SPRDA: a matrix completion approach based on the structural perturbation to infer disease-associated Piwi-Interacting RNAs

Kai Zheng¹, Zhu-Hong You^{2,*}, Lei Wang^{2,*}, Leon Wong²,
Zhao-hui Zhan³

¹School of Computer Science and Engineering, Central South University, Changsha, 410083, China

²Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Science, Urumqi, 830011, China

³Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

*Corresponding author

Email: zhuhongyou@ms.xjb.ac.cn; leiwang@ms.xjb.ac.cn

Keywords: PIWI-interacting RNA, disease, piRNA-disease association, structural perturbation, duplex network

Abstract

Emerging evidence suggests that PIWI-interacting RNAs (piRNAs) are one of the most influential small non-coding RNAs (ncRNAs) that regulate RNA silencing. piRNA and PIWI proteins have been confirmed for disease diagnosis and treatment as novel biomarkers due to its abnormal expression in various cancers. However, the current research is not strong enough to further clarify the functions of piRNA in cancer and its underlying mechanism. Therefore, how to provide large-scale and serious piRNA candidates for biological research has grown up to be a pressing issue. The main motivation of this work is tantamount to fill the gap in research on large-scale prediction of disease-related piRNAs. In this study, a novel computational model based on the structural perturbation method is proposed, called SPRDA. In detail, the duplex network is constructed based on the piRNA similarity network and disease similarity network extracted from piRNA sequence information, Gaussian interaction profile kernel similarity information and gene-disease association information. The structural perturbation method is then used to predict the potential associations on the duplex network, which is more predictive than other network structures in terms of structural consistency. In the five-fold cross-validation, SPRDA shows high performance on the benchmark dataset piRDisease, with an AUC of 0.9529. Furthermore, the predictive performance of SPRDA for 10 diseases shows the robustness of the proposed method. Overall, the proposed approach can provide unique insights into the pathogenesis of the disease and will advance the field of oncology diagnosis and treatment.

1 Introduction

PIWI interacting RNA (piRNA) is a distinct class of small non-coding RNA mainly expressed in germ cells [1, 2]. In 2006, Girard *et al.* discovered a previously uncharacteristic class of small RNAs that bind to murine Piwi proteins MIWI and are highly abundant in the testes. Therefore, these small RNAs are named Piwi-interacting RNAs (piRNAs) [3]. piRNA specifically binds to the PIWI subfamily of Argonaute proteins to form an effector complex to silence transposon activity in animal gonads, and maintain germline genome integrity through transcription or post-transcription mechanisms [4-7]. In addition, studies have demonstrated that piRNA can also regulate cellular genes and transmit "self" and

"non-self" memories during intergenerational inheritance, which implicates PIWI in establishment of transgenerational epigenetic states [8-10]. In metazoans, there are 3 types of small silencing RNA: microRNA (miRNA), piRNA and small interfering RNA (siRNA). The processing of both miRNAs and siRNAs need to cleave by Dicer endonuclease. Compare with other small RNAs, piRNAs arise by Dicer-independent mechanisms [9, 11-13]. Different biogenesis mechanisms led to the realization that piRNA is manifest from siRNA and miRNA.

As research continues, piRNA and PIWI proteins have been confirmed for disease diagnosis and treatment as novel biomarkers due to its abnormal expression in various cancers [14-19]. Therefore, exploring the mechanism of piRNA at the transcriptional or post-transcriptional level in pathological processes has become an important way to figure out the molecular roots of human disease. For example, Li *et al.* found that piR-32051 was positively correlated with high tumor stage and metastasis of Kidney cancer [20]. Subsequently, Mark *et al.* confirmed that the expression of piR-34871 and piR-52200 was positively correlated with RASSF1C expression in tumor tissue. In addition, knock-down of piR-34871 and piR-52200 was in a position to significantly reduce the proliferation of lung cancer cell lines (A549 and H1299) in 2017 [21]. In addition, Anna *et al.* Found that piR-651 expression was inhibited in serum samples from 94 Hematological malignancy patients. Interestingly, after complete remission, piR-651 levels returned to normal [22]. The piR-39980 in fibrosarcoma was first reported in 2019. Das *et al.* found that transient overexpression of piR-39980 can attenuate the proliferation and induce apoptosis of HT1080 fibrosarcoma cells, and has a strong antitumor effect [23]. Glioblastoma has also been linked to piRNA, Leng *et al.* found that overexpression of piR-DQ590027 increased the permeability of the blood-brain barrier under glioma conditions and promoted the transport of anti-glioma drugs to glioma tissue [24].

piRNA is accumulating rapidly as research continues. In order to homogenize complex and heterogeneous data, some databases were established to collect the features of piRNA such as piRNAQuest [25], piRNABank [26] and piRBase [27]. In 2019, Muhammad *et al.* developed the piRDisease database in order to promote research on piRNAs in disease diagnosis, prognosis, and assessment of treatment response [28]. The database provides comprehensive and proprietary data, including experimental support, sequence and location information. Although many predictors

have been proposed to infer the link between ncRNA and disease, piRNA-based calculation model is unexplored [29-33].

