genes

MDPI

Article

A Probabilistic Matrix Factorization Method for Identifying lncRNA-Disease Associations

Zhanwei Xuan [1,2], Jiechen Li [1,2], Jingwen Yu [1,2], Xiang Feng [1,2], Bihai Zhao [1] and Lei Wang [1,2,*]

[1] College of Computer Engineering & Applied Mathematics, Changsha University, Changsha 410001, China; Zhanwei_xuan@163.com (Z.X.); lijiechen39555@163.com (J.L.); jingwen.yu18@gmail.com (J.Y.); fengxiang@xtu.edu.cn (X.F.); bihaizhao@163.com (B.Z.)

[2] Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan 411105, China

* Correspondence: wanglei@xtu.edu.cn; Tel.: +86-151-1110-9999

Received: 29 December 2018; Accepted: 4 February 2019; Published: 8 February 2019

Abstract: Recently, an increasing number of studies have indicated that long-non-coding RNAs (lncRNAs) can participate in various crucial biological processes and can also be used as the most promising biomarkers for the treatment of certain diseases such as coronary artery disease and various cancers. Due to costs and time complexity, the number of possible disease-related lncRNAs that can be verified by traditional biological experiments is very limited. Therefore, in recent years, it has been very popular to use computational models to predict potential disease-lncRNA associations. In this study, we constructed three kinds of association networks, namely the lncRNA-miRNA association network, the miRNA-disease association network, and the lncRNA-disease correlation network firstly. Then, through integrating these three newly constructed association networks, we constructed an lncRNA-disease weighted association network, which would be further updated by adopting the KNN algorithm based on the semantic similarity of diseases and the similarity of lncRNA functions. Thereafter, according to the updated lncRNA-disease weighted association network, a novel computational model called PMFILDA was proposed to infer potential lncRNA-disease associations based on the probability matrix decomposition. Finally, to evaluate the superiority of the new prediction model PMFILDA, we performed Leave One Out Cross-Validation (LOOCV) based on strongly validated data filtered from MNDR and the simulation results indicated that the performance of PMFILDA was better than some state-of-the-art methods. Moreover, case studies of breast cancer, lung cancer, and colorectal cancer were implemented to further estimate the performance of PMFILDA, and simulation results illustrated that PMFILDA could achieve satisfying prediction performance as well.

Keywords: lncRNA; disease; miRNA; lncRNA-disease associations; identifying disease-related lncRNA

1. Introduction

Long non-coding RNAs (lncRNAs) are a class of important heterologous ncRNAs that differ in length from miRNAs by more than 200 nucleotides [1]. For a long time, lncRNAs have been considered to be transcriptional noise, and only recently have these views been changed by increasing evidence [2]. Related studies have shown that lncRNA plays an indispensable role in many biological processes, such as chromatin remodeling, gene transcription, protein transport and trafficking, and epigenetic regulation [3–9]. In addition, the dysregulation of lncRNA in coronary artery disease, autoimmune disease, neurological disorder, and various cancers suggests that lncRNA plays an important role in many complex diseases [10]. Recently, lncRNAs are increasingly attracting the attention of researchers in the field of bioinformatics [10–13].

With the rapid development of high-throughput sequencing technology, thousands of lncRNAs have been discovered in mammalian transcriptions. Numerous studies have also revealed the important role of lncRNA in biological processes and the significant effects in complex human diseases [1]. There is no doubt that lncRNAs are closely related to complex human diseases, and more importantly, some lncRNA-disease associations have been experimentally confirmed. For example, the expression of XIST is up-regulated in glioma tissues and GSCs. Functionally, XIST knockdown exerts tumor suppressor function by reducing cell proliferation, migration and invasion, and inducing apoptosis [14]. LncRNA HOTAIR is highly expressed in prostate cancer and is associated with the growth and aggressiveness of prostate cancer cells [15]. Hence, it is meaningful to identify as many potential lncRNA-disease associations as possible. However, up to now, due to the high costs of traditional biological experiments, the lncRNA-disease associations supported by biological experiments are still very limited. Therefore, it is highly desirable to develop effective computational models to predict potential lncRNA-disease associations. In recent years, some computational models have been developed already, and all these models can be approximately divided into three different categories such as the machine learning-based models, biological network-based models and the models without relying on known lncRNA-disease associations [16].

As for the machine learning-based models, Chen Xing et al. proposed a computational model called LRSLDA to predict potential lncRNA-diseases associations [17] through acquiring two different scores from lncRNA space and disease space simultaneously for the same lncRNA-disease pair. Huang et al. proposed a prediction model called ILNCSIM by combining the LRSLDA, lncRNA functional similarity and disease semantic similarity to calculate the probabilities of lncRNA-disease associations [18]. Zhao et al. developed a Bayesian classifier-based model to identify new cancer-associated lncRNAs by using known cancer-associated lncRNAs such as multivariate data, genome, regulatory protein and transcription data integration [19].

As for the biological network-based models, based on the assumption that lncRNAs with similar functions are often associated with phenotype-like diseases, Sun et al. proposed a model called RWRlncD based on the lncRNA-lncRNA function similarity network [20]. Through integrating known lncRNA expression profiles, lncRNA-disease associations, lncRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity, Chen et al. developed a prediction model called KATZLDA to discover potential lncRNA-disease associations [21].

Among these machine-learning-based models and biological network-based models mentioned above, one of their common features is that known lncRNA-diseases relationships are required during the implement of prediction. However, so far, due to the time complexity and high costs of traditional biological experiments, the experimentally identified known lncRNA-disease associations are still very limited. Hence, some computational models that do not rely on known lncRNA-disease associations have been proposed in recent years. For instance, Liu et al. proposed a model based on the intermediate node genes to predict the potential disease-related lncRNAs [22]. Chen et al. proposed a model called HGLDA based on integrating miRNA-disease associations and lncRNA-miRNA interactions to discover novel lncRNA-disease associations [23].

In this paper, unlike the most advanced prediction models described above, a new model based on probability matrix decomposition called PMFILDA is proposed to discover potential lncRNA-disease associations. At present, matrix decomposition has been widely used in the field of bioinformatics. For example, in the prediction of miRAN-disease correlation, Chen et al. proposed to predict miRNA-disease correlation based on induction matrix complementation, matrix decomposition and heterogeneous graphs [24,25]. Zhao et al. proposed a method based on symmetric non-negative matrix factorization and Kronecker regularized least squares to predict the correlation of miRNA-disease [26]. The difference between our PMFILDA method and above-mentioned models is that we first constructed three kinds of binary association networks based on experimentally validated lncRNA-miRNA associations, miRNA-disease associations, and lncRNA-disease associations separately. Then, based on these three newly constructed association networks, we constructed a weighted lncRNA-disease association network. Moreover, based on the semantic similarity of disease and the functional similarity of lncRNA, we further adopted the KNN algorithm [27] to update the weighted lncRNA-disease association network. Then, according to the updated weighted lncRNA-disease association network, we decomposed the weight matrix of lncRNA-disease into low-order characteristic matrices U and V of the lncRNAs and diseases based on the probability matrix factorization. Finally, the product of U and V would be used to predict the scores of lncRNA-disease pairs. The flowchart of our prediction model PMFIDLA is shown in the following Figure 1.

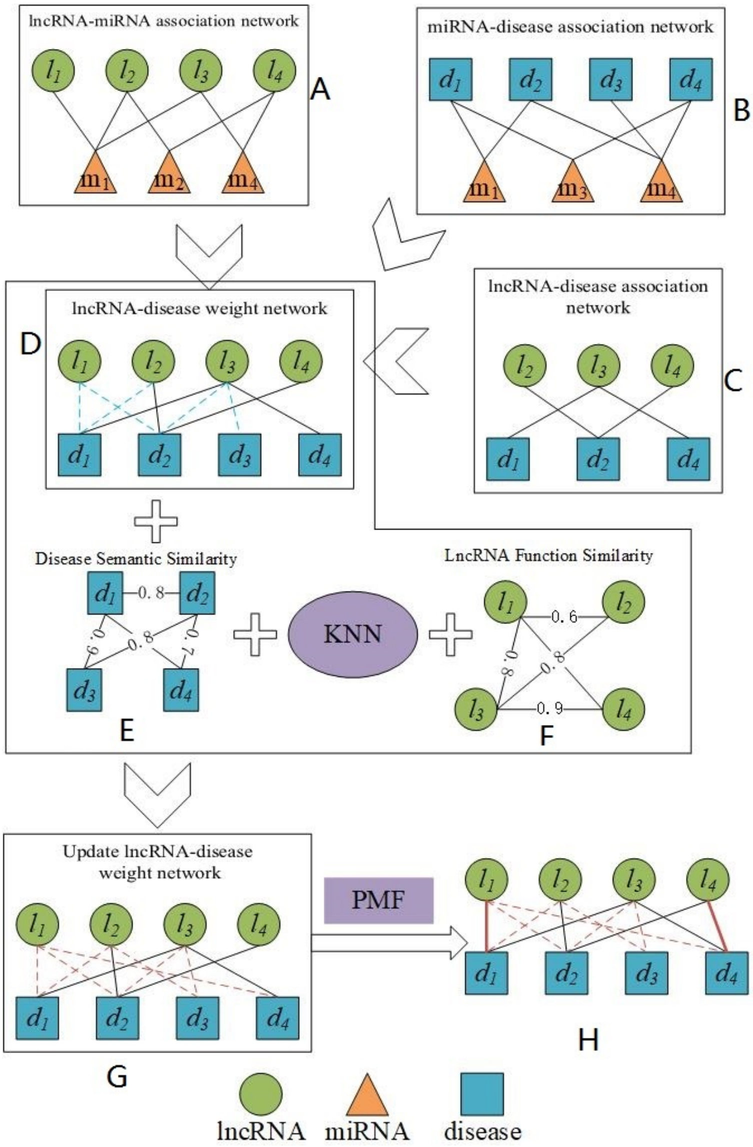
lncRNA-miRNA association network
l1 l2 l3 l4
m1 m2 m4
A
miRNA-disease association network
d1 d2 d3 d4
m1 m3 m4
B
lncRNA-disease weight network
D
l1 l2 l3 l4
d1 d2 d3 d4
lncRNA-disease association network
l2 l3 l4
d1 d2 d4
C
Disease Semantic Similarity
0.8
0.9
0.8
0.7
E
KNN
LncRNA Function Similarity
0.6
0.8
0.8
0.9
F
Update lncRNA-disease weight network
G
PMF
H
lncRNA
miRNA
disease

Figure 1. The flowchart of our prediction model of PMFILDA.

