BMC Bioinformatics

RESEARCH Open Access

Predicting virus-host association by Kernelized logistic matrix factorization and similarity network fusion

Dan Liu[1,2], Yingjun Ma[1,2], Xingpeng Jiang[1,2*] and Tingting He[1,2*]

From IEEE International Conference on Bioinformatics and Biomedicine 2018
Madrid, Spain. 3-6 December 2018

Abstract

Background: Viruses are closely related to bacteria and human diseases. It is of great significance to predict associations between viruses and hosts for understanding the dynamics and complex functional networks in microbial community. With the rapid development of the metagenomics sequencing, some methods based on sequence similarity and genomic homology have been used to predict associations between viruses and hosts. However, the known virus-host association network was ignored in these methods.

Results: We proposed a kernelized logistic matrix factorization with integrating different information to predict potential virus-host associations on the heterogeneous network (ILMF-VH) which is constructed by connecting a virus network with a host network based on known virus-host associations. The virus network is constructed based on oligonucleotide frequency measurement, and the host network is constructed by integrating oligonucleotide frequency similarity and Gaussian interaction profile kernel similarity through similarity network fusion. The host prediction accuracy of our method is better than other methods. In addition, case studies show that the host of crAssphage predicted by ILMF-VH is consistent with presumed host in previous studies, and another potential host *Escherichia coli* is also predicted.

Conclusions: The proposed model is an effective computational tool for predicting interactions between viruses and hosts effectively, and it has great potential for discovering novel hosts of viruses.

Keywords: Virus-host association, Logistic matrix factorization, Similarity network fusion, Oligonucleotide frequency, Gaussian interaction profile

Background

The two main components of human microbes are bacterial and viral communities, which play a vital role in human health and diseases. Bacterial communities have been proved to be associated with human diseases, including some skin conditions [1] and gastrointestinal diseases, such as inflammatory bowel disease [2], clostridium infection [3], and colorectal cancer [4]. Viral communities are also associated with diseases, such as periodontal disease [5] and antibiotic resistance [6]. Viruses are widespread in the environment and biological tissues, and they are the most abundant organisms on the planet [7]. Viruses cannot survive alone, they need to be parasitic in living cells to survive and produce offspring. Hosts infected by viruses include bacteria, archaea, eukaryotes, etc. Viruses produce DNA and proteins of offspring through hosts' replication mechanism. In particular, prokaryotic viruses have a significant impact on human health and ecosystem dynamics. Describing interactions between viruses and hosts is important for understanding hosts' effects on microbial communities.

The traditional approach for identifying viruses has been implemented by isolating from cultured host strains, because viruses are acquired from cultured host cells, we can directly know the host of the given virus. However, culturing a virus at a high enough concentration may be challenging in experiments because it may require appropriate growth conditions, such as temperature, growth media as

* Correspondence: xpjiang@mail.ccnu.edu.cn; tthe@mail.ccnu.edu.cn
[1]School of Computer, Central China Normal University, Wuhan, Hubei, China
Full list of author information is available at the end of the article

well as robust growth of target host strains [8], which is usually difficult to achieve in experiments. The isolation method based on culturing bacteria is inefficient to identify viruses, and it only identifies relatively fewer viruses. Nowadays, discoveries of unknown viruses have been greatly accelerated by metagenomic shotgun sequencing, but unlike viral isolation, viral sequences assembled from metagenomics usually fails to directly obtain hosts infected by them. For example, crAssphage is a highly abundant human enterovirus, which may play an important role in the human intestinal tract, but the cultivation of crAssphage in the laboratory is still not achievable, so its hosts and biological function has not yet been identified [9]. As more and more metagenomic sequencing datasets are available, it is urgent to propose effective culture-free methods to identify new viruses and their hosts.

Recognizing hosts infected by viruses is important for understanding dynamics of viruses and their effects on microbial communities. Recently, some computational methods have been used to infer associations between viruses and hosts. Edwards [10] et al. introduced three types of virus-host association prediction methods, including sequence homology [6, 11, 12], abundance profile co-occurrence [13] and sequence composition [14–16]. As for virus-host association prediction methods based on sequence homology, homologies between new viruses and potential hosts are limited, because they depend on whether hosts of the query virus exist in the host genome database. Abundance profile method is based on co-variation, but significant co-variation does not necessarily represent real interaction. Because there is usually a time delay in dynamic interactions between viruses and hosts, many interactions depending on time-scale sampling may not be detected. Sequence composition is based on codon usage or short pairs of nucleotides (k-mers) shared by viruses and hosts to predict which hosts the virus infects. Ahlgren et al. proposed 11 measurements of oligonucleotide frequency (ONF) such as d_2^* to calculate k-mers distances between viruses and hosts [17]. This method achieves good results in host prediction accuracy at the genus level, but less than 40% at the species level. In addition, previous human microbial community studies relied on independent bacterial and viral communities, i.e. they were divided into two separate network communities [2, 5], which could not capture complex dynamics of virus-host interactions.

In this paper, we propose a logistic matrix factorization algorithm based on integrating multi-information on the heterogeneous network to predict potential virus-host associations (ILMF-VH). The main differences from previous studies are that our proposed method combines information of three networks to form a virus-host heterogeneous network and applies similar network fusion (SNF) to integrate multiple host information for constructing the host-host similarity network. We used the benchmark data of viral and bacterial genomes in NCBI, and verified that ILMF-VH obtained best performance compared with recent five network-based methods under five-fold cross validation. Moreover, the host prediction accuracy is 63.66% which is 24.66 and 13.29% higher than two recently proposed virus-host association prediction methods respectively, and it is 0.49% higher than our previous approach [18]. In addition, the host of crAssphage inferred by our algorithm includes putative host *Bacteroides* obtained from previous studies [9, 19], and another potential host *Escherichia coli* is also suggested. Because previous studies have shown that *Escherichia coli* is associated with human intestinal diseases, such as diarrhea [20], our research indicates that crAssphage may be closely related to these diseases, and this proves that our approach is effective in predicting novel virus-host associations.

