genes

MDPI

*Article*

# DNILMF-LDA: Prediction of lncRNA-Disease Associations by Dual-Network Integrated Logistic Matrix Factorization and Bayesian Optimization

**Yan Li, Junyi Li and Naizheng Bian ***

College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
* Correspondence: nbian@hnu.edu.cn; Tel.: +86-135-7412-0024

Received: 27 June 2019; Accepted: 07 August 2019; Published: 12 August 2019


**Abstract:** Identifying associations between lncRNAs and diseases can help understand disease-related lncRNAs and facilitate disease diagnosis and treatment. The dual-network integrated logistic matrix factorization (DNILMF) model has been used for drug–target interaction prediction, and good results have been achieved. We firstly applied DNILMF to lncRNA–disease association prediction (DNILMF-LDA). We combined different similarity kernel matrices of lncRNAs and diseases by using nonlinear fusion to extract the most important information in fused matrices. Then, lncRNA–disease association networks and similarity networks were built simultaneously. Finally, the Gaussian process mutual information (GP-MI) algorithm of Bayesian optimization was adopted to optimize the model parameters. The 10-fold cross-validation result showed that the area under receiving operating characteristic (ROC) curve (AUC) value of DNILMF-LDA was 0.9202, and the area under precision-recall (PR) curve (AUPR) was 0.5610. Compared with LRLSLDA, SIMCLDA, BiwalkLDA, and TPGLDA, the AUC value of our method increased by 38.81%, 13.07%, 8.35%, and 6.75%, respectively. The AUPR value of our method increased by 52.66%, 40.05%, 37.01%, and 44.25%. These results indicate that DNILMF-LDA is an effective method for predicting the associations between lncRNAs and diseases.

**Keywords:** dual-network integrated logistic matrix factorization; Bayesian optimization; lncRNA and disease associations

## 1. Introduction

Long non-coding RNAs (lncRNAs) are a class of non-coding RNAs (ncRNAs) that are more than 200 nucleotides (nt) in length and do not encode proteins [1]. lncRNAs were originally thought to be genomic transcriptional noise without biological function [2]. Later, more and more evidence indicated that lncRNAs play an important role in many key biological processes, such as translation and post-translational regulation, cell differentiation, proliferation and apoptosis, and epigenetic regulation [3]. Meanwhile, mutations and dysregulation of lncRNAs can cause a variety of human diseases [4,5], including diabetes [6], AIDS [7], and many types of cancer, such as hepatocellular carcinoma [8], lung cancer [9], prostate cancer [10], breast cancer [11], and bladder cancer [12]. Therefore, predicting the potential associations between lncRNAs and diseases helps to explore the complex pathogenesis and etiology of disease at the molecular level and effectively improves the quality of disease diagnosis, treatment, and prevention.

In recent years, several lncRNAs function–disease relationship databases have been established. lncRNAdb [13], lncRNADisease [14], lnc2Cancer [15], and NONCODE [16] are some examples. However, the known lncRNA–disease relationship is still rare, and the use of biological experiments to explore lncRNA–disease associations is both time-consuming and expensive. Using computational

methods to infer the potential associations between lncRNAs and diseases has become an effective prior method for biological experiments.

Recently, many computational models have been proposed to predict potential lncRNA–disease associations, which can roughly be divided into three categories. The first class of methods is based on machine learning to predict potential associations. Chen et al. [17] proposed LRSLDA, a semi-supervised learning method based on Laplacian regular least squares. This method does not require a negative sample. However, the problem of parameter selection for combining two classifiers has not been well solved. LDAP [18] uses a support vector machine classifier to predict potential lncRNA–disease associations based on lncRNA similarity and disease similarity. Yu et al. [19] constructed a global quadruple network and a global tripartite network by integrating various biological information. Based on these two global networks, the novel probability model NBCLDAbased on the naive Bayesian classifier was proposed.

The second category is based on biological network models. Heterogeneous data have become a hot topic in recent years. These models tend to construct heterogeneous networks using disease-associated genes/miRNAs or predict new associations between lncRNAs and diseases using multi-data source information fusion. Liang et al. [20] proposed a new method, TPGLDA, for predicting lncRNA–disease associations using a lncRNA–disease–gene tripartite map. It integrates gene–disease associations and lncRNA–disease associations and can effectively identify potential lncRNA–disease associations. Chen et al. [21] proposed an improved restart random walk model IRWRLDA, which integrates multiple data sources including lncRNA expression similarity, functional similarity, Gaussian interaction profile kernel similarity, and disease semantic similarity to predict lncRNA–disease associations. Gu et al. [22] established a global network random walk model GrwLDA, which predicts potential lncRNA–disease associations by integrating disease semantic similarity, lncRNA functional similarity, and known lncRNA–disease associations. These data fusion-based methods have achieved significant improvements over methods that use a single data source.

The third category is some methods based on matrix completion. MFLDA [23] decomposes data matrices of heterogeneous data sources into low-rank matrices via matrix tri-factorization to explore and exploit their intrinsic and shared structure. However, it cannot predict lncRNAs that are not associated with any disease or diseases that are not associated with any lncRNA. SIMCLDA [24] models the lncRNA–disease associations' prediction problem as a recommended task and uses the induction matrix completion method to solve it.

The lncRNA–disease association matrix and the drug–target association matrix are generally sparse matrices with less known associations. The sparsity of the lncRNA–disease dataset used in this paper is 97.36%, which was obtained from 1-540/(115*178) (115 lncRNAs, 178 diseases, and 540 known associations, sparsity = 1-540/(115*178)), and the sparsity of the four benchmark datasets is 99.01%, 96.55%, 97.00%, and 93.59%, respectively, in drug–target interaction prediction [25]. With regard to the sparse characteristics of the drug–target matrix, neighborhood regularized logistic matrix factorization (NRLMF) was adopted in [26] to predict drug–target interactions, and the effect was significant. NRLMF has also been successfully applied to the prediction of the associations between miRNA–disease [27] and lncRNA–protein [28,29]. Based on NRLMF, dual-network integrated logistic matrix factorization (DNILMF) introduced a drug similarity network and target similarity network to improve the accuracy of prediction [26]. However, the DNILMF prediction effect was greatly affected by the parameter setting. The method of setting parameters based on experience had significant limits in [26]. Because the Gaussian process mutual information algorithm (GP-MI) [30], an advanced Bayesian optimization method, has been successfully applied to the parameter optimization of the logistic matrix factorization model and brings about positive results [31], this paper adopts the GP-MI algorithm to optimize the parameters for DNILMF.

The advantages of using DNILMF-LDA to predict lncRNA–disease associations are: (1) logistic matrix factorization, especially suitable for binary variables and sparsity problems, is used to model the interaction probability of each lncRNA–disease pair; (2) two different similarity kernel matrices of

lncRNAs and diseases are fused into a composite kernel matrix by nonlinear fusion technology, and then, the fused kernel matrices are integrated into the model; (3) the lncRNAs' and diseases' similarity networks are introduced in the model; the flowchart of DNILMF-LDA given in Figure 1.