In Subgraph A of above Figure 1, the lncRNA-miRNA association network is constructed based on known lncRNA-miRNA associations downloaded from starbase [28]. Nodes that are linked by solid lines indicate that they are associated. In Subgraph B, the miRNA-disease association network is constructed based on known miRNA-disease associations downloaded from HMDD [29]. Nodes that are linked by solid lines indicate that they are associated. In Subgraph C, the lncRNA-disease association network is constructed based on known lncRNA-disease associations downloaded from MNDR v2.0 [30]. Nodes that are linked by solid lines indicate that they are associated. In Subgraph D, the lncRNA-disease weight network is constructed based on Subgraph A, Subgraph B, and Subgraph C. Nodes that are linked by solid lines indicate that they are related. The blue dashed lines indicate the weight of the initial assignment between nodes. Subgraph E is a network of disease semantic similarity and the numbers in E are similarity scores. Subgraph F is a network of lncRAN functional similarity and the numerics in F are similarity scores. Subgraph G is a lncRNA-disease weighting network that has been updated and the red dashed line indicates weights having been redistributed between nodes. Subgraph H is the lncRNA-disease associations that are ultimately predicted by our method, and the solid red lines indicate the predicted lncRNA-disease associations with relatively high rankings. The KNN is a K-nearest neighbor algorithm used to find the most similar nodes. The PMF is a probability matrix factorization algorithm.

2. Materials

Since known lncRNA-disease associations were considered in our prediction model PMFIDLA, in this section, we download three kinds of gold standard datasets consisting of known lncRNA-miRNA associations, miRNA-disease associations, and lncRNA-disease associations from relevant authoritative databases, respectively.

2.1. Human LncRNA-MiRNA Associations and MiRNA-Disease Associations

Firstly, we downloaded the datasets of experimentally validated known miRNA-disease associations and lncRNA-miRNA associations from the two authoritative databases such as HMDD [29] and starbase [28] separately. Then, after having further unified the names of miRNAs in these two datasets, we could obtain 246 common miRNAs from both of these two datasets. For convenience, we denoted the set of these shared miRNAs as con_*M*. Thereafter, based on these 246 shared miRNAs in con_*M*, we finally downloaded 4704 different miRNA-disease associations and 9086 different lncRNA-miRNAs associations from above two authoritative databases. In addition, for convenience, we denoted the set of these 4704 different miRNA-disease associations as *MD* and the set of these 9086 different lncRNA-miRNAs associations as *LM* separately. Moreover, through statistics, there are 373 different diseases in *MD* and 1089 different lncRNAs in *LM* respectively (see Supplementary Materials Tables S1 and S2).

2.2. Human LncRNA-Disease Associations

Secondly, we downloaded the dataset of experimentally validated known lncRNA-disease associations from the MNDR v2.0 database [30], and for convenience, we denoted the dataset of these downloaded known lncRNA-disease associations as *LD*. Furthermore, to adapt the downloaded data to our prediction model, we would further process the original data as follows:

Step 1: Obtaining the set of different lncRNAs shared in both *LD* and *LM*. In addition, for convenience, we denoted the set of these shared lncRNAs as con_*L*.

Step 2: Obtaining the set of different diseases existed in *MD* and *LD*. In addition, for convenience, we denoted the set of these shared diseases as con_*D*.

Step 3: Obtaining the set of lncRNA-disease associations with both lncRNAs in con_*L* and diseases in con_*D* based on the set of *LD*.

And as a result, we finally obtained 407 different lncRNA-disease associations including 77 different lncRNAs and 95 different diseases(see Supplementary Materials Tables S3).

3. Methods

3.1. Construction of the lncRNA-miRNA Association Network and miRNA-Disease Association Network

Based on the set of *LM* and *MD*, we can construct the lncRNA-miRNA association network and miRNA-disease association network according to the following steps respectively:

Step 1: Supposing that there are n_l different lncRNAs in *LM*, and for convenience, we denote the set of these lncRNAs as $L = \{l_1, l_2, \ldots, l_{n_l}\}$.

Step 2: Supposing that there are n_d different diseases in *MD*, and for convenience, we denote the set of these diseases as $D = \{d_1, d_2, \ldots, d_{n_d}\}$.

Step 3: Supposing that there are n_m different common miRNAs existed in both *MD* and *LM*, and for convenience, we denote the set of these miRNAs as $con\_M = \{m_1, m_2, \ldots, m_{n_m}\}$.

Step 4: Hence, we can firstly obtain an lncRNA-miRNA association network $G_{LMN} = (L, con\_M, E_{lm})$, where E_{lm} denotes the set of experimentally verified known associations in *LM*. For $\forall l_i \in L, m_j \in M$, we define that there is an edge $e_{l_i - m_j}$ between l_i and m_j in E_{lm} if and only if there is an experimentally verified known associations between l_i and m_j in *LM*.

Step 5: Simultaneously, we can also obtain an miRNA-disease association network $G_{MDN} = (con\_M, D, E_{md})$, where E_{md} denotes the set of experimentally verified known associations in *MD*. For $\forall m_i \in con\_M, d_j \in D$, we define that there is an edge $e_{m_i - d_j}$ between m_i and d_j in E_{md} if and only if there is an experimentally verified known associations between m_i and d_j in *MD*.

3.2. Construction of the Weighted lncRNA-Disease Association Network

Based on the newly constructed association networks such as G_{LMN} and G_{MDN}, we can further obtain a weighted lncRNA-disease association network $G_{LDWN} = (L, D, E_{ld}, W_{ld})$, where E_{ld} denotes the set of edges between different lncRNAs in *L* and diseases in *D*. For $\forall l_i \in L, d_j \in D$, we define that there is an edge $e_{l_i - d_j}$ between l_i and d_j in E_{ld} if and only if there is at least one miRNA m_k in con_M with experimentally verified known associations with both l_i in *LM* and d_j in *MD* simultaneously. In addition, $W_{ld} = \{w_{l_i - d_j} | l_i \in L, d_j \in D, e_{l_i - d_j} \in E_{ld}\}$ denotes the set of weight of the edge $e_{l_i - d_j}$ in E_{ld}, and for $\forall l_i \in L, d_j \in D$, if there is $e_{l_i - d_j} \in E_{ld}$, then the weight $w_{l_i - d_j}$ corresponding to $e_{l_i - d_j}$ can be calculated according to the following steps:

Step 1: Supposing that there are *T* different miRNAs in *con_M* with experimentally verified known associations with both l_i in *LM* and d_j in *MD* simultaneously, and for convenience, we denote the set of these *T* different miRNAs as $CM = \{m_1, m_2, \ldots, m_T\}$.

Step 2: Supposing that $RM_{l_i} = \{m_{l_{i1}}, m_{l_{i2}}, \ldots, m_{l_{ip}}\}$ is a set consisting of all miRNAs that have experimentally verified known associations with l_i in *LM*, and $RM_{d_j} = \{m_{d_{j1}}, m_{d_{j2}}, \ldots, m_{d_{jq}}\}$ is a set consisting of all miRNAs that have experimentally verified known associations with d_j in *MD*.

Step 3: Let $RM = RM_{l_i \bigcup RM_{d_j}} = \{m_1, m_2, \ldots, m_S\}$, then we can calculate the weight of $e_{l_i - d_j}$ in G_{LDWN} according to the following Formula (1):

$$w_{l_i - d_j} = \begin{cases} 1 & \text{if } l_i \text{ associated with } d_j \text{ in } LD \\ T/(S+1) & \text{Otherwise} \end{cases} \tag{1}$$

3.3. Similarity Calculation

3.3.1. Disease Semantic Similarity Measure

Considering that the similarity between disease pairs can calculated by their directed acyclic graphs (DAGs) [31], while estimating the semantic similarity of diseases, for any given disease, we will firstly express it as its directed acyclic graph (DAG), and as illustrated in the following Figure 2, in its corresponding DAG, all annotated terms associated with this disease will be contained. For instance, in Figure 2 the DAGs of two different diseases such as Breast Neoplasms (d_1) and Liver Neoplasms (d_2) are shown, and it is obvious that the DAG of d_1 can be denoted as $DAG_{d_1} = (d_1, T_{d_1}, E_{d_1})$, where T_{d_1} denotes all the ancestor nodes of "d_1" and itself, and E_{d_1} represents the set of edges in DAG_{d_1}. Moreover, for any disease d' in DAG_{d_1}, its semantic contribution to d_1 can be calculated according to the following Formula (2):

$$D_{d_1}(d') = \begin{cases} 1 & \text{if there is } d' = d_1 \text{ in } DAG_{d_1} \\ max\{\Delta \times D_{d_1}(\mathrm{d}'') & \text{Otherwise. Here } d'' \in childrenofd'inDAG_{d_1} \end{cases} \tag{2}$$

where Δ will be set to 0.5 based on the suggestion proposed by the state-of-the-art literature [31]. Moreover, in the same way, it is easy to see that the DAG of d_2 can be denoted as $DAG_{d_2} = (d_3, T_{d_3}, E_{d_3})$, and then, for any given diseases d_1 and d_2, the semantic similarity between them can be measured according to the following Formula (3) obviously:

$$SD(d_1, d_2) = \frac{\sum_{t \in T_{d_1} \cap T_{d_2}} (D_{d_1}(t) + D_{d_2}(t))}{\sum_{t \in T_{d_1}} D_{d_1}(t) + \sum_{t \in T_{d_2}} D_{d_2}(t)} \tag{3}$$