Materials and methods

Data sets

We used the data adopted by Ahlgren et al. which collected accession numbers and taxonomies of 1427 viruses and 31,986 hosts. For the initial analysis, we selected a subset including 352 viruses whose hosts were at strain level [17]. In addition, we downloaded the benchmark datasets provided by Edwards et al. including accession numbers and taxonomies of 820 viruses and 2699 hosts [21]. Based on accession numbers of viruses and hosts, we have written scripts to obtain their whole genome sequences from NCBI. In terms of each virus, their known virus-host associations are obtained through the ‘isolate host = ‘or ‘host = ‘fields in the viral annotation file. The genome of crAssphage in the human intestinal metagenomic is downloaded from NCBI and the accession number is JQ995537.1 [19].

Methods

As for our model, the virus set and host set are represented by $V = \{v_1, v_2, \dots, v_{N_v}\}$ and $H = \{h_1, h_2, \dots, h_{N_h}\}$, where N_v and N_h represent the number of viruses and hosts, respectively. The associations between viruses and hosts are defined as an adjacency matrix $Y \in R^{N_v \times N_h}$, if a virus v_i is known to be associated with a host h_j, then y_{ij} is set to 1; otherwise, y_{ij} is set to 0. In terms of elements in the adjacency matrix Y, the negative and positive interactions between viruses and hosts are represented by 0 and 1, respectively. In this work, firstly, we define a set of viruses which are positively related to hosts as $V^+ = \{v_i | \sum_{i=1}^{N_v} y_{ij} > 0, \forall 1 \leq i \leq N_v\}$, then a set of viruses which are negatively related to hosts is defined as $V^- = V \backslash V^+$. Next, a set of hosts which are positively related to viruses is defined as $H^+ = \{h_j | \sum_{j=1}^{N_h} y_{ij} > 0, \forall 1 \leq j \leq N_h\}$, and a set of hosts which are negatively related to viruses is defined as $H^- = H \backslash H^+$. Finally, similarities between viruses are calculated by oligonucleotide frequency (ONF) measures and expressed by $S^v \in R^{N_v \times N_v}$; similarities between hosts are calculated by integrating ONF measures and Gaussian interaction profile (GIP) kernel similarity based on SNF model, and expressed by $S^h \in R^{N_h \times N_h}$.

Oligonucleotide frequency measures for viruses and hosts

Recently, dissimilarity measurements based on k-mer frequencies have been applied to infer relationships between genomic sequences [17]. Here, based on the hypothesis that similar viruses or hosts share similar k-mer patterns, we calculated k-mer similarities between viral genomic sequences to measure correlations between viruses. Similarly, k-mer similarities between hosts' genomic sequences are calculated to measure correlations between hosts. According to previous research [17], d_2^* [22] has a good performance in calculating k-mer similarity and k is set to 6, so we calculate the distance between k-mer frequency vectors of each pair of viruses or hosts. Finally, the virus-virus similarity matrix S^v and the host-host similarity matrix $S(onf)^h$ can be obtained.

Gaussian interaction profile kernel similarity for hosts

Zou et al. [23] calculated the GIP kernel similarity between microbes based on the known disease-microbe association matrix and achieved good results. Apart from sequence similarities of hosts, based on the assumption that similar hosts exhibit similar patterns with viruses, we apply GIP kernel similarity to measure associations between hosts. There are two steps to calculate GIP kernel similarity. First, the interaction profile $IP(h_i)$ of host $h_i$is the i-th column of the adjacency matrix Y, which is a binary relationship vector representing associations between a host $h_i$and each virus. The GIP kernel similarity between host $h_i$and h_j is calculated from their interaction profiles and defined as [24]:

$$S^h(h_i, h_j) = \exp\left(-\gamma_h \left\| IP(h_i) - IP(h_j) \right\|^2\right) \tag{1}$$

This is a kernel that represents the similarities between hosts. These kernels are called Gaussian kernels. The parameter γ_h is used to control the kernel bandwidth and defined as:

$$\gamma_h = r_h^{'} / \left(\frac{1}{N_h} \sum\nolimits_{k=1}^{N_h} \| IP(h_k) \|^2 \right) \tag{2}$$

Here, N_h is the number of hosts. According to the previous study [25], we simply set $r_h^{'}$ to 1.

Integrated similarity for hosts

The associations between hosts are measured by calculating ONF measures and GIP kernel similarity between hosts, respectively. Here, we introduce similar network fusion (SNF) [26] to integrate two host similarity networks. The SNF includes following three main steps. First, the edge weights of each host similarity network are represented by a $N_h \times N_h$ matrix S^h, respectively. Then, as for each similarity network, a normalized weight matrix p can be obtained by the following formula [26]:

$$P_{i,j} = \begin{cases} \dfrac{S(i,j)}{2\sum_{k \neq i} S(i,k)}, j \neq i \\ \dfrac{1}{2}, j = i \end{cases} \tag{3}$$

Here $S(i,j)$ is the matrix element of S^h. Then, k nearest neighbor (KNN) is used to measure the local relationship as follows:

$$KNN(i,j) \begin{cases} \dfrac{S(i,j)}{\sum_{k \epsilon N_i} S(i,k)}, j \epsilon N_i \\ 0, otherwise \end{cases} \tag{4}$$

N_i represents the number of neighbors in the host. This method filters out low-similar edges.