In this study, a piRNA-disease association predictor based on the structural perturbation method was proposed, called SPRDA. The SPRDA has following advances: (i) We construct a duplex network based on the piRNA similarity network and disease similarity network extracted from piRNA sequence information and gene-disease association information, which can effectively improve prediction performance. (ii) A new uniquely mapped disease characterization method is proposed and used to quantify the similarity. The results demonstrate that the method has lower computational cost under the same performance. (iii) We use of the structural consistency index to quantify the link predictability of the piRNA-disease network. The results show that the network consistency is much higher than other comparative networks. In the 5-fold cross-validation, the AUC of our method reaches 91.45% and the accuracy is 84.49%. In addition, we performed independent performance tests on 12 different diseases to evaluate the performance of SPRDA. From the results, it can effectively identify disease-related piRNAs. Overall, the proposed method can provide unique insights into the mechanisms of disease and will further the field of tumor diagnosis and treatment.

2 Materials and Methods

2.1 Benchmark dataset

Many public databases provide piRNA feature information, such as piRNAQuest [25], piRNABank [26] and piRBase [27]. However, only piRDisease [28] currently contains experimentally verified piRNA-disease associations. To enable large-scale analysis of disease-related piRNAs, we used piRDisease as the baseline dataset (<http://www.piwirna2disease.org/index.php>). Since there are a large number of nodes with *degree* = 1 in the piRNA-disease association network formed by this database, which seriously affects the predictability of the network. Therefore, we performed preprocessing on the original data by deduplication and filtering out points with *degree* = 1, and define the benchmark data set piRD, as shown in Table 1.

Table 1. The details of benchmark dataset piRD

Benchmark dataset	piRNAs	diseases	Associations
piRD	501	22	1212

2.2 Construct piRNA–disease bilayer network

The piRNA–disease duplex network is constructed by three networks, including the disease similarity network, piRNA similarity network, and piRNA-disease association network. The disease similarity network and piRNA similarity network are weighted by different subnets.

The piRNA similarity network. The piRNA sequence similarity network and the piRNA Gaussian interaction profile kernel similarity network (GIP) are important subnets for constructing the piRNA similarity network. These two similarity network are described in detail below. As a carrier of the genetic information and functions of RNA, sequences play a decisive role in many processes such as translation and catalysis. ncRNA silences or degrades other RNAs through complementary base pairing. Here, the k -mer algorithm is utilized to quantify the similarity of the sequences [34]. The core idea of the algorithm is to count the number of subsequences in the sequence (such as AAA, ACG, UCG, etc.) and calculate its Z-score. The calculation process is as follows:

$$Z_{p_i}(SQ_a) = \frac{Count(SQ_a) - \frac{1}{N_{SQ}} \sum_{k=1}^N Count(SQ_k)}{\sqrt{\frac{1}{N_{SQ}} \sum_{f=1}^{SQ} \left(SQ_f - \frac{1}{N_{SQ}} \sum_{k=1}^{N_{SQ}} Count(SQ_k) \right)^2}} \quad (1)$$

where p_i is the i -th piRNA. SQ_a is the a -th subsequence. N_{SQ} is the number of subsequences. $Count(SQ_a)$ is the number of SQ_a . Therefore, the sequence similarity PS_{seq} between piRNAs can be calculated:

$$PS_{seq}(p_i, p_j) = \frac{\sum(Z_{p_i} * Z_{p_j}) - \sum Z_{p_i} * \sum Z_{p_j}}{\sqrt{\left(\sum(Z_{p_i})^2 - \frac{(\sum Z_{p_i})^2}{N_{SQ}} \right) \left(\sum(Z_{p_j})^2 - \frac{(\sum Z_{p_j})^2}{N_{SQ}} \right)}} \quad (2)$$

Then, the piRNA gaussian interaction profile kernel similarity can be calculated from the piRNA-disease association network. Specifically, the gaussian interaction profile kernel similarity PS_{guss} between piRNA p_i and piRNA p_j is defined as follows:

$$PS_{guss}(p_i, p_j) = \exp(-\vartheta_p ||T(p_i) - T(p_j)||^2) \quad (3)$$

where $T(p_i)$ is the vector of the relationship between piRNA p_i and all diseases. If piRNA p_i is related to disease d , we set the value as 1, otherwise set as 0. ϑ_p is the kernel width coefficient and is defined:

$$\vartheta_p = \frac{1}{\frac{1}{N_p} \sum_{k=1}^{N_p} ||T(p_k)||^2} \quad (4)$$

where N_p is the number of piRNAs. According to the piRNA sequence similarity network and the piRNA Gaussian interaction profile kernel similarity network, the similar network of piRNA can be constructed:

$$PS(p_i, p_j) = \max(PS_{seq}(p_i, p_j), PS_{guss}(p_i, p_j)) \quad (5)$$

The disease similarity network. Disease functional similarity network and disease Gaussian interaction profile kernel similarity network (GIP) are important subnets for constructing disease similarity network. The two similarity networks are described in detail below. At present, there are few methods that can quantify the feature of diseases. Many methods use Medical Subject Headings (MeSH) to construct directed acyclic graphs, and calculate the semantic similarity between two diseases by the relative positions in the directed acyclic graphs [35]. The similarity matrix calculated by this method is very sparse as shown in Figure 1(A) and the coverage is limited. To this end, we propose unified embedding to the disease representation, called Dgene, provides unique 64-dimensional feature vectors for 24166 diseases. In detail, we downloaded 628684 associations about diseases and genes from DisGeNET, <http://www.disgenet.org>, which contains 24166 diseases and 17543 genes (Table 2). The specific calculation steps are as follows.