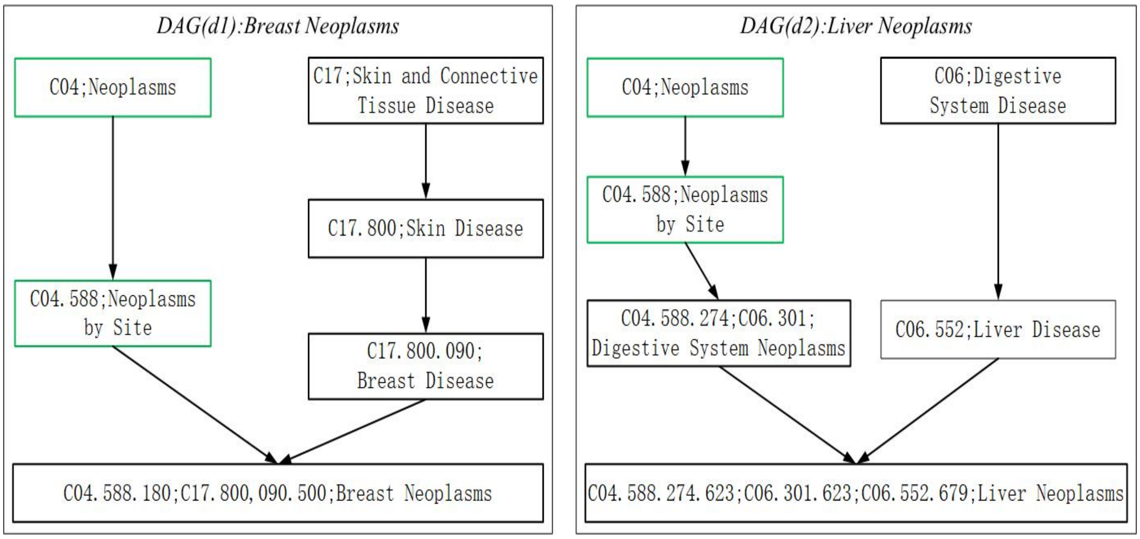
DAG(d1):Breast Neoplasms
C04;Neoplasms
C17;Skin and Connective Tissue Disease
C17.800;Skin Disease
C04.588;Neoplasms by Site
C17.800.090; Breast Disease
C04.588.180;C17.800,090.500;Breast Neoplasms
DAG(d2):Liver Neoplasms
C04;Neoplasms
C06;Digestive System Disease
C04.588;Neoplasms by Site
C04.588.274;C06.301; Digestive System Neoplasms
C06.552;Liver Disease
C04.588.274.623;C06.301.623;C06.552.679;Liver Neoplasms

Figure 2. The DAGs of the disease Breast Neoplasms and Liver Neoplasms. In addition, the disease term and its identification numbers are included in corresponding node. The common terms of the two diseases are illustrated by green nodes.

3.3.2. LncRNA Similarity Measure

The functional similarity between lncRNAs measures how similar their functions will be. In this section, based on the method proposed by the state-of-the-art literature [31], for any given lncRNAs l_i and l_j, Supposing that l_i and l_j have known associations with a group of diseases $GD(l_i) = d_{i1}, d_{i2}, \ldots, d_{ip}$ and $GD(l_j) = d_{j1}, d_{j2}, \ldots, d_{jq}$ in LD respectively, then the functional similarity between them can be measured according to the following Formula (4):

$$FS(l_i, l_j) = \frac{\sum_{t=1}^{p} S(d_{it}, GD(l_j)) + \sum_{t=1}^{q} S(d_{jt}, GD(l_i))}{p+q} \tag{4}$$

$$S(d_k, GD(l_i)) = \max_{t \in [1, |GD(l_i)|]} SD(d_k, d_t) \tag{5}$$

3.3.3. Weight Redistribution in G_{LDWN} Based on the KNN Algorithm

Based on the above descriptions, it is easy to see that we can represent the network G_{LDWN} with its weight matrix W_{ld}, where $W_{ld}[i][j] = w_{l_i - d_j}$. Moreover, considering that known lncRNA-disease associations are very sparse, which may cause that there exist some lncRNAs with no associations with any diseases, or some diseases with no associations with any lncRNAs. Hence, some potential associations between predicted lncRNAs and diseases will be invalid. Therefore, in this paper, we will rebuild the weight matrix W_{ld} to solve this kind of problem as follows:

Step 1: Firstly, representing the *ith* row of the weight matrix W_{ld} as $W_{ld}(l_i,:) = \{w_{l_i-d_1}, w_{l_i-d_2}, \ldots, w_{l_i-d_{nd}}\}$, and the *jth* column of the weight matrix W_{ld} as $W_{ld}(:,d_j) = \{w_{l_1-d_j}, w_{l_2-d_j}, \ldots, w_{l_{nl}-d_j}\}$.

Step 2: Then, for any given lncRNA l_q and any l_i in L other than l_q, based on above Formula (4), it is obvious that we can obtain the functional similarity $FS(l_i,l_q)$ easily, and moreover, after sorting these values of functional similarities between l_q and all remaining lncRNAs other than l_q in descending order, then we can obtain the corresponding lncRNAs from the first K elements in the sorted results. For convenience, let $l_1, l_2, \ldots, l_K$ denote these K lncRNAs, then the *qth* row of W_{ld} can be updated according to the following Formula (6):

$$W_{ld}(l_q,:) = \frac{1}{NL}\sum_{i\in[1,K]} \alpha^{i-1} * FS(l_i,l_q) * W_{ld}(l_i,:) \tag{6}$$

where $\alpha \in (0,1]$ is a decay factor, which means that a higher decay will be assigned to l_i if it is more dissimilar to l_q, and $NL = \sum_{i\in[1,K]} FS(l_i,l_q)$ is the normalization factor, which is used for normalization of the value of $W_{ld}(l_q,:)$. Additionally, in similar way, it is obvious that the *pth* column of W_{ld} can also be updated according to the following Formula (7):

$$W_{ld}(:,d_p) = \frac{1}{ND}\sum_{i\in[1,K]} \beta^{i-1} * SD(d_i,d_p) * W_{ld}(:,d_p) \tag{7}$$

where d_1 to d_K denote the top K diseases most similar to $d_p, \beta \in (0,1]$ is the decay factor, and $ND = \sum_{i\in[1,K]} SD(l_i,l_q)$ is the normalization factor.

3.4. Construction of Our Prediction Model PMFILDA Based on G_{LDWN}

3.4.1. Standard Matrix Factorization

Up to now, the matrix decomposition technology is widely used in the field of recommended systems, since not only the computational complexity can be reduced by matrix decomposition, but also good performance can be achieved in solving the matrix scarcity problem. The standard matrix decomposition aims to find two low-ranking, latent feature matrices whose products are used to fit the original matrix. Hence, for the weight matrix $W_{ld} \in R^{n_l \times n_d}$ constructed above, it is obvious that we can decompose W_{ld} into two different matrices $U \in R^{n_l \times k}$ and $V \in R^{n_d \times k}$ $(k \ll min(n_l, n_d))$, and there is $W_{ld} \approx UV^T$. Thereafter, the problem of disease-related lncRNA prediction can be further expressed by the following Formulas (8) and (9):

$$\arg\min_{U,V} \sum_{i=1}^{n_l}\sum_{j=1}^{n_d}(W_{ld}(i,j) - \widehat{W}_{ld}(i,j))^2 \tag{8}$$

$$\widehat{W}_{ld}(i,j) = \sum_k U_{ik} * V_{jk} = \sum_k U_{i,k} * V_{kj}^T = U_i V_j^T \tag{9}$$

where the row vectors U_i and is V_j represent the ith lncRNA-specific and jth disease-specific latent feature vectors respectively. In addition, obviously, the above Formulas (8) and (9) form a convex optimization problem, which can be solved by some existing optimization algorithms such as the iterative update algorithm [32] easily.

3.4.2. Probabilistic Matrix Factorization

Since the probability matrix decomposition is based on the decomposition of the standard matrix, supposing that W_{ld} is a positive distribution with Gaussian noise, then we can define the conditional distribution over the $W_l d$ as:

$$p(W_{ld}|U,V,\sigma^2)=\prod_{i=1}^{n_l}\prod_{j=1}^{n_d}[N(W_{ld}(i,j)|U_iV_j^T,\sigma^2)]^{I_{ij}} \tag{10}$$

where $N(W_{ld}(i,j)|U_iV_j^T,\sigma^2)$ is the probability distribution function of the normal distribution and $I_{ij}=\begin{cases}1 & W_{ld}(i,j)\neq 0\\ 0 & \text{Otherwise.}\end{cases}$ Obviously, $p(W_{ld}|U,V,\sigma^2)$ is the likelihood function (i.e., the product of all the weights).