Let $P^{(\nu)}$ and $KNN^{(\nu)}$ represent similar matrices of the above two hosts, respectively. The process of SNF is an iterative update of similarity matrices, which corresponds to each data type as follows [26]:

$$\begin{aligned} P^{(\nu)} &= KNN^{(\nu)} \left(\frac{\sum_{k \neq \nu} P^{(k)}}{m-1} \right) \left(KNN^{(\nu)} \right)^T, \nu \\ &= 1, 2 \ldots, m \end{aligned} \tag{5}$$

This step updates the matrix $P^{(\nu)}$ when m parallel exchange diffusion processes are generated on m host networks. In this paper, we have two types of host similar matrices, so m is set to 2. The final similarity matrix that integrates all data types is defined as follows:

$$P = \frac{1}{2}\left(P^1 + P^2\right) \tag{6}$$

Construction of heterogeneous networks

The construction of heterogeneous networks is mainly divided into three steps. First, based on known virus-host associations, we can construct a virus-host relationship network, where nodes in the network include viruses and hosts and if a virus and a host are known to be related, their edge weights are set to 1, otherwise, they are set to 0. Then, we calculate similarities between viruses based on ONF measures to construct the virus network, and calculate similarities between hosts by integrating ONF measure and GIP kernel similarity based on SNF model to construct the host network. Finally, the virus network and the host network are connected through known virus-host associations to construct a heterogeneous network between viruses and hosts.

Kernelized logistic matrix factorization

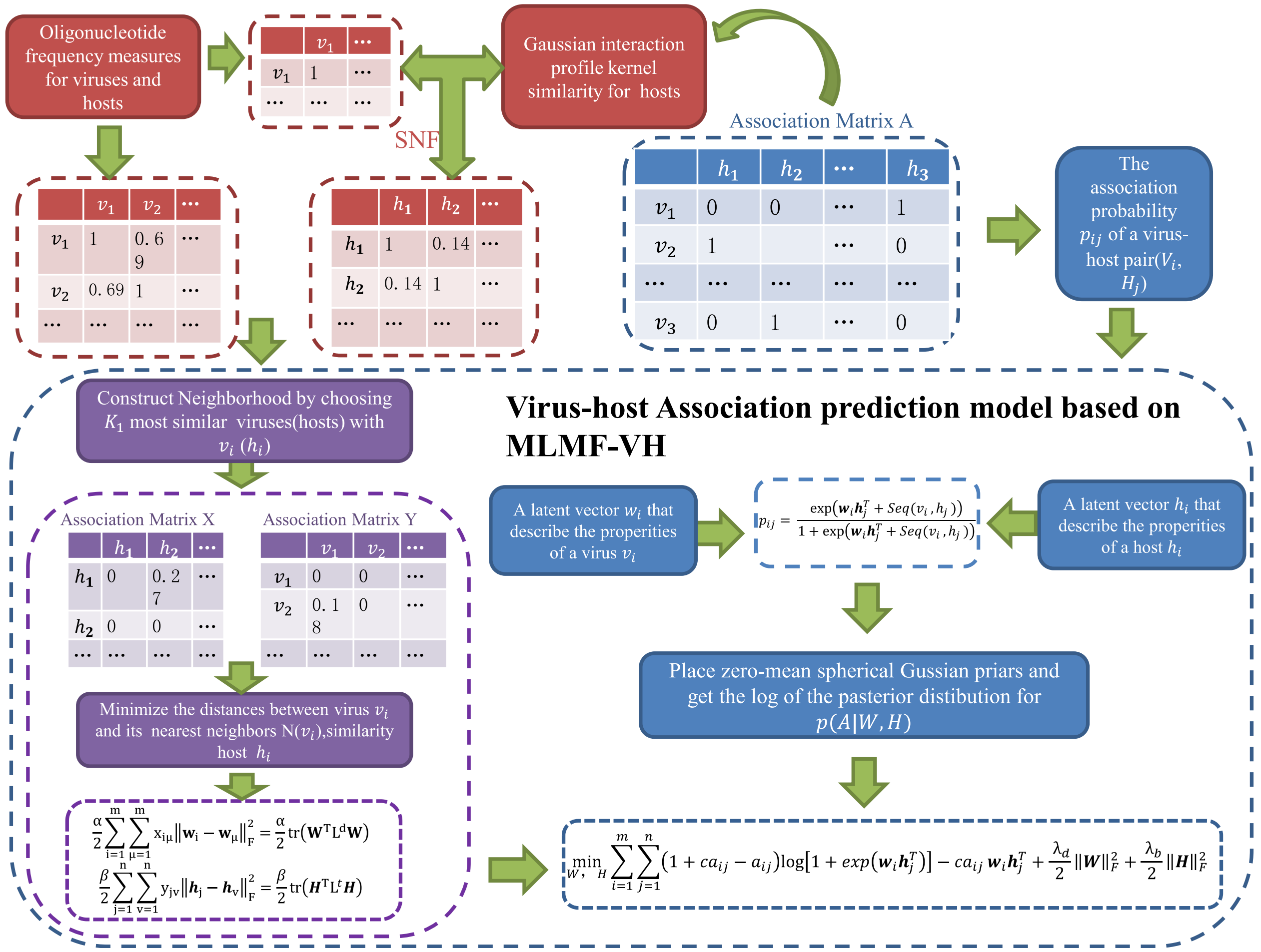
We developed a kernelized logistic matrix factorization algorithm based on network similarity fusion for predicting virus-host associations, and the flowchart of ILMF-VH model is shown in the Fig. 1. First, the binary matrix Y is decomposed into $W \in R^{N_v \times k}$ and $\in R^{N_h \times k}$, so viruses and hosts are mapped to the shared potential low-dimensional space. $Seq(v_i, h_j)$ represents the ONF similarity between each pair of virus and host, and we integrate this sequence similarity information into the associated probability p_{ij}, which represents association probability of virus-host pair (v_i, h_j) and is defined as the logistic function:

$$p_{ij} = \frac{\exp\left(\boldsymbol{w}_i \boldsymbol{h}_j^T + Seq\left(v_i, h_j\right)\right)}{1 + \exp\left(\boldsymbol{w}_i \boldsymbol{h}_j^T + Seq\left(v_i, h_j\right)\right)} \tag{7}$$

It is hypothesized that known relationships between viruses and hosts provide useful information for virus-host association prediction. Current importance weighting methods have been proven to be effective for personalized recommendations and drug-target interaction predictions [27, 28]. We apply the weight constant c to control the level of importance between each known and unknown associations. According to previous studies, c is set to 5. The conditional probability of Y is defined as:

$$p(Y|W,H) = \prod_{i=1}^{N_v} \prod_{j=1}^{N_h} p_{ij}^{cy_{ij}} \left(1-p_{ij}\right)^{\left(1-y_{ij}\right)} \tag{8}$$