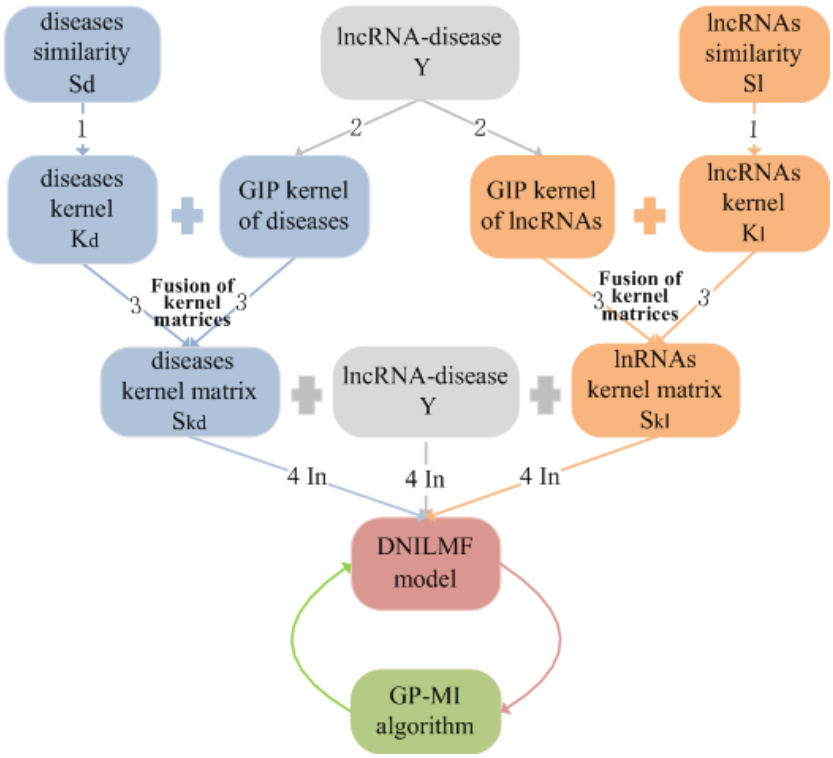


**Figure 1.** The flowchart of dual-network integrated logistic matrix factorization-lncRNA–disease association (DNILMF-LDA). **Step 1**: converting the calculated lncRNAs' similarity matrix and the diseases' similarity matrix to the corresponding kernel matrix; **Step 2**: calculating the Gaussian interaction profile kernel matrix of lncRNAs and diseases, respectively; **Step 3**: fusing two kernel matrices corresponding to the lncRNAs and the diseases respectively into one kernel matrix; **Step 4**: constructing the DNILMF model with the lncRNA–disease associations matrix, lncRNAs, and diseases kernel matrices as the input data. In order to ensure the optimal performance of the algorithm, the Gaussian process mutual information (GP-MI) algorithm is used to select parameters. GIP, Gaussian interaction profile.

## 2. Materials

### 2.1. lncRNA–Disease Associations Matrix

The original lncRNA–disease association dataset was downloaded from the lncRNADisease [14] database, which integrated 687 experimentally-validated lncRNA–disease associations between 246 diseases and 369 lncRNAs. The diseases without disease ontology (http://disease-ontology.org/) and lncRNAs without expression profiles in ArrayExpress [32] (http://www.ebi.ac.uk/arrayexpress/) were filtered out, and 540 experimentally-validated lncRNA–disease associations between 115 lncRNAs and 178 diseases were obtained. The lncRNA–disease association matrix is represented by $Y$.

### 2.2. lncRNA Expression Similarity Matrix and Disease Semantic Similarity Matrix

More than 60,000 expression profiles from 16 human tissues were downloaded from ArrayExpress [32]. The Spearman correlation coefficient between any two lncRNAs in 115 lncRNAs was calculated and was used as the expression similarity for this pair of lncRNAs [17]. The expression similarity matrix of all lncRNAs is represented by $S_l$.

The semantic similarity of diseases is often used to predict potential lncRNA–disease associations. The semantic similarity of the disease in this paper was calculated with the method in paper [33]. Each disease was represented by a directed acyclic graph (DAG) containing all relevant annotated items, which came from the National Library of Medicine (http://www.nlm.nih.gov/mesh). The semantic similarity of two diseases is based on both the addresses of these diseases in DAG graphs and their semantic relations with their ancestor diseases. The DOSE package provided us with the method to calculate semantic similarities among diseases [34]. The semantic similarity matrix of the disease is represented by $S_d$.

### 2.3. Similarity Kernel Matrices

Kernel matrices of lncRNAs and diseases were constructed for nonlinear kernel fusion. The construction of kernel matrices consisted of two steps.

The first step was to convert the lncRNAs' and diseases' similarity matrices into kernel matrices. In this step, $S_l$ and$S_d$ were converted to kernel matrices by:

(1) converting $S_l$ and$S_d$ to the corresponding symmetric matrices, $S_{sym} = \left(S + S^T\right)/2$;

(2) transforming the symmetric matrices obtained in the first step into semi-positive definite matrices by adding multiple small identity matrices [35]. The transformed lncRNAs' and diseases' kernel matrices are represented by $K_l$ and $K_d$, respectively.

The second step was to calculate the Gaussian interaction profile (GIP) kernel matrix of lncRNAs and diseases. $Y_{li}$ and $Y_{lj}$ represent the interaction profile of lncRNA iand lncRNA j, which are the $i$th row and $j$th vector of association matrix $Y$. The distance between these two vectors was computed as their GIP kernel. In this step, for a given lncRNA–disease associations matrix $Y$, the GIP kernel $K^l_{gip}$ between lncRNAs was calculated according to Formula (1) [35]:

$$K_{gip}\left(l_i, l_j\right) = \exp\left(-\frac{\left\|Y_{li} - Y_{lj}\right\|^2}{\sigma}\right) \quad (1)$$

where $\|\cdot\|$ represents the Euclidean distance and $\sigma$ represents the kernel bandwidth of the Gaussian spectrum. In our work, the value of $\sigma$ was set to one. GIP kernel $K^d_{gip}$ between diseases was calculated using the same method.

### 2.4. Fusion of Similarity Kernel Matrices

The purpose of similarity kernel matrices' fusion is to merge $K^l_{gip}$ and $K_l$ into a kernel matrix and merge $K^d_{gip}$ and $K_d$ into another kernel matrix. The steps of kernel fusion [36,37] are:

(1) Normalize and symmetrize the above four kernel matrices. Taking the fusion steps between $K^l_{gip}$ and $K_l$ as an example, the resulting matrices are denoted by $P^{(1)}$ and $P^{(2)}$.