In addition, supposing that the matrices U and V satisfy the Gaussian prior to a mean of 0, then the priors of U and V can be denoted as follows:

$$p(U|\sigma_U^2=\Pi_{i=1}^{n_l}N(U_i|0,\sigma_U^2\boldsymbol{I})) \tag{11}$$

$$p(V|\sigma_V^2=\Pi_{i=1}^{n_d}N(V_j|0,\sigma_V^2\boldsymbol{I})) \tag{12}$$

Here, the matrix I is a $T\times T$ dimensional unit diagonal matrix. Assuming that U and V are independent of each other, then the posterior distribution of U and V can be obtained by following Formula (13):

$$\begin{aligned}p(U,V|W_{ld},\sigma^2,\sigma_U^2,\sigma_V^2)&=\frac{p(W_{ld}|U,V,\sigma^2,\sigma_U^2,\sigma_V^2)\times p(U,V)}{p(W_{ld}|U,V,\sigma^2,\sigma_U^2,\sigma_V^2)}\\ \sim p(W_{ld}|U,V,\sigma^2,\sigma_U^2,\sigma_V^2)\times p(U,V)&=p(W_{ld}|U,V,\sigma^2,\sigma_U^2,\sigma_V^2)\times p(U)\times P(V)\\ =\Pi_{i=1}^{n_l}\Pi_{j=1}^{n_d}[N(W_{ld}(i,j)|U_iV_j^T,\sigma^2)]^{I_{ij}}&*\Pi_{i=1}^{n_l}N(U_i|0,\sigma_U^2\boldsymbol{I})*\Pi_{j=1}^{n_d}N(V_j|0,\sigma_V^2\boldsymbol{I})\end{aligned} \tag{13}$$

Then the log of the posterior distribution over the features of lncRNAs and diseases can be calculated as follows:

$$\begin{aligned}\ln p(U,V|W_{ld},\sigma^2,\sigma_U^2,\sigma_V^2)&=\sum_{i=1}^{n_l}\sum_{j=1}^{n_d}I_{ij}\ln N(W_{ld}(i,j)|U_iV_j^T,\sigma^2)+\sum_{i=1}^{n_l}\ln N(U_i|0,\sigma_U^2I)\\ +\sum_{i=1}^{n_d}N(V_j|0,\sigma_V^2\boldsymbol{I})&=-\frac{1}{2\sigma^2}\sum_{i=1}^{n_l}\sum_{j=1}^{n_d}I_{ij}(W_{ld}(i,j)-U_iV_j^T)^2-\frac{1}{2\sigma_U^2}\sum_{i=1}^{n_l}U_iU_i^2-\frac{1}{2\sigma_V^2}\sum_{j=1}^{n_d}V_jV_j^2\\ &-\frac{1}{2}((\sum_{i=1}^{n_l}\sum_{j=1}^{n_d}I_{ij})\ln\sigma^2+Tn_l\ln\sigma_U^2+Tn_d\ln\sigma_V^2)+C\end{aligned} \tag{14}$$

Here, C is a constant factor. In addition, as for $N(U_i|0,\sigma_U^2\boldsymbol{I})$, since there is:

$$N(U_i|0,\sigma_U^2I)=-\frac{1}{(2\pi)^{\frac{T}{2}}|\sigma_U^2I|^{\frac{1}{2}}}exp(-\frac{1}{2}U_i(\sigma_U^2\boldsymbol{I})^{-1}U_i^T) \tag{15}$$

Hence, considering that the matrix $\boldsymbol{I}$ is an unit diagonal matrix, which means that there is $(\sigma_U^2\boldsymbol{I})^{-1}=\frac{1}{\sigma_U^2}\boldsymbol{I}$, then we have:

$$lnN(U_i|0,\sigma_U^2\boldsymbol{I})=ln(-\frac{1}{(2\pi)^{\frac{T}{2}}|\sigma_U^2\boldsymbol{I}|^{\frac{1}{2}}})-\frac{U_iU_i^T}{2\sigma_U^2}=-\frac{T}{2}ln(\sigma_U^2)-\frac{U_iU_i^T}{2\sigma_U^2}+C_U \tag{16}$$

Here, C_U is a constant factor. Similar to the above analyses, we can also have:

$$lnN(V_j|0,\sigma_V^2\boldsymbol{I})=-\frac{T}{2}ln(\sigma_V^2)-\frac{V_iV_j^T}{2\sigma_V^2}+C_V \tag{17}$$

$$lnN(W_{ld}(i,j)|U_iV_j^T,\sigma^2) = -\frac{1}{2}ln(\sigma^2) - \frac{(W_{ld}(i,j)-U_iV_j^T)^2}{2\sigma^2} + C_W \tag{18}$$

Therefore, it is obvious that maximizing the log-posterior on U and V with hyper-parameters being kept fixed in Formula (13) will be equivalent to minimizing the following objective function:

$$\arg\min_{U,V} \frac{1}{2}||\boldsymbol{I} \bigodot (W_{ld} - UV^T)||_F^2 + \frac{\lambda_U}{2}\sum_{i=1}^{n_l}||U_i||_F^2 + \frac{\lambda_V}{2}\sum_{i=1}^{n_d}||V_j||_F^2 \tag{19}$$

where $\bigodot$ is the Hadamard product, $\lambda_U = \frac{\sigma^2}{\sigma_U^2}, \lambda_V = \frac{\sigma^2}{\sigma_V^2}$ and $||\cdot||_F$ represents the Frobenius norm.

3.4.3. Optimization

Based on the properties of Frobenius norm, the Formula (19) can be rewritten as the form of the LaGrangian function as follows:

$$\begin{aligned} L_f(U,V) = \frac{1}{2}Tr(I\bigodot(W_{ld}W_{ld}^T - 2W_{ld}VU^T + UV^TVU^T)) \\ + \frac{\lambda_U}{2}Tr(UU^T) + \frac{\lambda_V}{2}Tr(VV^T) \end{aligned} \tag{20}$$

Based on above Formula (20), we can further obtain its partial derivatives with respect to U and V as follows:

$$\frac{\partial L_f}{\partial U} = I\bigodot(-W_{ld}V + UV^TV) + \lambda_U U \tag{21}$$

$$\frac{\partial L_f}{\partial V} = I\bigodot(-W_{ld}^TU + VU^TU) + \lambda_V V \tag{22}$$

Therefore, we can construct the update rules based on the gradient descent algorithm as follows:

$$U \leftarrow \lambda_m U - \lambda(I\bigodot(-W_{ld}V + UV^TV)) \tag{23}$$

$$V \leftarrow \lambda_m V - \lambda(I\bigodot(-W_{ld}^TU + VU^TU)) \tag{24}$$

where λ_m is the momentum parameter, which can accelerate the convergence speed of U and V, the parameter λ denotes the learning rate, and based on the suggestion proposed by the state-of-the-art literature [31] (Wang et al., 2010), λ_m and λ will be set to 0.8 and 0.005 respectively [33].

Hence, based on above update rules illustrated in Formulas (23) and (24), we can update the lncRNA-specific and disease-specific latent feature matrix U and V until they become converged. Then, we can finally obtain the predicted lncRNA-disease association matrix $\widehat{W}_{ld} = FS \times UV^T \times SD$. In addition, as for any column d_i in W_{ld}, we can sort the elements (i.e., lncRNAs) in d_i in descending order, then the top-ranked lncRNAs in d_i can be predicted as d_i-related lncRNAs, while the bottom-ranked lncRNAs in d_i can be predicted as d_i- disrelated lncRNAs at the same time.

4. Results and Discussion

4.1. Performance Evaluation Metrics

To evaluate the robustness and prediction performance of PMFILDA, in this section, the Leave One Out Cross-Validation (LOOCV) was implemented based on the experimentally verified lncRNA-disease associations. In LOOCV, each pair of known lncRNA-disease associations is used as a validation set, while other known lncRNA-disease associations are used as training sets. Moreover, all the lncRNA-disease pairs without experimentally verify are used as candidate samples. The ranking of the test sample relative to the candidate sample needs to be evaluated after the implementation of

PMFILDA. When a threshold is given, if the test sample ranks above the given threshold, then we will regard that a correctly positive sample has been predicted by PMFILDA, otherwise we will regard that a correctly negative sample has been predicted by PMFILDA. Moreover, while different thresholds are set, a series of True Positive Rate (TPR) and False Positive Rate (FPR) can also be obtained according to the following formulas:

$$TPR = \frac{TP}{TP + FN} \tag{25}$$

$$FPR = \frac{FP}{TN + FP} \tag{26}$$

where TP and TN denote the number of positive and negative samples that have been correctly identified, while FP and FN represent the number of positive and negative samples that have been incorrectly identified. Hence, the Receiver Operating Characteristic (ROC) curve can be drawn by plotting TPRs versus FPRs, and the area under ROC curve (AUC) can be further calculated to measure the global performance of PMFILDA. Obviously, the closer the value of AUC is to 1, the more robust the prediction model would be.

Moreover, during simulation, to eliminate the random errors caused by the random initialization of U and V, we repeated our experiments 100 times and took the mean and variance of AUCs as our final results, which were shown in the following Figure 3. In addition, from Figure 3, it is easy to see that our newly proposed prediction model PMFILDA can achieve the mean AUC of 0.8794 and the standard deviation of 0.0011.

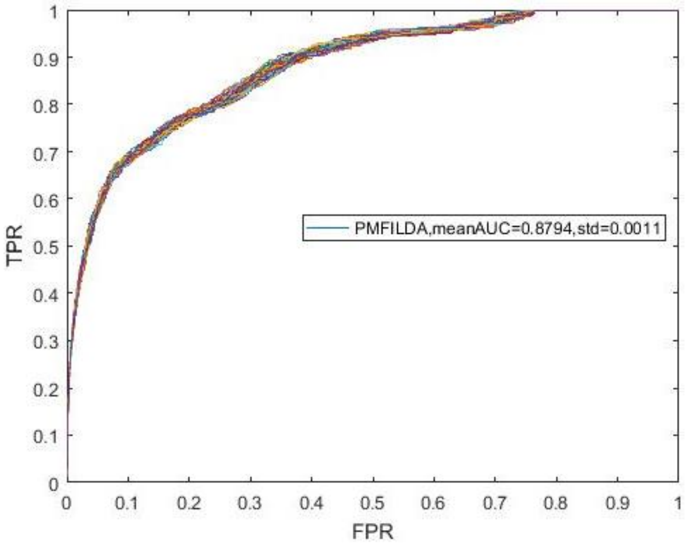

Figure 3. ROC curves for PMFILDA.

Next, to further evaluate the performance of PMFILDA, based on the framework of LOOCV, we compared PMFILDA with some state-of-the-art models such as NBCLDA [34], HGLDA [23], and the method proposed by Yang et al. [35]. Similarly, during simulation, to eliminate the random errors caused by the random initialization of U and V, we repeated our experiments 50 times and took the mean of AUCs as our final results, which were shown in the following Table 1.

Table 1. Comparison of AUCs of PMFILDA with state-of-the-art methods.

Methods	AUCs	Methods	AUCs	Methods	AUCs
PMFILDA	0.8793	PMFILDA	0.9169	PMFILDA	0.9090
$NBCLDA\_GN_1$	0.8519	HGLDA	0.8519	Method of Yang et al.	0.8568

In addition, while comparing PMFILDA with the NBCLDA, considering that we did not consider the genes in our method, then we compared PMFILDA with the $NBCLDA\_GN_1$ only, and the simulation results are shown in Table 1 and the following Figure 4. Obviously, from Table 1 and Figure 4, it is easy to see that our newly proposed prediction model PMFILDA can achieve a reliable AUC of 0.8793 that is much higher than the AUC of 0.8519 achieved by $NBCLDA\_GN_1$.

