frontiers in Genetics

ORIGINAL RESEARCH
published: 04 February 2020
doi: 10.3389/fgene.2020.00005

Pathogenic Gene Prediction Algorithm Based on Heterogeneous Information Fusion

Chunyu Wang[1], Jie Zhang[1], Xueping Wang[1], Ke Han[2] and Maozu Guo[3,4]**

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [2] School of Computer and Information Engineering, Harbin University of Commerce, Harbin, China, [3] School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China, [4] Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing University of Civil Engineering and Architecture, Beijing, China

OPEN ACCESS

Edited by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS), China

Reviewed by:
Xiucai Ye,
University of Tsukuba, Japan
Jialiang Yang,
Geneis (Beijing) Co. Ltd, China

***Correspondence:**
Chunyu Wang
chunyu@hit.edu.cn
Maozu Guo
guomaozu@bucea.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *26 November 2019*
Accepted: *06 January 2020*
Published: *04 February 2020*

Citation:
Wang C, Zhang J, Wang X, Han K and Guo M (2020) Pathogenic Gene Prediction Algorithm Based on Heterogeneous Information Fusion. Front. Genet. 11:5. doi: 10.3389/fgene.2020.00005

Complex diseases seriously affect people's physical and mental health. The discovery of disease-causing genes has become a target of research. With the emergence of bioinformatics and the rapid development of biotechnology, to overcome the inherent difficulties of the long experimental period and high cost of traditional biomedical methods, researchers have proposed many gene prioritization algorithms that use a large amount of biological data to mine pathogenic genes. However, because the currently known gene–disease association matrix is still very sparse and lacks evidence that genes and diseases are unrelated, there are limits to the predictive performance of gene prioritization algorithms. Based on the hypothesis that functionally related gene mutations may lead to similar disease phenotypes, this paper proposes a PU induction matrix completion algorithm based on heterogeneous information fusion (PUIMCHIF) to predict candidate genes involved in the pathogenicity of human diseases. On the one hand, PUIMCHIF uses different compact feature learning methods to extract features of genes and diseases from multiple data sources, making up for the lack of sparse data. On the other hand, based on the prior knowledge that most of the unknown gene–disease associations are unrelated, we use the PU-Learning strategy to treat the unknown unlabeled data as negative examples for biased learning. The experimental results of the PUIMCHIF algorithm regarding the three indexes of precision, recall, and mean percentile ranking (MPR) were significantly better than those of other algorithms. In the top 100 global prediction analysis of multiple genes and multiple diseases, the probability of recovering true gene associations using PUIMCHIF reached 50% and the MPR value was 10.94%. The PUIMCHIF algorithm has higher priority than those from other methods, such as IMC and CATAPULT.

Keywords: pathogenic gene prediction, induction matrix completion, compact feature learning, PU-Learning, mean percentile ranking

INTRODUCTION

The discovery of disease-causing genes plays an important role in understanding the causes of diseases, clinically diagnosing diseases, and achieving early prevention and treatment (Cheng et al., 2016; Zeng et al., 2017; Cheng et al., 2019). It is also an important goal of human genome research, with great scientific and social significance. Prioritization of potentially pathogenic genes is an important step in the discovery of disease-causing genes and obtaining an understanding of genetic diseases.

Early studies of gene–disease associations were based on clinical and biological experiments, which are expensive and time-consuming. Owing to the inherent difficulties and delays in the study of human genetic diseases, there are very few known identified gene–disease links in public databases, such as the widely used Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2015) and Genetic Association Database (Becker et al., 2004). Because of the specificity of the study of disease-causing genes, we do not know the genes that are not related to a particular disease. We only know the few genes that have been proven to be related to it. Against this background, with the emergence of bioinformatics, researchers have begun to focus on and study genetic disease prioritization algorithms, and use computer technology to mine candidate pathogenic genes from massive data (Liu et al., 2020; Wang et al., 2018; Zeng et al., 2018; Zhang et al., 2019; Zeng et al., 2019; Pan et al., 2019). The selected genes are more likely to be related to diseases, and gene sorting algorithms with better predictive performance would be more helpful to conduct targeted biological experiments and understand pathogenic mechanisms.

Early gene sorting algorithms based on network similarity focused on local information in the gene–disease network, namely, nodes adjacent to gene or disease nodes; an example of these is the molecular triangulation method (Krauthammer et al., 2004). It has been found that the global topology of a network can improve the performance in predicting disease-causing genes (Pan et al., 2019; Chen et al., 2019). Kohler et al. (Kohler et al., 2008) used the random walk (RWR) algorithm to analyze candidate disease-causing genes, which further improved the predictive performance.

