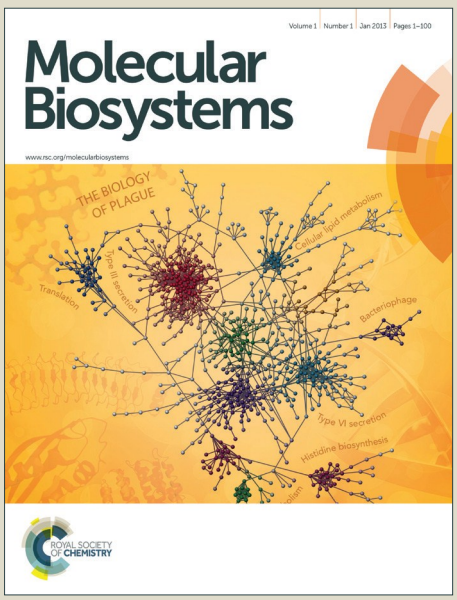
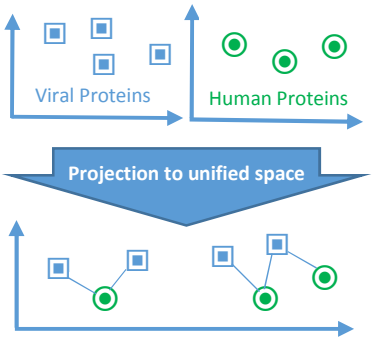
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems



Computational Prediction of Virus-Human Protein-Protein Interactions using Embedding Kernelized Heterogeneous Data

Table of Contents

Molecular BioSystems

ARTICLE

Computational Prediction of Virus-Human Protein-Protein Interactions using Embedding Kernelized Heterogeneous Data

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Esmaeil Nourani[a], Farshad Khunjush[a,b,†] and Saliha Durmuş[c]

Pathogenic microorganisms exploit host cellular mechanisms and evade host defense mechanisms through molecular pathogen-host interactions (PHIs). Therefore, comprehensive analysis of these PHI networks should be an initial step for developing effective therapeutics against infectious diseases. Computational prediction of PHI data is gaining increasing demand because of scarcity of experimentally-found data. Prediction of protein-protein interactions (PPIs) within PHI systems can be formulated as a classification problem, which requires the knowledge of non-interacting protein pairs. This is a restricting requirement, since we lack datasets which report non-interacting protein pairs. In this study, we formulated the "computational prediction of PHI data" problem using embedding kernelized heterogeneous data. This eliminates the above-mentioned requirement and enables us to predict new interactions without randomly labeling protein pairs as non-interacting. Domain-domain associations are used to filter the predicted results leading to 175 novel PHIs between 170 human proteins and 105 viral proteins. To compare our results with the state of the art studies whose approach is using a binary classification formulation, we modified our settings to consider the same formulation. Detailed evaluations are conducted and our results provide more than 10 percent improvements for accuracy and AUC (area under the receiving operating curve) results in comparison with the state of the art methods.

Introduction

A global rise of human infectious disease outbreaks brings the necessity of facing these major health threats more than ever[1]. Emerging viral diseases such as MERS (Middle East Respiratory Syndrome), Hepatitis C, H1N1 influenza, Ebola and other major viral infections like HIV cause serious morbidity and mortality rates worldwide. United States recently warned about growing hepatitis C outbreak in several states. Currently, South Korea is experiencing a deadly outbreak of MERS. The similarities between MERS and SARS (Severe Acute Respiratory Syndrome) which occurred in Hong Kong and Singapore in 2003[2] emphasize the importance of a thorough understanding of the underlying mechanisms to prevent their recurrence in the future.

Systems biology is widely accepted as a promising approach to reveal characteristics of diseases through molecular interaction networks. Interactions between proteins of pathogens and hosts are the crucial parts of the infection mechanisms[3]. This motivates researchers to focus on studying pathogen-host interactions (PHIs) whose experimental verification is challenging and time consuming. Therefore, rather than evaluating all possible PHIs, which is extremely unjustifiable, computational approaches can pave the way for these experiments by predicting high potential PHIs. Computational studies for evaluating protein interactions and their associated networks have been initiated more than a decade ago[4]. Most of the corresponding studies primarily focus on protein-protein interactions (PPIs) within a single organism (intra-species PPI prediction). Inter-species PPI prediction including PHIs has not gained significant share of the research yet. However, a few studies have been conducted in this field on different PHI systems using various machine learning techniques[5].

The current PHI knowledge suffers from scarcity of available experimentally-verified PHI data. Therefore, development of efficient computational approaches to predict PHI data is required urgently. The experimental PHI data are collected within a number of databases like PATRIC[6], VirusMentha[7], VirHostNet[8] and PHISTO[9]. These data can be used to discover the interaction patterns between pathogen and host proteins in order to use in computational prediction processes. Discovering pathogen-host protein binaries as the most probable pairs to interact can be formulated as a binary classification problem, in which interacting and non-interacting pairs should be distinguished. To train the model, both positive and negative samples are required. In the case of negative samples there is a serious obstacle, since the PHI databases report only positive interactions without listing non-interacting pathogen-host pairs.

Random sampling within all possible pathogen-host protein pairs (after discarding known interacting pairs) is a common approach to construct negative samples. Consequences of

[a] Department of Computer Science and Engineering, School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran.
[b] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran.
[c] Computational Systems Biology Group, Department of Bioengineering, Gebze Technical University, Kocaeli, Turkey.
†Correspondence: Farshad Khunjush, Department of Computer Science & Engineering, School of Electrical & Computer Engineering, Zand Avenue, Shiraz 71348 - 51154, Iran, khunjush@shirazu.ac.ir

selecting random samples as non-interacting pairs can be unpredictable and the results will be dependent on selected negative class. Here, we propose a novel approach for computational PHI prediction, which learns interaction pattern based on available positive samples. In this approach, latent features are extracted for each of pathogen and host protein, integrated with explicit features in the form of kernel matrices. The proposed approach is based on a method presented by Gonen[10] for embedding heterogeneous data. We consider each protein of pathogen and host in a K dimensional space and learn these features using PHI interaction patterns and kernel matrices designed by similarities among protein features. Training process tries to map these R dimensional coordinates from both pathogen and host to an embedded space, in which close proximity of proteins means higher probability of interaction. We compare our approach with various state of the art methods for PHI predictions[11,12,13]. Various experiment scenarios and evaluation metrics verify outperformance and validity of our method for diverse PHI systems.

This paper is organized as follows, first we present our method for predicting protein-protein interactions between viral and human proteins. Then, we illustrate various features and how they are designed and used as kernel matrices. Finally, results of the experiments on various PHI systems and biological assessments of predicted PHIs are reported at the final section of this study.

Methods

We address the problem of PHI prediction by projecting pathogen and host latent features into a unified embedding space. This embedding is based on available samples of PHIs, integrated with multiple kernels designed over pathogen and host protein features.

Figure 1 shows each viral (V_i) and human protein (H_i) represented by a K dimensional latent vector. Adjacent proteins in the projected unified space are considered to have an interaction.

As shown in Figure 2, an interaction matrix includes the known PHIs, set to be '1' and others marked as unknown. It is clear that relying exclusively on the interaction matrix, could not reveal the interaction pattern and consequently could not accurately predict the missing entries. This is due to the fact that in most of the pathogenic systems, currently we have access only to a small fraction of interactions, that is, one of the used datasets [11] includes 1035 available PHIs, whose ratio to all possible interactions, 106720, is about 0.0096. This leads to an extremely sparse matrix.

