algorithm, and introduction of the calculation method based on an improved PageRank algorithm, which ensured that CFMM can be applied to predict essential proteins without relying entirely on known protein–domain associations.

Simulation results showed that CFMM can achieve reliable prediction accuracies of 92.16, 83.14, 71.37, 63.87, 55.84, and 52.43% in the top 1, 5, 10, 15, 20, and 25% predicted candidate key proteins based on the DIP database, which are remarkably higher than 14 competitive state-of-the-art predictive models as a whole, and in addition, CFMM can achieve satisfactory predictive performances based on different databases with various evaluation measurements, which further indicated that CFMM may be a useful tool for the identification of essential proteins in the future.

Keywords: essential proteins, collaborative filtering model, PDI network, data integration, prediction model

INTRODUCTION

Researches show that essential proteins are not only important for survival of organisms but also play critical roles in the development of life processes. Hence, it is of practical significance to identify potential essential proteins (Meng et al., 2021). With the development of biotechnologies, some essential proteins have been identified successively by traditional biological experiments such as single gene knockouts (Giaever et al., 2002), RNA interference (Cullen and Arndt, 2005), and so on. However, since these traditional biological experiments are quite time consuming and expensive, it has become a hot topic to predict essential proteins by developing computational models (Wang et al., 2013). Up to now, a large number of computational models have been developed to detect essential proteins based on protein–protein interaction (PPI) networks, which can be roughly...
classified into two major categories. Among them, the first category of models focuses on adopting only topological features of PPI networks to predict essential proteins. For instance, based on the rule of centrality–lethality proposed (Jeong et al., 2001), a series of models, such as DC (Degree Centrality) (Hahn and Kern, 2005), SC (Subgraph Centrality) (Estrada and Rodriguez-Velázquez, 2005), BC (Betweenness Centrality) (Joy et al., 2005), EC (Eigenvector Centrality) (Bonacich, 1987), IC (Information Centrality) (Stephenson and Zelen, 1989), CC (Closeness Centrality) (Wuchty and Stadler, 2003), and NC (Neighbor Centrality) (J. Wang et al., 2012), have been designed in succession for inferring essential proteins based on topological features of PPI networks. Except for these models, Li et al. (2011) proposed a novel model called LAC to predict potential essential proteins based on neighborhoods of protein nodes in PPI networks. B. Xu et al. (2019) developed a model to detect essential proteins by applying random walks on PPI networks. Wang et al. (2011) presented a model called SoECC based on edge clustering coefficients to infer essential proteins. Qin et al. (2016) designed a method called LBCC based on characteristics of PPI networks to predict essential proteins. However, due to the incompleteness of PPI networks, all these first category of models cannot achieve satisfactory prediction accuracies of potential essential proteins.

In order to overcome the incompleteness of PPI networks, in recent years, another category of models have been proposed by integrating topological features of PPI networks and some biological information of proteins to infer essential proteins. For example, Chen et al. (2017) developed a computational model to infer essential proteins by combining PPI networks with gene ontology and KEGG pathway. Zhang X. et al. (2018) presented a prediction model by combing gene expression data with PPI networks to predict essential proteins. W. Peng et al. (2015a) proposed a prediction model called UDoNC by integrating protein domains with PPI networks to infer essential proteins. Jiang et al. (2015) developed a method called IEW to detect key essentials by combining domain interactions and topological features of PPI networks. Zhao et al. (2019) put forward a prediction model called RWTHN to infer key proteins by integrating PPI networks with protein domains and some other biological information. Lei et al. (2018) put forward a prediction model named RSG by integrating subcellular localization and GO data of proteins with PPI networks to infer key proteins. Y. Fan et al. (2016) proposed a novel prediction model by adopting Pearson correlation coefficients and subcellular localization to update the PPI network. Qin et al. (2017) put forward a method for recognizing essential proteins based on the topological information of PPI networks and orthologous information of proteins. Peng et al. (2012) proposed an advanced iterative algorithm named ION for identifying key proteins based on the topological information of PPI networks and homologous information of proteins. Li et al. (2012) put forward a novel prediction method called Pec through integrating the PPI network with the gene expression of proteins to improve the accuracy of the prediction model. Zhang et al. (2013) presented a novel calculation model named CoEWC by combining PPI networks with the gene expression profiles of proteins to recognize potential key proteins. Liu et al. (2020) proposed a novel prediction model named DEP-MSB by integrating biological features of proteins and topological features of PPI networks. Zhao et al. (2014) put forward an advanced iterative algorithm named POEM for detecting key proteins through utilizing improved particle swarm optimization. Zhang W. et al. (2018) presented a computational model called TEGS to recognize key proteins by combining biological information of proteins and topological features of PPI networks. S. Li et al. (2020) developed a novel prediction model called CVIM by combining PPI networks and orthologous information of proteins for inferring essential proteins. Z. Chen et al. (2020) presented a novel strategy named NPRI by combining various biological data of proteins and the topological features of PPI networks to infer key proteins. Although the second category of methods can greatly improve the predictive accuracy of potential essential proteins, it remains to be a challenging work to scientifically integrate topological features of PPI networks and biological features of proteins to effectively improve the accuracy of essential protein prediction.