(2) Construct local similarity matrix $L^{(1)}$ and $L^{(2)}$ of $K^l_{gip}$ and $K_l$ by Formula (2):

$$L^{(1)}(i,j) = \begin{cases} \frac{P^{(1)}(i,j)}{\sum_{k \in N_i} P^{(1)}(i,k)}, j \in N_i \\ 0, \text{ others} \end{cases} \quad (2)$$

where $P^{(1)}(i,j)$ represents the $i$th row and $j$th column element in matrix $P^{(1)}$. $N_i$ denotes the nearest neighbors of the current target $i$. The number of nearest neighbors was set to 3 according to experience. The similarity between lncRNAi and non-nearest neighbors was zero. Finally, $L^{(1)}$ and $L^{(2)}$ can be obtained;

(3) Update $P^{(1)}$ and $P^{(2)}$ iteratively by Formulas (3) and (4). Iteration step $t$ was set to two by experience.

$$P_t^{(1)} = L^{(1)} P_{t-1}^{(2)} \left(L^{(1)}\right)^T \quad (3)$$

$$P_t^{(2)} = L^{(2)} P_{t-1}^{(1)} \left(L^{(2)}\right)^T \quad (4)$$

(4) After the iterations, $P_t^{(1)}$ and $P_t^{(2)}$ were averaged and normalized as the final kernel matrix of diseases, denoted as $S_d^k$. $S_l^k$ was calculated using the same method.

## 3. Methods

### 3.1. Problem Formalization

In this paper, the collection of lncRNAs is represented by $L = \{l_i\}_1^m$, and the collection of diseases is represented by $D = \{d_j\}_1^n$, where $m$ and $n$ are the number of lncRNAs and diseases, respectively. The associations between lncRNAs and diseases are represented by a binary matrix $Y \in R^{m \times n}$. When lncRNA $l_i$ was experimentally verified to be associated with disease $d_j$, $y_{ij} = 1$, otherwise, $y_{ij} = 0$. $L^+ = \left\{l_i | \sum_{j=1}^n y_{ij} > 0, \forall 1 \le i \le m\right\}$ is the collection of positive lncRNAs, and $D^+ = \left\{d_i | \sum_{i=1}^m y_{ij} > 0, \forall 1 \le j \le n\right\}$ is the collection of positive diseases. Thus, $L^- = L \setminus L^+$ is the collection of lncRNAs with no known association with all diseases. $D^- = D \setminus D^+$ is the collection of diseases with no known association with all lncRNAs. $S_l^k \in R^{m \times m}$ is the final similarity kernel matrix of lncRNAs, and $S_d^k \in R^{n \times n}$ is the final similarity kernel matrix of diseases. The purpose of this paper is to predict lncRNA–disease interaction probabilities and rank candidate lncRNA–disease pairs based on predicted probabilities. The higher ranked lncRNA–disease pairs are most likely to be correlated.

### 3.2. Prediction of lncRNA–Disease Associations Using the DNILMF Model

The lncRNAs' kernel matrix $S_l^k$, diseases' kernel matrix $S_d^k$, and lncRNA–disease association matrix $Y$ are the input data for the DNILMF model to infer potential lncRNA–disease associations. lncRNAs and diseases were mapped to the r-dimensional shared potential space, where r < min(m, n). Latent vectors $u_i \in R^{1 \times r}$ and $v_j \in R^{1 \times r}$ represent the characteristics of lncRNA $l_i$ and disease $d_j$, respectively. U $\in R^{m \times r}$ and V $\in R^{n \times r}$ are potential vectors for all lncRNAs and diseases. Then, the probabilities $P$ of all lncRNAs and diseases were modeled by the following logistic function:

$$P = \frac{\exp\left(UV^T\right)}{1 + \exp\left(UV^T\right)} \quad (5)$$

What needs to be emphasized is the calculation of $P$ depends on the lncRNA–disease association network $Y$. Based on the hypothesis that similar diseases are always associated with functionally similar lncRNAs, the interaction probability of lncRNA–disease is affected not only by the lncRNA–disease association network $Y$, but also by lncRNAs' similarity network $S_l^k$ and diseases' similarity network $S_d^k$. Hence, $Y$ is combined with $S_l^k$ and $S_d^k$ for matrix factorization. The interaction probabilities of lncRNAs and diseases are:

$$P = \frac{\exp\left(\alpha UV^T + \beta S_l^k UV^T + \gamma UV^T S_d^k\right)}{1 + \exp\left(\alpha UV^T + \beta S_l^k UV^T + \gamma UV^T S_d^k\right)} \quad (6)$$

where $\alpha, \beta, \gamma$ are the corresponding weight of $Y$, $S_l^k$ and $S_d^k$. Their sum is 1, and $\beta = \gamma$.

Since the known lncRNA–disease associations are more important than the unknown lncRNA–disease associations, we set the weight of the known lncRNA–disease pairs to $c$ ($c \ge 1$ )

and that of the unknown lncRNA–disease pairs to 1. By assuming all samples are independent, the probability $p(Y|U,V)$ can be calculated by:

$$p(Y|U,V) = \prod_{i=1}^{m}\prod_{j=1}^{n} P_{ij}^{cY_{ij}} \left(1 - P_{ij}\right)^{1-Y_{ij}} \tag{7}$$

where $P_{ij}$ is the interaction probability between lncRNA $l_i$ and disease $d_j$. Setting the zero-mean spherical Gaussian prior in lncRNAs' and diseases' potential vectors is done as follows:

$$p\left(U|\sigma_l^2\right) = \prod_{i=1}^{m} N\left(u_i|0,\sigma_l^2 I\right), p\left(V|\sigma_d^2\right) = \prod_{j=1}^{n} N\left(v_j|0,\sigma_d^2 I\right) \tag{8}$$

where $\sigma_l^2$ and $\sigma_d^2$ are the parameters that control the variance of the Gaussian distribution and $I$ represents the identity matrix. According to Bayesian inference:

$$p\left(U,V|Y,\sigma_l^2,\sigma_d^2\right) \propto p(Y|U,V)\, p\left(U|\sigma_l^2\right) p\left(V|\sigma_d^2\right) \tag{9}$$