To overcome this limitation, we feed the model with similarity kernels, which give different similarity scores for proteins in each domain.

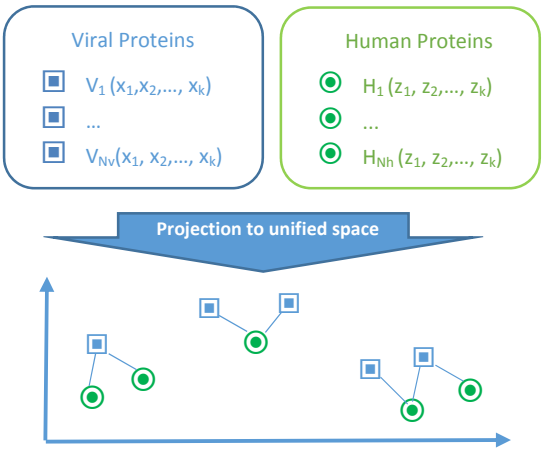


Figure 1- Projecting latent features of human and viral proteins into unified space

We call viral and human proteins, as different domains of the problem. The idea here is that the similar proteins have similar interaction behaviour. Similarity measures are designed for viral and human proteins using different features. Our approach can be easily extended to consider multiple kernels for each domain, however, to keep the formulation as simple as possible, we consider one kernel for each domain.

Viral and human proteins are converted into *K*-dimensional vectors in the Euclidean space represented by

$$E_v = \{e_{v,i} \in \mathbb{R}^k\}_{i=1}^{N_v} \quad , \quad E_h = \{e_{h,i} \in \mathbb{R}^k\}_{i=1}^{N_h} \qquad (1)$$

N_v and N_h are the number of viral and human proteins, respectively. Learning projected coordinates are based on approximating three scoring functions. As shown in Figure 2, S_v indicates similarity score between viral proteins, S_h indicates similarity score between human proteins and S_{vh} as follows:

$$S_{vh} = \begin{cases} 1 & if\ v_i\ and\ h_j\ interacting \\ NA & otherwise \end{cases} \qquad (2)$$

S_v, S_h and S_{vh} are approximated by three kernel functions K_v, K_h and K_{vh}, respectively. Radial basis functions with kernel width σ are used to approximate similarity scores using projected coordinates.

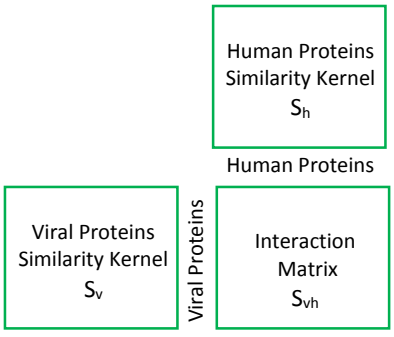


Figure 2- Integrating multiple kernels and interaction matrix to obtain unified space

$$K_v = exp\left(-\frac{||\, e_{v,i} - e_{v,j}||_2^2}{\sigma^2}\right) \forall i,j\ ,$$

$$K_h = exp\left(-\frac{||\, e_{h,i} - e_{h,j}||_2^2}{\sigma^2}\right) \forall i,j\ \ ,$$

$$K_{vh} = exp\left(-\frac{||\, e_{v,i} - e_{h,j}||_2^2}{\sigma^2}\right) \forall i,j \qquad (3)$$

The loss function to simultaneously approximate similarity functions S_v, S_h and the interaction function S_{vh} are formulated as follows:

$$\mathcal{L} = \frac{\lambda_{vh}}{|\phi_{vh}|}\sum_{\phi_{vh}}(K_{vh} - S_{vh})^2 + \frac{\lambda_v}{|\phi_v|}\sum_{\phi_v}(K_v - S_v)^2 + \frac{\lambda_h}{|\phi_h|}\sum_{\phi_h}(K_h - S_h)^2 \qquad (4)$$

The optimization function is formulated as follows: Minimize $\mathcal{L}$ with respect to E_v, E_h and σ, where ϕ_{vh}, ϕ_v and ϕ_h are index sets indicating available indices of S_{vh}, S_v and S_h. Considering index set ϕ_{vh} is crucial since S_{vh} represents all possible interactions between N_v viral and N_h human proteins. However experimentally detected interactions include a very small fraction of all N_v*N_h indices. Therefore, ϕ_{vh} is considered to indicate which indices are available to be used by training. S_v contains similarity values between viral proteins and ϕ_v indicates available indices of S_v. Accordingly, S_h and ϕ_h are considered for similarity between human proteins. Kernel width σ is not a constant parameter and will be learned during the optimization process. Our proposed method, *adapted multiple kernel preserving embedding* (AMKPE) is an adapted version of *(MKPE)* method originally presented by [10] and produces projected coordinates E_v and E_h in the target unified space. K_{vh} contains the desired output, since it is trained as an approximation for the interaction matrix (S_{vh}) and represents interaction scores for all virus-human protein pairs.

Feature Sets

We use a minimal set of explicit features to show the advantage of exploiting latent features, extracted from known interaction patterns. Regarding the scarcity of available feature values, especially for pathogen proteins, using limited number of features heals the problem of confronting considerable missing features. Biological features utilized for PHI prediction are summarized in [5]. We use some of them and add new features, which have not utilized yet. We make use of latent features along with explicit features, which lead to a minimal set of required explicit features in comparison to other approaches. This is important, since using a large set of features with a small set of available samples, may end up with overfitting due to insufficient training data.

Despite we used a collection of features, we report only the results of features, which contribute most for each pathogenic system and dataset. In the following section, we introduce utilized features and explain how they are prepared for the experiments.

Sequence Features

There are contradictory reports about the efficiency of features extracted from protein sequences. Shen [14] uses only sequence features to predict PPIs, whereas other studies[15] claim that solely depending on sequence features cannot provide promising results. We should notice that protein sequence is rather complete and ubiquitous representation of protein which can shed light on protein functions. In this study, we create a kernel based feature from protein sequence using the spectrum kernel [16]. Sequence kernels can be described as follows:

$$K(x,y) = \sum_{m\in\mathcal{M}} N(m,x).N(m,y) \qquad (5)$$

Where x and y are protein sequences, $\mathcal{M}$ contains all possible k-mers, m is one pattern of feature space $\mathcal{M}$ and $N(m,x)$ is the number of occurrences of pattern m in sequence x. The kernel is computed by the dot product of the two-feature vectors for the sequences, x and y. Kernels are normalized and feature vectors are scaled to the unit sphere, which represents the cosine similarity between feature vectors. Kernel normalization is performed as follows:

$$k'(x,y) = \frac{k(x,y)}{\sqrt{k(x,x).k(y,y)}} \qquad (6)$$

We use mixed spectrum kernel [17] in which multiple length for k-mers are simultaneously taken into account to compute the similarity value. We create our mixed kernel ($K_{SequenceK\text{-}mers}$) for k=1, 2, 3 and 4, then combine using equally weighted sum of four kernels.