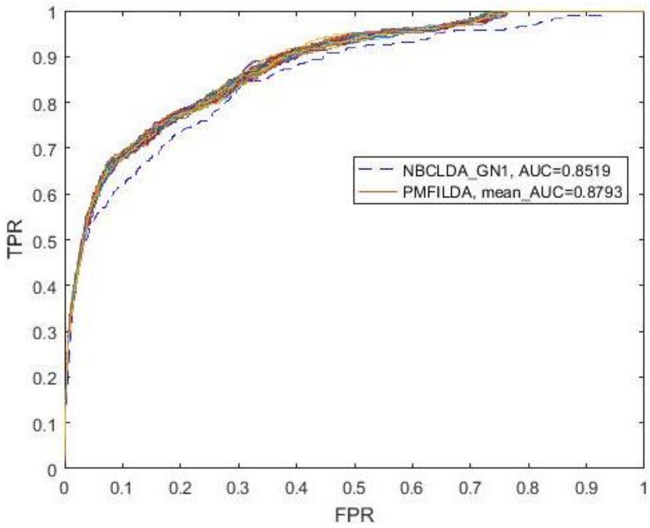

Figure 4. ROC curves and AUC value for $NBCLDA_{GN1}$ and PMFILDA.

Moreover, while comparing PMFILDA with the HGLDA, in order to make a fair comparison, we implemented LOOCV on both PMFILDA and HGLDA based on the same dataset, i.e., we used the same 183 known lncRNA-disease associations proposed by HGLDA in the comparison simulation, and the simulation results are shown in Table 1 and the following Figure 5. Obviously, from Table 1 and Figure 5, it is easy to see that our newly proposed prediction model PMFILDA is superior to HGLDA.

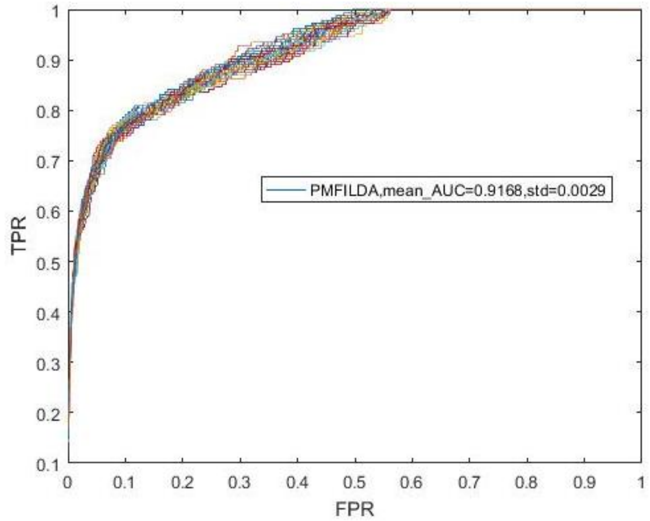

Figure 5. The ROC curve and AUCs of PMFILDA based on the same known 183 lncRNA-disease associations proposed by HGLDA.

Finally, while comparing PMFILDA with the method proposed by Yang et al, in order to make a fair comparison, we implemented LOOCV on both PMFILDA and method proposed by Yang et al. based on the same dataset also, i.e., we used the same 319 known lncRNA-disease associations between 37 lncRNAs and 52 diseases proposed by Yang et al in the comparison simulation, and the simulation results are shown in Table 1 and the following Figure 6. Obviously, from Table 1 and Figure 6, it is easy to see that our newly proposed prediction model PMFILDA can achieve a reliable AUC of 0.9090 that is much higher than the AUC of 0.8568 achieved by Yang et al.

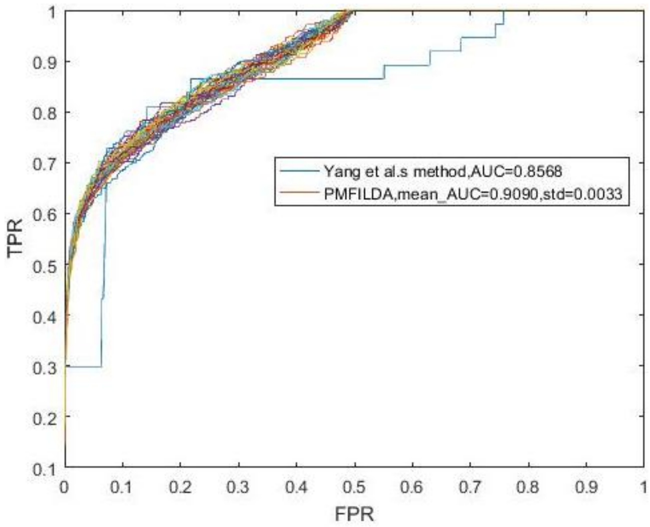

Figure 6. ROC curves and AUCs of PMFILDA and the method proposed by Yang et al.

4.2. Contribution Analysis of lncRNA-Disease Associated Network

In our method, we constructed a weighted lncRNA-disease association networks based on the known lncRNA-disease, microRNA-disease and lncRNA-microNA association networks. It may be useful to discuss the contribution of lncRNA-disease associations separately here. Hence, without considering the relationship between lncRNA and disease, we constructed a weighted network of lncRNA-disease through using known lncRNA-microRNA associations and microRNA-disease associations only. Based on the steps in Section 3.2, we finally obtained 304 lncRNA-disease associations including 60 lncRNAs and 73 diseases. Thereafter, we further obtained the corresponding weight matrix W_{ld}, and then performed LOOCV 100 times on the PMFILDA. The results were shown in the following Figures 7 and 8, obviously, the AUC value achieved by PMFILDA based on three association networks can be increased by 0.0763 than the AUC value achieved by PMFILDA based on two association networks.

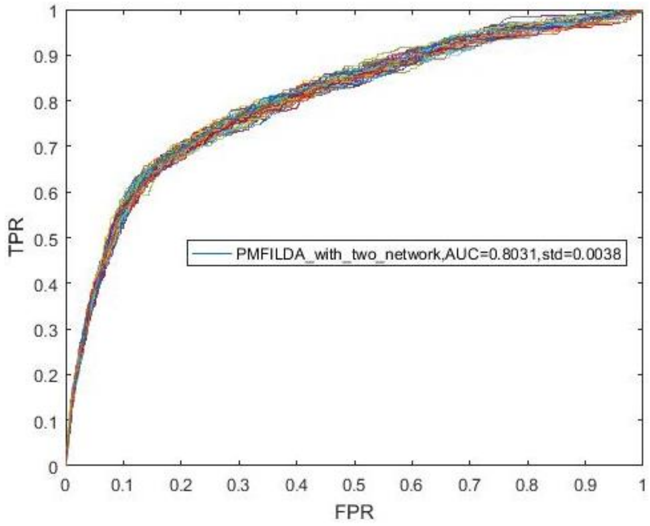

Figure 7. ROC curves and AUCs achieved by PMFILDA based on two association networks.

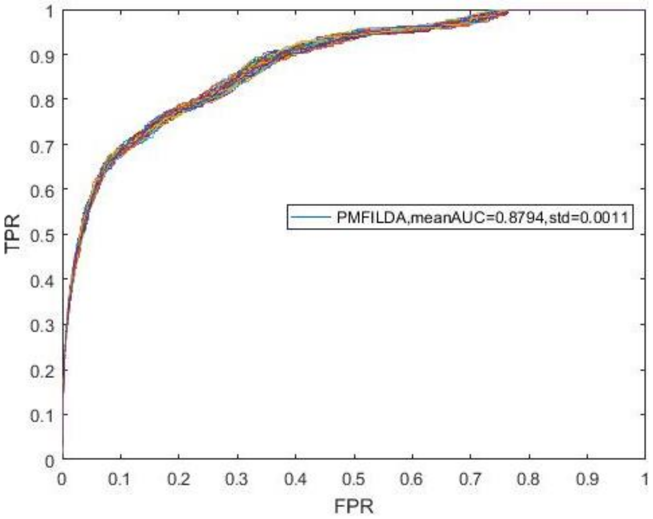

Figure 8. ROC curves and AUCs achieved by PMFILDA based on three association networks.

4.3. The Effects of KNN on Performance

Considering that known lncRNA-disease associations are very sparse, there may exist some lncRNAs with no associations with any diseases, or some diseases with no associations with any lncRNAs. Hence, some potential associations between predicted lncRNAs and diseases will be invalid. Therefore, in this paper, we rebuilt the weight matrix W_{ld} based on KNN algorithm to solve this kind of problem.

Here, we also investigated the influence of KNN algorithm on our method from two aspects. One is that we do not use KNN algorithm to deal with the weighted network G_{LDMN}, the other is to update the weighted network G_{LDMN} with other algorithms, such as K-means algorithm. When PMFILDA is directly executed without using KNN algorithm to process the weighted network G_{LDMN}, the result is shown in Figure 9. It is easy to see that PMFILDA could achieve an average AUC of 0.8794 while the weight of G_{LDMN} was reallocated; however, while the weight of G_{LDMN} was not reallocated, PMFILDA can achieve an average AUC of 0.8042 only, which demonstrated that it can improve the performance of our model through adopting the KNN algorithm to re-allocate the weight of G_{LDMN}.