Complex biological systems cannot always meet the needs of analysis with single network data (Chen et al., 2019). The continuous growth of biological data, such as high-throughput sequencing, also brings opportunities to study new predictive methods. The more commonly used databases include the gene expression database GEO (Barrett et al., 2007), the cancer gene information TCGA database (Cancer Genome Atlas Research et al., 2013), the protein interaction network database STRING (Szklarczyk et al., 2017), the Gene Ontology (GO) database (Ashburner et al., 2000), and Disease Ontology (DO) (Schriml et al., 2012). Recently, there has been increasing interest in studying gene sorting algorithms and starting to integrate a large amount of biological data and analyze heterogeneous networks (Gomez-Cabrero et al., 2014; Jiang, 2015; Zhang et al., 2019; Deng et al., 2019). In 2008, the CIPHER algorithm (Wu et al., 2008) was proposed by Wu et al., which combines protein interaction and disease-like networks but only considers local information in the network and lacks global topology. In 2010, Vanunu et al. (Vanunu et al., 2010) proposed the PRINCE algorithm, based on the idea of global network information and network dissemination. In the same year, Yongjin Li et al. (Li and Patra, 2010) proposed the restarted random walk algorithm (RWRH) that fused a gene similarity network, a disease phenotypic similarity network, and a large heterogeneous network composed of a disease phenotype–gene relationship network. In addition, Singh-Blom et al. (Singh-Blom et al., 2013) further improved the predictive performance in 2013 using the Katz method commonly used in the field of social networks for the task of predicting gene–disease relationships.

With the rapid development of machine learning and artificial intelligence in recent years, new algorithms based on machine learning have been applied to predict candidate pathogenic genes; they have shown good predictive performance (Zou et al., 2018; Peng et al., 2018; Liao et al., 2018; Zhang et al., 2018; Xiong et al., 2018; He et al., 2018; Cheng et al., 2018; Cheng et al., 2018; Zeng et al., 2019; Ding et al., 2019; Liu, 2019; Liu et al., 2019a; Zhu et al., 2019). In 2011, Mordelet et al. (Mordelet and Vert, 2011) considered the problem of genetic prediction as a supervised machine learning problem and proposed the ProDiGe method. Moreover, in 2013, Singh-Blom et al. (Kohler et al., 2008) proposed the supervised machine-learning method CATAPULT using a variety of data sources. Then, Natarajan et al. (Natarajan and Dhillon, 2014) applied the inductive matrix completion algorithm (IMC) in the recommendation system to predict pathogenic genes. This algorithm can not only predict existing genes and diseases but also predict new genes and diseases that have not previously been shown to be related. To compensate for the impact of a data sparseness and the PU problem, the PUIMCHIF algorithm is proposed in this paper. Specifically, on the basis of the original IMC algorithm, the main innovations and contributions of this paper can be summarized as follows: (1) owing to the sparsity of gene–disease association data, we used a variety of data sources to construct the characteristics of genes and diseases, and added a STRING data set for the compact feature learning of genes, which contained the physical relationships and other interactions that were not in the original data set. (2) For the gene–gene network and the disease–disease network (Li et al., 2019), we used the RWR method to obtain the diffusion state of each node in the network under a steady state in accordance with the network topology, used diffusion component analysis (DCA) to reduce the dimensions of the data, and finally obtained the network characteristics of genes or diseases. One advantage of this approach is the ability to analyze both HumanNet and STRING networks. (3) Self-encoders in machine learning can learn efficient representations of data for dimensionality reduction. Combined with the characteristics of biological data, the work described in this paper used denoising self-encoding to reduce the dimensionality of high-dimensional data features of genes and diseases. (4) Considering the sparse disease–gene association data and the prior knowledge that most unknown associations are negative cases, we adopted the PU-Learning strategy to treat unlabeled data as negative cases for biased learning, so as to replace the IMC method involving learning for only positive cases. (5) To verify the effectiveness of the PUIMCHIF method proposed in this paper, we

used two commonly used evaluation indexes, Precision and Recall. On this basis, we added the MPR index of mean percentile ranking to further analyze the experimental results comprehensively.

INTRODUCTION TO METHODS

We are interest in kinds of associations between the genes and diseases, but only part of them are known. So we want to make a prediction about the unknown pare from the known ones. As shown in **Figure 1**, our goal was to predict these unknown associations based on the constructed low-dimensional characteristics of the genes and diseases, and some known items in the gene–disease association matrix *P*, that is, to predict candidate genes potentially involved in the pathogenicity of the disease.

First, we constructed a low-dimensional eigenvector of genes and diseases from different biological sources (compact feature learning). We proposed different methods for learning compact features based on different forms of data. For the network data of genes and diseases, the random walk with restart algorithm (RWR) was first used to extract the diffusion state of each node in the network, and then DCA was used for dimensionality reduction to obtain the similarity of each gene (or disease) node in the heterogeneous network encoded by low-dimensional feature vectors. This is because genes (or diseases) with similar topological properties in the network are more likely to be functionally related.

Second, for common feature matrix data, to reduce the influence of high noise and data loss of biological data, we used denoising autoencoder (DAE) to reduce the dimensions of features.