Network Topology Measures

Systems biology exploits a lot of features from networks of biological nodes. Here, the network is formed by the interactions between pathogen and host proteins, resulting in a bipartite graph. Within this network, pathogens tend to target vital host proteins to maximize their ability to manipulate host cell mechanisms. Node centrality measures can be an appropriate measure for identifying such vital proteins. We use degree centrality of human proteins as the most important topology feature. Gaussian kernel can be used as follows to create a kernel matrix using degree values for human proteins ($Degrees_h$).

$$K_{Degree} = exp\left(-\frac{||\, Degrees_h||_2^2}{{\sigma_d}^2}\right) \qquad (7)$$

Where σ_d is the kernel width and can be estimated using Jaakkola's heuristic [18] as an initial guess. We consider all pairs of protein degrees and compute the differences between all pairs. The median of these differences is used as a guess for σ and can be computed as follows:

$$\sigma_{Jaakkola} = median(||degree_{h,i} - degree_{h,j}||\ \forall i,j) \qquad (8)$$

Domain Information

Domains are building blocks of proteins and act as the mediators of interactions. A large number of studies were performed using domain-domain interaction knowledge to predict PHIs [5]. In this study, we use domain information in two manners. Firstly, we collect domains for each protein and create similarity kernels based on the occurrence of same domains between protein pairs. The number of same domain occurrence in each protein pair is computed and normalized to create a similarity kernel (K_{Domain}). Due to the scarcity of available domain information, resulting matrix is sparse. However, it gives an effective similarity measure for proteins. Secondly, we count the occurrence of domains in virus-human interacting protein pairs and find out which domain pairs frequently occurred in interacting pairs. This can be used as a measure for evaluating each candidate virus-human protein pair. If the candidate contains domain pairs, which are frequently observed in PHIs, it can be considered as a high probable interaction. The similar idea is used in [11] by inferring virus-host, domain-domain associations from interacting protein pairs. They report domain-domain association as one of the best features for PHI prediction. We use this feature for filtering semi-final prediction results into a limited list of candidate PHIs. We obtained the list of domains and protein families from InterPro[I] database[19]. InterPro tries to classify protein families and domains by integrating different protein family databases with different biological focuses and approaches.

Pathway Membership

Biological pathway is a graph connecting proteins, which are involved in a certain biological process. Pathogens tend to target specific biological processes to use host functions for their own advantage. Each human protein may be member of several pathways. We compute a similarity measure between human proteins based on their membership in different biological pathways.

We collect the list of pathways for human proteins from *Reactome*[II] database [20]. For each protein we consider a binary feature vector (P) with length of pathways count and set to be '1' each entry that the human protein takes part in the corresponding pathway.

$$K_{pathway}(h_i, h_j) = count\left(and(p_i, p_j)\right) \quad (9)$$

Where hi and hj are two human proteins, $\boldsymbol{K_{pathway}(h_i, h_j)}$ gives an un-normalized pathway membership similarity value for two proteins. And (x,y) does a logical 'and' between two binary vectors and set to be one for each entry of returned vector, where both proteins participate in the corresponding pathway, and count (x) returns the number of non-zero elements of vector x, indicating un-normalized similarity value between two proteins. Kernel values are normalized using Equation (6).

Gene Ontology

The Gene Ontology (GO) is a standard for the annotation of gene products. Extracted features from GO have provided significant results for pathogen-host and interspecies interaction prediction [21]. Terms of three orthogonal ontologies including 'BP' (biological process), 'MF' (molecular function) and 'CC' (cellular component) can be used as a basis of computing quantitative semantic similarity. In this paper, we compute three similarity kernels based on semantic comparisons of GO annotations using Wang's measure [22] which is a graph based method and uses topology of GO graph structure. This method exploits both the locations of GO terms in the graph and the relations with their ancestor terms. We use GOSemSim[23] implementation of this method to compute multiple kernels for human proteins, as K_{GO_BP}, K_{GO_MF} and K_{GO_CC}.

Datasets

We use three different datasets for evaluation of our method with state of the art studies for PHI prediction. First virus-host PPI dataset[11], obtained from VirusMINT[III] [24]. The dataset includes several medically significant viral pathogens including human immunodeficiency virus 1 (HIV-1), simian virus 40 (SV40), hepatitis B virus (HBV), hepatitis C virus (HCV) and papilloma virus. After eliminating redundant interactions and those have not any InterPro domain hit, 1035 virus-human interactions remained out of 2707 initial interactions. This dataset covers 160 viral proteins and 667 host proteins. Second dataset[12], contains HCV and human interactions from the IntAct database[IV] [25] including PHIs which are annotated as 'physical association' or 'direct interaction'. It contains 657 interactions between 52 HCV proteins and 420 human proteins. Third dataset[26], contains retroviridae-human interactions from PHISTO[V] [9]. It includes 9439 interactions between 292 viral proteins and 1108 human proteins. Viral pathogens include 12 retroviruses, including Abelson murine leukemia virus, Avian myeloblastosis virus (AMV), Bovine leukemia virus, Equine infectious anemia virus, HIV-1, HIV-2, Mason-Pfizer monkey virus (MPMV), Murine leukemia virus, Primate T-lymphotropic virus 1, Rous sarcoma virus, Simian foamy virus, Simian immunodeficiency virus (SIV) and Y73sarcoma virus. Properties of the datasets are summarized in Table 1.

Table **1**. Summary of the employed datasets

Pathogen	Host	Number of pathogen Proteins	Number of host proteins	All possible interactions	Available Interactions	Sparsity
1- Medically Significant Viruses	Human	160	667	106720	1035	0.96%
2-HCV	Human	52	420	21840	657	3.00%
3-Retroviridae Viruses	Human	292	1108	323536	9439	2.92%

Experiments

We conduct different sets of experiments to evaluate our PHI prediction method. The method has the flexibility to formulate the problem as a classification problem using both negative and positive labels, and also formulate the problem without using negative samples. Most of the previous studies for PHI prediction are formulated as a classification problem, which requires negative samples. Unfortunately most of the databases lack protein pairs which do not have interaction. Nevertheless, in the first set of experiments we formulate the problem as a classification problem on the same dataset used in the state of the art studies [11,12] to compare the results of our method. For prediction of new PHIs, we use a second set of experiments, in which no negative samples are used. The second formulation was also compared with the study proposed in [26] which similarly do not use negative samples by means of formulating the problem as matrix factorization.
For both settings interaction matrix is reconstructed using Euclidean distance between projected coordinates of proteins.

Since different evaluation metrics are used for the state of the art methods, we use same measures to evaluate the accuracy of our method. The list of the metrics and their computation formula is presented in Table 2. Parameter set used in the experiments are as follows, (λ_{vh}, λ_v, λ_h, *K, Iteration*) set to (1, 0.1, 0.1, 10, 100). Magnitude of λ_{vh} in comparison with λ_v and λ_h is reasonable, since the essential part of the objective function is focusing on S_{vh} which is prior knowledge about available PHIs. To reconstruct the interaction matrix and predict new PHIs, we can use computed interaction scores of K_{vh}, since K_{vh} represents scores for all virus-human protein pairs, including hided or previously unknown entries.