Inspired by the above methods, in this paper, a novel Collaborative Filtering Model-Based Method (CFMM) was proposed to predict potential essential proteins, in which, an original protein–domain interaction (PDI) network was constructed first, and then, considering that the number of known interactions between domains and proteins was quite limited, an updated PDI network was built by applying the collaborative filtering algorithm on the original PDI network. Next, based on the updated PDI network, some key topological features and biological features of proteins were extracted, which would be further integrated together to infer potential essential proteins based on an improved PageRank algorithm. Finally, in order to estimate the performance of CFMM, it was compared with 14 competitive prediction models such as DC (Hahn and Kern, 2005), SC (Estrada and Rodriguez-Velázquez, 2005), BC (Joy et al., 2005), EC (Bonacich, 1987), IC (Stephenson and Zelen, 1989), CC (Wuchty and Stadler, 2003), NC (J. Wang et al., 2012), ION (Peng et al., 2012), Pec (Li et al., 2012), CoEWC (Zhang et al., 2013), POEM (Zhao et al., 2014), TEGS (Zhang W. et al., 2018), CVIM (Li et al., 2020), and NPRI (Z. Chen et al., 2020) based on three kinds of well-known public databases. And as a result, CFMM can achieve better prediction accuracies than all these competing methods.

**MATERIALS**

In this section, in order to construct the original PPI network, we first downloaded known PPI data from the DIP database ( Xenarios et al., 2002), the Krogan database (Krogan et al., 2006) and the Gavin database (Gavin et al., 2006) separately.
After removing self-interactions and repeated interactions, we finally obtained 1,167 essential proteins, 3,926 nonessential proteins, and 24,743 known interactions between 5,093 proteins from the DIP database, 14,317 known interactions between 3,672 proteins from the Krogan database, and 7,669 known interactions between 1,855 proteins from the Gavin database, respectively. Moreover, we downloaded the dataset of 1,107 different domains from the Pfam database (Bateman et al., 2004). The subcellular localization data from the COMPARTMENTS databases (X. Peng et al., 2015b), (Binder et al., 2014), which consists of 4,865 proteins involved in 11 kinds of subcellular localizations, including the cytoskeleton, mitochondrion, nucleus, peroxisome, plasma, extracellular, endosome, vacuole, endoplasmic, cytosol, and Golgi. Additionally, the gene expression data were provided by Tu et al. (2005), which include 6,777 gene expressions products and 36 samples. The dataset of orthologous information of proteins are from the InParanoid database (Östlund et al., 2010), which includes a collection of pairwise comparisons between 100 whole genomes. Finally, in order to verify the accuracy of CFMM, we further downloaded a set of 1,293 essential genes from four diverse databases such as MIPS (Mewes et al., 2004), DEG (Zhang and Lin, 2009), SGD (Cherry et al., 1998), and SGDP (Saccharomyces Genome Deletion Project, 2012) separately. The detailed information of datasets downloaded from the DIP, Krogan, and Gavin databases are shown in the following Table 1.

### 3 METHOD

As illustrated in Figure 1, CFMM consists of the following three major steps:

**Step 1:** First, an original PDI network will be constructed based on known protein–domain interactions downloaded from given public databases, and then, a recommendation matrix will be obtained by applying the collaborative filtering algorithm on the original PDI network.

**Step 2:** Next, based on known PPI data and biological information of proteins downloaded from public databases, key topological features and biological features of proteins will be extracted separately, and then, an improved entropy weight method will be applied to effectively integrate all these features.
Step 3: Finally, based on a newly designed distribution rate matrix, an iterative algorithm will be proposed to infer potential essential proteins based on an improved PageRank algorithm.

Construction of Protein–Domain Interaction
Based on known protein–domain interactions downloaded above, we can first construct an original network PDI as follows: for any given protein node $p_i$ and domain node $d_j$, if and only if there is a known interaction between them, there is an edge between $p_i$ and $d_j$ in PDI. Then we can further obtain an adjacency matrix $AM_{pd}$ as follows: for any given protein $p_i$ and domain $d_j$, if and only if there is a known interaction between $p_i$ and $d_j$, there is $AM_{pd}(p_i, d_j) = 1$; otherwise, there is $AM_{pd}(p_i, d_j) = 0$. Due to limited known PDI, obviously, $AM_{pd}$ is a sparse matrix. Hence, in order to improve the density of $AM_{pd}$, we will apply the collaborative filtering algorithm on $AM_{pd}$ according to the following steps:

Step 1: Applying the protein-based collaborative filtering algorithm on PDI as follows:
First, based on $AM_{pd}$ and PDI, we will construct a novel co-occurrence matrix $CM_{pp}$ as follows: for any two given proteins $p_i$ and $p_j$, there is $CM_{pp}(p_i, p_j) = 1$, if and only if there is at least one common domain node existing between them; otherwise, there is $CM_{pp}(p_i, p_j) = 0$. Hence, a similarity matrix $SMPP$ between protein and protein can be calculated after normalizing $CM_{pp}$ as follows:

$$SMPP(p_i, p_j) = \begin{cases} \frac{|N(p_i) \cap N(p_j)|}{\sqrt{|N(p_i)| \times |N(p_j)|}}, & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases}$$ (1)

Here, $|N(p_i)|$ denotes the number of known domains associated to $p_i$ in PDI; in other words, it denotes the sum of elements equaling to one in the $i$th row of $AM_{pd}$. $|N(p_i) \cap N(p_j)|$ represents the number of known domains related to both $p_i$ and $p_j$ simultaneously.