In this work, we also use the neighborhood regularization method to regularize the logistic matrix factorization algorithm [28]. The nearest neighbors of virus v_i and host h_i are defined as $N(v_i) \in V \backslash v_i$ and $N(h_j) \in H \backslash h_i$, $N(v_i)$ and $N(h_i)$ represent the K_1 neighbors of the virus v_i or the host h_i, respectively. K_1 is set to 5 according to the experiment. The neighborhood information of viruses and hosts is represented by the adjacency matrices A and B, respectively. In terms of matrix A, if virus $v_m \in N(v_i)$, $a_{im} = s_{im}^v$, otherwise $a_{im} = 0$; in terms of


Oligonucleotide frequency measures for viruses and hosts
SNF
Gaussian interaction profile kernel similarity for hosts
Association Matrix A
The association probability p_{ij} of a virus-host pair(V_i, H_j)
Construct Neighborhood by choosing K_1 most similar viruses(hosts) with v_i (h_i)
Virus-host Association prediction model based on MLMF-VH
Association Matrix X
Association Matrix Y
Minimize the distances between virus v_i and its nearest neighbors N(v_i),similarity host h_i
A latent vector w_i that describe the properities of a virus v_i
A latent vector h_i that describe the properities of a host h_i
Place zero-mean spherical Gussian priars and get the log of the pasterior distibution for $p(A|W,H)$

	v_1	···
v_1	1	···
···	···	···

	v_1	v_2	···
v_1	1	0. 69	···
v_2	0. 69	1	···
···	···	···	···

	h_1	h_2	···
h_1	1	0. 14	···
h_2	0. 14	1	···
···	···	···	···

Association Matrix A

	h_1	h_2	···	h_3
v_1	0	0	···	1
v_2	1		···	0
···	···	···	···	···
v_3	0	1	···	0

Association Matrix X

	h_1	h_2	···
h_1	0	0. 27	···
h_2	0	0	···
···	···	···	···

Association Matrix Y

	v_1	v_2	···
v_1	0	0	···
v_2	0. 18	0	···
···	···	···	···

$$\frac{\alpha}{2}\sum_{i=1}^{m}\sum_{\mu=1}^{m} x_{i\mu}\left\|\mathbf{w}_i - \mathbf{w}_\mu\right\|_F^2 = \frac{\alpha}{2}\mathrm{tr}\left(\mathbf{W}^T L^d \mathbf{W}\right)$$

$$\frac{\beta}{2}\sum_{j=1}^{n}\sum_{v=1}^{n} y_{jv}\left\|\boldsymbol{h}_j - \boldsymbol{h}_v\right\|_F^2 = \frac{\beta}{2}\mathrm{tr}\left(\boldsymbol{H}^T L^t \boldsymbol{H}\right)$$

$$p_{ij} = \frac{\exp\left(\boldsymbol{w}_i \boldsymbol{h}_j^T + Seq(v_i, h_j)\right)}{1 + \exp\left(\boldsymbol{w}_i \boldsymbol{h}_j^T + Seq(v_i, h_j)\right)}$$

$$\min_{W, \, H} \sum_{i=1}^{m}\sum_{j=1}^{n}(1 + ca_{ij} - a_{ij})\log\left[1 + exp\left(\boldsymbol{w}_i \boldsymbol{h}_j^T\right)\right] - ca_{ij}\,\boldsymbol{w}_i \boldsymbol{h}_j^T + \frac{\lambda_d}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_b}{2}\|\boldsymbol{H}\|_F^2$$

Fig. 1 The flow chart of ILMF-VH model

matrix B, if host $h_n \in N(h_j)$, $b_{jn} = s_{jn}^h$, otherwise $b_{jn} = 0$. The main purpose of virus-host association prediction is to minimize distances between v_i/h_i and nearest neighbors $N(v_i)/N(h_i)$. We should try to minimize the following objective formula:

$$\frac{\alpha}{2}\sum\nolimits_{i=1}^{N_v}\sum\nolimits_{m=1}^{N_v} a_{im}\|w_i - w_m\|_F^2 = \frac{\alpha}{2} tr\left(W^T L^v W\right) \tag{9}$$

$$\frac{\beta}{2}\sum\nolimits_{i=1}^{N_h}\sum\nolimits_{m=1}^{N_h} b_{jn}\|w_j - w_n\|_F^2 = \frac{\beta}{2} tr\left(W^T L^h W\right) \tag{10}$$

Where tr(·) is the trace of the matrix, $L^v = (D^v + \widetilde{D^v}) - (A + A^T)$, and the diagonal element of D^v is $D_{ii}^v = \sum_{m=1}^{N_v} a_{im}$; $L^h = (D^h + \widetilde{D^h}) - (B + B^T)$, the diagonal elements of D^h are $D_{jj}^h = \sum_{n=1}^{N_h} b_{jn}$. Our goal is to find the minimum of the following objective functions:

$$\begin{aligned}\min_{W,H} & \sum\nolimits_{i=1}^{N_v}\sum\nolimits_{j=1}^{N_h}\left(1 + cy_{ij} - y_{ij}\right) \ \ln\left[1 + \ exp\left(\boldsymbol{w}_i \boldsymbol{h}_j^T\right)\right] - cy_{ij}\boldsymbol{w}_i\boldsymbol{h}_j^T \\ & + \frac{1}{2} tr\left[\boldsymbol{W}^T\left(\lambda_v \boldsymbol{I} + \alpha \boldsymbol{L}^v\right)\boldsymbol{W}\right] \\ & + \frac{1}{2} tr\left[\boldsymbol{H}^T\left(\lambda_h \boldsymbol{I} + \beta \boldsymbol{L}^h\right)\boldsymbol{H}\right]\end{aligned} \tag{11}$$

Where $\lambda_v = \frac{1}{\sigma_v^2}, \lambda_h = \frac{1}{\sigma_h^2}$, σ_v and σ_h are expressed as the variance of Gaussian distribution of viruses and hosts, respectively. $\|\cdot\|_F$ represents the Frobenius norm of the matrix, and W and H are randomly initialized using a Gaussian distribution with a mean of 0 and standard deviation of $\frac{1}{\sqrt{r}}$. We use the AdaGrad algorithm [29] to solve the optimization problem of Eq. (11).