Table **2**. Evaluation metrics used for PHI prediction

Metric	Formula
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$
AUC	The area under the ROC curve
True Positive (TP) - True Negative (TN) False Positive (FP) - False Negative (FN)	

Formulating PHI Prediction as a Classification Problem

To formulate the problem as a classification problem, we should modify equation (2) to take into account the possibility of representing non-interacting protein pairs. Samples of negative class are denoted by '0', which is in equation (10).

$$S_{vh} = \begin{cases} 1 & if\ v_i\ and\ h_j\ interacting \\ 0 & if\ v_i\ and\ h_j\ not\ interacting \\ NA & otherwise \end{cases} \quad (10)$$

Since non-interacting PPIs are not available for different PHI systems, providing a negative class is one of the challenges for PHI prediction. Random sampling from all pathogen-host protein pairs is a well-accepted method to construct a negative class. [27] removed sub-cellular, co-localized pairs from negative samples and reported better results in comparison to random sampling. However, this method can be criticized for additional bias through dominating localization information.
The next issue is selecting the ratio of negative to positive samples. It should be noted that using very unbalanced ratio may introduce a biased model. In this study, we chose 1:1 and 1:4 for Dataset 1 and Dataset 2, respectively. This is equal to the ratios used in [11] and [12] to provide a fare comparison.

Medically Significant Viruses

The first set of experiments was conducted on several medically significant viral pathogens. We compare our results with the study, which used the same dataset. Results of our 5-fold cross validation experiments with combination of different kernels are shown in Table3 . Results of the proposed method are significantly better than those of [11].

Table**3** . Classification results for medically significant viruses

Pathogen Proteins Kernel	Human Proteins Kernel	Accuracy (%)	AUC
-	-	60.6	0.656
K_{Domain}	$K_{SequenceK\text{-}mers}$	**79.6**	**0.832**
K_{Domain}	K_{Degree}	74.6	0.817
K_{Domain}	K_{GO_CC}	77.3	0.827
$K_{SequenceK\text{-}mers}$	$K_{SequenceK\text{-}mers}$	75.7	0.812
$K_{SequenceK\text{-}mers}$	K_{degree}	73.9	0.815
$K_{SequenceK\text{-}mers}$	K_{GO_CC}	76.5	0.820
Results of [11]		72.4	0.76

Hepatitis C Virus- Human PPIs

Hepatitis C virus (HCV) causes mostly chronical liver infection which leads to long-term serious health problems and even death. About 3% of the world population are chronically infected by HCV and thousands of deaths are reported annually caused by HCV [12]. Dataset 2 contains available HCV-human PPIs and are used in our experiment. We compare the accuracy of our results with the last reported study [12]. It should be noted that the reported accuracy metric is sensitive to the ratio of positive to negative class. For a fare comparison, we use the same ratio of 1:4. Results of our 10-fold cross validation experiments with combination of different kernels are shown in Table 4. The reported accuracy value in [12] is 0.83 .

Table **4**. Classification results for HCV

Pathogen Proteins Kernel	Human Proteins Kernel	Accuracy (%)	AUC
-	-	65.1	0.643
K_{Domain}	$K_{SequenceK\text{-}mers}$	87.6	0.870
K_{Domain}	K_{Degree}	88.0	0.857
K_{Domain}	K_{GO_CC}	**88.3**	**0.873**
$K_{SequenceK\text{-}mers}$	$K_{SequenceK\text{-}mers}$	87.7	0.872
$K_{SequenceK\text{-}mers}$	K_{degree}	88.4	0.868
$K_{SequenceK\text{-}mers}$	K_{GO_CC}	88.1	0.872
Results of [12]		83	-

Formulating PHI Prediction using Positive and Unlabelled Data

The superiority of our method is presented in the classification formulation. However, as discussed before, considering the obstacles for selecting dependable negative samples, we put aside random negative samples to predict new interactions. We change the formulation as shown in equation (11).

$$S_{vh} = \begin{cases} 1 & if\ v_i\ and\ h_j\ interacting \\ NA & otherwise \end{cases} \quad (11)$$

We sort the non-interacting pairs based on their Euclidean distances resulted in the embedding space and select the top N pair with the smallest distances. To evaluate the performance of our method in the new formulation, we compare the results using another evaluation method. [26] uses 'Hit Rate' to measure performance of their model. 'Hit Rate' can be defined as equation (12).

$$hitrate = \frac{|D_{cand} \cap D_{test}|}{|D_{test}|} \quad (12)$$

Where D_{test} contains indices of hided actual PHIs from S_{vh}. We try to check if the method can distinguish hided actual PHIs from other unknown indices. Entries of reconstructed interaction matrix will be sorted based on the computed interaction score. The method is expected to place indices of hided actual PHIs within the top ranks by giving them a large score. We pick $|D_{test}|$ top ranked entries, called as D_{cand} to see which fraction of hided actual PHIs are discovered among the top ranked indices. It should be noted that the top results may contain actual PHIs, which are not discovered yet.The model counts only the hided actual PHIs which place at the top. However, this method of evaluation is also implemented for similar areas like drug-target interaction prediction [28].

Furthermore, for each pathogen protein, we order all of the human proteins based on the predicted value of interaction probability. Then, for every pathogen–human interaction pairs of hidden part, we record the rank of the human proteins. Having a rank 1 for the target human protein between 1108 proteins means an ideal prediction result. We set the threshold *r* and computed the percentage of hided interactions which gained the rank lower than *r*.

This gives the performance of the method for recovering a hided interaction in the *top-r* predictions for a specific pathogen protein. We compute same results for each human protein. Considering different sample sizes for pathogen and human proteins, we set *r* to be 15 and 50 respectively. Clearly small values of *r* will be valuable for biologists. Recently [29] used this measure for evaluation of gen-disease association prediction. We apply different sets of kernels on Data set 3 and report the results in Table5 .

We apply two kinds of kernels for pathogen proteins, where K_{Domain} performs better by taking an average value of all performance measures presented in Table 5.

Table**5** . Performance results for Data Set 3 using different kernels

Pathogen Proteins kernel	Human Proteins kernel	Hit among top15 pathogen (%)	Hit among top50 host (%)	AUC	Hit Rate
K_{Domain}	$K_{SequenceK\text{-}mers}$	41	85	0.93	0.82
K_{Domain}	K_{Degree}	41	84	0.79	0.76
K_{Domain}	K_{GO_CC}	42	85	0.85	0.77
K_{Domain}	K_{GO_MF}	41	83	0.74	0.76
K_{Domain}	K_{GO_BP}	41	86	0.92	0.81
K_{Domain}	$K_{Pathway}$	40	85	**0.93**	**0.82**
K_{Domain}	K_{Domain}	41	86	0.93	0.81
Average		**41**	**85**	**0.87**	**0.79**
$K_{SequenceK\text{-}mers}$	$K_{SequenceK\text{-}mers}$	40	85	**0.91**	**0.82**
$K_{SequenceK\text{-}mers}$	K_{Degree}	41	82	0.67	0.70
$K_{SequenceK\text{-}mers}$	K_{GO_CC}	40	85	0.75	0.75
$K_{SequenceK\text{-}mers}$	K_{GO_MF}	40	82	0.56	0.66
$K_{SequenceK\text{-}mers}$	K_{GO_BP}	40	86	0.89	0.82
$K_{SequenceK\text{-}mers}$	$K_{Pathway}$	39	85	0.90	0.82
$K_{SequenceK\text{-}mers}$	K_{Domain}	40	85	0.90	0.82
Average		**40**	**84**	**0.80**	**0.77**

Hit rate results of [26]: PMF is 0.73 ,similarity based PMF is 0.61

We chose thresholds in such a way that, it indicates about 5% of all sample sizes. In other words, 15 and 50 are about 5% of pathogen (292) and human (1108) proteins. To be clear, 85% of hided interactions were discovered in the top 50 predictions between 1108 samples (top 5%). This measure for human proteins is about 41%. It means that 41% of hided interactions for human proteins are discovered in the top 15 predictions between 292 pathogen proteins (top 5%). As illustrated in Table 5, pathway kernels and similarities created using biological process terms clearly outperform other kernels.