Next, we applied the partial inductive matrix completion algorithm to predict the relationship between genes and diseases by combining the characteristics of multiple diseases and genes. One of the main advantages of this method is that it is generalized and can be applied to diseases that are not present during training, which cannot be predicted by traditional matrix completion methods. This allows us to take advantage of previous knowledge of known gene–disease interactions to predict unknown gene–disease interactions. Because we added an unbiased learning scheme for the unknown association relationship as a negative example, we finally adopted the PUIMC method for disease-causing gene prediction. The details of the PUIMCHIF algorithm are described below.

Compact Feature Learning

In machine learning, the data are more important than the algorithm because the generalization of machine learning algorithm is about the ability from known data to the unknown data. Therefore, when we choose the prediction method based on machine learning to predict the disease-causing gene. First, we need to use high-quality data. Second, we need to conduct feature processing on the data to obtain more favorable data features for the prediction task.

We integrated a variety of biological data to extract characteristics of genes or diseases. Moreover, our goal was to obtain a low-dimensional effective data feature matrix, where one row of the feature matrix refers to a gene or disease, and the columns of the matrix represent different characteristics. The different compact feature learning methods that we used are described below.

RWR

Closely linked or functionally similar genes are more likely to cause the same or similar diseases. Random walk provides an effective framework for exploring relationships in networks.

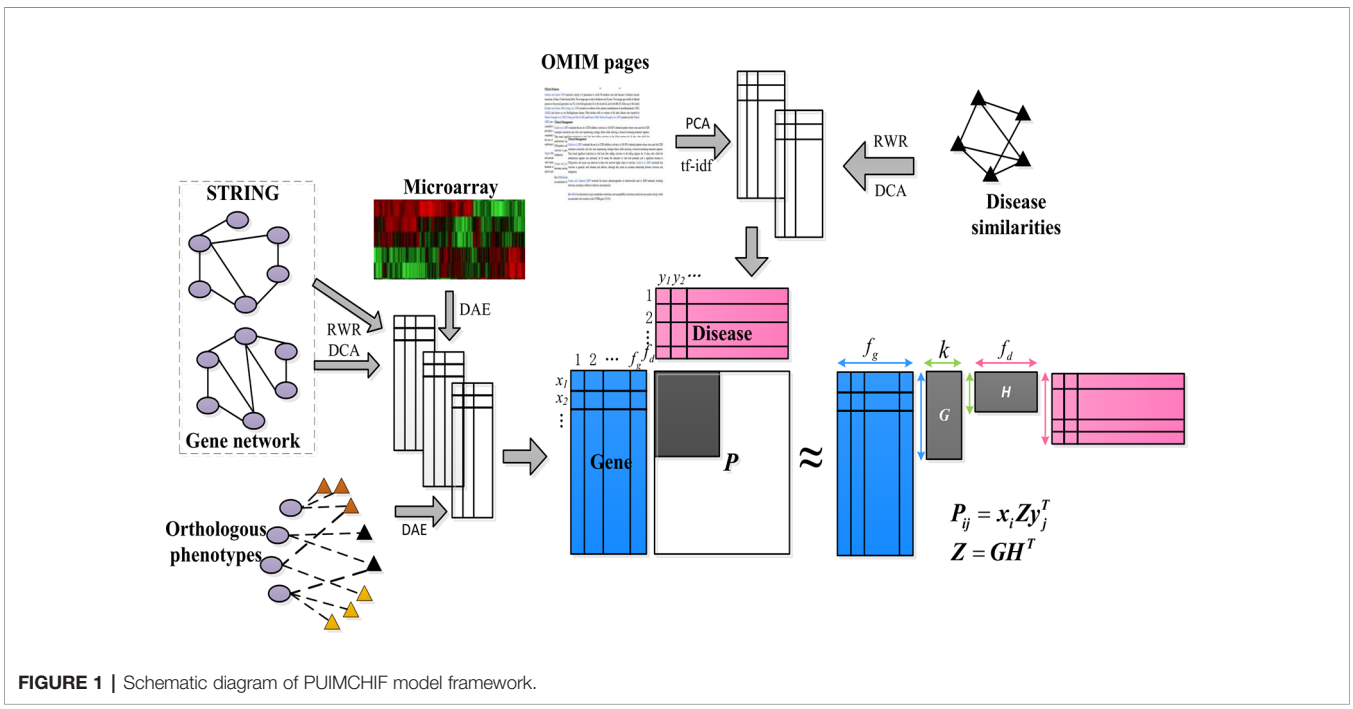

FIGURE 1 | Schematic diagram of PUIMCHIF model framework.

Random walk with restart is referred to as RWR, which is a network diffusion algorithm widely used in the analysis of complex biological network data (Navlakha and Kingsford, 2010; Cao et al., 2014). Different from the traditional random walk method, each iteration of RWR introduces a predefined restart probability at the initial node, which can consider both local and global topological connection patterns within the network and take full advantage of direct or indirect relationships between nodes.