Then, learn the model parameters $U$ and$V$ by maximizing the logarithm of the posterior distribution. The objective function $L$ is:

$$\begin{aligned} L = \max_{U,V} \sum_{i,j} \Big(cY \odot \Big(\alpha UV^T + \beta S_l^k UV^T + \gamma UV^T S_d^k\Big) - (1 + cY - Y) \odot \ln\big[1 + \exp\big(\alpha UV^T \\ + \beta S_l^k UV^T + \gamma UV^T S_d^k\big)\big]\Big) - \frac{\lambda_u}{2}\|U\|_F^2 - \frac{\lambda_v}{2}\|V\|_F^2 \end{aligned} \tag{10}$$

where $\lambda_u = \frac{1}{\sigma_l^2}$, $\lambda_v = \frac{1}{\sigma_d^2}$, $\lambda_u$, and $\lambda_v$ are regularization coefficient of $U$ and $V$, $\|\cdot\|_F^2$ is the Frobenius norm, and $\odot$ is the Hadamard product. Starting from the above objective function, the gradient descent algorithm was used to solve $U$ and $V$, and the gradient variables of $U$ and $V$ are as follows:

$$\frac{\partial L}{\partial U} = c\left(\alpha I + \beta\left(S_l^k\right)^T\right) YV + \gamma\left(cY - Q\right)\left(S_d^k\right)^T V - \left(\alpha I + \beta\left(S_l^k\right)^T\right) QV - \lambda_u U \tag{11}$$

$$\frac{\partial L}{\partial V} = c\left(\alpha I + \gamma\left(S_d^k\right)^T\right) Y^T U + \beta\left(cY^T - Q^T\right) S_l^k U - \left(\alpha I + \gamma\left(S_d^k\right)^T\right) Q^T U - \lambda_v V \tag{12}$$

where $Q = (1 + cY - Y) \odot \frac{1}{\exp\left(-\left(\alpha UV^T + \beta S_l^k UV^T + \gamma UV^T S_d^k\right)\right)+1}$, $Q^T$ is the transposed matrix of $Q$. This work uses the AdaGrad algorithm [38] to accelerate the convergence of $U$ and $V$.

Based on the matrices $U$ and $V$, the interaction probabilities of any unknown lncRNA–disease pairs can be calculated by Formula (6). Due to the uncertainty of lncRNA $l_i \in L^-$ and disease $d_j \in D^-$, their potential vectors $u_i$ and $v_j$ obtained by gradient descent cannot accurately describe their characteristics, so k-nearest neighbor sets $N^+(l_i)$ and $N^+(d_j)$ of $l_i$ and $d_j$ were constructed (k was empirically set to 5). Then, replace potential vector $u_i$ and $v_j$ with the linear combination of the k-nearest neighbors [25,26]. The modified interaction probability is:

$$\hat{p}_{ij} = \frac{\exp\left(\hat{u}_i \hat{v}_j^T\right)}{1 + \exp\left(\hat{u}_i \hat{v}_j^T\right)} \tag{13}$$

where:

$$\hat{u}_i = \begin{cases} u_i,\ l_i \in L^+ \\ \frac{1}{\sum_{u \in N^+(l_i)} S_{iu}^l} \sum_{u \in N^+(l_i)} S_{iu}^l u_u,\ l_i \in L^- \end{cases} \tag{14}$$

$$\hat{v}_j = \begin{cases} v_j,\ d_j \in D^+ \\ \frac{1}{\sum_{v \in N^+(d_j)} S_{jv}^d} \sum_{v \in N^+(d_j)} S_{jv}^d u_v,\ d_j \in D^- \end{cases} \quad (15)$$

$S_{iu}^l$ denotes the similarity between unknown lncRNA $l_i$ and known lncRNA $l_u$, and $u_u$ denotes the latent variable of $l_u$.

The selection of model parameter $r, \alpha, \beta, \gamma, \lambda_u, \lambda_v$ can affect the performance of the model somehow. It is difficult to ensure the best performance of the model by using empirical parameter values. In order to improve the performance of the model, the Bayesian optimization algorithm was adopted to optimize the setting of parameter values in this work.

### *3.3. Bayesian Optimization*

The Gaussian process mutual information algorithm (GP-MI) was used to optimize the setting of the parameter values. The optimization process of GP-MI for the DNILMF model parameters is shown in Figure 2.

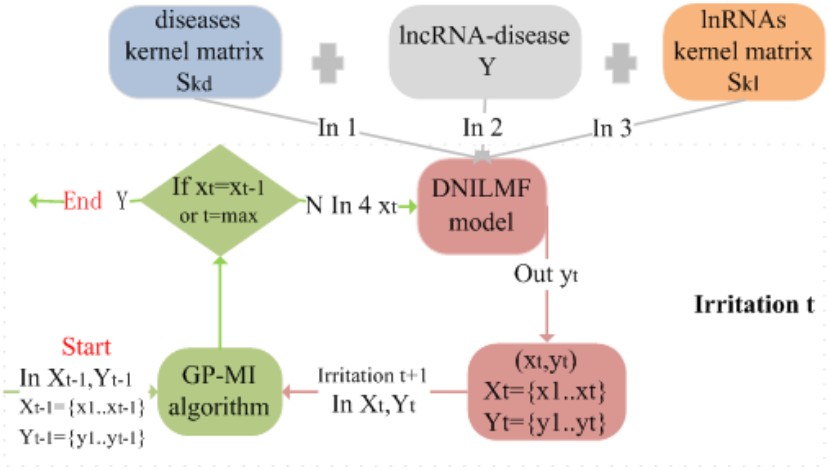


**Figure 2.** The optimization process of GP-MI for the DNILMF model parameters. At irritation t: **Step 1**: get $x_t$ according to the previous query $\chi_{t-1}$ and observations $\mathcal{Y}_{t-1}$; **Step 2**: if $x_t$=$x_{t-1}$ or t is equal to the max value, exit the program; if not, put $x_t$, the disease kernel matrix, lncRNA–disease association matrix, and lncRNA kernel matrix into the DNILMF model, and we can get output $x_t$. Then, take $x_t$ and $y_t$ as the start of the next irritation.

(1) Bayesian optimization

For function $f : \chi \to \mathbb{R}$, $f$ is an unknown function to be optimized, and $\chi \subset \mathbb{R}^n (n \in \mathbb{N})$, a tight convex set. In this paper, the DNRLMF model is $f$, and $\mathbb{R}^n$ is the parameter search space. The purpose of Bayesian optimization is to find the optimal solution for $f$ through continuous queries x $(x_1, x_2, ... \in \chi)$. At iteration t, the new query $x_t$ is selected from $\chi$ according to the previous query $\chi_{t-1} = \{x_1, x_2, ...x_{t-1}\}$ and observations $\mathcal{Y}_{t-1} = \{y_1, y_2, ..., y_{t-1}\}$. The relationship between $y_t$ and $x_t$ is $y_t = f(x_t) + \epsilon_t$, where $\epsilon_t$ is the noise variable, $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$.

(2) Gaussian process

Suppose the function $f$ follows Gaussian process $GP(m, k)$ [30], where $m : \chi \to \mathbb{R}$ is a mean function and $k : \chi \times \chi \to \mathbb{R}$ is a kernel function. Let the mean function be zero, that is $m : \chi \to 0$, the kernel function is a square exponential kernel.