Similarity based probabilistic matrix factorization (SPMF) is used by Li[26] and compared with PMF and standard matrix factorization for evaluating the results. We compare our results with their two best methods including PMF and SPMF (Figure 3). Similarity-based PMF outperforms PMF when the considerable fraction of training data is hided (i.e. more than 80%). While AMKPE is constantly superior to both methods for all ratios of hided data, for this comparison we selected $K_{SequenceK\text{-}mers}$ kernels for pathogen (K_v) and human proteins (K_h). Li[26] uses similarity matrices computed based on sequence alignment for both pathogen and host proteins.

To show the superiority of our method using the same similarity matrices, we apply $K_{SequenceK\text{-}mers}$ kernels to recompute the results for SPMF and PMF. Based on the results in Figure 3, AMKPE clearly outperforms two other methods, especially when the training set is very sparse.

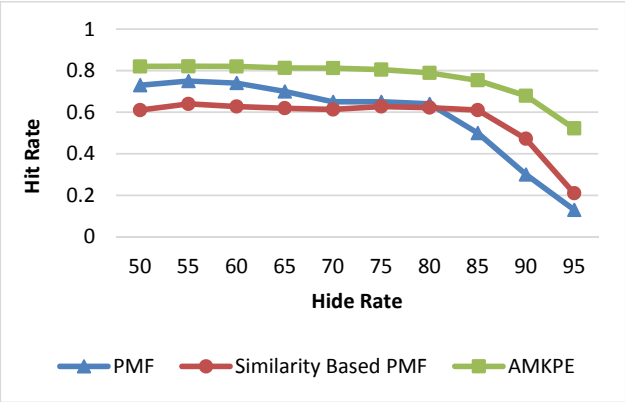


Figure 3- Hit rate comparison for probabilistic matrix factorization (PMF), similarity based PMF and the proposed method (AMKPE)

Validating reported AUC measures

To evaluate stability of the reported AUC values, we use repeated K-fold cross validation along with bootstrapping. Totally, 50 iterations of K-fold cross validations are conducted. For every fold, 95% confidence interval is computed using 500 bootstrap resamples. We conduct this set of experiments for best kernel of each data set. Figure 4 depicts the results for data set 1, where 50*5-fold experiments are conducted and AUC(base) refers to the result achieved by[11].
Summary of the sampling study are represented by Table 6.

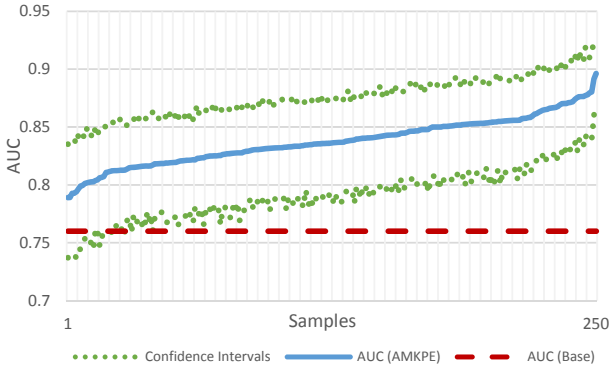


Figure 4. Sorted Confidence Intervals and estimated AUCs of various samples for medically significant viruses

Table **6**. Summary of sampling study for AUC metric

Data Set	Mean Confidence Interval Lower Bound	Mean Confidence Interval Upper Bound	Mean AUC
Medically Significant Viruses	0.791	0.876	0.837
HCV	0.787	0.925	0.872
Retroviridae Viruses	0.906	0.929	0.919

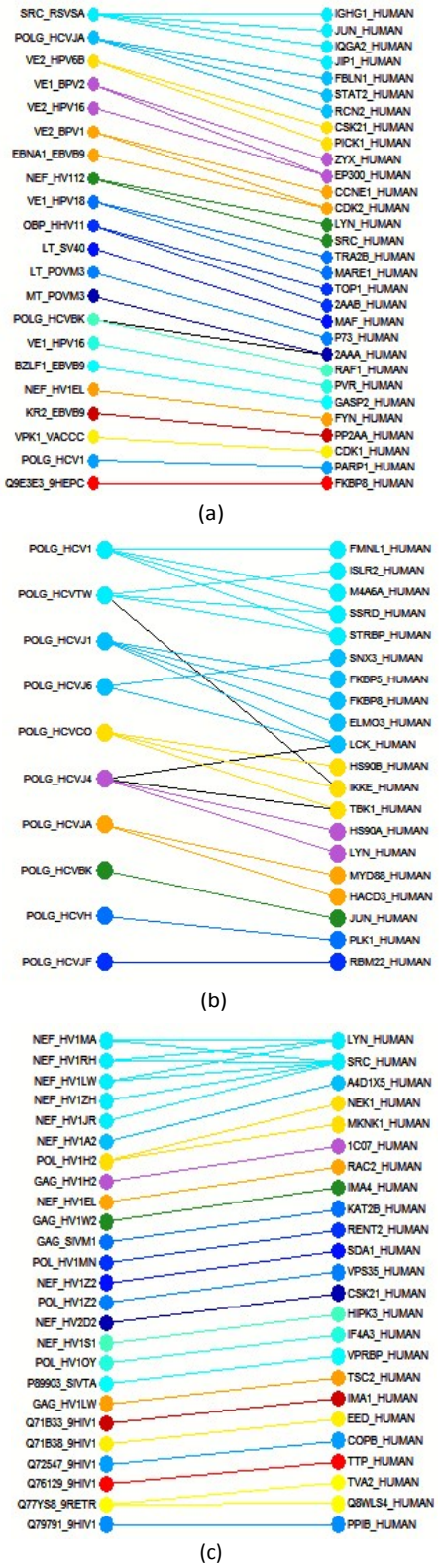


Figure 5- Sample predicted Interactions between viral proteins (Left nodes) and human proteins for Medically significant viruses (a), HCV (b) and Retroviridae Viruses (c)

Prediction Results

We make use of the available PHI samples beside the unlabelled data to discover interaction patterns. This method of positive unlabelled (PU) learning was also conducted within a previous study [29] for predicting gene-disease associations. To predict new PHIs, we use all available interactions without hiding any sample. We construct a prediction matrix for each data set using their positive available PHIs integrated with all human proteins extracted from PHISTO. The final prediction matrix includes 7529 human proteins against viral proteins of each data set. This is an interesting task, since most of the previous studies consider a limited set of human proteins for PHI prediction. To report predicted samples, we extract the top five probable human proteins, which gained the highest scores for viral proteins within the reconstructed interaction score matrix (K_{vh}). Final candidate PHIs are extracted by integrating different matrices computed using various kernels.

These lists are further filtered based on the domain-domain associations observed within experimentally verified PHIs.