When learning vectors W and H, vectors of the negative virus group or host group are learned only based on negative associations in the training process. However, some unknown virus-host associations may exist potential correlations. Based on previous studies, we replaced the vector of a negative virus/host with a linear combination of its neighbors in the positive set [28]. Here, we build K_2 nearest neighbor sets for each virus and host separately and K_2 is set to 5, according to the experimental study. We use $N^+(v_i)/N^+(h_j)$ to express K_2 nearest neighbors of $v^i \in V^-/h_j \in H^-$ in V^+/H^+ . Therefore, $\boldsymbol{w}_i$ and $\boldsymbol{h}_j$ in Eq. (7) are corrected to:

$$\tilde{\boldsymbol{w}}_i = \begin{cases} \boldsymbol{w}_i & if \ v_i \in V^+ \\ \dfrac{1}{\sum_{\mu \in N^+(v_i)} s_{im}^v} \sum\nolimits_{m \in N^+(v_i)} s_{im}^v \boldsymbol{w}_m & if \ v_i \in V^- \end{cases} \tag{12}$$

$$\tilde{\boldsymbol{h}}_j = \begin{cases} \boldsymbol{h}_j \ if \ h_j \in H^+ \\ \dfrac{1}{\sum_{v \in N^+(h_j)} s_{jn}^h} \sum\nolimits_{n \in N^+(h_j)} s_{jn}^h \boldsymbol{h}_n \ if \ h_j \in H^- \end{cases} \tag{13}$$

Evaluation metrics

Based on the heterogeneous network constructed by the above method, we compare the AUC [30] and AUPR [31] of ILMF-VH and recent five network-based algorithms by five times five-fold cross-validation to evaluate their performances. Then, based on previous studies [10, 17], we evaluated our virus-host association prediction methods by host prediction accuracy on a benchmark dataset including 820 viruses genomes. The host prediction accuracy refers to the percentage of the virus which is predicted to have the same host taxonomy level as known hosts of the query virus.

Results and discussion

Performance evaluation of different based-network methods

In order to assess the performance of our model, we trained datasets including 352 viruses and 71 hosts to obtain model parameters and tested our model on benchmark datasets including 820 viruses and 2699 hosts. In addition, we compare ILMF-VH model with five recently proposed network-based methods (LMFH-VH [18], NetLapRLS [29], KBMF2K [32], BLM -NII [33], CMF [34]) through five-fold cross-validation in the dataset containing 352 viruses. In each round of five-fold cross-validation, one-fifth of the virus-host associations are set to test data, and corresponding elements in the adjacency matrix Y are set to 0, the other four subsets are used as training data. It should be noted that in each round of five-fold cross-validation experiment, when virus-host relationships are set to 0, the Y matrix has been changed, so each time we need to recalculate GIP kernel similarities between hosts, and then kernel similarities can be fused with ONF similarities of hosts by applying SNF model to obtain updated host-host similarities. In addition, according to previous studies [28, 32–35], the range of parameter settings for each method is shown in the Table 1. Here, we use a random search strategy [36] for each model to select optimal parameters.

Table 2 shows the AUC values and AUPR values obtained by six methods in the data sets including 352 viruses. The results showed that ILMF-VH achieved the best performance and AUC value and AUPR value are 0.9202 and 0.6243, respectively. This result demonstrates the effectiveness of our model in virus-host association prediction.

Table 1 Experimental setup for the five network-based modeling process

Method	Parameter	Range of Parameter
ILMF/LMFH-VH	α	$\{2^{-5}, 2^{-4}\ldots, 2^{2}\}$
	β	$\{2^{-5}, 2^{-4}\ldots, 2^{0}\}$
	λ_v, λ_h	$\{2^{-5}, 2^{-4}\ldots, 2^{1}\}$
	γ	$\{2^{-3}, 2^{-2}\ldots, 2^{0}\}$
	k	{20,40,60…100}
KBMF	r	{50,100}
NetLapRLS	gamma_d	$\{10^{-3}, 10,^{-2}\ldots, 2^{0}\}$
	beta_d, beta_t	$\{2^{-3}, 2,^{-2}\ldots, 2^{0}\}$
BLM-NII	combination weight α	{0,0.1,0.2, …, 0.9,1}
CMF	k	{50,100}
	regularization coefficient lambda_l	$\{2^{-2}, \ldots 2^{1}\}$
	lambda_d, lambda_t	$\{2^{-3}, 2^{-2}, \ldots 2^{5}\}$

Sensitivity analysis of parameter values

As seen in Additional file 1: Figure S1-Figure S4, these figures show AUPR values obtained by ILMF-VH model corresponding to different parameter settings. We also tested effects of different K value (the number of neighbors of KNN) of SNF model on AUPR values (Additional file 1: Figure S5). So, we mainly analyze five parameters of ILMF-VH and the number of neighbors K of SNF model.