Here, matrix A and B are defined. Matrix A represents the weighted adjacency matrix of the interaction network of genes (or diseases). And in matrix B as shown in equation (1), each element B_{ij} describes the probability of transition from node i to node j. s_i^t represents an n-dimensional distribution vector, and each element stores the probability that a node is accessed after iterating t times from node i during the random walk. The formula for calculating RWR is shown in equation (2).

$$B_{ij} = \frac{A_{ij}}{\sum_{j'} A_{ij'}} \tag{1}$$

$$s_i^{t+1} = (1 - p_r)s_i^t B + p_r \delta_i \tag{2}$$

In equation (2), δ_i represents an n-dimensional standard basis vector and $\delta_i(i) = 1$, $\delta_i(j) = 0$, for $\forall j \neq i$. And p_r is a predefined restart probability that controls the relative influence of local structure and global structure in the diffusion process. With a higher value, more attention is paid to the local structure in the network.

For a node in the iterative process, we can obtain a stable distribution s_i^∞, so we define s_i as the "diffusion state" of node i, that is $s_i = s_i^\infty$. The jth element s_{ij} of s_i represents the probability that the RWR starts from node i and ends at node j in equilibrium. When two nodes have similar diffusion states, it generally means that they are more similar than other nodes in the network and may have similar functions. This discovery provides a basis for predicting unknown gene–disease associations.

Diffusion Component Analysis

Although the diffuse states generated by the above RWR process represent the underlying topological environment and intrinsic connectivity spectrum of each gene or disease node in the network, they may not be completely accurate due to the low-quality and high-dimensional nature of biological data. For example, a small number of missing or false interactions in the network can significantly affect the outcome of the diffusion process (Kim and Leskovec, 2011). It is often inconvenient to directly use high-dimensional diffusion states as topological features in prediction tasks.

To solve this problem, we used a dimensionality reduction method called DCA to reduce the dimensions of the feature space and obtain important topological features from the diffusion state. In addition, for multi-omics networks, DCA also performs very well. The key idea of DCA is to obtain an informative but low-dimensional vector representation. Similar to principal component analysis (PCA), which seeks the inherent low-dimensional linear structure of data to best interpret variances, DCA learns the low-dimensional vector representation of all nodes to best interpret their patterns of connection in heterogeneous networks. We will briefly describe the DCA framework below.

To achieve the purposes of noise reduction and dimensionality reduction, DCA uses the polynomial logic model represented by a low-dimensional vector to approximate the obtained diffusion state distribution, and it has far fewer dimensions than the original n-dimensional vector representing the diffusion state. Specifically, the probability of assigning node i to node j in the diffusion state is modeled as:

$$\hat{s}_{ij} = \frac{\exp\{x_i^T w_j\}}{\sum_{j'} \exp\{x_i^T w_{j'}\}} \tag{3}$$

In equation (3), x_i, $w_i \in \mathbb{R}^d$, $d \ll n$. We take w_i as the context feature and x_i as the node feature of node i, both of which describe the topological properties of the network. If x_i and w_i point in similar directions, we obtain a larger inner product. This means that node j may be frequently visited in a random walk starting from node i. DCA uses the obtained diffusion state $S=\{s_1,\cdots, s_n\}$ as input to optimize w and x of all nodes. The optimization method uses KL divergence, as shown in equation (4).

$$\min_{w,\,x} C(s,\ \hat{s}) = \min_{w,x} \frac{1}{n} \sum_{i=1}^{n} D_{KL}(s_i || \hat{s}_i) \tag{4}$$

$D_{KL}(\cdot||\cdot)$ is the KL divergence between the two distributions. We use w and x to represent this formula according to the definition of KL invergence and $\hat{s}$.

$$C(s, \hat{s}) = \frac{1}{n} \sum_{i=1}^{n} \left[H(s_i) - \sum_{j=1}^{n} s_{ij} \left(w_i^T x_j - \log\left(\sum_{j'=1}^{n} \exp\{w_i^T, x_{j'}\} \right) \right) \right] \tag{5}$$

In equation (5), $H(\cdot)$ represents entropy. The objective function can find the low-dimensional vector representation of w and x using the standard quasi-Newton L-BFGS method. Although the obtained low-dimensional vector can effectively capture the network structure, we found that this optimization method is time-consuming.