Based on matrices $AM_{pd}$ and $SMPP$, we can further obtain a novel recommendation matrix $RMPD$ as follows:

$$RMPD = SMPP \times AM_{pd}$$ (2)

Next, for any given protein node $p_i$ and domain node $d_j$ in PDI, if the interaction between $p_i$ and $d_j$ is associated already, then for a protein node $p_k$ other than $p_i$, it is no doubt that the higher the similarity between $p_k$ and $p_i$, the more possibility that there may exist a potential association between $p_k$ and $d_j$. Thereafter, we can define the recommendation standard between protein $p_k$ and $d_j$ based on the similarities between proteins as follows:

$$\text{Std}_{pk, d_j} = \frac{1}{N} \sum_{i=1}^{N} RMPD(p_i, d_j)$$ (3)

Here, $N$ denotes the number of proteins in PDI. Based on the above Eq. 3, for any given domain node $d_j$, if there is a protein node $p_k$ satisfying $RMPD(p_k, d_j) > \text{Std}_{pk, d_j}$, then we will further recommend the protein $p_k$ to the domain $d_j$. Thereafter, we will add a new association edge between $p_k$ and $d_j$ and obtain an update protein–domain adjacency matrix $UAM_{pd}$.

Step 2: Applying the domain-based collaborative filtering algorithm
Similarly, we can also obtain an original adjacency matrix $AM_{dp}$ and a co-occurrence matrix $CM_{dd}$. Obviously, as for the matrix $AM_{dp}$, there is $AM_{dp} = AM_{pd}^T$. However, as for the matrix $CM_{dd}$, for any two given domains $d_i$ and $d_j$, there is $CM_{dd}(d_i, d_j) = 1$, if and only if there is at least one common protein node existing between them; otherwise, there is $CM_{dd}(d_i, d_j) = 0$. After normalizing $CM_{dd}$, we can calculate the similarity between $d_i$ and $d_j$ as follows:

$$\text{SMDD}(d_i, d_j) = \begin{cases} \frac{|N(d_i) \cap N(d_j)|}{\sqrt{|N(d_i)| \times |N(d_j)|}}, & \text{if } k \neq r \\ 0, & \text{otherwise} \end{cases}$$ (4)

where $|N(d_i)|$ represents the number of known proteins associated with $d_i$ in PDI, and $|N(d_i) \cap N(d_j)|$ represents the number of known proteins related to $d_i$ and $d_j$ simultaneously.

We can as well define the recommended standard and recommendation matrix as follows:

$$RMDP = \text{SMDD} \times AM_{dp}$$ (5)

$$\text{Std}_{pk, d_j} = \frac{1}{M} \sum_{i=1}^{M} RMDP(d_i, p_k)$$ (6)

Here, $M$ means the number of domains in PDI. In particular, if there exists a domain node $d_k$ in the $i$th column of $RMDP$ satisfying $RMDP(d_k, p_i) > \text{Std}_{pk, d_j}$, then we will recommend the protein $d_k$ to domain $p_i$. Thereafter, we also add a new association edge between $d_k$ and $p_j$ in $AM_{dp}$ and obtain an update association $UAM_{dp}$.

Step 3: Mutual recommendation between proteins and domains
Based on the updated matrix $UAM_{pd}$ and $UAM_{dp}$, the $UAM_{pd}$ is $N \times M$ dimension matrix, and $UAM_{dp}$ is $M \times N$ matrix. By transposing the matrix $AM_{dp}$, it is obvious that we can construct the mutual recommendation matrix $MRM$ as follows:

$$\text{MRM}(p_i, d_j) = \begin{cases} UAM_{pd}(p_i, d_j) + UAM_{dp}^T(p_i, d_j), & \text{otherwise} \\ 1, & \text{if } UAM_{pd}(p_i, d_j) = 1 \text{ and } UAM_{dp}^T(p_i, d_j) = 1 \end{cases}$$ (7)

For instance, according to Figure 1 and the given matrix $AM_{pd}$, we can obtain its corresponding matrices $CM_{pp}$, $SMPP$, and $RMPD$ as follows:
mutual recommendation matrix MRM. On PDI. Finally, as shown in Figure 2 as an instance, then it is obvious that there are two protein nodes \( p_1 \) and \( p_2 \) associated with \( d_1 \) from the matrix \( AM_{pd} \). In addition, according to Eq. 2, we can as well obtain the recommended standard \( RMPD(p_3,d_1) = 0.71 > Std_{p3,d1} = 0.44 \). Hence, we will recommend the protein node \( p_3 \) to \( d_1 \). In the same way, the protein node \( p_4 \) will be recommended to \( d_1 \) as well. On the contrary, \( RMPD(p_2,d_2) = 0.5 \) and \( RMPD(p_5,d_2) = 0.5 \) are less than the recommended standard \( Std_{p2,d2} = Std_{p5,d2} = 1.01 \). So there is no need to recommend the protein node \( p_2 \) and \( p_5 \) to \( d_2 \). In addition, according to a previous description, it is obvious that these novel edges between \( p_3 \) and \( d_1 \), \( p_4 \) and \( d_1 \), \( p_1 \) and \( d_2 \), \( p_3 \) and \( d_2 \) will be added to the original protein–domain association matrix \( AM_{pd} \) in the same time. Similarly, we can apply the domain-based collaborative filtering algorithm. Thereafter, we can obtain a recommendation protein–domain adjacency matrix based on PDI. Finally, as shown in Figure 2. We can get the mutual recommendation matrix MRM.