More specificity, we analyse the change trend of AUPR values with different factorization factor k used for matrix factorization. As shown in Additional file 1: Figure S1, the optimal value of k is 100 and average AUPR value of ILMF-VH is 0.6305 under five-fold cross validation. In addition, we also study impacts of regularization parameters α and β used for neighborhood smoothing in the prediction procedure. Additional file 1: Figure S2 shows the change trend of AUPR values under different α and β. The optimal values of α and β are 0.0625 and 0.25, respectively. When $\alpha > 0.0625$ and $\beta > 0.25$, corresponding AUPR values begin to decrease. These results emphasize that neighbor regularization has a certain impact on the virus-host prediction model. Moreover, we also analyse effects of λ on the prediction procedure. Here, $\lambda = \lambda_v = \frac{1}{\sigma_v^2} = \lambda_h = \frac{1}{\sigma_h^2}$, σ_v and σ_h represent the variance of Gaussian distribution of viruses and hosts, respectively. As shown in Additional file 1: Figure S3, the AUPR value becomes larger gradually with the increase of λ, and when λ equals 2, AUPR reaches optimal value. Additional file 1: Figure S4 shows the variation trend of AUPR when learning rate parameters γ is set to different values. When γ equals 0.25, AUPR takes the optimal value; when γ increases, the AUPR value begins to decrease, so γ is set to 0.25. Furthermore, we also analyzed influences of different neighbor parameter K of SNF model on AUPR values. As shown in Additional file 1: Figure S5, the AUPR value reaches the optimal value when K is set to 5; when K increases again, the AUPR value begins to decrease, so the optimal value of K is 5.

Table 2 The AUC and AUPR obtained by ILMF-VH and other five network-based methods

AUC						
	ILMF-VH	LMFH-VH	KBMF	BLM-NII	NetLapRLS	CMF
352 virus	**0.9202**	0.8568	0.7934	0.8201	0.6711	0.8286
AUPR						
	ILMF-VH	LMFH-VH	KBMF	BLM-NII	NetLapRLS	CMF
352 virus	**0.6243**	0.5560	0.3408	0.10054	0.2749	0.3100

Comparison of ILMF-VH and previous virus-host prediction studies

In this work, we apply the ILMF-VH method to the benchmark dataset including 820 viruses and 2699 complete bacterial genomes. First, we calculate scores between each virus and candidate hosts. The higher the predicted score, the more likely the virus is infected by the host. Here, the highest ranked host is identified as the predicted result of the given virus, and if the predicted host is the same as known host of the given virus at the species level, the predicted host is considered as a correct one. Figure 2 shows the host prediction accuracy of four types of methods include abundance profile co-occurrence, sequence homology, sequence composition, and network-based. The result shown that ILMF-VH achieved the highest host prediction accuracy (58.90%) compared with other three types of methods.

In order to further improve the host prediction accuracy, we apply a consensus method [17] to our method. We believe that the most frequent host species in the top n predicted hosts of a virus can be classified as the host taxon of the given virus. The prediction accuracy is highest at $n = 5$, therefore, we selected the most frequent classification among the top 5 hosts as the host taxon of the query virus. As shown in Fig. 2, when a consensus strategy is applied to our model, the host prediction accuracy can be increased to 63.66%, which is 24.66%, 13.29 and 0.49% higher than three proposed virus-host prediction methods [10, 17, 18], respectively.

As for the general situation, when a new virus lacks host information, we can use the ILMF-VH method to predict its potential hosts. First, we constructed a virus-host network based on known virus-host associations; then GIP kernel similarities between hosts based on known virus-host associations can be calculated, and these GIP kernel similarities and ONF similarities of hosts are integrated through SNF model, so the host similarity network can be constructed. At the same time, we can calculate ONF similarities of whole genome sequences between the new virus and other viruses in the

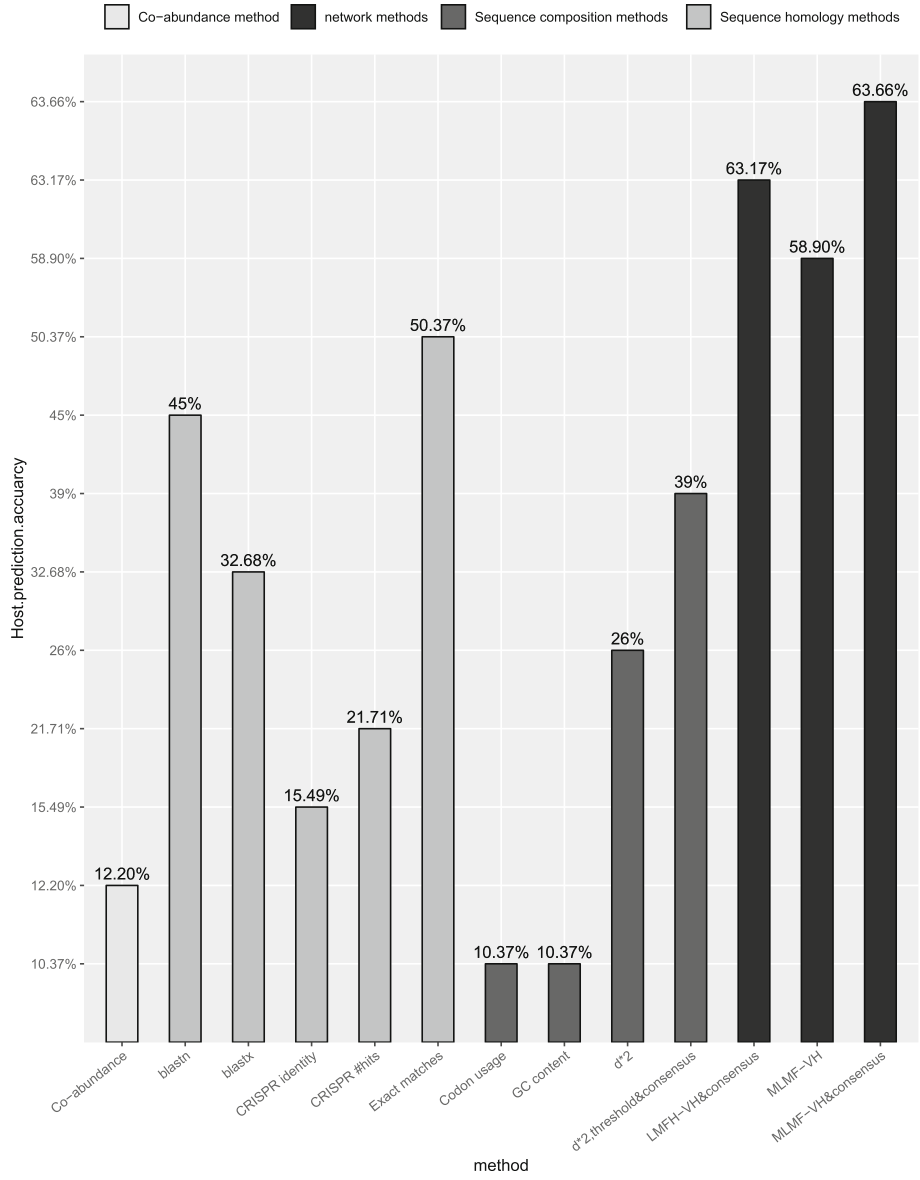


Fig. 2 The host prediction accuracy of four types of methods for benchmark datasets including 820 viruses and 2699 hosts