According to the previous $t-1$ times queries $\chi_{t-1}$ and observations $\mathcal{Y}_{t-1}$, the posterior distribution at iteration $t$ is a Gaussian process with expectation as $\mu_t(x)$ and variance as $\sigma_t^2(x)$ by Bayesian inference.

(3) GP-MI algorithm

The most critical aspect of the GP-MI algorithm is the choice of the next query $x_t \in \chi$ using $\mu_t(x)$ and variance $\sigma_t^2(x)$.

$$x_t = \underset{x \in \chi}{argmax} \mu_t(x) + \phi_t(x) \tag{16}$$

where $\phi_t : \chi \to \mathbb{R}$ is the increment function of $\sigma_t^2(x)$:

$$\phi_t(x) = \sqrt{\log \frac{2}{\delta}} \left( \sqrt{\sigma_t^2(x) + \hat{\gamma}_{t-1}} - \sqrt{\hat{\gamma}_{t-1}} \right) \tag{17}$$

$\hat{\gamma}_{t-1} \leftarrow \hat{\gamma}_{t-2} + \sigma_{t-1}^2(x_{t-1})$; $\delta > 0$ is a hyperparameter; and the iteration ending condition is $x_{t+1} = x_t$. The pseudocode of the GP-MI is shown in Algorithm 1:

**Algorithm 1** GP-MI.

$\hat{\gamma}_0 \leftarrow 0$

**for** $t = 1, 2, \ldots$ **do**

Compute $\mu_t$ and $\sigma_t^2$ by $\chi_t = \{x_1..x_{t-1}\}$ and $\mathcal{Y}_t = (y_1..y_{t-1})$ *// Bayesian inference*

$\phi_t(x) \leftarrow \sqrt{\log \frac{2}{\delta}} \left( \sqrt{\sigma_t^2(x) + \hat{\gamma}_{t-1}} - \sqrt{\hat{\gamma}_{t-1}} \right)$ *// Definition of $\phi_t(x)$ for all $x \in \chi$*

$x_t \leftarrow \underset{x \in \chi}{argmax} \mu_t(x) + \phi_t(x)$ *// Selection of the next query location*

$\hat{\gamma}_t \leftarrow \hat{\gamma}_{t-1} + \sigma_t^2(x_t)$ *// Update $\hat{\gamma}_t$*

get $y_t$ by the DNILMF model and $x_t$ *// Query $(x_t, y_t)$*

**end for**

## 4. Experimental Results

### 4.1. Evaluation of Prediction Performance

In this paper, the prediction performance of the detection model was verified by 10-fold cross-validation (CV). AUC and the area under precision-recall (PR) curve (AUPR) were used as the performance evaluation indexes of the model. AUC is an important index to evaluate the classification model. If AUC = 1, the model has perfect performance; if AUC = 0.5, this means random performance. The higher the values of AUC and AUPR, the better the prediction performance.

During the 10-fold CV process, lncRNA–disease pairs (including known pairs and unknown pairs) were randomly divided into ten groups with almost the same data size by setting random seeds. Each time, one of the ten groups was used as the test data, and the values of the test data in the adjacency matrix $Y$ were set to zero. The resulting matrix was the training data $Y_{train}$. In each iteration of 10-fold CV, firstly, calculate the kernel matrix and the GIP kernel matrix of lncRNAs and diseases. Secondly, fuse the kernel matrices of lncRNAs and diseases to get two composite kernel matrices. Then, take the fused kernel matrices and $Y_{train}$ as the model input and update the value of the potential vectors $U$, $V$ through gradient descent until the optimal value of the model is achieved. Finally, the AUC and AUPR values were obtained by using the trained model to predict and evaluate the test data. After ten iterations, the AUC values of 10 test sets were obtained, and their mean value was taken as the AUC value of one time 10-fold CV. Under 10-fold CV, the AUC value of the model reached 0.9202, and the AUPR value reached 0.5610.

### *4.2. Comparison with Other Methods*

To further evaluate the performance of our model, we compared it with LRLSLDA, BiwalkLDA, SIMCLDA, and TPGLDA under 10-fold CV. The prediction result of the five models using the same dataset is shown in Table 1. The result showed that both AUC and AUPR values of DNILMF-LDA were the highest among five models, indicating that the performance of our model was better than the others. Figures 3 and 4 respectively show the receiver operating characteristic (ROC) curve and precision-recall (PR) curve of the five models.

**Table 1.** AUC and area under precision-recall (PR) curve (AUPR) values of the five models.

| Method | AUC | AUPR |
|---|---|---|
| LRLSLDA | 0.5321 | 0.0344 |
| SIMCLDA | 0.7895 | 0.1605 |
| BiwalkLDA | 0.8367 | 0.1909 |
| TPGLDA | 0.8527 | 0.1185 |
| DNILMF-LDA | 0.9202 | 0.5610 |