Table 2. The details of dataset DisGeNET.

Dataset	Disease	Gene	Association
DisGeNET	24166	17543	628684

We defined $V(d_i)$ is the vector of the relationship between disease d_i and all diseases. The d_i is the i -th disease. If disease d_i is related to gene g , we set the value as 1, otherwise set as 0. Then, the original feature vectors of all diseases are put

into a stacked autoencoder to extract deeper abstract information and generate 64-dimensional feature vector $V'(d_i)$. This operation has the following advantages: reducing the redundancy of the original data and the subsequent calculation costs. According to $V'(d_i)$, we can calculate the disease functional similarity network:

$$DS_{gene}(d_i, d_j) = \frac{\sum (V'(d_i) * V'(d_j)) - \sum V'(d_i) * \sum V'(d_j)}{\sqrt{\left(\sum V'(d_i)^2 - \frac{(\sum V'(d_i))^2}{N_d}\right) \left(\sum V'(d_j)^2 - \frac{(\sum V'(d_j))^2}{N_d}\right)}} \quad (6)$$

where N_d is the number of diseases in piRD. The visualized disease function similarity matrix network is shown in Figure 1 (B). Compared with disease semantic similarity, Dgene-based disease functional similarity covers more diseases. At the same time, in Dgene, each disease has a unique corresponding 64-dimensional feature vector, and the time-cost of Dgene-based method is much lower than that of MeSH-based method when dealing with different prediction tasks (Table 3).

Table 3. Comparison of different types of disease similarity.

Method	Benchmark dataset	CPU time/pair (seconds) ^a	Size ^b
MeSH-based	piRD	378.4498	484
Dgene-based	piRD	0.2919	484

^aThe average CPU time (seconds) required to calculate the disease similarity matrix.

^bThe size of the disease similarity matrix

Figure 1. (A) The disease similarity network based on Medical Subject Headings. (B) The disease similarity network based on disease-gene association

Then, the disease gaussian interaction profile kernel similarity can be calculated. In detail, the gaussian interaction profile kernel similarity DS_{guss} between disease d_i and disease d_j is defined as follows:

$$DS_{guss}(d_i, d_j) = \exp(-\vartheta_d ||T(d_i) - T(d_j)||^2) \quad (7)$$

where $T(d_i)$ is the vector of the relationship between piRNA d_i and all diseases. If disease d_i is related to piRNA p , we set the value as 1, otherwise set as 0. ϑ_d is the kernel width coefficient and is defined:

$$\vartheta_d = \frac{1}{\frac{1}{N_d} \sum_{k=1}^{N_d} ||T(d_k)||^2} \quad (8)$$

where N_d is the number of disease. According to the disease functional similarity network and the disease Gaussian interaction profile kernel similarity network, the similar network of disease can be figured:

$$DS(d_i, d_j) = \alpha * DS_{gene}(d_i, d_j) + \beta * DS_{guss}(d_i, d_j) \quad (9)$$

where α and β are contribution coefficients, and $\alpha + \beta = 1$. Finally, based on the calculated DS and PS , and the piRNA-disease association network DP constructed from the benchmark dataset, the duplex network Adj can be defined:

$$Adj = \begin{bmatrix} PS & DP \\ DP^T & DS \end{bmatrix} \quad (10)$$

2.3 Structural consistency index

In 2015, Liu *et al.* proposed a method for assessing the predictability of complex network links without any prior knowledge of the network, called the "structural consistency" index [36]. Due to its excellent performance and wide versatility in prediction work in other fields[37], this method applied to piRNA-disease duplex networks. In the previous section, we introduced the duplex network Adj , which is composed of adjacency matrix DP , disease similarity matrix DS , and piRNA similarity matrix PS . Therefore, the edges in the duplex network composed of piRNA and disease are weighted. The duplex network can be defined as $DN(N, E, W)$. Where N represents the set of nodes composed of piRNA and disease in the network, E represents the set of edges in the network, and W represents the weight of each edge. The perturbation set E^p consists of randomly selected edges of the duplex network, and the remaining edges $E - E^p$ are defined as E^r . The duplex network made up of the perturbed set E^p is denoted as Adj^p , and the duplex network made up of the remaining edges set E^r is denoted as Adj^r . In particular, $Adj = Adj^p +$