In other words, filtered results include only candidates which contain a verified domain-domain association within the domain pairs of candidate PHIs. Sample filtered predicted interactions are depicted in Figure 5.

Assessment of Predicted Interactions

We perform GO enrichment analysis for virus-targeted human proteins within the predicted PHIs, using PANTHER tool[30]. The top enriched process terms for human proteins, which are predicted to interact with viral proteins, are presented in Table 7-9.

We combine all predicted human proteins for our three data sets and classify them using PANTHER classification system which is shown in Figure 6. The results of GO enrichment analysis and PANTHER classification of the predicted interacting human proteins, based on their functional properties, reflect some of the known facts about the virus-targeted human proteins obtained from the available experimental virus-human PHI data[31,32,33]. The largest fraction of the predicted human proteins function as "nucleic acid binding", "transcription factor", and "enzyme modulator" within the human cellular processes related to cell cycle and metabolism (Figure 6). The following classified terms (cytoskeletal protein, hydrolase, signalling protein, etc.) for the predicted interacting human proteins give insights on the viral infection strategies, i.e. through functional properties of the human proteins viruses attack. Metabolic processes and protein transport related terms are the top enriched GO processes for HCV-targeted human proteins (Table 8), whereas RNA processing related terms are for Retroviruses-targeted ones (Table 9). On the other hand, specific metabolic and cellular processes related terms were enriched for the predicted human proteins interacting with medically important viruses (Table 7).

Table **7**. The enriched GO terms in human proteins predicted to interact with medically significant viruses

Biological process term	P value
positive regulation of macromolecule metabolic process	1.4E-14
regulation of protein modification process	1.1E-12
positive regulation of gene expression	2.1E-12
positive regulation of nitrogen compound metabolic process	2.9E-12
positive regulation of transcription	3.6E-12
positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.8E-11
cell cycle	2.0E-11
positive regulation of cellular biosynthetic process	2.0E-11
positive regulation of biosynthetic process	3.2E-11
negative regulation of macromolecule metabolic process	4.7E-11

Table **8**. Enriched GO terms in human proteins predicted to interact with HCV

Biological process term	Pvalue
response to organic substance	7.1E-5
negative regulation of cellular protein metabolic process	3.8E-4
negative regulation of protein metabolic process	4.9E-4
intracellular transport	6.6E-4
response to virus	6.9E-4
regulation of cellular protein metabolic process	7.4E-4
response to cytokine stimulus	1.0E-3
membrane organization	1.3E-3
regulation of cell proliferation	1.7E-3
protein localization	2.3E-3

Table **9**. Enriched GO terms in human proteins predicted to interact with Retroviruses

Biological process term	P value
mRNA processing	2.0E-14
RNA processing	5.2E-14
RNA splicing	1.3E-13
mRNA metabolic process	2.4E-13
cell cycle	4.1E-12
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	1.8E-11
nuclear mRNA splicing, via spliceosome	1.8E-11
RNA splicing, via transesterification reactions	1.8E-11
cell cycle process	6.3E-10
mitotic cell cycle	2.6E-9

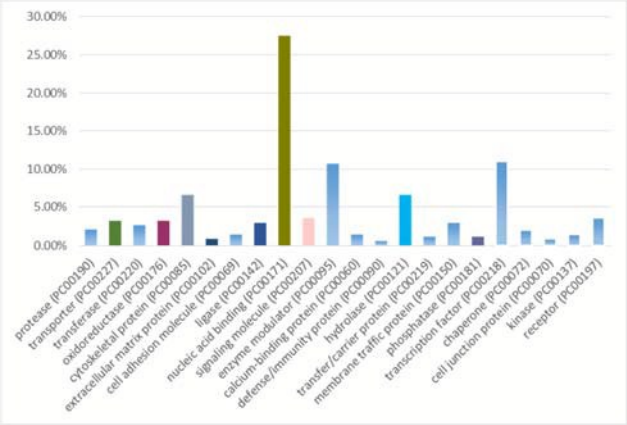


Figure 6- Classification of viruses-targeted human proteins in the predicted PHIs

A final assessment is conducted by comparison of the enriched GO terms in the sets of human proteins within the predicted and experimental PHI data. It should be noted that, in this comparison we use the predicted interacting human proteins which are new, i.e. they are not included within the experimentally-found PHIs. This comparison was performed to show the functional similarities between the predicted new virus-targeted human proteins and the previously reported ones within the experimental data. About 50% of the enriched terms in the set of predicted human proteins are within the results for the set of experimentally-verified interacting human proteins with the HCV. This value for medically significant viruses and Retroviruses are about 80 and 59 per cent, respectively. We select the top 10 common enriched terms, which have the smallest p-values for comparing the proportion of genes annotated to the term for the sets of predicted and experimental interacting human proteins.

To perform the same comparison after filtering, we use only the filtered results. Surprisingly, 95% of enriched terms for the set of predicted human proteins are within the enriched terms for experimentally interacting human proteins.

Discussion

Computational methods for PHI prediction utilize the known interactions and information on protein sequences and network topology measures. Some studies revealed the usual behaviors of pathogen proteins such as having a tendency to target hub and bottleneck proteins in the human PPI network [34, 31, 35, 32]. However, they are not the sole targeted human proteins [36] and some pathogenic systems such as HIV tends to target the peripheral human proteins [21]. We generate a specific kernel using degree values of proteins in the human intranetwork of PPIs to make use of the diverse viral behaviors.

Binary classification is the usual approach for PHI prediction, in which both positive and negative samples are required. PHI databases report only interacting proteins as positive samples without listing non-interacting pairs. Consequences of selecting random samples as non-interacting pairs may be unpredictable and the results are dependent on the selected negative class. This challenge motivated researchers to overcome this problem by removing the dependence on the negative data[37–41]. They integrate bi-clustering with association rule mining, utilizing only positive samples to predict virus-human interactions.

Same challenge is addressed by Li[26] which uses similarity matrices to enrich sparse interaction matrix and consequently discover the interaction pattern using matrix factorization. They use a primitive sequence based similarity matrix which only shows its effect when a significant fraction of data set is hided. Our approach learns the interaction pattern based on available positive samples without the need for negative samples, exploiting effective genomic features. According to the presented results in Figure 3, our approach shows significant improvements in comparison to the results achieved in [26].

Domains, as building blocks of proteins and mediators of interactions, have crucial roles for predicting intraspecies PPIs [42,43]. As one of the initial approaches, [44] makes use of protein domain profiles for prediction of PHIs. Domain-domain interactions by means of Pfam domains are evaluated by Dyer[45] to predict and rank bacteria–human PPIs. Here, we make use of domain profiles to create similarity kernels based on the occurrence of same domains between protein pairs. This kernel is created for both pathogen and human proteins. As we can see in Table5 , the efficiency of domain kernels for prediction of PHIs may be significant in some pathogen systems. Furthermore domain-domain associations are used as the metric for filtering predicted candidate PHIs.

From the biological point of view, the findings about the predicted PHIs presented in Figures 6-9 and in Tables 7-9, give additional support for the reliability of the computational PHI prediction results. First of all, the observation in Figure 6 is that nucleic acid binding proteins, transcription factors and enzyme modulators constitute the largest fraction of the virus-targeted human proteins within the predicted results. These human proteins were extensively reported as the main targets of viral pathogens within the experimental PHI data [31–33]. For the case of GO enrichment results (Tables 7-9), viruses are observed to attack human metabolic and cellular processes for exploitation, since they lack their own metabolism and machineries for viral genetic material transcription and translation. Specifically for RNA viruses/Retroviruses, RNA processing and intracellular transport/localization are obtained as the top enriched GO terms. The GO enrichment results coincide considerably for the virus-targeted human proteins within the predicted and experimental PHIs (Figures 7-9).