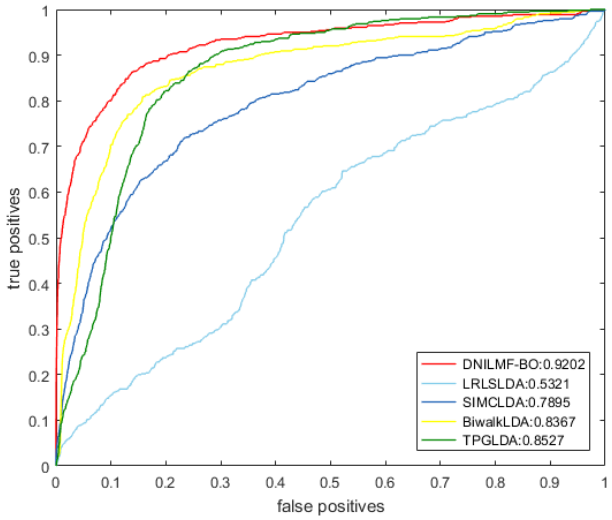


**Figure 3.** ROC curve of the five models.

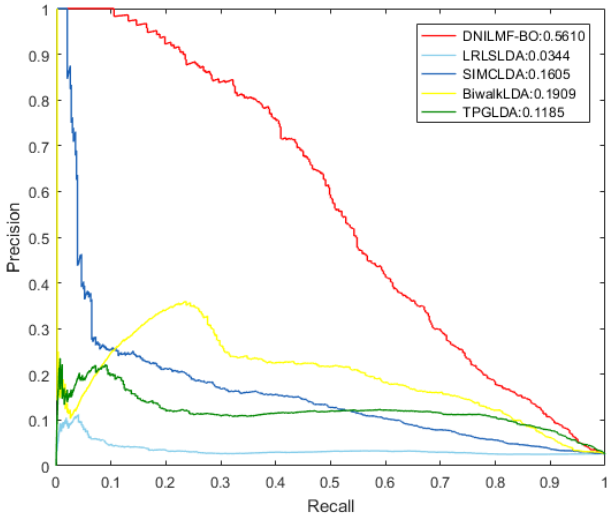


**Figure 4.** PR curve of the five models.

### *4.3. Parameter Analysis*

For DNILMF-LDA, the dimension $r$ of shared potential space was from 50–100 with a step length of 10 [31]. The coefficient of the potential matrix product ranged from 0–1 with a step length of 0.1, $\beta = \gamma = (1 - \alpha)/2$; the regularization coefficients $\lambda_u$ and $\lambda_v$ for potential variables of lncRNAs and diseases ranged from 1–10, with a step size of one [25]. The number of neighbors to construct the neighbor set of unknown lncRNAs and diseases was set to five. The weight of known interaction pair was set to five. According to the results of the literature [31], when $\delta = 10^{-100}$, the Bayesian optimization was very close to the prediction accuracy of the grid search, but the calculation time decreased by 8.94-times on average. Therefore, we set the value of $\delta$ and the noise variance of the Gaussian process kernel function $\sigma^2$ to $10^{-100}$ and 0.1, respectively. In summary, $r = \{50, 100\}$, $\alpha = \{0.1, 1\}$, $\lambda_u = \{1, 10\}$, $\lambda_v = \{1, 10\}$, $K = 5$, $c = 5$, $\delta = 10^{-100}$, and $\sigma^2 = 0.1$.

The parameter optimization results of the DNILMF model by the GP-MI algorithm showed that the prediction performance of the model was good when the model parameters $r$ took any value in the range of $\{50, 100\}$, $\alpha = 0.1$, $\beta = \gamma = 0.45$, $\lambda_u = 1$, $\lambda_v = 1$. When $r = 90$, the AUC value of the model reached its highest at 0.9202. The AUC value is shown in Figure 5 when $r$ took different values. The weight of $\beta$ and $\gamma$ was greater than that of $c$, which indicated the importance of the lncRNAs' and diseases' similarity network and also indicated the effectiveness of adding the lncRNA–disease associations network and similarity networks into the model.

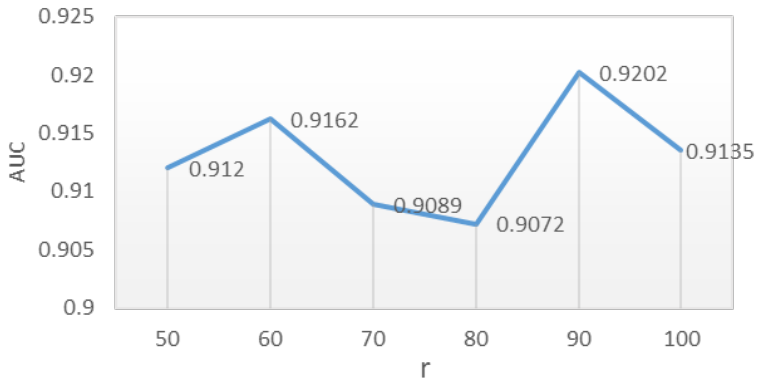


**Figure 5.** Influence of $r$ on AUC value when $\alpha = 0.1$, $\beta = \gamma = 0.45$, $\lambda_u = 1$, $\lambda_v = 1$.

### *4.4. Case Studies on Breast, Lung, and Colon Cancer*

We further evaluated the role of the DNILMF-LDA model in predicting lncRNA–disease associations by studying three common and typical cancers: breast cancer, lung cancer, and colon cancer. The top ten candidate lncRNAs calculated by DNILMF-LDA for three cancers and their evidence are listed in Tables 2–4. The verification of the prediction results was supported by the lncRNADisease and lnc2Cancer databases [14,15].

**Table 2.** The top ten lncRNA candidates for lung cancer.

| Top | lncRNA | Evidence | Description |
|---|---|---|---|
| 1 | CCAT2 | 26729200 | lncRNADisease |
| 2 | CDKN2B-AS1 | 26729200 | lncRNADisease |
| 3 | PVT1 | 28731781 | lnc2Cancer |
| 4 | UCA1 | 29731641 | lnc2Cancer |
| 5 | CCAT1 | 27212446 | lncRNADisease |
| 6 | SPRY4-IT1 | 26302345 | lncRNADisease |
| 7 | GAS5 | 26634743 | lncRNADisease |
| 8 | HULC | unconfirmed | unconfirmed |
| 9 | SRA1 | unconfirmed | unconfirmed |
| 10 | XIST | unconfirmed | unconfirmed |

**Table 3.** The top ten lncRNA candidates for colon cancer.

| Top | lncRNA | Evidence | Description |
|---|---|---|---|
| 1 | SPRY4-IT1 | 28099409 | lnc2Cancer |
| 2 | HOTTIP | 26617875 | lnc2Cancer |
| 3 | GHET1 | 27931286 | lnc2Cancer |
| 4 | MINA | unconfirmed | unconfirmed |
| 5 | HIF1A-AS2 | 29278853 | lnc2Cancer |
| 6 | ADAMTS9-AS2 | 27596298 | lnc2Cancer |
| 7 | TUG1 | 28302487 | lnc2Cancer |
| 8 | LINC00152 | 29180678 | lnc2Cancer |
| 9 | PANDAR | 28176943 | lnc2Cancer |
| 10 | BC040587 | unconfirmed | unconfirmed |

**Table 4.** The top ten lncRNA candidates for breast cancer.

| Top | lncRNA | Evidence | Description |
|---|---|---|---|
| 1 | MNX1-AS1 | unconfirmed | unconfirmed |
| 2 | CCAT1 | 26464701 | lnc2Cancer |
| 3 | TUSC7 | 23558749 | lnc2Cancer |
| 4 | BANCR | 29565494 | lnc2Cancer |
| 5 | DNM3OS | unconfirmed | unconfirmed |
| 6 | TUG1 | 27791993 | lncRNADisease |
| 7 | RPL34-AS1 | unconfirmed | lncRNADisease |
| 8 | MINA | 25586347 | lnc2Cancer |
| 9 | GHET1 | 29843220 | lnc2Cancer |
| 10 | PTENP1 | 29085464 | lnc2Cancer |

Lung cancer is one of the most common and deadly cancers in the world. Among the top 10 candidate lncRNAs calculated by DNILMF-LDA, seven lncRNAs were experimentally verified to be associated with lung cancer. For example, the lncRNA-CDKN2B-AS1 promotes NSCLC cell proliferation and inhibits apoptosis by suppressing KLF2 and P21 expression [39]. In addition, a recent study has shown that upregulated lncRNA-UCA1 plays an important role in the development of lung cancer, and it has great application prospects in clinical diagnosis [40].