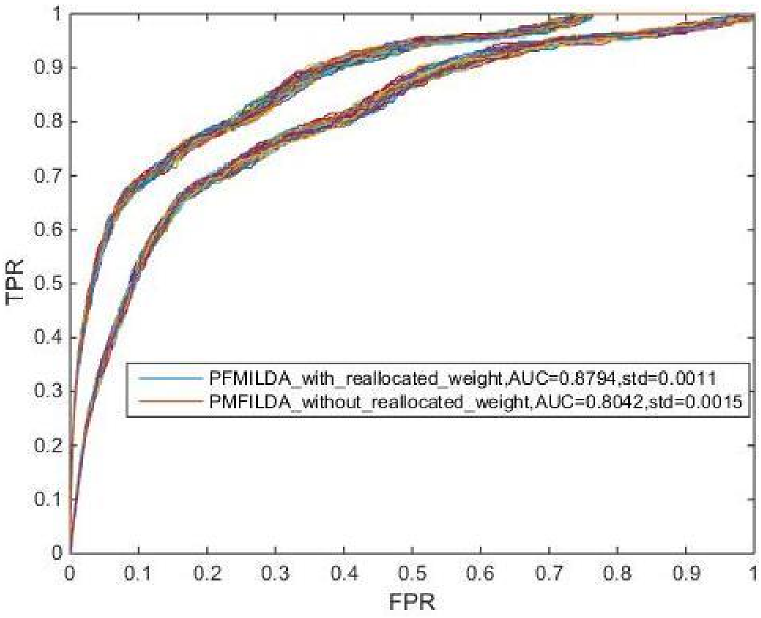

Figure 9. AUCs achieved by PMFILDA in LOOCV while the weight of G_{LDMN} was reallocated or not reallocated respectively.

And in addition, to estimate the impacts of other algorithms, we selected the K-means algorithm for further testing. After performing LOOCV 100 times, we presented the simulation results in the following Table 2, and from observing the results in Table 2, it is easy to see that the performance of KNN is better than K-means.

Table 2. Comparison of the effects of KNN and K-means on PMFILDA.

	KNN	K-Means
Mean_AUC	0.8794	0.8589
STD	0.0278	0.0011

4.4. Parameter Sensitivity Analysis

From above descriptions, it is easy to see that to improve the prediction performance of PMFILDA, some parameters have been introduced in the model construction of PMFILDA, whose values will need to be finalized by the training of the prediction model. For example, how to choose the value of the parameter K while adopting the algorithm of K-nearest neighbor? How to choose the attenuation coefficients α and β given in the Formulas (6) and (7)? How to choose the value of the parameter T while adopting Formula (1) to implement the matrix decomposition? and so on. Hence, firstly, to evaluate the impacts of the parameters K, α and β to the performance of our model PMFILDA, during simulation, we will set K from 2 to 10 and α from 0.1 to 0.9, respectively. Moreover, we will set $\alpha = \beta$ for convenience. The detailed simulation results were shown in the Supplementary Table S4. In addition, through experimental results, it is easy to know that our model PMFILDA can achieve the highest AUC of 0.9204 while $K = 10$ and $\alpha = \beta = 0.8$ in the LOOCV framework. Next, to estimate the impacts of the parameter T to the performance of our model PMFILDA, during simulation, we will set T from 10 to 50 and the step size to 10. In addition, through experimental results, it is easy to know that our model PMFILDA can achieve the highest AUC of 0.9210 while $T = 20$ in the LOOCV framework.

4.5. Case Studies

In this section, we implemented case studies based on the optimal settings of above parameters to further verify the prediction performance of PMFILDA. During simulation, for each given disease, its potentially relevant lncRNAs predicted by PMFILDA will be sorted according to their predicted scores in descending order. In addition, as a result, the top 20 predicted lncRNAs related to the disease potentially will be recorded in the Supplementary Table S5, and then, two public databases such as MNDR V2.0 and LncRNA-Disease database will be used to confirm these potential associations between the given disease and each of these 20 predicted lncRNAs. In this section, we selected three kinds of common diseases such as breast cancer, lung cancer, and colorectal cancer as the targets of our case studies.

As for breast cancer, according to the reports of relevant literatures, it is very common in the group of women [36,37] and may be caused by a variety of molecular alterations. For example, studies have shown that the formation of breast tumors is closely related to lncRNA [38,39]. Hence, predicting breast cancer-associated lncRNA and identifying lncRNA markers are important for the diagnosis and treatment of breast cancer [39]. In this section, we will implement PMFILDA to discover the potential breast cancer-associated lncRNAs. In addition, as shown in the following Table 3, it is easy to see that 12 of the top 20 breast cancer-related lncRNAs predicted by our model PMFILDA have been confirmed in authoritative databases. For example, MALAT1, HOTAIR and H19 ranked the 1st, 2nd and 3rd in the list of our predicted results respectively, and among them, it is proved that MALAT1 has functional and prognostic significance as a metastasis driver in ER negative and lymph node negative breast cancer [40], HOTAIR will be overexpressed in approximately one quarter of human breast cancers and increased in expression in primary breast tumors and metastases [41], and the down-regulation of H19 will significantly reduce colony formation and anchorage-independent growth of breast and lung cancer cells [42].

Moreover, in recent years, lung cancer is a leading cause of cancer-related deaths worldwide, regardless of gender. According to the disease patterns and treatment strategies, it can be roughly divided into non-small cell lung cancer (NSCLC) and small cell lung cancer [43]. To diagnose and treat lung cancers more effectively, researchers have paid lots of attention to the deregulation of protein-coding genes in the past few decades to identify oncogenes and tumor suppressors [43–45]. However, recent studies have shown that lncRNAs play a significant role in the development and progression of lung cancers [43,45]. Hence, in this section, we will implement PMFILDA to infer the potential lung cancer-related lncRNAs. In addition, as illustrated in the following Table 3, it is easy to see that 14 of the top 20 potential lung cancer-related lncRNAs predicted by our model PMFILDA have been confirmed by authoritative biological experiments. For instance, MALAT1, HOTTIP and MEG3 ranked the 3rd, 4th and 5th in the list of our predicted results respectively, and among them, it is identified that MALAT1 is highly correlated with lung cancer metastasis [46,47], will promote lung cancer cell movement by regulating motor-related gene expression [48], and can be an important biomarker for the development of lung cancer metastasis [49]. Additionally, it is demonstrated that through knocking out HOXA13 by RNA interference (siHOXA13), HOTTIP can promote lung cell proliferation, migration, and inhibition of apoptosis, which could serve as a new biomarker and a therapeutic target for NSCLC intervention [50]. Moreover, as for MEG3, it is proved that the down-regulation of MEG3 will enhance cisplatin resistance of lung cancer cells through activation of the WNT/β-catenin signaling pathway [51]. Additionally, the colorectal cancer (CRC) has a high incidence in Western countries in recent years [52], and more and more research indicates that lncRNAs play a significant role in the formation of CRC [44,45]. Hence, in this section, we will implement PMFILDA to predict the potential CRC-related lncRNAs. In addition, as shown in the following Table 2, it is easy to see that eight of the top 20 CRC-related lncRNAs predicted by our model PMFILDA have been confirmed by authoritative biological experiments. For instance, MALAT1, NEAT1 and TUG1 ranked the 2nd, 8th and 10th in the list of our predicted results respectively, and among them, it is identified that MALAT1 may be a potential predictor of tumor metastasis and prognosis, and that the interaction between MALAT1 with SFPQ may be a new therapeutic target for CRC [53]. In addition, it is proved that NEAT1 can be used as an indicator of tumor recurrence and colorectal cancer prognosis [54], and the expression of NEAT1 in CRC may play a carcinogenic role in the differentiation, invasion, and metastasis of CRC, hence, the whole blood NEAT1 expression can be used as a new diagnostic and prognostic biomarker for overall survival in CRC [55]. Moreover, it is demonstrated that the tumor expression of TUG1 plays an important role in colorectal cancer metastasis, and TUG1 can be used as a biomarker or therapeutic target for potential CRC [56].

Table 3. The experimentally confirmed lncRNAs in the top 20 potential lncRNAs predicted by PMFILDA in three kinds of case studies.

Diseases	lncRNAs	Evidence (PUBMED)
Breast Cancer	MALAT1	22492512, 22996375, 24499465, 27250026, 27777857, 27191888
Breast Cancer	HOTAIR	24499465, 20930520, 21925379, 20393566, 19182780, 21903344
Breast Cancer	H19	22996375, 21489289, 14729626, 16707459, 21748294, 18794369
Breast Cancer	MEG3	27166155, 14602737, 22393162, 22487937
Breast Cancers	GAS5	27034004, 18836484, 20673990, 22487937, 22664915, 26662314
Breast Cancer	PTPRG-AS1	26409453
Breast Cancer	NEAT1	25417700, 27147820, 21532345, 27556296
Breast Cancer	PVT1	24780616, 17908964, 25122612, 26889781
Breast Cancer	CDKN2B-AS1	17440112, 20956613, 20453838, 20956613
Breast Cancer	TUG1	27791993
Breast Cancer	XIST	17545591, 27248326, 18006640, 19440381, 24141629, 26637364
Breast Cancer	ZFAS1	21460236
Lung Cancer	H19	27186394, 26729200
Lung Cancer	HOTAIR	27186394, 26729200, 24757675, 23668363, 27270317
Lung Cancer	MALAT1	25217850, 20937273, 20937273, 27777857

Table 3. *Cont.*

Diseases	lncRNAs	Evidence (PUBMED)
Lung Cancer	HOTTIP	27347311, 26265284
Lung Cancer	MEG3	14602737, 26059239
Lung Cancer	CDKN2B-AS1	27307748, 26729200, 26453113, 25964559, 25889788
Lung Cancer	GAS5	27631209, 26634743, 24357161
Lung Cancer	CCAT1	25129441
Lung Cancer	XIST	27501756, 26339353
Lung Cancer	CASC2	26790438
Lung Cancer	PVT1	26908628, 26729200, 25400777
Lung Cancer	ZNRD1-AS1	27166266
Lung Cancer	NEAT1	27351135, 27270317, 25889788
Lung Cancer	TUG1	24853421, 27485439
Colorectal Cancer	H19	8564957, 22427002, 11120891, 26989025, 19926638, 26068968
Colorectal Cancer	HOTTIP	26617875, 26678886, 27546609
Colorectal Cancer	XIST	17143621
Colorectal Cancer	NEAT1	26314847, 26552600
Colorectal Cancer	MEG3	25636452, 26934323
Colorectal Cancer	TUG1	26856330, 27421138
Colorectal Cancer	PVT1	26990997, 24196785
Colorectal Cancer	CCAT1	23416875, 26064266, 26823726, 24594601,23594791,26752646