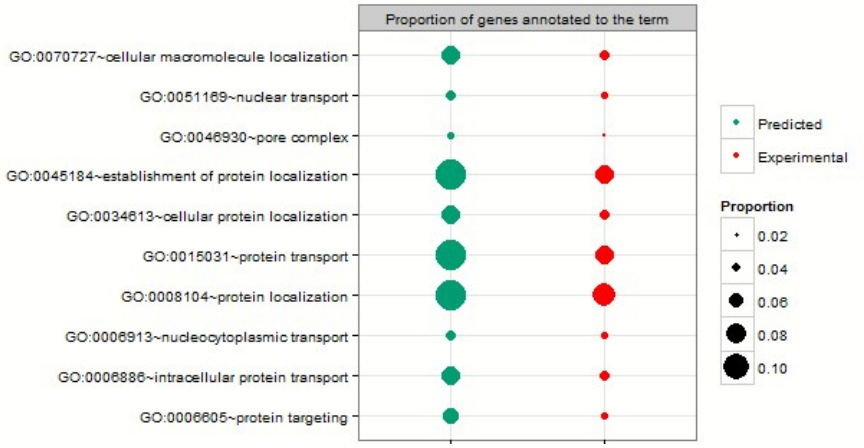


Figure 7. Proportion of enriched genes for the experimental and predicted interacting human proteins with HCV

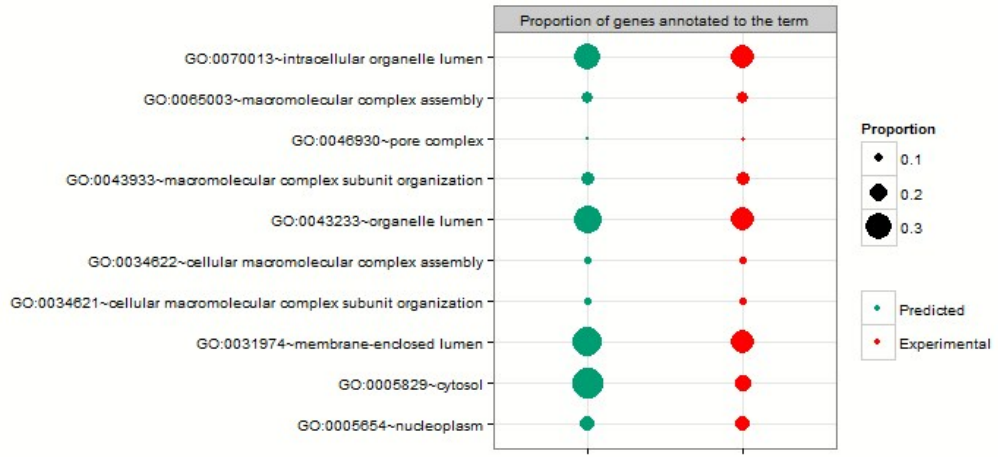


Figure 8. Proportion of enriched genes for the experimental and predicted interacting human proteins with medically significant viruses

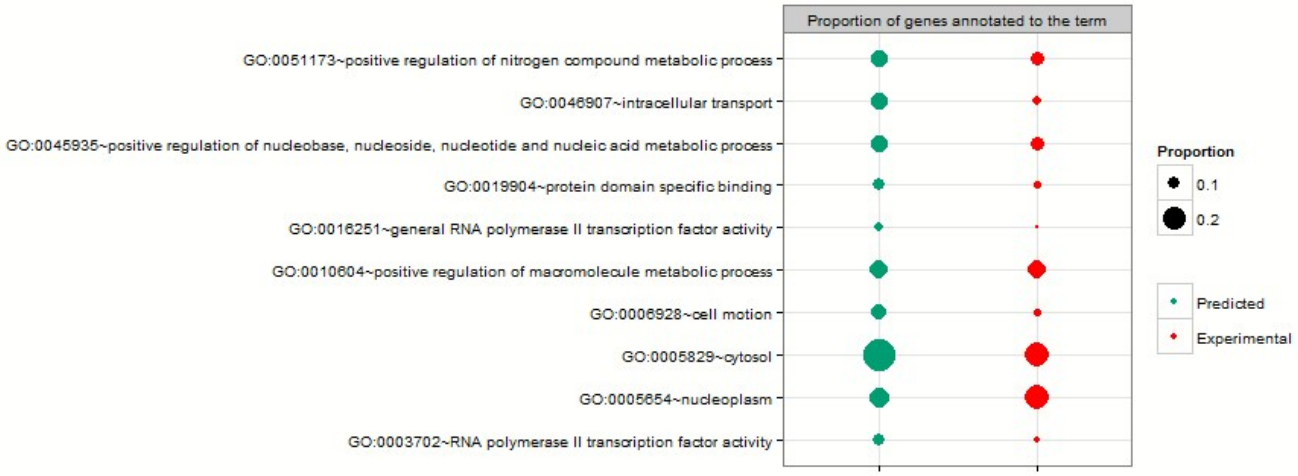


Figure 9. Proportion of enriched genes for the experimental and predicted interacting human proteins with Retroviruses

Conclusions

In this paper, we present a new approach for computational prediction of pathogen-host protein-protein interactions. Distinguished aspect of the method is relieving the need for samples of non-interacting protein pairs. Furthermore, the presented approach needs a minimum number of features, as it requires only one similarity kernel for both pathogen and host proteins generated using single features. This is significant, since we would encounter numerous missing feature values along with increasing number of features. This is due to the fact that, except sequence-based features, values for most of protein features do not exist especially for pathogen proteins. We make use of interacting samples enriched with genomic and topological similarity kernels to discover interaction patterns within virus-human protein-protein interactions. Pathway membership and domain similarity kernels are effective features for PHIs as illustrated based on different experiments. Besides the negative-sample-free approach, we formulate the problem as a binary classification to compare the performance of the method with the state of the art approaches mostly presented in this formulation. Performance results of both formulations, present at least 10% improvements in comparison with the state of the art methods.

Notes and references

[I] http://www.ebi.ac.uk/interpro/

[II] http://www.reactome.org/pages/download-data/

[III] ftp://mint.bio.uniroma2.it/pub/virusmint/MITAB/current/2012 -10-26-mint-viruses-binary.mitab26.txt

[IV] https://www.ebi.ac.uk/intact/downloads

[V] http://www.phisto.org/

1 K. F. Smith, M. Goldberg, S. Rosenthal, L. Carlson, J. Chen, C. Chen and S. Ramachandran, *J. R. Soc. Interface*, 2014, **11**, 20140950.