Adj^r . As the duplex network is undirected, Adj , Adj^p , and Adj^r are all real symmetric matrices. Therefore, we can diagonalize Adj^r as follow:

$$Adj^r = \sum_{j=1}^N \lambda_j x_j x_j^T \quad (11)$$

where λ_j is the eigenvalue of the duplex network Adj^r , and x_j is the corresponding orthogonal and normalized eigenvector. Then, transforming the perturbation set E^p into a perturbation matrix through first-order approximation. In the case of repeated eigenvalues, there will be degradation eigenvalues. Therefore, we discuss two situations that can arise. The first is the case where there are nonduplicate eigenvalues. After perturbing Adj^r , the eigenvalue changes from λ_j to $\lambda_j + \lambda_j^p$, and the corresponding eigenvector changes from x_j to $x_j + x_j^p$. Then, left-multiplying eigenfunction:

$$(Adj^r + Adj^p)(x_j + x_j^p) = (\lambda_j + \lambda_j^p)(x_j + x_j^p) \quad (12)$$

According to the above formula, when the second-order terms $x_j^T Adj^p x_j^p$ and $\lambda_j^p x_j^T x_j^p$ are ignored, we can obtain the following results.

$$\lambda_j^p \approx \frac{x_j^T Adj^p x_j}{x_j^T x_j} \quad (13)$$

Under the condition that the eigenvector is kept unchanged, the matrix after the perturbation can be represented by the eigenvalue of Adj^p corresponding to perturbation set E^p and the eigenvalue of Adj^r .

$$Adj' = \sum_{j=1}^N (\lambda_j + \lambda_j^p) x_j x_j^T \quad (14)$$

The second is the case where there are duplicate eigenvalues. This means that there are multiple eigenvectors corresponding to the same eigenvalue. Therefore, the eigenvalue is defined as λ_{ji} , where j refers to the j th eigenvalue and i refers to the i th eigenvector corresponding to the eigenvalue. According to previous research, we assume that any linear combination of eigenvectors corresponding to the same eigenvalue is still a eigenvector [38]. After perturbing Adj^r , the symmetry of the node will be improved. For this reason, the degenerate eigenvalues needs to be selected to continuously convert them into nondegenerate eigenvalues [39]. In detail, the selected eigenvectors is defined as $\bar{x}_{ji} = \sum_{k=1}^M \beta_{jk} x_{jk}$. Where M represents the number of all eigenvectors corresponding to the same eigenvalue. Then, the eigenfunction can be defined as follows:

$$(Adj^r + Adj^p)\bar{x}_{ji} = (\lambda_{ji} + \bar{\lambda}_{ji}^p)\bar{x}_{ji} \quad (15)$$

giving us

$$\bar{\lambda}_{ji}^p \sum_{k=1}^M \beta_{jk} x_{jk} = \sum_{k=1}^M \beta_{jk} Adj^p x_{jk} \quad (16)$$

Then, left-multiplying x_{jn}^T where $n \in [1, \dots, n]$:

$$\bar{\lambda}_{ji}^p \beta_{jn} = \sum_{k=1}^M \beta_{jk} x_{jn}^T Adj^p x_{jk} \quad (17)$$

The above equation transformed into matrix form can be written as:

$$QV_j = \bar{\lambda}_j^p V_j \quad (18)$$

where Q is a matrix of size $M \times M$. The element Q_{nk} can be calculated by $Q_{nk} = x_{jn}^T Adj^p x_{jk}$. V_j is the column vector of β_{jk} . Finally, the matrix after the perturbation Adj' can be obtained by replacing x_j and λ_j^p with $\bar{x}_j$ and $\bar{\lambda}_j^p$. Since the eigenvector can quantify the network structure, if the eigenvectors of Adj and Adj' tend to be the same, then the perturbation cannot obviously change the matrix structure, that is, Adj has high structural consistency. Structural consistency τ can be defined as follows:

$$\tau = \frac{|E^s \cap E^p|}{s} \quad (19)$$

where s is the number of edges in E^p , and E^s is the s edges with the highest predicted score in Adj' . In Table 4, we calculated the structural consistency τ of the five related networks mentioned above, where E^p is 10% of E . It can be seen from the table that the structural consistency scores of DP+PS and duplex network are higher than DP, indicating that including more information can improve the structural consistency of the matrix. There is a gap between DP+MeSH, DP+DS and DP scores, which shows that too little added information will decrease the structural consistency score. In addition, DP+DS has a higher structural consistency score than DP + MeSH, which shows that DP+DS can improve the predictability of the matrix.

Table 4 Comparison of the structural consistency of the five networks.