5. Discussion

Increasing research has shown that lncRNAs play a crucial role in the occurrence, formation, diagnosis, treatment, and prognosis of diseases. The discovery of complex disease-associated lncRNAs as biomarkers based on existing biological experiments is not only costly but also requires a large amount of clinical data. Therefore, it is a future trend to integrate potential biological data resources and use developed computers to develop efficient and accurate computational models to predict potential new disease-related lncRNAs. In this paper, we proposed a novel computational model called PMFILDA to predict potential disease-associated lncRNAs. In this model, we first integrated known lncRNA-miRNA associations, miRNA-disease associations, and a small number of known lncRNA-disease associations into a new weighted lncRNA-disease association network. Then, based on the newly constructed association network, through adopting the semantic similarity of the disease, the functional similarity of lncRNA and the KNN algorithm to update the weight network, an lncRNA-disease association matrix W_{ld} can be obtained. Hence, through adopting the probability matrix decomposition scheme to decompose the matrix W_{ld} into the feature matrix U of lncRNA and the feature matrix V of the disease, we can finally construct our model PMFILDA based on the two feature matrices to predict the potential associations between lncRNAs and diseases. Compared to existing state-of-the-art models, simulation results have demonstrated that our model PMFILDA has better prediction performance. Moreover, case studies of breast cancer, lung cancer and colorectal cancer also indicated that PMFILDA can be used as a superior computational model to predict potential lncRNA-disease associations. However, it is obvious that there are still some biases in our model. When we only use lncRNA-disease associations and regardless of any miRNAs, the performance of PMFILDA may be reduced. To illustrate this situation, we did the following experiment. After processing the data, we obtained 246 pairs of lncRNA-disease associations, including 44 lncRNAs, 68 diseases. Then we performed 100 LOOCVs on the PMFILDA method, and the average AUC value was 0.8111, and the standard deviation was 0.0073. When we used miRNA, the average AUC value was 0.8794 and the standard deviation was 0.0011. The reason for this difference is that when we don′t consider miRNAs, the information we use for lncRNA-disease may be incomplete. There may be some important associations that do not exist in the lncRNA-disease data set. When the miRNA node is added, these important relationships can be re-established. Therefore, in our model, we need to consider not only the lncRNA-disease relationship, but also the nodes that can improve the lncRNA-disease relationship.

6. Conclusions

In this study, our major contributions are as follows: Firstly, we constructed a novel weighted lncRNA-disease association network through integrating the known lncRNA-miRNA association network, the known miRNA-disease association network and the known lncRNA-disease association network. Secondly, based on the semantic similarity of disease and the similarity of lncRNA function, we adopted the KNN algorithm to update the newly constructed weighted lncRNA-disease association network. Thirdly, based on the probability matrix decomposition model, we proposed a novel computational model called PMFILDA to predict potential lncRNA-disease associations, which cannot only predict the potential associations between lncRNAs and disease contained in the experimentally validated lncRNA-disease associations, but also predict the potential associations of its elements in unknown datasets. To improve the efficiency of our model, in the future, we plan to integrate more intermediate nodes such as genes to update the weighted lncRNA-disease association network. In addition, we also believe that the results [25,57–63] of the miRNA-disease association prediction field will promote the development of lncRNA-disease correlation prediction. Moreover, while studying the association prediction of lncRNA-disease, focusing on the research results in other fields will also broaden our horizons.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/10/2/126/s1.

Author Contributions: Conceptualization, Z.X. and L.W.; Methodology, Z.X., J.Y. and L.W.; Validation, J.L., X.F. and B.Z.; Formal Analysis, Z.X. and L.W.; Investigation, J.Y. and J.L.; Resources, X.F. and B.Z.; Data Curation, Z.X. and J.Y.; Writing-Original Draft Preparation, Z.X.; Writing-Review and Editing, L.W. and X.F.; Supervision, L.W.; Project Administration, X.F.; Funding Acquisition, L.W.

Funding: This research was funded by the National Natural Science Foundation of China (No. 61873221, No. 61772089, No. 61672447), the Natural Science Foundation of Hunan Province (No.2018JJ4058, No. 2017JJ5036), and the CERNET Next Generation Internet Technology Innovation Project (No. NGII20160305, No. NGII20170109).

Acknowledgments: The authors thank the anonymous referees for suggestions that helped improve the paper substantially.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Guttman, M.; Russell, P.; Ingolia, N.; Weissman, J.; Lander, E. Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell* **2013**, *154*, 240–251. [CrossRef] [PubMed]
2. Vakul, M.; Yesim, G.P.; Sunil, B.; Sarath Chandra, J. Role of lncRNAs in health and disease-size and shape matter. *Brief. Funct. Genom.* **2015**, *14*, 115–129.
3. Zhao, W.; Luo, J.; Jiao, S. Comprehensive characterization of cancer subtype associated long non-coding RNAs and their clinical implications. *Sci. Rep.* **2014**, *4*, 6591. [CrossRef] [PubMed]
4. Lu, Q.; Ren, S.; Lu, M.; Zhang, Y.; Zhu, D.; Zhang, X.; Li, T. Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genom.* **2013**, *14*, 651. [CrossRef] [PubMed]
5. Li, J.; Xuan, Z.; Liu, C. Long non-coding RNAs and complex human diseases. *Int. J. Mol. Sci.* **2013**, *14*, 18790–18808. [CrossRef] [PubMed]
6. Bussemakers, M.J.; Bokhoven, A.V.; Verhaegh, G.W.; Smit, F.P.; Karthaus, H.F.; Schalken, J.A.; Debruyne, F.M.; Ru, N.; Isaacs, W.B. DD3: A new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res.* **1999**, *59*, 5975–5979. [PubMed]
7. Managadze, D.; Rogozin, I.B.; Chernikova, D.; Shabalina, S.A.; Koonin, E.V. Negative Correlation between Expression Level and Evolutionary Rate of Long Intergenic Noncoding RNAs. *Genome Biol. Evol.* **2011**, *3*, 1390–1404. [CrossRef] [PubMed]
8. Nicole, S.; Harvey, R.P.; Mattick, J.S. Long noncoding RNAs in cardiac development and pathophysiology. *Circ. Res.* **2012**, *111*, 1349–1362.
9. Deeksha, B.; Shruti, K.; Saakshi, J.; Satish, S.; Kriti, K.; Chetana, S.; Sridhar, S.; Vinod, S. Conceptual approaches for lncRNA drug discovery and future strategies. *Expert Opin. Drug Discov.* **2012**, *7*, 503–513.

10. Liao, Q.; Liu, C.; Yuan, X.; Kang, S.; Miao, R.; Xiao, H.; Zhao, G.; Luo, H.; Bu, D.; Zhao, H.; et al. Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network. *Nucleic Acids Res.* **2011**, *39*, 3864–3878. [CrossRef]
11. Chen, X.; Yan, C.C.; Luo, C.; Ji, W.; Zhang, Y.; Dai, Q. Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci. Rep.* **2015**, *5*, 11338. [CrossRef] [PubMed]
12. Chen, X.; Sun, Y.Z.; Guan, N.; Qu, J.; Huang, Z.A.; Zhu, Z.X.; Li, J.Q. Computational models for lncRNA function prediction and functional similarity calculation. *Brief. Funct. Genom.* **2018**, ely031. [CrossRef] [PubMed]
13. Ping, P.; Wang, L.; Kuang, L.; Ye, S.; Mfb, I.; Pei, T. A Novel Method for LncRNA-Disease Association Prediction Based on an lncRNA-disease Association Network. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2018**, 1. [CrossRef] [PubMed]
14. Yao, Y.; Ma, J.; Xue, Y.; Wang, P.; Li, Z.; Liu, J.; Chen, L.; Xi, Z.; Teng, H.; Wang, Z. Knockdown of long non-coding RNA XIST exerts tumor-suppressive functions in human glioblastoma stem cells by up-regulating miR-152. *Cancer Lett.* **2015**, *359*, 75–86. [CrossRef] [PubMed]
15. Zhu, Y.; Ri-Kao, Y.U.; A-Lin, J.I.; Yao, X.L.; Fang, J.J.; Jin, X.D.; Urology, D.O. Effects of long non-coding RNA-HOTAIR on the cell cycle and invasiveness of prostate cancer. *Zhonghua Nan Ke Xue* **2015**, *21*, 792–796. [PubMed]
16. Chen, X.; Yan, C.C.; Zhang, X.; You, Z.H. Long non-coding RNAs and complex diseases: From experimental results to computational models. *Brief. Bioinform.* **2017**, *18*, 558–576. [CrossRef] [PubMed]
17. Chen, X.; Yan, G.Y. Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinformatics* **2013**, *29*, 2617–2624. [CrossRef]
18. Huang, Y.A.; Chen, X.; You, Z.H.; Huang, D.S.; Chan, K.C.C. ILNCSIM: Improved lncRNA functional similarity calculation model. *Oncotarget* **2016**, *7*, 25902–25914. [CrossRef]
19. Zhao, T.; Xu, J.; Liu, L.; Bai, J.; Xu, C.; Xiao, Y.; Li, X.; Zhang, L. Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features. *Mol. Biosyst.* **2015**, *11*, 126–136. [CrossRef]
20. Sun, J.; Shi, H.; Wang, Z.; Zhang, C.; Liu, L.; Wang, L.; He, W.; Hao, D.; Liu, S.; Zhou, M. Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. *Mol. Biosyst.* **2014**, *10*, 2074–2081. [CrossRef]
21. Chen, X. KATZLDA: KATZ measure for the lncRNA-disease association prediction. *Sci. Rep.* **2014**, *5*, 16840. [CrossRef]
22. Liu, M.X.; Chen, X.; Chen, G.; Cui, Q.H.; Yan, G.Y. A computational framework to infer human disease-associated long noncoding RNAs. *PLoS ONE* **2014**, *9*, e84408. [CrossRef] [PubMed]
23. Chen, X. Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. *Sci. Rep.* **2015**, *5*, 13186. [CrossRef] [PubMed]
24. Chen, X.; Wang, L.; Qu, J.; Guan, N.; Li, J.Q. Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* **2018**, *34*, 4256–4265. [CrossRef]
25. Chen, X.; Yin, J.; Qu, J.; Huang, L.; Wang, E. MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction. *PLoS Comput. Biol.* **2018**, *14*, e1006418. [CrossRef] [PubMed]
26. Zhao, Y.; Chen, X.; Yin, J. A novel computational method for the identification of potential miRNA-disease association based on symmetric non-negative matrix factorization and Kronecker regularized least square. *Front. Genet.* **2018**, *9*, 324. [CrossRef] [PubMed]
27. Chen, X.; Wu, Q.F.; Yan, G.Y. RKNNMDA: Ranking-based KNN for MiRNA-Disease Association prediction. *RNA Biol.* **2017**, *14*, 952–962. [CrossRef] [PubMed]
28. Li, J.H.; Liu, S.; Zhou, H.; Qu, L.H.; Yang, J.H. starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* **2014**, *42*, D92–D97. [CrossRef] [PubMed]
29. Li, Y.; Qiu, C.; Tu, J.; Geng, B.; Yang, J.; Jiang, T.; Cui, Q. HMDD v2.0: A database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.* **2014**, *42*, D1070–D1074. [CrossRef]