Colon cancer is the third most common cancer and the second leading cause of cancer death in men and women [41]. Of the top 10 candidate lncRNAs calculated by DNILMF-LDA, eight lncRNAs were experimentally demonstrated to be associated with colon cancer. Studies have shown that inhibiting the expression of lncRNA-TUG 1 can significantly inhibit the migration ability of colon cancer cells, and the overexpression of TUG 1 may promote the proliferation and migration of colon cancer cells [42].

Breast cancer is the most common cancer in women and the most common cancer in the world. Among the top ten candidate lncRNAs calculated by DNILMF-LDA, seven lncRNAs were experimentally demonstrated to be associated with breast cancer. Studies have shown that upregulated lncRNA-CCAT 1, second in our list of breast cancer, participates in various cellular processes related to cancer occurrence [43].

These case studies reconfirmed the potential of DNILMF-LDA in identifying potential lncRNA–disease associations.

## 5. Discussion

Studies have shown that lncRNAs play an essential role in biological processes and in the diagnosis, prevention, and treatment of complex diseases. It has become an extraordinary method to combine multiple different similarity matrices in the computational model, and using matrix factorization to predict the potential lncRNA–disease associations is also a hot topic. In this paper, the dual-network integrated logistic matrix factorization model was used to predict the potential

lncRNA–disease associations, and the GP-MI algorithm of Bayesian optimization was applied for parameter optimization to ensure the optimal performance of the model.

The main advantages of DNILMF-LDA are: (1) Logistic matrix factorization, especially suitable for binary variables and sparsity problems, was used to model the associations probability of each lncRNA–disease pair. (2) The GIP kernel matrix and similarity matrix of lncRNAs and diseases were obtained, and the nonlinear fusion method was adopted in the process of similarity kernel fusion to reduce the difference between similarity matrices. (3) lncRNAs' and diseases' similarity networks were introduced in the model. In this paper, 10-fold CV was used to evaluate the prediction performance of our model. The results showed that compared with the LRLSLDA, BiwalkLDA, SIMCLDA, and TPGLDA models, the AUC value of DNILMF-LDA was higher and the prediction performance of DNILMF-LDA better. In addition, case studies of lung cancer, colon cancer, and breast cancer also suggested that DNILMF-LDA was a better computational method to predict the potential lncRNA–disease associations.

Although DNILMF-LDA has obtained reliable experimental results, there are still some biases. For example, the known experimentally-verified lncRNA–disease associations are still limited, and the predictive performance of DNILMF-LDA will be improved by a more comprehensive dataset.

## 6. Conclusions

In this paper, our major contributions were as follows: First, logistic matrix factorization was used to model the interaction probability of each lncRNA–disease pair. Second, lncRNA and disease similarity networks were introduced into the model. Third, the imbalance between known and unknown interaction pairs was balanced by giving higher weights to known interactions in the model. Fourth, the method of neighborhood information was used to deal with the problems of new lncRNAs and diseases in the process of prediction. Fifth, multiple source similarity fusion was used to improve the prediction accuracy. We obtained the Gaussian kernel matrix and similarity kernel matrix of lncRNAs and diseases, adopted nonlinear fusion to weaken the differences between similar matrices, and extracted the most important information from different similarity data. Sixth, the GP-MI algorithm in Bayesian optimization was adopted in this paper for parameter optimization.

In the future, we expect to acquire new multi-source datasets and explore better kernel fusion methods. Then, we can improve the prediction performance by fully exploiting multi-source data and advanced fusion technology.

**Author Contributions:** Conceptualization, Y.L.; methodology, Y.L.; software, Y.L.; writing, original draft preparation, Y.L., J.L., and N.B.; writing, review and editing, Y.L., J.L., and N.B.; supervision, N.B.

**Funding:** This research received no external funding.

**Acknowledgments:** The authors thank Xiaofang Xiao for assistance with the experiments.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Chang, H.Y. Abstract IA02: Genome regulation by long noncoding RNAs. *Cancer Res.* **2016**, *76*, IA02.
2. Yanofsky, C. Establishing the triplet nature of the genetic code. *Cell* **2007**, *128*, 815–818. [CrossRef] [PubMed]
3. Merry, C.R.; Niland, C.; Khalil, A.M. *Diverse Functions and Mechanisms of Mammalian Long Noncoding RNAs*; Springer: New York, NY, USA, 2015; pp. 1–14.
4. Cheetham, S.; Gruhl, F.; Mattick, J.; Dinger, M. Long noncoding RNAs and the genetics of cancer. *Br. J. Cancer* **2013**, *108*, 2419. [CrossRef] [PubMed]
5. Taft, R.J.; Pang, K.C.; Mercer, T.R.; Dinger, M.; Mattick, J.S. Non-coding RNAs: Regulators of disease. *J. Pathol.* **2010**, *220*, 126–139. [CrossRef] [PubMed]
6. Pasmant, E.; Sabbagh, A.; Vidaud, M.; Bièche, I. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J.* **2011**, *25*, 444–448. [CrossRef]