To be specific, as illustrated in Figure 1, if taking the domain node \( d_1 \) as an instance, then it is obvious that there are two protein nodes \( p_1 \) and \( p_2 \) associated with \( d_1 \) from the matrix \( AM_{pd} \). In addition, according to Eq. 2, we can as well obtain the recommended standard \( RMPD(p_3,d_1) = 0.71 > Std_{p3,d1} = 0.44 \). Hence, we will recommend the protein node \( p_3 \) to \( d_1 \). In the same way, the protein node \( p_4 \) will be recommended to \( d_1 \) as well. On the contrary, \( RMPD(p_2,d_2) = 0.5 \) and \( RMPD(p_5,d_2) = 0.5 \) are less than the recommended standard \( Std_{p2,d2} = Std_{p5,d2} = 1.01 \). So there is no need to recommend the protein node \( p_2 \) and \( p_5 \) to \( d_2 \). In addition, according to a previous description, it is obvious that these novel edges between \( p_3 \) and \( d_1 \), \( p_4 \) and \( d_1 \), \( p_1 \) and \( d_2 \), \( p_3 \) and \( d_2 \) will be added to the original protein–domain association matrix \( AM_{pd} \) in the same time. Similarly, we can apply the domain-based collaborative filtering algorithm. Thereafter, we can obtain a recommendation protein–domain adjacency matrix based on PDI. Finally, as shown in Figure 2. We can get the mutual recommendation matrix MRM.

### Construction of the Weighted Protein–Protein Interaction Network

For any two given protein \( p_i \) and \( p_j \), we estimate the relationship between \( p_i \) and \( p_j \) by applying the Gaussian kernel interaction profile (van Laarhoven et al., 2011) and further obtain an \( N \times N \) dimensional weight matrix between proteins WBP based on the mutual recommendation matrix MRM. \( WBP(p_i, p_j) \) represents the relationship between protein \( p_i \) and \( p_j \), and it can be defined as follows:

\[
WBP(p_i, p_j) = \exp(-\delta_p \|IP(d_i) - IP(d_j)\|^2)
\]

where

\[
\delta_p = \frac{1}{N} \sum_{i=1}^{N} \|IP(d_i)\|^2
\]

Here, \( IP(d_i) \) and \( IP(d_j) \) represents the vector at the \( i^{th} \) and \( j^{th} \) column of the mutual recommendation matrix MRM separately. \( \delta_p \) is an adjustment coefficient, which controls kernel bandwidth based on normalizing the new bandwidth parameter \( \delta_p' \).

### Calculate the Score of Multiple Features of Protein

Previous research has indicated that with similar functions, co-expressed and complex topologies are more likely to be essential proteins. Inspired by them, in this paper, we combine biological and topological features to detect potential proteins by subcellular localizations, gene expression data, and orthologous information and PPI networks.

It is obvious that the location information of a protein in a cell is an important characteristic of essential proteins. First, we analyze the 11 kinds of subcellular location relationship between the known essential proteins, and the Figure 3 statistical distribution of each subcellular location is shown in Figure 4. We can find that essential proteins are not randomly distributed in different subcellular locations, and essential proteins appear more often in the nucleus and...
mitochondrion, which means that proteins in the nucleus and mitochondrion are more possible to be essential proteins. What is more, from Figure 4, there are more essential proteins in the nucleus and mitochondrion and a few essential proteins in the peroxisome and extracellular, which provides us with convenience.

In order to distinguish the importance of different subcellular locations, let $N_{sub}$ mean the number of all subcellular localizations and $N_{sub}(i)$ represent the number of proteins associated with the $i^{th}$ subcellular localization. Then $Ave_{sub}$ denotes the average number of proteins related to each subcellular localization. The score of the $i^{th}$ subcellular localization $Eve_{sub}(i)$ can be expressed as follows:

\[
Ave_{sub} = \frac{\sum_{i=1}^{N_{sub}} N_{sub}(i)}{N_{sub}} \quad (10)
\]

\[
Eve_{sub}(i) = \frac{N_{sub}(i)}{Ave_{sub}} \quad (11)
\]

Let $Sub_{pk}$ represent the set of subcellular localizations associated with the protein $p_k$. Therefore, for a given protein $p_k$, its subcellular localization score $Pro_{sub}(p_k)$ is computed as the sum of the scores of all subcellular locations where it appears.

\[
Pro_{sub}(p_k) = \sum_{i \in Sub_{pk}} Eve_{sub}(i) \quad (12)
\]

Similar to describing subcellular scores, for any given protein $p_k$, let $Pro_{ort}(p_k)$ mean the score of orthologous information. Hence, we can define its feature of orthology information score for $p_k$ as follows:

\[
Pro_{ort}(p_k) = \frac{Ort(p_k)}{\max_{p \in PPI} Ort(p)} \quad (13)
\]