30. Cui, T.; Zhang, L.; Huang, Y.; Yi, Y.; Tan, P.; Zhao, Y.; Hu, Y.; Xu, L.; Li, E.; Wang, D. MNDR v2.0: An updated resource of ncRNA-disease associations in mammals. *Nucleic Acids Res.* **2018**, *46*, D371–D374. [CrossRef] [PubMed]
31. Wang, D.; Wang, J.; Lu, M.; Song, F.; Cui, Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* **2010**, *26*, 1644–1650. [CrossRef]
32. Lee, D.; Seung, H. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. [CrossRef]
33. Mnih, A.; Salakhutdinov, R.R. Probabilistic Matrix Factorization. *Adv. Neural Inf. Process. Syst.* **2007**, *20*, 1257–1264.
34. Yu, J.; Ping, P.; Wang, L.; Kuang, L.; Li, X.; Wu, Z. A Novel Probability Model for LncRNA–Disease Association Prediction Based on the Naïve Bayesian Classifier. *Genes* **2018**, *9*. [CrossRef]
35. Yang, X.; Gao, L.; Guo, X.; Shi, X.; Wu, H.; Song, F.; Wang, B. A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PLoS ONE* **2014**, *9*, e87797. [CrossRef]
36. Donahue, H.J.; Genetos, D.C. Genomic approaches in breast cancer research. *Brief. Funct. Genom.* **2013**, *12*, 391–396. [CrossRef]
37. Karagoz, K.; Sinha, R.; Arga, K.Y. Triple Negative Breast Cancer: A Multi-Omics Network Discovery Strategy for Candidate Targets and Driving Pathways. *Omics J. Integr. Biol.* **2015**, *19*, 115–130. [CrossRef]
38. Jin, M.; Li, P.; Zhang, Q.; Yang, Z.; Shen, F. A four-long non-coding RNA signature in predicting breast cancer survival. *J. Exp. Clin. Cancer Res.* **2014**, *33*, 84.
39. Xu, N.; Wang, F.; Lv, M.; Cheng, L. Microarray expression profile analysis of long non-coding RNAs in human breast cancer: A study of Chinese women. *Biomed. Pharmacother.* **2015**, *69*, 221–227. [CrossRef]
40. Jadaliha, M.; Zong, X.; Malakar, P.; Ray, T.; Singh, D.K.; Freier, S.M.; Jensen, T.; Prasanth, S.G.; Karni, R.; Ray, P.S. Functional and prognostic significance of long non-coding RNA MALAT1 as a metastasis driver in ER negative lymph node negative breast cancer. *Oncotarget* **2016**, *7*, 40418–40436. [CrossRef]
41. Gupta, R.A.; Nilay, S.; Wang, K.C.; Jeewon, K.; Horlings, H.M.; Wong, D.J.; Miao-Chih, T.; Tiffany, H.; Pedram, A.; Rinn, J.L. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **2010**, *464*, 1071–1076. [CrossRef]
42. Dalia, B.L.; Lau, S.K.; Boutros, P.C.; Fereshteh, K.; Igor, J.; Andrulis, I.L.; Ming, S.T.; Penn, L.Z. The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis. *Cancer Res.* **2006**, *66*, 5330–5337.
43. White, N.M.; Cabanski, C.R.; Silva-Fisher, J.M.; Dang, H.X.; Govindan, R.; Maher, C.A. Transcriptome sequencing reveals altered long intergenic non-coding RNAs in lung cancer. *Genome Biol.* **2014**, *15*, 429. [CrossRef]
44. Prensner, J.R.; Chinnaiyan, A.M. The emergence of lncRNAs in cancer biology. *Cancer Discov.* **2011**, *1*, 391–407. [CrossRef]
45. Tony, G.; Sven, D. The hallmarks of cancer: A long non-coding RNA point of view. *RNA Biol.* **2012**, *9*, 703–719.
46. Tony, G.; Monika, H.; Moritz, E.; Jeff, H.; Youngsoo, K.; Alexey, R.; Gayatri, A.; Marion, S.; Matthias, G. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res.* **2013**, *73*, 1180–1189.
47. Ji, P.; Diederichs, S.; Wang, W.; Boing, S.; Metzger, R.; Schneider, P.M.; Tidow, N.; Brandt, B.; Buerger, H.; Bulk, E.; et al. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **2003**, *22*, 8031. [CrossRef]
48. Tano, K.; Mizuno, R.; Okada, T.; Rakwal, R.; Shibato, J.; Masuo, Y.; Ijiri, K.; Akimitsu, N. MALAT-1 enhances cell motility of lung adenocarcinoma cells by influencing the expression of motility-related genes. *FEBS Lett.* **2010**, *584*, 4575–4580. [CrossRef]
49. Hrdlickova, B.; Almeida, R.C.D.; Borek, Z.; Withoff, S. Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *BBA Mol. Basis Dis.* **2014**, *1842*, 1910–1922. [CrossRef]
50. Sang, Y.; Zhou, F.; Wang, D.; Bi, X.; Liu, X.; Hao, Z.; Li, Q.; Zhang, W. Up-regulation of long non-coding HOTTIP functions as an oncogene by regulating HOXA13 in non-small cell lung cancer. *Am. J. Transl. Res.* **2015**, *8*, 2022.

51. Xia, Y.; He, Z.; Liu, B.; Wang, P.; Chen, Y. Downregulation of Meg3 enhances cisplatin resistance of lung cancer cells through activation of the WNT/β-catenin signaling pathway. *Mol. Med. Rep.* **2015**, *12*, 4530–4537. [CrossRef]
52. Berger, F.G. Interview: Screening and treatment for colorectal cancer. *Colorectal Cancer* **2013**, *2*, 117–120. [CrossRef]
53. Ji, Q.; Zhang, L.; Liu, X.; Zhou, L.; Wang, W.; Han, Z.; Sui, H.; Tang, Y.; Wang, Y.; Liu, N. Long non-coding RNA MALAT1 promotes tumour growth and metastasis in colorectal cancer through binding to SFPQ and releasing oncogene PTBP2 from SFPQ/PTBP2 complex. *Br. J. Cancer* **2014**, *111*, 736. [CrossRef]
54. Li, Y.; Li, Y.; Chen, W.; He, F.; Tan, Z.; Zheng, J.; Wang, W.; Zhao, Q.; Li, J. NEAT expression is associated with tumor recurrence and unfavorable prognosis in colorectal cancer. *Oncotarget* **2015**, *6*, 27641–27650. [CrossRef]
55. Wu, Y.; Yang, L.; Zhao, J.; Li, C.; Nie, J.; Liu, F.; Zhuo, C.; Zheng, Y.; Li, B.; Wang, Z. Nuclear-enriched abundant transcript 1 as a diagnostic and prognostic biomarker in colorectal cancer. *Mol. Cancer* **2015**, *14*, 191. [CrossRef]
56. Sun, J.; Ding, C.; Yang, Z.; Liu, T.; Zhang, X.; Zhao, C.; Wang, J. The long non-coding RNA TUG1 indicates a poor prognosis for colorectal cancer and promotes metastasis by affecting epithelial-mesenchymal transition. *J. Transl. Med.* **2016**, *14*, 42. [CrossRef]
57. Chen, X.; Xie, D.; Wang, L.; Zhao, Q.; Liu, H. BNPMDA: Bipartite Network Projection for MiRNA-Disease Association prediction. *Bioinformatics* **2018**, *34*, 3178–3186. [CrossRef]
58. Chen, X.; Huang, L. LRSSLMDA: Laplacian Regularized Sparse Subspace Learning for MiRNA-Disease Association prediction. *PLoS Comput. Biol.* **2017**, *13*, e1005912. [CrossRef]
59. Chen, X.; Huang, L.; Xie, D.; Zhao, Q. EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association prediction. *Cell Death Dis.* **2018**, *9*, 3. [CrossRef]
60. Zhao, H.; Kuang, L.; Feng, X.; Wang, L. Inferring microRNA-disease associations based on Weighted Interactive Network. *Int. J. Mol. Sci.* **2019**, *20*, 110. [CrossRef]
61. Zhao, H.; Kuang, L.; Wang, L.; Ping, P.; Xuan, Z.; Pei, T.; Wu, Z. Prediction of microRNA-disease associations based on distance correlation set. *BMC Bioinform.* **2018**, *19*, 141. [CrossRef]
62. Zou, Q.; Li, J.; Song, L.; Zeng, X.; Wang, G. Similarity computation strategies in the microRNA-disease network: A survey. *Brief. Funct. Genom.* **2015**, *15*, 55–64,. [CrossRef]
63. Zeng, X.; Zhang, X.; Zou, Q. Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Brief. Bioinform.* **2016**, *17*, 193–203. [CrossRef]