Network	Structural consistency	Additional information
<i>DP</i>	0.3680 ± 0.0160	N/A
<i>DP + PS</i>	0.5900 ± 0.0331	piRNA similarity
<i>DP + MeSH</i>	0.0165 ± 0.0077	Disease similarity
<i>DP + DS</i>	0.3589 ± 0.0167	Disease similarity
<i>Duplex network(Adj)</i>	0.6023 ± 0.0297	piRNA and disease similarity

2.4 Method overview

In this paper, a novel method based on duplex network and structural perturbations is proposed to identify biologically significant, yet unmapped piRNA-disease associations, called SPRDA. The SPRDA has three main steps as shown in Figure 2. First, disease function information, piRNA sequence information and original piRNA-disease association information were used to construct the disease functional similarity network DS_{gene} , the piRNA sequence similarity network PS_{seq} , the disease Gaussian interaction profile kernel similarity network DS_{guss} , and the piRNA Gaussian interaction profile kernel similarity network PS_{guss} , respectively. Based on the four similar networks obtained above, the disease similar network DS and the piRNA similar network PS can be integrated. Second, the duplex network Adj is built based on the calculated disease similarity network DS , piRNA similarity network PS , and known piRNA-disease associations PD . Finally, the predicted scores of the unknown associations in the duplex network Adj are calculated through structural perturbation.

Figure 2. The main step of SPRDA to identify biologically significant, yet unmapped piRNA-disease associations

3 Experimental Results

3.1 Parameter optimization

Parameters often affect the performance of the model. To obtain the optimal model, two parameters were used to study its effect on the prediction performance of SPRDA. They are a and t , where a is the percentage of E^p in E and t is the number of perturbations, that is, the number of times that Equation 14 is executed. In order to more fully evaluate the performance of the model, five-fold cross-validation was used. It is worth noting that this model is different from the traditional machine learning model, so the evaluation index is different. According to previous studies [40], we calculated the evaluation indexes of the model under different precision ($Pre.$). The precision ($Pre.$) is defined as:

$$Pre. = \frac{TP}{TP + FP} \quad (20)$$

Firstly, we explored the effect of the parameter t on the prediction performance of SPRDA based on duplex network when the parameter a is fixed at 10%, and shown in Table 5. We can see from the table that as the number of perturbations

increases, the various evaluation indexes of the model also increase. In addition, the performance of SPRDA is optimal when $t = 10$, and it stabilizes at $t = 12$. Therefore, in the following experiments, t is set to 10. Secondly, the influence of parameter a on the prediction performance of SPRDA when the parameter t is fixed at 10 is discussed, as shown in Table 6. From the table, we can see that with the increase of parameter a , the various evaluation indexes of the model tend to increase first and then decrease. When $a = 40\%$, the performance of SPRDA is optimal. When $a = 50\%$, the performance decreases. It is worth noting that when $a = 30\%$ and $a = 40\%$, the performance of the proposed model is very close, but when $a = 40\%$, the calculation time increases significantly. Therefore, in order to balance the efficiency of the model, a is set to 30% in the following experiments.

Table 5. The effect of the parameter t on the performance of SPRDA when fixed parameter $a = 10\%$.

Parameter	<i>Pre.</i>	AUC	Accuracy	Recall	MCC
$t=2$	90%	0.9007±0.0171	0.8284±0.0204	0.7918±0.0285	0.7137±0.0285
	95%	0.8969±0.0044	0.7987±0.0269	0.7331±0.0303	0.6657±0.0384
$t=4$	90%	0.9097±0.0127	0.8622±0.0108	0.8417±0.0173	0.7621±0.0161
	95%	0.9102±0.0149	0.8013±0.0808	0.7442±0.0900	0.6723±0.1155
$t=6$	90%	0.9117±0.0152	0.8556±0.0143	0.8316±0.0222	0.7525±0.0209
	95%	0.9157±0.0082	0.8322±0.0411	0.7751±0.0515	0.7152±0.0608
$t=8$	90%	0.9139±0.0091	0.8677±0.0324	0.8526±0.0498	0.7712±0.0472
	95%	0.9145±0.0126	0.8380±0.0518	0.7842±0.0660	0.7245±0.0771
$t=10$	90%	0.9158±0.0107	0.8684±0.0260	0.8531±0.0427	0.7720±0.0391
	95%	0.9140±0.0102	0.8231±0.0593	0.7664±0.0716	0.7029±0.0866
$t=12$	90%	0.9153±0.0085	0.8680±0.0239	0.8521±0.0378	0.7713±0.0353
	95%	0.9154±0.0091	0.8305±0.0355	0.7721±0.0420	0.7123±0.0514

Table 6. The effect of the parameter α on the performance of SPRDA when fixed parameter $t = 10$.