To make the DCA framework more suitable for large biological networks, we use a more efficient method, clusDCA (Wang et al., 2015), which is based on matrix factorization, to decompose the diffusion states and obtain their low-dimensional vector representations. According to the definition, the following formula can be obtained:

$$\log \hat{s}_{ij} = x_i^T w_j - \log \sum_{j'} \exp\{x_i^T w_{j'}\} \tag{6}$$

The first term corresponds to the low-dimensional approximation of $\hat{s}_{ij}$. The second term is a normalization factor, ensuring that $\hat{s}_i$ is a well-defined distribution. By removing the second term, we relax the constraint that the elements in $\hat{s}_{ij}$ must add up to 1. Although the obtained low-dimensional

approximation of the diffusion state is no longer a strictly valid probability distribution, it is found that these approximations are very close to the true distribution, and the effects of relaxation are negligible. Therefore, it can be simplified as:

$$\log \hat{s}_{ij} = x_i^T w_j . \tag{7}$$

In addition, we use the sum of squared errors as the objective function, instead of minimizing the relative entropy between the original diffusion state and the approximate diffusion state.

$$\min_{w,\ x} C(s,\ \hat{s}) = \min_{w,x} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(w_i^T x_j - \log s_{ij} \right)^2 \tag{8}$$

The obtained objective function can be optimized by singular value decomposition (SVD). To avoid taking the logarithm of 0, we add a small positive number $\frac{1}{n}$ to s_{ij}. The calculation formula of the logarithm diffusion state matrix L is as follows:

$$L = \log(S + Q) - \log(Q) . \tag{9}$$

In equation (9), $S \in \mathbb{R}^{n \times n}$, $Q \in \mathbb{R}^{n \times n}$ and $Q_{ij} = \frac{1}{n}$, for $\forall i, j$. Using the singular value decomposition method, we decompose L into three matrices:

$$L = U \Sigma V^T \tag{10}$$

In equation (10), $U \in \mathbb{R}^{n \times n}$, $V \in \mathbb{R}^{n \times n}$, $\Sigma \in \mathbb{R}^{n \times n}$ and Σ is a diagonal singular value matrix. To obtain the low-dimensional vectors w_j and x_i in d dimensions, we simply select the first d singular vectors U_d, V_d, and Σ_d. Each row of matrix $X = [x_1, \ldots, x_n]^T$ represents the low-dimensional eigenvector corresponding to each node in the network. In matrix $W = [w_1, \ldots, w_n]^T$, each row represents the corresponding vector of the context feature. The formulas for calculating X and W are as follows:

$$X = U_d \Sigma_d^{1/2}, \quad W = V_d \Sigma_d^{1/2} . \tag{11}$$

Denoising Autoencoder

Autoencoder is an unsupervised neural network model. It learns the implicit features of input data, which is called "coding." At the same time, the original input data can be reconstructed with the learned new features, which is called "decoding." Intuitively, autoencoder can be used for reducing feature dimensionality, like principal components analysis (PCA), but with stronger performance than PCA because the neural network model can extract more effective new features.

The denoising autoencoder adds noise to the input x to obtain $\tilde{x}$, and after training, it obtains a noiseless output z, as shown in **Figure 2**.

This prevents the autoencoder from simply copying the input to the output, so as to extract useful patterns in the data and improve the weight robustness. Noise can be either pure gaussian noise added to the input or randomly discarding a feature at input layer, similar to dropout. The specific equation for calculating z is as follows:

$$\begin{aligned} y &= f(\tilde{x} W_1 + b_1) \\ z &= g(y W_2 + b_2) \end{aligned} \tag{12}$$

In addition, network parameters are trained to minimize reconstruction errors, namely:

$$\min L_H(x, z) = \min ||x - z_p . \tag{13}$$

Pathogenic Gene Prediction Method

Standard Inductive Matrix Completion

In the gene–disease association matrix $P \in \mathbb{R}^{N_g \times N_d}$, each row represents a gene ID and the number of genes is N_g. Each column represents a disease phenotype and the number of diseases is N_d. If $P_{ij} = 1$, this means that gene i is related to disease j, and $P_{ij} = 0$ means that the relationship between gene i and disease j is uncertain. Based on the most successful and deeply studied matrix completion method in the recommender systems, the IMC algorithm was used to complete the task of learning gene–disease associations. The advantage of this is that this method is inductive, and it can achieve the prediction of new genes or diseases that have rarely been studied.

IMC assumes that the association matrix has a low rank, with the goal of recovering Z using the observed values of P and the eigenvectors of genetic diseases, as shown in **Figure 3**.

The eigenvector matrix of N_g genes is represented by $X \in \mathbb{R}^{N_g \times f_g}$, and the eigenvector of gene i is represented by $x_i \in \mathbb{R}^{f_g}$.

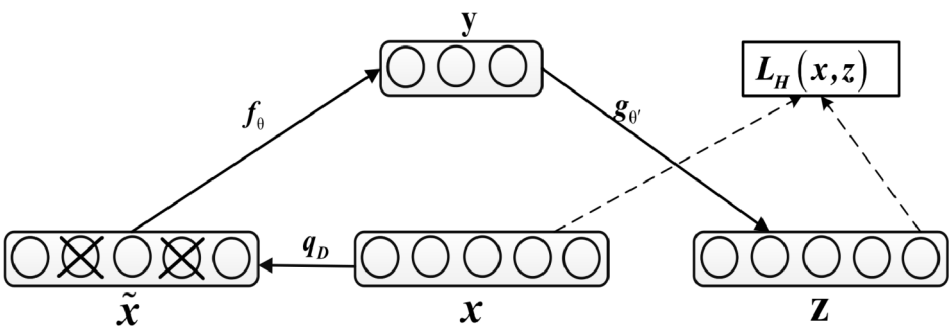

FIGURE 2 | Diagram of Denoising Autoencoder.