We use the Pearson correlation coefficient (Priness et al., 2007) as a similarity measure of gene expression profiles to calculate the expression intensity of two genes.
\[ PCC(p_k, p_r) = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{Exp(p_k, i) - Exp(pk)}{\sigma(p_k)} \right) \times \left( \frac{Exp(p_r, i) - Exp(pr)}{\sigma(p_r)} \right) \]  

(14)

Here \( Exp(p_k, i) \) represents the expression level of \( p_k \) at the \( i^{th} \) time node. \( Exp(p_k) \) is the average gene expression value of protein \( p_k \), and \( \sigma(p_k) \) is the standard deviation of protein \( p_k \). Thereafter, let \( NG(p_k) \) denote the set of neighbors of protein \( p_k \). So we can compute its new functional score of protein \( p_k \) as follows:

\[ Pro_{Exp}(p_k) = \frac{exp(p_k)}{\max_{p_i \in PPI}\{exp(p_i)\}} \]  

(15)

where

\[ exp(p_k) = \sum_{p_i \in NG(p_k)} PCC(p_k, p_i) \]  

(16)

It is a fact that essential proteins are more likely products of complex functions (Dezso et al., 2003). In addition, it is obvious that triangles have stable characteristics. Inspired by this, we further utilize the major triangle topological feature calculated by the original PPI network for obtaining each protein topological feature score. Therefore, for a given protein \( p_k \), we can calculate the topological feature score as follows:

\[ Pro_{Tri}(p_k) = \frac{\sum_{p_i \in NG(p_k)} NG(p_k) \cap NG(p_i)}{NG(p_k)} \]  

(17)

Based on the above formulas for any given protein \( p_k \), we can obtain the main topological and biological feature scores.

In order to effectively solve the problem of multifeature integration, we apply an improved entropy weight method (Dastbaz et al., 2018) to automatically generate the best parameters to integrate biological features. Based on the protein characteristics we have normalized, let \{\( BF_{i1}, BF_{i2}, ..., BF_{iM} \}\} represent all features; then we can further construct an \( N \times M \) dimensional matrix \( BF \) and an \( M \times 1 \) dimensional matrix \( PM \) as follows:

\[
BF = \begin{bmatrix}
BF_{11} & \cdots & BF_{iM} \\
\vdots & \ddots & \vdots \\
BF_{N1} & \cdots & BF_{NM}
\end{bmatrix}
\]  

(18)

\[
PM = \begin{bmatrix}
p_1 \\
\vdots \\
P_M
\end{bmatrix}
\]  

(19)

Next, based on our normalized biological features, we can obtain the entropy value of each feature separately as follows:

\[ e_i = \frac{1}{\ln N} \sum_{j=1}^{N} BF_{ij} \ln(BF_{ij}) \]  

(20)

Therefore, for the \( i^{th} \) protein biological feature, we can calculate the entropy weight of each feature by the following formula:

\[ w_j = \frac{(1 - e_i)}{\sum_{i=1}^{M} (1 - e_i)} \]  

(21)

Based on the above formula, for a given protein \( p_k \), we can further calculate its integrated biological score as follows:

\[ pro_{bio}(p_k) = \sum_{k=1}^{M} w_k BF_{kj} \]  

(22)

Finally, according to the above Eq. 18, for any given protein \( p_k \), we can further obtain its initial score as follows.

\[ pro_{score}(p_k) = \lambda \times pro_{bio}(p_k) + (1 - \lambda) \times Pro_{tri}(p_k) \]  

(23)

Here, \( \lambda \) is a proportion parameter with a value between 0 and 1.

**Construction of the Prediction Model**

**Collaborative Filtering Model-Based Method**

According to \( WBP \), our prediction model CFMM can apply improved PageRank to identify potential proteins. Let \( WP(p_k, p_r) = \frac{1}{1 + \text{max}(WP(p_k, p_r))} \) and for any two given proteins \( p_k \) and \( p_r \), we can define the distribution rate possibility matrix as follows:

\[
DRPM(p_n, p_r) = \begin{cases} 
WP(p_n, p_r) \times \frac{pro_{score}(p_r)}{\sum_{k \in NG(p_n)} pro_{score}(p_k)} & \text{if } WP(p_n, p_r) \neq 0 \\
0 & \text{Otherwise}
\end{cases}
\]  

(24)

Based on the above distribution rate matrix \( DRPM \), let a possibility vector \( pro_{score}(t) \), \( pro_{score}(t+1) \) mean the score vector of protein at the \( t^{th} \) and \( t+1^{th} \) time separately; therefore, we can iteratively compute the protein ranks as follows:

\[
pro_{score}(t+1) = \alpha \times pro_{score}(t) \times DRPM + (1 - \alpha) \times pro_{score}(0)
\]  

(25)

Here the parameter \( \alpha \in (0, 1) \) in order to adjust the proportion \( pro_{score}(t) \) and initial score \( pro_{score}(0) \). Based on the above descriptions, our prediction method CFMM can be concisely described as follows.