Parameter	<i>Pre.</i>	AUC	Accuracy	Recall	MCC
$\alpha=5\%$	90%	0.8904±0.0200	0.8437±0.0345	0.8155±0.0483	0.7361±0.0482
	95%	0.8896±0.0128	0.7621±0.0675	0.6981±0.0681	0.6153±0.0941
$\alpha=10\%$	90%	0.9158±0.0107	0.8684±0.0260	0.8531±0.0427	0.7720±0.0391
	95%	0.9140±0.0102	0.8231±0.0593	0.7664±0.0716	0.7029±0.0866
$\alpha=20\%$	90%	0.9293±0.0112	0.8792±0.0274	0.8712±0.0458	0.7884±0.0416
	95%	0.9433±0.0092	0.8726±0.0464	0.8292±0.0464	0.7765±0.0509
$\alpha=30\%$	90%	0.9524±0.0085	0.9002±0.0183	0.9073±0.0335	0.8207±0.0290
	95%	0.9529±0.0088	0.8734±0.0452	0.8325±0.0671	0.7788±0.0715
$\alpha=40\%$	90%	0.9566±0.0080	0.9064±0.0166	0.9187±0.0313	0.8306±0.0267
	95%	0.9563±0.0074	0.8936±0.0329	0.8612±0.0507	0.8101±0.0528
$\alpha=50\%$	90%	0.9587±0.0089	0.9171±0.0120	0.9389±0.0237	0.8480±0.0198
	95%	0.9589±0.0091	0.8792±0.0480	0.8415±0.0688	0.7881±0.0748

3.2 The performance of SPRDA in predicting piRNA-disease associations on piRDisease

According to the parameter optimization in the previous section, in this part, the parameter α and t of the experimental model are 30% and 10. As shown in Figure 3, when *Pre.* = 90%, the average predicted AUC of SPRDA is 0.9524 +/- 0.0085. The AUCs for the five experiments were 0.9603, 0.9493, 0.9624, 0.9435, and 0.9460. When *Pre.* = 95%, the average predicted AUC of SPRDA is 0.9529 +/- 0.0088. The AUCs of the five experiments were 0.9609, 0.9549, 0.9610, 0.9456, and 0.9419. Table 7 lists that when *Pre.* = 90%, the average accuracy, recall, Precision, and Matthews correlation coefficient (MCC) are 0.9002, 0.9073, 0.8927, and 0.8207, respectively. Meanwhile, when *Pre.* = 95%, the average accuracy, recall, accuracy, and Matthews correlation coefficient (MCC) were 0.8734, 0.8325, 0.9422, and 0.7788, respectively. From the above evaluation indexes, it is not hard to see that the outstanding performance of SPRDA may promote research in the field of piRNA.

Figure 3. ROC curves performed by SPRDA on piRD (*Pre.*=90%)

Figure 4. ROC curves performed by SPRDA on piRD (*Pre.*=95%)

Table 7. Five-fold cross-validation results performed by SPRDA on piRD dataset.

<i>Pre.</i>	Testing set	Accuracy	Recall	Precision	MCC
90%	1	0.9174	0.9391	0.8926	0.8482
	2	0.8802	0.8710	0.8926	0.7890
	3	0.9070	0.9191	0.8926	0.8313
	4	0.9153	0.9351	0.8926	0.8448
	5	0.8811	0.8720	0.8934	0.7905
	Average	0.9002±0.0183	0.9073±0.0335	0.8927±0.0004	0.8207±0.0290
95%	1	0.8988	0.8669	0.9421	0.8173
	2	0.8512	0.7972	0.9421	0.7426
	3	0.9339	0.9268	0.9421	0.8765
	4	0.8678	0.8201	0.9421	0.7680
	5	0.8156	0.7516	0.9426	0.6895
	Average	0.8734±0.0452	0.8325±0.0671	0.9422±0.0002	0.7788±0.0715

3.3 Comparison with other four related networks

In Section 2.3, the structural consistency of five related networks is compared. In order to compare these networks more comprehensively, in this section, we compare their performance in predicting piRNA-disease associations. It can be seen from Table 8 that the performance of these related networks tends to be consistent with the structural consistencies. The AUCs of DP+MeSH network and DP+DS network are lower than DP network, which indicates that too little information contained in DS and MeSH reduces the structural consistency, which leads to differences in their AUC. In addition, since the DP+DS network has a higher structural consistency than the DP+MeSH network, the DP+DS network has relatively better prediction performance. what's more, the duplex network we build has the highest structural consistency due to the integration of information from multiple networks. Similarly, its performance is the highest among the five related networks. This shows that adding DS and PS can effectively improve the prediction performance of the model.

Table 8. The AUC across all tested diseases of the four related networks

Method	Five-fold cross-validation					Average (std)
	1st	2nd	3rd	4th	5th	
<i>DP</i>	0.7955	0.8364	0.8004	0.8256	0.7808	0.8077 (0.0227)
<i>DP+MeSH</i>	0.7031	0.7014	0.6978	0.7466	0.7479	0.7194 (0.0255)
<i>DP+DS</i>	0.7624	0.7768	0.7909	0.7487	0.7659	0.7689 (0.0158)
<i>DP+PS</i>	0.9541	0.9413	0.9626	0.9413	0.9479	0.9495 (0.0091)
<i>Duplex network(Adj)</i>	0.9603	0.9493	0.9624	0.9435	0.9460	0.9523 (0.0085)

3.4 Performance for predicting disease specific associations

In order to verify that SPRDA has sufficient recognition ability in predicting piRNAs corresponding to a single disease, we separately predicted piRNAs related to ten diseases. Specifically, when predicting the specific disease, all associations related to the disease in the benchmark matrix are removed, and the remaining associations are used to construct a duplex network for prediction. In Table 8, we sort these ten diseases from high to low in number. It is not difficult to see that when the proposed model predicts 10 specific disease-related piRNAs, most AUCs are higher than 0.80. In detail, Figure 5 lists the ROC curves and recalls for three diseases. As k increases, recall rises. This phenomenon indicates that the higher the score, the more likely there is a association. The above experiment shows that the proposed model has sufficient recognition ability in this task.