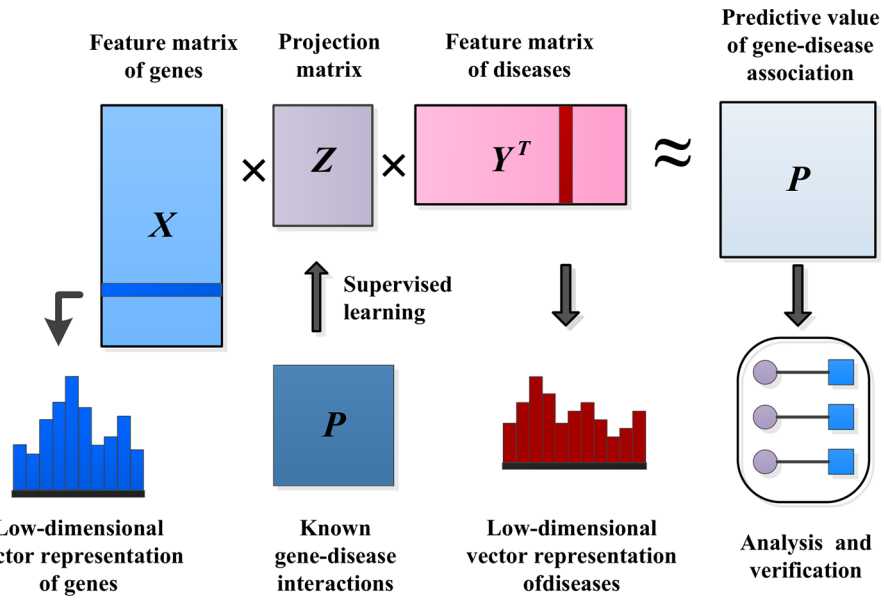

FIGURE 3 | Methods of predicting pathogenic genes.

Similarly, $Y \in \mathbb{R}^{N_d \times f_d}$ is used to represent the eigenvector matrix of N_d diseases, and $y_i \in \mathbb{R}^{f_d}$ is used to represent the eigenvector of disease j. The inductive matrix completion problem is to recover a low-rank matrix Z by using the known association Ω^+ from the gene–disease association matrix P. We established a bilinear function to learn the projection matrix Z between the gene space and the disease space to predict the interaction between unknown genes and diseases. We modeled the matrix P as $XZY^T \approx P$. Then, we used the following formula to measure the probability of pairwise interaction score between gene i and disease j, and the higher the score(i, j) value, the more likely gene i and disease j interact.

$$score(i,j) = x_i Z y_j^T \tag{14}$$

There is usually a significant correlation between spatially close eigenvectors of genes or diseases, which can greatly reduce the number of effective parameters needed to model gene–disease interactions in Z. To consider this problem, we applied a low-rank constraint on Z and learned only a few potential factors. Let $Z = GH^T$, where $G \in \mathbb{R}^{f_g \times k}$, $H \in \mathbb{R}^{f_d \times k}$, and k is small. This low-rank constraint not only alleviates the overfitting problem, but also facilitates the process of optimizing the calculation (Wang et al., 2015). The optimization problem of low-rank constraint is NP-hard on the original matrix Z. One standard method of relaxing the low-rank constraint is to minimize the trace norm, that is, the sum of the singular values. Minimizing the trace norm of $Z = GH^T$ is equivalent to minimizing $\frac{1}{2}(\|G\|_F^2 + \|H\|_F^2)$. The decomposition of Z into G and H solves the following optimization problems by alternating minimization. A common choice for the loss function ℓ is the square loss function. λ is the regularization parameter.

$$\min \sum_{(i,j)\in\Omega^+} \ell\left(P_{ij}, x_i^T G H^T y_i\right) + \frac{\lambda}{2}\left(\|G\|_F^2 + \|H\|_F^2\right) \tag{15}$$

Improved Inductive Matrix Completion

To optimize the objective function, we introduce the idea of PU-Learning. Although we predicted positive examples from unknown relationships, that is, candidate disease-causing genes, it was undeniable that these unknown genes-disease pairs may be unrelated. Therefore, unknown association relationship information was added to the learning process as a negative example, and the objective function was as follows:

$$\begin{aligned}\min \sum_{(i,j)\in\Omega^+} \ell\left(P_{ij}, x_i^T G H^T y_i\right) + \alpha \sum_{(i,j)\in\Omega^-} \ell\left(P_{ij}, x_i^T G H^T y_i\right) \\ + \frac{\lambda}{2}\left(\|G\|_F^2 + \|H\|_F^2\right)\end{aligned} \tag{16}$$

We represent the unknown association in the gene–disease association matrix P as Ω^-. The key parameter $\alpha < 1$ because the penalty weight of the known relationship must be greater than the unknown relationship. Finally, equation (14) was still used to calculate the interaction score between gene i and disease j. The scores are sorted in descending order, and the first k genes were selected as candidate pathogenic genes for the corresponding disease.