2 B. J. Cowling, M. Park, V. J. Fang, P. Wu, G. M. Leung and J. T. Wu, *Eurosurveillance*, 2015, **20**.

3 S. Durmuş, T. Çakır, A. Özgür and R. Guthke, *Front. Microbiol.*, 2015, **6**, 235.

4 P. Aloy and R. B. Russell, *Bioinformatics*, 2003, **19**, 161–2.

5 E. Nourani, F. Khunjush and S. Durmuş, *Front. Microbiol.*, 2015, **6**, 94.

6 A. R. Wattam, D. Abraham, O. Dalay, T. L. Disz, T. Driscoll, J. L. Gabbard, J. J. Gillespie, R. Gough, D. Hix, R. Kenyon, D. Machi, C. Mao, E. K. Nordberg, R. Olson, R. Overbeek, G. D. Pusch, M. Shukla, J. Schulman, R. L. Stevens, D. E. Sullivan, V. Vonstein, A. Warren, R. Will, M. J. C. Wilson, H. S. Yoo, C. Zhang, Y. Zhang and B. W. Sobral, *Nucleic Acids Res.*, 2014, **42**, D581–91.

7 A. Calderone, L. Licata and G. Cesareni, *Nucleic Acids Res.*, 2014, gku830–.

8 V. Navratil, B. de Chassey, L. Meyniel, S. Delmotte, C. Gautier, P. André, V. Lotteau and C. Rabourdin-Combe, *Nucleic Acids Res.*, 2009, **37**, D661–8.

9 S. Durmuş Tekir, T. Çakır, E. Ardiç, A. S. Sayılırbaş, G. Konuk, M. Konuk, H. Sarıyer, A. Uğurlu, İ. Karadeniz, A. Özgür, F. E. Sevilgen and K. Ö. Ülgen, *Bioinformatics*, 2013, **29**, 1357–8.

10 M. Gonen, in *ECAI 2014: 21st European Conference on Artificial Intelligence*, 2014, p. 381.

11 R. K. Barman, S. Saha and S. Das, *PLoS One*, 2014, **9**, e112034.

12 A. Emamjomeh, B. Goliaei, J. Zahiri and R. Ebrahimpour, *Mol. Biosyst.*, 2014, **10**, 3147–54.

13 B. Y. S. Li, L. F. Yeung and G. Yang, in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Ieee, 2014, pp. 357–362.

14 J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li and H. Jiang, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 4337–41.

15 J. Yu, M. Guo, C. J. Needham, Y. Huang, L. Cai and D. R. Westhead, *Bioinformatics*, 2010, **26**, 2610–4.

16 C. Leslie, E. Eskin and W. S. Noble, *Biocomput. 2002 - Proc. Pacific Symp.*, 2001, **575**, 564–575.

17 J. Palme, S. Hochreiter and U. Bodenhofer, *Bioinformatics*, 2015, 1–3.

18 T. Jaakkola and M. Diekhans, in *Seventh International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, 1999, pp. 149–158.

19 A. Mitchell, H.-Y. Chang, L. Daugherty, M. Fraser, S. Hunter, R. Lopez, C. McAnulla, C. McMenamin, G. Nuka, S. Pesseat, A. Sangrador-Vegas, M. Scheremetjew, C. Rato, S.-Y. Yong, A. Bateman, M. Punta, T. K. Attwood, C. J. a Sigrist, N. Redaschi, C. Rivoire, I. Xenarios, D. Kahn, D. Guyot, P. Bork, I. Letunic, J. Gough, M. Oates, D. Haft, H. Huang, D. a Natale, C. H. Wu, C. Orengo, I. Sillitoe, H. Mi, P. D. Thomas and R. D. Finn, *Nucleic Acids Res.*, 2014, 1–9.

20 D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob,

L. Stein and P. D'Eustachio, *Nucleic Acids Res.*, 2014, **42**, D472–7.

21 S. Mei, *PLoS One*, 2013, **8**, e79606.

22 J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu and C.-F. Chen, *Bioinformatics*, 2007, **23**, 1274–81.

23 G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu and S. Wang, *Bioinformatics*, 2010, **26**, 976–8.

24 A. Chatr-aryamontri, A. Ceol, D. Peluso, A. Nardozza, S. Panni, F. Sacco, M. Tinti, A. Smolyar, L. Castagnoli, M. Vidal, M. E. Cusick and G. Cesareni, *Nucleic Acids Res.*, 2009, **37**, D669–73.

25 S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del-Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni and H. Hermjakob, *Nucleic Acids Res.*, 2014, **42**, D358–63.

26 B. Y. S. Li, L. F. Yeung and G. Yang, *2014 IEEE Int. Conf. Bioinforma. Biomed.*, 2014, 357–362.

27 A. Ben-Hur and W. S. Noble, *BMC Bioinformatics*, 2006, **7 Suppl 1**, S2.

28 M. C. Cobanoglu, C. Liu, F. Hu, N. Oltvai and I. Bahar, *J. Chem. Inf. Model.*, 2013, **53**, 3399–3409.

29 N. Natarajan and I. S. Dhillon, *Bioinformatics*, 2014, **30**, i60–i68.

30 H. Mi, a. Muruganujan and P. D. Thomas, *Nucleic Acids Res.*, 2013, **41**, D377–D386.

31 M. D. Dyer, T. M. Murali and B. W. Sobral, *PLoS Pathog.*, 2008, **4**, e32.

32 S. Durmuş Tekir, T. Cakir and K. Ö. Ulgen, *Front. Microbiol.*, 2012, **3**, 46.

33 A. Pichlmair, K. Kandasamy, G. Alvisi, O. Mulhern, R. Sacco, M. Habjan, M. Binder, A. Stefanovic, C.-A. Eberle, A. Goncalves, T. Bürckstümmer, A. C. Müller, A. Fauster, C. Holze, K. Lindsten, S. Goodbourn, G. Kochs, F. Weber, R. Bartenschlager, A. G. Bowie, K. L. Bennett, J. Colinge and G. Superti-Furga, *Nature*, 2012, **487**, 486–90.

34 S. Schleker and M. Trilling, *Front. Microbiol.*, 2013, **4**, 51.

35 L.-L. Zheng, C. Li, J. Ping, Y. Zhou, Y. Li and P. Hao, *Biomed Res. Int.*, 2014, **2014**, 867235.

36 K.-C. Chen, T.-Y. Wang and C. Chan, *PLoS One*, 2012, **7**, e34240.

37 A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay and R. Eils, in *2010 International Conference on Systems in Medicine and Biology*, Kharagpur, 2010, pp. 344–348.

38 A. Mukhopadhyay, U. Maulik and S. Bandyopadhyay, *PLoS One*, 2012, **7**, e32289.

39 A. Mukhopadhyay, S. Ray and U. Maulik, *BMC Bioinformatics*, 2014, **15**, 26.

40 K. C. Mondal, N. Pasquier, A. Mukhopadhyay, C. Pereira, U. Maulik and A. G. B. Tettamanzi, in *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms*, vilamoura, 2012, pp. 164–173.

41 S. Ray, A. Mukhopadhyay and U. Maulik, in *Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on.*, Kolkata, 2012, pp. 3–6.

42 J. Wojcik and V. Schächter, *Bioinformatics*, 2001, **17**, S296–S305.

43 P. Pagel, P. Wong and D. Frishman, *J. Mol. Biol.*, 2004, **344**, 1331–46.

44 M. D. Dyer, T. M. Murali and B. W. Sobral, *Bioinformatics*, 2007, **23**, i159–66.

45 R. Arnold, K. Boonen, M. G. F. Sun and P. M. Kim, *Methods*, 2012, **57**, 508–18.