Figure 4. ROC curves and recall for 3 diseases performed by SPRDA on piRD. (A) The ROC curve and recall for Head and neck (squamous cell) carcinoma. (B) The ROC curve and recall for Alzheimer disease. (C) The ROC curve and recall for Ovarian cancer.

Table 9. SPRDA predicts performance at 10 disease-specific associations and the number of associations for each disease on the benchmark dataset piRD

Disease	Number of associations	AUC
Alzheimer disease	221	0.8654
Ovarian cancer	139	0.9034
Gastric cancer (recurrence)	89	0.8551
Male infertility	61	0.8382
Breast cancer	60	0.8625
Dysplastic liver nodules and Hepatocellular carcinoma	58	0.7825
Rheumatoid arthritis	52	0.9267
Heart stroke	42	0.827
Head and neck (squamous cell) carcinoma	37	0.9474
Endometrial carcinogenesis	10	0.8548

4 Conclusion

In this study, we proposed a piRNA-disease association prediction based on the structural perturbation method to predict biologically significant, yet unmapped piRNA-disease associations. In particular, we propose a novel disease characterization method (Dgene) and use it to quantify similarity. In detail, the duplex network is constructed based on the piRNA similarity network and disease similarity network extracted from piRNA sequence information, Gaussian interaction profile kernel similarity information and gene-disease association information. Then the structure perturbation method is utilized to predict potential associations on the duplex network. In terms of structural consistency, the duplex network is more predictable than other network structures.

The duplex network and Dgene not only offers excellent prediction performance, but also reduce the time needed for calculation. Experimental results are shown the performance and efficacy of SPRDA in predicting piRNA-disease associations. In the five-fold cross-validation, SPRDA's AUC, ACC and MCC are 0.9529, 0.8734 and 0.7788, respectively (pre=0.95). To further evaluate the robustness of the proposed model, we tested its performance in predicting 10 disease-specific associations, most AUCs are higher than 0.8000. Overall, the proposed method can provide unique

insights into the mechanisms of disease and will further the field of tumor diagnosis and treatment.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work is supported is supported in part by Awardee of the NSFC Excellent Young Scholars Program, under Grants 61722212, in part by the National Science Foundation of China, under Grants 61873212, 61702444, 61572506, in part by the Pioneer Hundred Talents Program of Chinese Academy of Sciences, in part by the Chinese Postdoctoral Science Foundation, under Grant 2019M653804, in part by the West Light Foundation of The Chinese Academy of Sciences, under Grant 2018-XBQNXZ-B-008. The authors would like to thank all anonymous reviewers for their constructive advices.

Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

1. Yin, H. and H. Lin, *An epigenetic activation role of Piwi and a Piwi-associated piRNA in Drosophila melanogaster*. Nature, 2007. **450**(7167): p. 304.
2. Gou, L.-T., *et al.*, *Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis*. Cell research, 2014. **24**(6): p. 680.
3. Girard, A., *et al.*, *A germline-specific class of small RNAs binds mammalian Piwi proteins*. Nature, 2006. **442**(7099): p. 199.
4. Iwasaki, Y.W., M.C. Siomi, and H. Siomi, *PIWI-interacting RNA: its biogenesis and functions*. Annual review of biochemistry, 2015. **84**: p. 405-433.
5. Grimson, A., *et al.*, *Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals*. Nature, 2008. **455**(7217): p. 1193.

6. Malone, C.D., *et al.*, *Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary*. Cell, 2009. **137**(3): p. 522–535.
7. Iwasaki, Y.W., M.C. Siomi, and H. Siomi, *PIWI-interacting RNA: its biogenesis and functions*. Annual review of biochemistry, 2015. **84**.
8. Brennecke, J., *et al.*, *Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila*. Cell, 2007. **128**(6): p. 1089–1103.
9. Siomi, M.C., *et al.*, *PIWI-interacting small RNAs: the vanguard of genome defence*. Nature reviews Molecular cell biology, 2011. **12**(4): p. 246.
10. Luteijn, M.J. and R.F. Ketting, *PIWI-interacting RNAs: from generation to transgenerational epigenetics*. Nature Reviews Genetics, 2013. **14**(8): p. 523–534.
11. Leslie, M., *The immune system's compact genomic counterpart*. 2013, American Association for the Advancement of Science.
12. Houwing, S., *et al.*, *A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish*. Cell, 2007. **129**(1): p. 69–82.
13. Moazed, D., *Small RNAs in transcriptional gene silencing and genome defence*. Nature, 2009. **457**(7228): p. 413.
14. Zou, A.E., *et al.*, *The non-coding landscape of head and neck squamous cell carcinoma*. Oncotarget, 2016. **7**(32): p. 51211.
15. Chu, H., *et al.*, *Identification of novel piRNAs in bladder cancer*. Cancer letters, 2015. **356**(2): p. 561–567.
16. Cheng, J., *et al.*, *piRNA, the new non-coding RNA, is aberrantly expressed in human cancer cells*. Clinica chimica acta, 2011. **412**(17–18): p. 1621–1625.
17. Assumpcao, C.B., *et al.*, *The role of piRNA and its potential clinical implications in cancer*. Epigenomics, 2015. **7**(6): p. 975–984.