**PERFORMANCE EVALUATION**

**Comparison Between Collaborative Filtering Model-Based Method and 14 Representative Methods**

In order to further evaluate the performance of CFMM in this section, two different datasets, the DIP database and the Krogan database, are adopted to compare CFMM with 14 competitive detection models, which include DC (Hahn and Kern, 2005), SC (Estradra and Rodriguez-Velazquez, 2005), BC (Joy et al., 2005), EC (Bonach et al., 1987), IC (Stephenson and Zelen, 1989), CC (Wuchty and Stadler, 2003), NC (J. Wang, 2012), ION (Peng et al., 2012), Pec (Li
et al., 2012), CoEWC (Zhang et al., 2013), POEM ((Zhao et al., 2014), TEGS (Zhang W. et al., 2018), CVIM (S. Li et al., 2020), and NPRI (Z. Chen et al., 2020). For the purpose of observing the accuracy of the experiment more intuitively, we chose to use a bar graph to compare the 1, 5, 10, 15, 20, and top 25% of each method. Figure 5 shows the comparison of the identifying results of different algorithms on the DIP and Krogan database separately. From Figure 5A, the newly put forward CFMM method detected a larger number of essential proteins in the top 1–25% compared with 14 other competitive methods. It is obvious that CFMM can reach the accuracy of 92.16, 83.14, 71.37, 63.87, 55.84, and 52.43% in the top 1, 5, 10, 15, 20, and 25% predicted candidate key proteins based on the DIP database. Among the top 25% proteins predicted by the CFMM method, there are 668 proteins correctly detected, which indicates that the CFMM method has superior advantages over other methods. From Figure 5B, we can see that CFMM can reach the accuracy of 94.59, 75.54, 70.03, 65.34, 60.08, and 54.68% in the top 1, 5, 10, 15, 20, and 25%, which are superior to all 14 advanced methods, except that in the top 10% CFMM-predicted 257 proteins, they are a little lower than NPRI. Therefore, we can make a conclusion that CFMM always obtains the better prediction accuracy from the top 1% to the top 25%.

**Validated by Jackknife Methodology**

Due to the jackknife methodology (Holman et al., 2009) that can evaluate the advantages and disadvantages of the prediction model, in this section, we will apply the jackknife method to assess the predictive effect of our proposed model CFMM. Figures 6, 7 show the experimental comparisons between CFMM and 14 advanced competitive methods based on the first 1,000 candidate proteins. By observing Figure 6A, it is obvious that CFMM can achieve better performance than the seven network topology-based methods including DC, SC, BC, EC, IC, CC, and NC. What is more, Figure 6B shows that the performance of CFMM is better than the other seven methods that are based on the combination of biological information of proteins and PPI networks including Pec, CoEWC, POEM, ION, TEGS, CVIM, and NPRI. From Figure 7A, we can easily conclude that the CFMM is advanced than these centrality-based methods including DC, IC, EC, BC, CC, SC, and NC. Although the performance curves of CFMM and NPRI overlap partially, as the number of candidate proteins increases to 450, the predictive performance of CFMM will be significantly higher than that of

Algorithm CFMM

Input: original protein–domain network, original PPI network subcellular data, orthologous data, expression data, the iteration termination condition $\varepsilon$, and adjustment parameter $\alpha$.

Output: the final score of proteins.

Step 1: Apply the protein-based collaborative filtering algorithm by Eqs 1–3.

Step 2: Apply the domain-based collaborative filtering algorithm by Eqs 4–6.

Step 3: Calculate the weights between proteins based on the MRM based on Eqs 7–9.

Step 4: Compute the protein feature score based on Eqs 10–23.

Step 5: Establishing distribution network based on Eq. 24.

Step 6: Let $t = t + 1$, calculate $\text{proscore}(t + 1)$ according to Eq 26.

Step 7: Repeat step 6 until $\text{proscore}(t + 1) - \text{proscore}(t) < \varepsilon$.

Step 8: Sorting the proteins scores $\text{proscore}(t + 1)$ through descending order.
NPRI. Therefore, based on the above description, we can make a conclusion that the performance of CFMM is not only superior to the first category of methods, such as DC, IC, EC, SC, BC, CC, and NC, but also better than these multiple biological data methods including Pec, CoEWC, POEM, ION, TEGS, CVIM, and NPRI.

**Differences Between Collaborative Filtering Model-Based Method and Competitive Methods**

In order to further prove the accuracy of the CFMM model, we will analyze the differences between CFMM and other models based on the top 100 predicted proteins under the DIP database and the Krogan database separately, and comparison results are shown in Tables 2, 3, respectively. Here ME denotes one of the 14 competitive methods. \( |\text{CFMM} \cap \text{ME}| \) represents the number of essential proteins predicted by both CFMM and ME. \( |\text{CFMM} - \text{ME}| \) denotes the number of essential proteins recognized by the CFMM but not by ME, and \( |\text{ME} - \text{CFMM}| \) means the number of key proteins predicted by ME but ignored by CFMM. In addition, \( |\text{CFMM} - \text{ME}| \) represents the set of key proteins recognized by CFMM but not by ME. \( |\text{ME} - \text{CFMM}| \) means the set of essential proteins predicted by ME but not by CFMM. Hence, Tables 2, 3 show the difference between the 14 competitive methods and CFMM under the DIP and Krogan datasets separately. Figure 8 indicates that CFMM can achieve much better predictive performance than all these competing methods as a whole.