virus-host association network, so the virus similarity network can be established; finally, the ILMF-VH method is executed on the virus-host heterogeneous network, thereby the predicted scores between the new virus and all candidate hosts can be obtained.

Case study

In this work, we evaluated the performance of ILMF-VH model through case studies. In a recent study, Dutil et al. used viral metagenomic sequencing data in human fecal samples to find [19] a highly abundant phage crAssphage and utilized co-occurrence profile of crAssphage and 404 potential human intestinal bacterial hosts from 151 human gut genomes in the human microbiome program to predict hosts of crAssphage, their results indicates that the host of crAssphage belongs to *Bacteroidetes*. At the same time, Ahlgren et al. [17] predicted potential hosts of crAssphage based on sequence similarities between crAssphage and candidates hosts; WANG et al. [9] used the Markov random field integration network to predict potential hosts of crAssphage. They all suggested that bacteria belonging to *Bacteroidetes* are the host of crAssphage. According to previous study [19], crAssphage is a virus that is widely found in the human gut genome, but we know very little about its biological significance and hosts of crAssphage, due to the difficulty of culturing crAssphage. Different methods have been proposed to predict hosts of the given virus, our information integration algorithm validates the host of crAssphage which was found in previous studies and also predicts another potential host *Escherichia coli*.

As for each virus, the candidate host is ordered according to predicted association scores obtained by ILMF-VH algorithm. In this paper, we assume that if the known candidate host of a virus v_j is h_i, another new host h_k at the same taxon level as the host h_i may be a potential host of the virus v_j. At the same time, the higher the predicted score of the candidate host h_k, the more likely it is to have a potential correlation with the query virus. In the case study, we added the whole genome sequence of crAssphage to the similarity network containing 820 viruses, that is, similarities between the crAssphage and 820 virus sequences can be calculated based on ONF measurement, thus a new virus-virus similarity network can be constructed. Apart from that, we also add links between crAssphage and 2699 hosts to the virus-host network to build a new virus-host association network. Based on ONF measurement and known associations between viruses and hosts, we used our algorithm to obtain predicted scores between crAssphage and 2699 candidate hosts.

Our approach supports the previous conclusion that candidate hosts belonging to *Bacteroides* are potential hosts of crAssphage. As for the top 50 predicted hosts of crAssphage, there were three hosts belonging to phylum *Bacteroidetes* and were ranked 4th, 44th and 50th: *Cardinium endosymbiont of Encarsia pergandiella*, *Weeksella virosa*, and *Tannerella forsythia*. Our prediction model also inferred that *Escherichia coli* belonging to phylum *Proteobacteria* is the potential host of crAssphage, and *Escherichia coli* ranks highest among 2699 hosts. A possible explanation for its highest predicted score is that the alignment-free similarity score between crAssphage and *Escherichia coli* is 0.6568, which is higher than the average score (0.6096) between the virus and all candidate hosts. Therefore, sequence alignment is an important part of extracting virus-host association signal, and it provides an efficient contribution indicator for this prediction result.

Our algorithm predicted host of crAssphage based on the metagenomics sequencing data, which is identical to the putative host at phylum level in previous studies. In addition, another potential host *Escherichia coli* is also inferred. Recent studies have shown that [20] most *Escherichia coli* strains grow harmlessly in the gut and rarely cause diseases in healthy individuals. However, many pathogenic strains can cause diarrhea or extraintestinal disease in both healthy and immunocompromised individuals. Our experimental results suggest that crAssphage may play an important role in these diseases. In general, our algorithmic model is effective in predicting potential hosts of new viruses.

Conclusion and outlook

Viral infection usually results in changes in the ecosystem function of host cells. Virus-host association studies can reveal complex virus-host network interactions and are important for understanding of microorganism diversity. Despite this, although some methods for virus-host association prediction have been proposed, the host prediction accuracy at the species level cannot be achieved very well and these methods need to be improved.

We present an effective method ILMF-VH for predicting virus-host associations. We performed the best performance compared to recent five network-based methods by five-fold cross-validation. Secondly, we compared the host prediction accuracy with several recently proposed virus-host association prediction methods [10, 17]. Our method obtained the highest host prediction accuracy (63.66%). Finally, we analyzed our method's abilities to predict potential hosts for the given virus. As for the crAssphage, our predicted hosts are corresponding to previous studies, and predicted another host *Escherichia coli* is associated with intestinal diseases. In general, it is important to study virus-host associations. Our research not only has potential to predict hosts of viruses, but also can be applied to predict virus-host associations.

Although some results have been achieved so far, there are still some problems that can be further studied in the future. First, the biology characteristics of viruses and hosts are abundant and varied. Apart from whole genome sequences, protein, amino acid, abundance profile and other related information might also have contribution to the prediction model. It needs further research to study what information provides a reliable basis for virus-host association prediction, and extracting appropriate characteristics of viruses and hosts are important for predicting results. Here, we integrate genome sequence information and known virus-host associations. In the future researches, we will consider adding different information sources of viruses and hosts to analyze impacts of different characteristics on prediction results.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-019-3082-0.