18. Ng, K.W., *et al.*, *Piwi-interacting RNAs in cancer: emerging functions and clinical utility*. *Molecular cancer*, 2016. **15**(1): p. 5.
19. Romano, G., *et al.*, *Small non-coding RNA and cancer*. *Carcinogenesis*, 2017. **38**(5): p. 485-491.
20. Li, Y., *et al.*, *Piwi-interacting RNAs (piRNAs) are dysregulated in renal cell carcinoma and associated with tumor metastasis and cancer-specific survival*. *Molecular Medicine*, 2015. **21**(1): p. 381-388.
21. Reeves, M.E., *et al.*, *Identification and characterization of RASSF1C piRNA target genes in lung cancer cells*. *Oncotarget*, 2017. **8**(21): p. 34268.
22. Cordeiro, A., *et al.*, *PiwiRNA-651 as marker of treatment response and survival in classical Hodgkin lymphoma*. *Oncotarget*, 2016. **7**(29): p. 46002.
23. Das, B., *et al.*, *Tumor suppressive activity of PIWI - interacting RNA in human fibrosarcoma mediated through repression of RRM2*. *Molecular carcinogenesis*, 2019. **58**(3): p. 344-357.
24. Leng, X., *et al.*, *Mechanism of piR-DQ590027/MIR17HG regulating the permeability of glioma conditioned normal BBB*. *Journal of Experimental & Clinical Cancer Research*, 2018. **37**(1): p. 246.
25. Sarkar, A., *et al.*, *piRNAQuest: searching the piRNAome for silencers*. *BMC genomics*, 2014. **15**(1): p. 555.
26. Sai Lakshmi, S. and S. Agrawal, *piRNABank: a web resource on classified and clustered Piwi-interacting RNAs*. *Nucleic acids research*, 2007. **36**(suppl_1): p. D173-D177.
27. Wang, J., *et al.*, *piRBase: a comprehensive database of piRNA sequences*. *Nucleic acids research*, 2018. **47**(D1): p. D175-D180.
28. Muhammad, A., *et al.*, *piRDisease v1. 0: a manually curated database for piRNA associated diseases*. *Database*, 2019. **2019**.
29. Wang, L., *et al.*, *Predicting Protein-Protein Interactions from Matrix-Based Protein Sequence Using Convolution Neural Network and*

- Feature-Selective Rotation Forest*. Scientific reports, 2019. **9**(1): p. 9848.
30. Zheng, K., *et al.* *MISSIM: Improved miRNA-Disease Association Prediction Model Based on Chaos Game Representation and Broad Learning System*. in *International Conference on Intelligent Computing*. 2019. Springer.
 31. Zheng, K., *et al.*, *MLMDA: a machine learning approach to predict and validate MicroRNA -disease associations by integrating of heterogenous information sources*. Journal of translational medicine, 2019. **17**(1): p. 1-14.
 32. Wang, L., *et al.*, *LMTRDA: Using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities*. PLoS computational biology, 2019. **15**(3): p. e1006865.
 33. Li, Y., *et al.*, *An Ensemble Classifier to Predict Protein - Protein Interactions by Combining PSSM-based Evolutionary Information with Local Binary Pattern Model*. International journal of molecular sciences, 2019. **20**(14): p. 3511.
 34. Chikhi, R. and P. Medvedev, *Informed and automated k-mer size selection for genome assembly*. Bioinformatics, 2013. **30**(1): p. 31-37.
 35. Lord, P.W., *et al.*, *Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation*. Bioinformatics, 2003. **19**(10): p. 1275-1283.
 36. Lü, L., *et al.*, *Toward link predictability of complex networks*. Proceedings of the National Academy of Sciences, 2015. **112**(8): p. 2325-2330.
 37. Zeng, X., *et al.*, *Prediction of potential disease-associated microRNAs using structural perturbation method*. Bioinformatics, 2018. **34**(14): p. 2425-2432.
 38. Zeng, X., *et al.*, *Prediction of potential disease-associated microRNAs using structural perturbation method*. Bioinformatics, 2018. **34**(14): p. 2425-2432.

39. MacArthur, B. D. and R. J. Sánchez-García, *Spectral characteristics of network redundancy*. Physical Review E, 2009. **80**(2): p. 026117.
40. Zhang, T., *et al.*, *LPGMMF: predicting long non-coding RNA and protein interaction using graph regularized nonnegative matrix factorization*. IEEE/ACM transactions on computational biology and bioinformatics, 2018.