DATA SETS AND FEATURES

The data sets used in this paper can be divided into three categories: gene–disease association data, gene characteristic data, and disease characteristic data.

Gene–Disease Associations

The known gene–disease association data that we used were from the OMIM database, which contained 12,331 genes, 3,209 diseases, and 3,954 known gene–disease associations (the total number of nonzero elements in the gene–disease association matrix). It can be seen that the data in the incidence matrix are very sparse, with more than 90% of the columns having only one nonzero item and 70% of the rows having no nonzero elements.

Gene Characteristics

Gene characteristics were obtained by processing four different data sources through compact feature learning (*Compact Feature Learning*). The first source of gene characteristics was gene microarray data, which contained 8,755 genes and 4,536 characteristics. First, we linearly transformed the expression range of each gene to between 0 and 1. Because these characteristics are highly correlated, we used four layers of denoising autoencoder to reduce the dimensionality of the data, and the number of cells in each hidden layer was 3,000-800-300-100, respectively. Moreover, gaussian noise with a noise factor of 0.2 was added to the input data, and sigmoid was used to activate each layer. The model was optimized with Adam, and epoch was 100.

The second source of gene characteristics was from homologous gene phenotypic associations in eight other species, which were more abundant than in studies of human genetic diseases. The data used in the experiment are shown in **Table 1**. The features were extracted by two-layer denoising autoencoder with the following specific parameters: the number of nodes in each layer is "200–100," the corruption level of data is 0.2, the activation function is sigmoid function, the batch size is set as 150, and the model is optimized by Adam.

In addition, the data on interactions between genes can also be used as a part of the characteristics of genes. We integrated two networks, HumanNet (Lee et al., 2011) and STRING (Szklarczyk et al., 2017), for unified analysis. These two sets of data represent gene–gene interaction networks, but there are differences between them (Kuang et al., 2018). The integrated analysis of different sets of data can verify each set, and they can help to validate each other and expand understanding the potential rules. We used the RWR and DCA methods to fuse two networks to extract gene features. We set the restart probability to 0.05 and extracted the 600-dimensional gene characteristics. Finally, the gene characteristics used in the model were 800 dimensions.

TABLE 1 | Species Details.

Number	Species name	Number of disease phenotypes	Number of associations
1	Human	3209	3954
2	Arabidopsis thaliana	1137	12010
3	Worm	744	30519
4	Drosophila	2503	68525
5	Zebrafish	1143	4500
6	Escherichiacoli	324	72846
7	Gallus	1188	22150
8	Mouse	4662	75199
9	Saccharomyce	1243	73284

Disease Characteristics

The disease characteristics are mainly derived from two data sources: the disease similarity network MimMiner and clinical manifestation data of the disease, as well as a large amount of data from analysis of the medical literature.

MimMiner data are processed by literature (van Driel et al., 2006) and are freely available online. This data set has been applied in gene prioritization methods (Vanunu et al., 2010; Singh-Blom et al., 2013; Natarajan and Dhillon, 2014). RWR and DCA were used to extract 100-dimension disease features in the disease similarity network, and the restart probability was set as 0.05.

Another disease feature that we incorporated was from the OMIM disease webpage. We paid special attention to the clinical features and clinical management of webpages. We obtained disease features through text mining. We used PCA to reduce the dimensions of feature space and retained the first 100 principal components. Finally, we obtained 200-dimension disease characteristics.

EXPERIMENT

Evaluation Indexes and Methods introduces the evaluation indexes and methods of the experiment. *Parameter Settings* describes the influence of important parameters in the experiment. In *Global Performance*, the global performance of the experiment is compared. *Prediction of New Genes and New Diseases* compares the ability to predict new genes and new diseases. *Newly Discovered Genes* compares the ability to predict newly discovered associations.

Evaluation Indexes and Methods

In the experiment, to quantitatively evaluate our method and compare it with the most advanced disease-causing gene prioritization methods, we used a cross-validation strategy to measure gene recovery. We divided the known gene–disease pairs into three groups of the same size. The associations in one group were hidden, and the associations in the remaining two groups were used as training data, repeated three times to ensure that each group was hidden only once. For each disease in our data set, we ranked all of the genes according to the degree to which they were associated with the disease. The first r genes were taken as candidate pathogenic genes for corresponding diseases; namely, the top-r ranking method was used. The performance of the algorithm was analyzed by comparing the recall and precision of each method under different thresholds r, usually $r \leq 100$. The formula for calculating this was as follows:

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{17}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{18}$$

Recall rate refers to the proportion of positive cases correctly judged by the model relative to all positive cases (TP+FN) in the

data set. FN represents the data that are mistaken as negative cases by the model but are actually positive cases. The precision rate is the proportion of true positive cases (TP) relative to all positive cases (TP+FP) judged by the model (Xiong et al., 2012; Xu et al., 2017; Cheng et al., 2019; Cheng et al., 2019).