7. Zhang, Q.; Chen, C.Y.; Yedavalli, V.S.; Jeang, K.T. NEAT1 long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression. *MBio* **2013**, *4*, e00596-12. [CrossRef] [PubMed]
8. Wang, J.; Liu, X.; Wu, H.; Ni, P.; Gu, Z.; Qiao, Y.; Chen, N.; Sun, F.; Fan, Q. CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic Acids Res.* **2010**, *38*, 5366–5383. [CrossRef]
9. Wapinski, O.; Chang, H.Y. Long noncoding RNAs and human disease. *Trends Cell Biol.* **2011**, *21*, 354–361. [CrossRef]
10. Cui, Z.; Ren, S.; Lu, J.; Wang, F.; Xu, W.; Sun, Y.; Wei, M.; Chen, J.; Gao, X.; Xu, C.; et al. The prostate cancer-up-regulated long noncoding RNA PlncRNA-1 modulates apoptosis and proliferation through reciprocal regulation of androgen receptor. *Urol. Oncol. Semin. Orig. Investig.* **2013**, *31*, 1117–1123. [CrossRef]
11. Sun, J.; Chen, X.; Wang, Z.; Guo, M.; Shi, H.; Wang, X.; Cheng, L.; Zhou, M. A potential prognostic long non-coding RNA signature to predict metastasis-free survival of breast cancer patients. *Sci. Rep.* **2015**, *5*, 16553. [CrossRef]
12. Ma, Z.; Xue, S.; Zeng, B.; Qiu, D. lncRNA SNHG5 is associated with poor prognosis of bladder cancer and promotes bladder cancer cell proliferation through targeting p27. *Trends Cell Biol.* **2018**, *15*, 1924–1930. [CrossRef] [PubMed]
13. Quek, X.C.; Thomson, D.W.; Maag, J.L.; Bartonicek, N.; Signal, B.; Clark, M.B.; Gloss, B.S.;Dinger, M.E. lncRNAdb v2. 0: Expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.* **2014**, *21*, D168–D173.
14. Chen, G.; Wang, Z.; Wang, D.; Qiu, C.; Liu, M.; Chen, X.; Zhang, Q.; Yan, G.; Cui, Q. lncRNADisease: A database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* **2012**, *41*, D983–D986. [CrossRef] [PubMed]
15. Gao, Y.; Wang, P.; Wang, Y.; Ma, X.; Zhi, H.; Zhou, D.; Li, X.; Fang, Y.; Shen, W.; Xu, Y.; et al. Lnc2Cancer v2.0: Updated database of experimentally supported long non-coding RNAs in human cancers. *Nucleic Acids Res.* **2018**, *47*, D1028–D1033. [CrossRef] [PubMed]
16. Zhao, Y.; Li, H.; Fang, S.; Kang, Y.; Wu, W.; Hao, Y.; Li, Z.; Bu, D.; Sun, N.; Zhang, M.Q.; et al. NONCODE 2016: An informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.* **2015**, *44*, D203–D208. [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. Lan, W.; Li, M.; Zhao, K.; Liu, J.; Wu, F.X.; Pan, Y.; Wang, J. LDAP: A web server for lncRNA-disease association prediction. *Bioinformatics* **2016**, *33*, 458–460. [CrossRef]
19. 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*, 345. [CrossRef]
20. Ding, L.; Wang, M.; Sun, D.; Li, A. TPGLDA: Novel prediction of associations between lncRNAs and diseases via lncRNA-disease-gene tripartite graph. *Sci. Rep.* **2018**, *8*, 1065. [CrossRef]
21. Chen, X.; You, Z.H.; Yan, G.Y.; Gong, D.W. IRWRLDA: Improved random walk with restart for lncRNA-disease association prediction. *Oncotarget* **2016**, *7*, 57919. [CrossRef]
22. Gu, C.; Liao, B.; Li, X.; Cai, L.; Li, Z.; Li, K.; Yang, J. Global network random walk for predicting potential human lncRNA-disease associations. *Sci. Rep.* **2017**, *7*, 12442. [CrossRef]
23. Fu, G.; Wang, J.; Domeniconi, C.; Yu, G. Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. *Bioinformatics* **2017**, *34*, 1529–1537. [CrossRef]
24. Lu, C.; Yang, M.; Luo, F.; Wu, F.X.; Li, M.; Pan, Y.; Li, Y.; Wang, J. Prediction of lncRNA-disease associations based on inductive matrix completion. *Bioinformatics* **2018**, *34*, 3357–3364. [CrossRef]
25. Hao, M.; Bryant, S.H.; Wang, Y. Predicting drug–target interactions by dual-network integrated logistic matrix factorization. *Sci. Rep.* **2017**, *7*, 40376. [CrossRef]
26. Liu, Y.; Wu, M.; Miao, C.; Zhao, P.; Li, X.L. Neighborhood regularized logistic matrix factorization for drug–target interaction prediction. *PLoS Comput. Biol.* **2016**, *12*, e1004760. [CrossRef]
27. Yan, C.; Wang, J.; Ni, P.; Lan, W.; Wu, F.; Pan, Y. DNRLMF-MDA: Predicting microRNA-disease associations based on similarities of microRNAs and diseases. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**, *16*, 233–243. [CrossRef]

28. Zhao, Q.; Zhang, Y.; Hu, H.; Ren, G.; Zhang, W.; Liu, H. IRWNRLPI: Integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interaction prediction. *Front. Genet.* **2018**, *9*, 239. [CrossRef]
29. Liu, H.; Ren, G.; Hu, H.; Zhang, L.; Ai, H.; Zhang, W.; Zhao, Q. LPI-NRLMF: lncRNA-protein interaction prediction by neighborhood regularized logistic matrix factorization. *Oncotarget* **2017**, *8*, 103975. [CrossRef]
30. Contal, E.; Perchet, V.; Vayatis, N. Gaussian process optimization with mutual information. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 253–261.
31. Ban, T.; Ohue, M.; Akiyama, Y. Efficient hyperparameter optimization by using Bayesian optimization for drug–target interaction prediction. In Proceedings of 2017 IEEE 7th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), Orlando, FL, USA, 19–21 October 2017; pp. 1–6.
32. Parkinson, H.; Kapushesky, M.; Shojatalab, M.; Abeygunawardena, N.; Coulson, R.; Farne, A.; Holloway, E.; Kolesnykov, N.; Lilja, P.; Lukk, M.; et al. ArrayExpress—A public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* **2006**, *35*, D747–D750. [CrossRef]
33. 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]
34. Yu, G.; Wang, L.G.; Yan, G.R.; He, Q.Y. DOSE: An R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* **2014**, *31*, 608–609. [CrossRef]
35. van Laarhoven, T.; Nabuurs, S.B.; Marchiori, E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* **2011**, *27*, 3036–3043. [CrossRef]
36. Hao, M.; Wang, Y.; Bryant, S.H. Improved prediction of drug–target interactions using regularized least squares integrating with kernel fusion technique. *Anal. Chim. Acta* **2016**, *909*, 41–50. [CrossRef]
37. Wang, B.; Mezlini, A.M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **2014**, *11*, 333. [CrossRef]
38. Duchi, J.; Hazan, E.; Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **2011**, *12*, 2121–2159.
39. Nie, F.Q.; Sun, M.; Yang, J.S.; Xie, M.; Xu, T.P.; Xia, R.; Liu, Y.W.; Liu, X.H.; Zhang, E.B.; Lu, K.H.; et al. Long noncoding RNA ANRIL promotes non-small cell lung cancer cell proliferation and inhibits apoptosis by silencing KLF2 and P21 expression. *Mol. Cancer Ther.* **2015**, *14*, 268–277. [CrossRef]
40. Wang, H.M.; Lu, J.H.; Chen, W.Y.; Gu, A.Q. Upregulated lncRNA-UCA1 contributes to progression of lung cancer and is closely related to clinical diagnosis as a predictive biomarker in plasma. *Int. J. Clin. Exp. Med.* **2015**, *8*, 11824.
41. Prenner, S.; Levitsky, J. Comprehensive review on colorectal cancer and transplant. *Am. J. Transplant.* **2017**, *17*, 2761–2774. [CrossRef]
42. Zhai, H.Y.; Sui, M.H.; Yu, X.; Qu, Z.; Hu, J.C.; Sun, H.Q.; Zheng, H.T.; Zhou, K.; Jiang, L.X. Overexpression of long non-coding RNA TUG1 promotes colon cancer progression. *Med. Sci. Monit.* **2016**, *22*, 3281. [CrossRef]
43. Zhang, X.F.; Liu, T.; Li, Y.; Li, S. Overexpression of long non-coding RNA CCAT1 is a novel biomarker of poor prognosis in patients with breast cancer. *Int. J. Clin. Exp. Pathol.* **2015**, *8*, 9440.