**Validation by Receiver Operating Characteristic Curve**

The receiver operating characteristic (ROC) curve and precision recall curve (PR) are used to scientifically prove the performance of the
TABLE 2 | The connection and difference between CFMM and 14 competing methods based on the top 100 ranked proteins in the DIP database.

| Different methods (ME) | |CFMM ∩ ME| |CFMM – ME| Percentage of key proteins in (%)|CFMM – ME| Percentage of key proteins in (%)|ME – CFMM |
|-----------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| DC                    | 6               | 94              | 88.30           | 42.55           |
| IC                    | 6               | 94              | 88.30           | 40.43           |
| EC                    | 6               | 94              | 88.30           | 32.98           |
| SC                    | 6               | 94              | 88.30           | 32.98           |
| BC                    | 5               | 95              | 88.42           | 41.05           |
| CC                    | 5               | 95              | 88.42           | 37.89           |
| NC                    | 35              | 65              | 89.23           | 36.92           |
| Pec                   | 46              | 54              | 87.04           | 59.26           |
| CoEWC                 | 47              | 53              | 84.91           | 54.72           |
| POEM                  | 56              | 44              | 84.09           | 65.91           |
| ION                   | 38              | 62              | 88.71           | 70.97           |
| TEGS                  | 58              | 42              | 80.95           | 64.29           |
| CVIM                  | 44              | 56              | 85.71           | 83.93           |
| NPRI                  | 76              | 24              | 91.67           | 87.50           |

TABLE 3 | The connection and difference between CFMM and 14 competing methods based on the top 100 ranked proteins in the Krogan database.

| Different methods (ME) | |CFMM ∩ ME| |CFMM – ME| Percentage of key proteins in (%)|CFMM – ME| Percentage of key proteins in (%)|ME – CFMM |
|-----------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| DC                    | 17              | 83              | 84.34           | 42.17           |
| IC                    | 12              | 88              | 85.23           | 44.32           |
| EC                    | 5               | 95              | 86.32           | 38.95           |
| SC                    | 5               | 95              | 86.32           | 38.95           |
| BC                    | 8               | 92              | 85.87           | 40.22           |
| CC                    | 5               | 95              | 86.32           | 43.16           |
| NC                    | 48              | 52              | 88.46           | 50.00           |
| Pec                   | 43              | 57              | 77.19           | 56.14           |
| CoEWC                 | 41              | 59              | 77.97           | 52.54           |
| POEM                  | 45              | 55              | 85.45           | 58.18           |
| ION                   | 30              | 70              | 82.86           | 65.71           |
| TEGS                  | 58              | 42              | 80.95           | 52.38           |
| CVIM                  | 67              | 33              | 75.76           | 72.73           |
| NPRI                  | 61              | 39              | 76.92           | 53.85           |

FIGURE 8 | The X-axis represents different protein predicted methods. The Y-axis represents the proportion of essential proteins in (ME–CFMM) or (CFMM–ME).
prediction model. The area under the curve (AUC) is used to evaluate the performance of the prediction method. The closer the AUC value is to 1, the better the prediction performance of the method. The curve can be plotted by the ratio of true positive rate (TPR) to false positive rate (FPR) according to different thresholds (Peng et al., 2020). Hence, we will further utilize the ROC curves to compare CFMM with other advanced models. Figures 9, 10 indicate that the ROC curves and PR curves of CFMM and other competitive models are based on the DIP and Krogan databases separately. It is obvious that CFMM has a higher AUC curve than other competitive models. Although we can see that the ROC curve of CFMM and the NPRI ROC curves overlap slightly, the AUC value of CFMM is higher than NPRI. Finally, in order to prove the applicability of CFMM, we will further test it in the Gavin database and compare with other methods. The experimental results are shown in Tables 4, 5.

**The Analysis of Parameter**

In this section, we discuss the effect of the two self-defined parameters $\alpha$ and $\lambda$ on the prediction results of CFMM. We set the parameter $\alpha$ to vary from 0.1 to 0.9, then the CFMM algorithm is run nine times from $\alpha = 0.1$ to $\alpha = 0.9$ separately. Finally, the number of true essential proteins identified by CFMM based on the DIP and Krogan databases are shown in Tables 6, 7 separately. Here we select from the top 1% to the top 25% of the proteins identified by CFMM. The prediction accuracy is based on the number of essential proteins that are truly identified. It is obvious that the closer $\alpha$ value is to 1, the higher the prediction accuracy CFMM can achieve. So, we consider that the parameter $\alpha$ on all the databases is 0.9, which can achieve the best performance. When $\alpha$ is set to 0.9, and $\lambda$ is set to 0.65, the amount of true essential protein is closest to its average level. Therefore, as a result, we will set $\alpha$ and $\lambda$ on the DIP and Krogan databases to 0.9 and 0.65 separately, while for the Gavin database, the optimum parameters $\alpha$ and $\lambda$ will be set to 0.9 and 0.8, respectively.