Additional file 1: Figure S1. The trend chart of AUPR values vary with the factorization factor *k*. **Figure S2.** The trend chart of AUPR values vary with regularization parameters *α* and *β*. **Figure S3.** The trend chart of AUPR values vary with the inverse of the variance *λ*. **Figure S4.** The trend chart of AUPR values vary with the learning rate parameter *γ*. **Figure S5.** The trend chart of AUPR values vary with the neighbor number parameter *K*.

Acknowledgements

Not applicable.

About the Supplement

This article has been published as part of BMC Bioinformatics Volume 20 Supplement 16, 2019: Selected articles from the *IEEE BIBM International*

Conference on Bioinformatics and Biomedicine (BIBM) 2018: bioinformatics and systems biology. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-16

Authors' contributions

DL designed virus-host association prediction algorithm under XJ's guidance. DL implemented the model and designed experiment to compare and analysis results. YM contributed to train and find optimal parameters. XJ and DL contributed to writing the manuscript. TH supervised and helped conceive the study. All authors read and approved the final manuscript.

Funding

The research was supported by the National Key Research and Development Program of China (2017YFC0909502), the National Natural Science Foundation of China (61532008, 61872157).
Specifically, the publication costs are funded by the National Key Research and Development Program of China (2017YFC0909502).

Availability of data and materials

Our dataset is primarily derived from virus-host metagenomic sequencing data and known virus-host associations in NCBI. There is also the sequence of crAssphage in the human gut metagenomic. The data and code for this article is available at https://github.com/liudan111/ILMF-VH.git.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Computer, Central China Normal University, Wuhan, Hubei, China. [2]Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University, Wuhan, Hubei, China.

Published: 2 December 2019

References

1. Hannigan GD, Grice EA. Microbial ecology of the skin in the era of metagenomics and molecular microbiology. Cold Spring Harb Perspect Med. 2013;3(12):a015362.
2. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, Kambal A, Monaco CL, Zhao G, Fleshner P, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell. 2015;160(3):447–60.
3. Seekatz AM, Rao K, Santhosh K, Young VB. Dynamics of the fecal microbiome in patients with recurrent and nonrecurrent Clostridium difficile infection. Genome Med. 2016;8(1):47.
4. Zackular JP, Rogers MA, MTt R, Schloss PD. The human gut microbiome as a screening tool for colorectal cancer. Cancer Prev Res (Phila). 2014;7(11): 1112–21.
5. Ly M, Abeles SR, Boehm TK, Robles-Sikisaka R, Naidu M, Santiago-Rodriguez T, Pride DT. Altered oral viral ecology in association with periodontal disease. MBio. 2014;5(3):e01133–14.
6. Modi SR, Lee HH, Spina CS, Collins JJ. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. Nature. 2013;499(7457):219–22.
7. Delwart EL. Viral metagenomics. Rev Med Virol. 2007;17(2):115–31.
8. Akhter S, Aziz RK, Edwards RA. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Res. 2012;40(16):e126.
9. Wang W, Ren J, Ahlgren NA, et al. A network-based integrated framework for predicting virus-host interactions with applications[J]. bioRxiv. 2018. https://doi.org/10.1101/505768.
10. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE, Smith M. Computational approaches to predict bacteriophage–host relationships. FEMS Microbiol Rev. 2016;40(2):258–72.
11. Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. Science. 2010;327(5962):167–70.
12. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015;3:e985.
13. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32(8):822–8.
14. Roux S, Hallam S J, Woyke T, et al. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. Elife. 2015;4:e08490.
15. Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–70.
16. Pride DT, Wassenaar TM, Ghose C, Blaser MJ. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics. 2006;7:8.
17. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Res. 2017;45(1):39–53.
18. Liu D, Hu X, Jiang X. Virus-host association prediction by using Kernelized logistic matrix factorization on heterogeneous networks. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2018 IEEE. p. 108–13.
19. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GG, Boling L, Barr JJ, Speth DR, Seguritan V, Aziz RK, et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun. 2014;5:4498.
20. Gomes TA, Elias WP, Scaletsky IC, Guth BE, Rodrigues JF, Piazza RM, Ferreira LC, Martinez MB. Diarrheagenic Escherichia coli. Braz J Microbiol. 2016; 47(Suppl 1):3–30.
21. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage-host relationships. FEMS Microbiol Rev. 2016;40(2): 258–72.
22. Reinert G, et al. Alignment-free sequence comparison (I): statistics and power. J Comput Biol. 2009;16(12):1615–34.
23. Zou S, Zhang J, Zhang Z. A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network. PLoS One. 2017;12(9):e0184394.
24. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics. 2011;27(21): 3036–43.
25. Huang ZA, Chen X, Zhu Z, et al. PBHMDA: path-based human microbe-disease association prediction. Front Microbiol. 2017;8:233.
26. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333–7.
27. Hu Y, Koren Y, Volinsky C. Collaborative filtering for implicit feedback datasets. Eighth IEEE International Conference on Data Mining. Pisa: IEEE; 2008:263–272.
28. Liu Y, Wu M, Miao C, Zhao P, Li XL. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. PLoS Comput Biol. 2016;12(2):e1004760.
29. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res. 2011;12(Jul): 2121–59.
30. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
31. Davis J, Goadrich M. The relationship between precision-recall and ROC curves: Proceedings of the 23rd international conference on Machine learning. New York: ACM; 2006. p. 233–240.
32. Gonen M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. Bioinformatics. 2012;28(18):2304–10.
33. Mei JP, Kwoh CK, Yang P, Li XL, Zheng J. Drug-target interaction prediction by learning from local information and neighbors. Bioinformatics. 2013;29(2):238–45.
34. Zheng X, Ding H, Mamitsuka H, et al. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions: Proceedings of the 19th ACM SIGKDD international

conference on Knowledge discovery and data mining. Chicago:ACM; 2013. p. 1025–1033.
35. Xia Z, Wu LY, Zhou X, et al. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Syst Biol. 2010;4(2):S6.
36. Bergstra JBY. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13(Feb).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.