**FIGURE 9** | The precision recall (PR) curves and receiver operating characteristic (ROC) curves between CFMM and other advanced methods based on the DIP database. (A) The PR curves and the ROC curves of DC, BC, SC, NC, EC, IC, and CC. (B) The PR curves and the ROC curves of Pec, CoEWC, POEM, ION, TEGS, CVIM, and NPRI.
DISCUSSION

Accumulating evidence have shown that prediction of essential proteins is important for the development of an organism in biological process, complex disease diagnoses, and drug design. However, the requirement of identifying key protein prediction accuracy is not satisfied only through biological experiments and relying on the topological characteristics of the PPI network. In this manuscript, we constructed an original protein–domain network by combining protein and domain associations first. Then we formulated the prediction of potential essential proteins as a problem of the recommendation system and obtained an updated recommendation network through applying a novel mutual recommendation between protein and domain to the original association network. Next, after we integrate the biological features, we combine with the major topological features to obtain the initial protein score. Finally, we design a
novel distribution rate matrix and apply an iterative algorithm based on the improved PageRank algorithm to calculate protein scores iteratively. In addition, we apply the CFMM method on the DIP database, Krogan database, and Gavin database to testify the performance, respectively. Experiments show that CFMM can achieve better performance than other advanced methods. In future work, we will use multi-information fusion method to integrate various information related to proteins and machine learning methods to further improve the prediction performance (Peng et al., 2017; Zhou et al., 2019).

TABLE 5 | The number of key proteins recognized by CFMM and other methods based on the Gavin database.

| Methods | Top 1% (19) | Top 5% (93) | Top 10% (196) | Top 15% (279) | Top 20% (371) | Top 25% (464) |
|---------|------------|-------------|--------------|--------------|--------------|--------------|
| DC      | 7          | 36          | 101          | 158          | 222          | 264          |
| IC      | 16         | 55          | 119          | 163          | 213          | 254          |
| CC      | 11         | 45          | 93           | 136          | 180          | 221          |
| BC      | 9          | 40          | 85           | 122          | 162          | 201          |
| SC      | 0          | 17          | 87           | 130          | 190          | 240          |
| EC      | 0          | 38          | 94           | 134          | 166          | 209          |
| NC      | 11         | 51          | 123          | 170          | 213          | 259          |
| CoEWC   | 16         | 69          | 136          | 190          | 237          | 275          |
| Pec     | 15         | 69          | 142          | 193          | 238          | 285          |
| ION     | 17         | 73          | 150          | 207          | 263          | 312          |
| POEM    | 17         | 74          | 148          | 199          | 249          | 296          |
| CVIM    | 16         | 80          | 160          | 219          | 271          | 322          |
| NPRI    | 18         | 75          | 153          | 221          | 278          | 322          |
| CFMM    | 19         | 84          | 162          | 222          | 280          | 332          |

TABLE 6 | Effects of the parameter α to CFMM based on the DIP database.

|       | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Rank  |     |     |     |     |     |     |     |     |     |
| Top 1% (51) | 47  | 47  | 47  | 47  | 47  | 47  | 47  | 47  | 47  |
| Top 5% (255) | 206 | 208 | 207 | 208 | 209 | 209 | 210 | 213 | 212 |
| Top 10% (610) | 357 | 357 | 358 | 361 | 361 | 359 | 358 | 360 | 364 |
| Top 15% (764) | 469 | 473 | 474 | 476 | 480 | 483 | 485 | 485 | 488 |
| Top 20% (1,019) | 572 | 574 | 573 | 573 | 571 | 575 | 576 | 573 | 569 |
| Top 25% (1,274) | 650 | 653 | 657 | 656 | 658 | 661 | 665 | 667 | 668 |

TABLE 7 | Effects of the parameter α to CFMM based on the Krogan database.

|       | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Rank  |     |     |     |     |     |     |     |     |     |
| Top 1% (51) | 36  | 36  | 36  | 36  | 36  | 35  | 35  | 35  | 35  |
| Top 5% (255) | 141 | 140 | 140 | 139 | 140 | 140 | 140 | 138 | 139 |
| Top 10% (610) | 255 | 255 | 255 | 254 | 254 | 253 | 254 | 256 | 257 |
| Top 15% (764) | 366 | 365 | 364 | 364 | 363 | 363 | 363 | 360 | 360 |
| Top 20% (1,019) | 442 | 443 | 442 | 444 | 444 | 443 | 441 | 441 | 441 |
| Top 25% (1,274) | 497 | 496 | 497 | 496 | 498 | 498 | 499 | 499 | 500 |

Author Contributions

XZ and XH conceived the study. XZ, XH, LK, and ZC improved the study based on the original model. XZ and XH implemented the algorithms corresponding to the study. ZC and LK supervised the study. XZ and XH wrote the manuscript. All authors including CL reviewed and improved the manuscript.

Funding

This research is partly sponsored by the Research Foundation of Education Bureau of Hunan Province (No. 20B080), the Natural Science Foundation of Hunan Province (No. 2019J70010), the Hunan Provincial Natural Science Foundation of China (2020J4152), and the Science and Technology Plan Project of Hunan Province (2016TP1020). The Hunan Province Science and Technology Project Funds (2018TP1036), the National Scientific Research Foundation of Hunan Province Education Commission (18B367).

Acknowledgments

The authors sincerely thank all the teachers and students who participated in this study for their guidance and help.

Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.763153/full#supplementary-material