To further confirm the value of our approach, we also used the mean percentile ranking (MPR), an evaluation index based on recall, to evaluate the performance of the algorithm. This evaluation index has been applied in recommendation algorithm and analyses of the performance for predicting drug-targets (Hu et al., 2008; Johnson, 2014; Li et al., 2015; Ding et al., 2017; Hao et al., 2019; Liu et al., 2019b; Liu et al., 2019c; Zeng et al., 2019) and disease biomarkers (Chen et al., 2016; Zeng et al., 2016; Hong et al., 2019; Xu et al., 2019). For each disease, the genes were ranked in descending order according to the calculated gene–disease predictive value. The average ranking of the true and established associations among them is the final result. Here, rank_{ji} can be used to represent the percentile ranking (PR) of gene j and disease i. $\text{rank}_{ji} = 0\%$ indicates that disease i is most likely to interact with gene j. Similarly, $\text{rank}_{ji} = 100\%$ indicates that disease i has the lowest probability of interacting with gene j. Therefore, the definition of MPR is as follows:

$$MPR = \frac{\sum_{i=1}^{N_D^t} R_i}{N_D^t} \tag{19}$$

N_T^t represents the number of diseases in the test set, and the formula for calculating R_i is as follows:

$$R_i = \frac{\sum_{j=1}^{N_T^t} \text{rank}_{ji}}{N_T^t} \tag{20}$$

N_T^t represents the number of genes in the test set for current disease i. It is important to emphasize that lower MPR values are preferable because they indicate that our approach has a higher probability, which means that the model works better. Conversely, a higher MPR indicates a lower likelihood of gene interactions with disease. Clearly, the randomly generated list is expected to have an MPR of 50%. Using this measure, we can obtain a list of recommended candidate pathogenic genes, where the recommended optimal prediction is used for higher priority experimental validation.

Parameter Settings

The key parameters of PUIMCHIF are the rank k of matrix $Z \in \mathbb{R}^{800\times200}$, the regularized parameter λ, and the penalty weight α for the unknown relation. As can be seen from **Figure 4**, the performance of the PUIMCHIF method increases with the increase of k. When k = 100, 150, and 200, the three curves are very close. In the following experiment, the PUIMCHIF method uniformly set parameters as follows: k = 200, λ = 0.02, and α=0.0035.

As mentioned earlier, our approach features four improvements over the original IMC approach. For **Figure 5**, recall, precision, and MPR were used to analyze the effect of our improved method. The four experimental results in the figure represent (a) the initial experimental results of the original IMC method, (b) the results of extracting features by using RWR and DCA, instead of PCA, for the network data of diseases and genes, (c) the prediction results of adding STRING data to the gene interaction network, and (d) the experimental results of each index of the PUIMCHIF method.

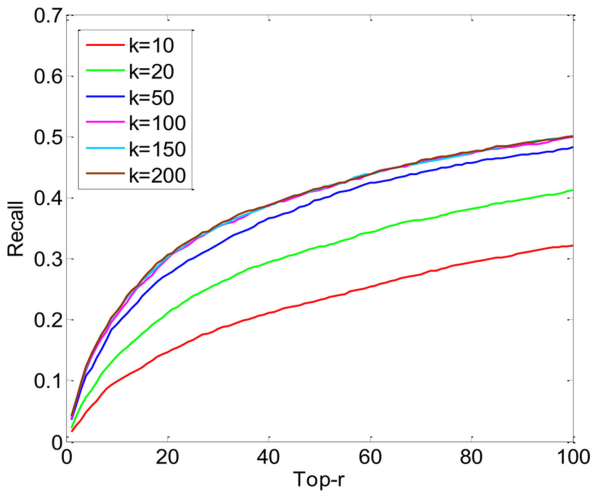

FIGURE 4 | Performance Comparison of PUIMCHIF with Different k Values.

We found that using RWR and DCA can better extract the gene–gene and disease–disease relationships, and helps to improve the prediction of candidate pathogenic genes. Meanwhile, it was also found that the protein interaction network STRING improved the prediction recall rate to 47.45%, and the MPR value also decreased significantly. Using denoising autoencoder to represent the characteristics of genes and diseases, and introducing the idea of PU-Learning into the inductive matrix completion can further improve the predictive performance.

Global Performance

In this experiment, the threefold cross-validation method was used to compare the overall performance of the proposed method with CATAPULT, Katz, and IMC. As shown in **Figure 6A**, the vertical axis gives the probability of recovering the true gene association in the top-r prediction of different r values on the horizontal axis. The experimental results show that the PUIMCHIF algorithm proposed in this paper has a much higher probability of recovering true gene associations under different thresholds than the other methods. **Figure 6B** presents the precision–recall curve.

In addition, **Table 2** shows the results of three evaluation indexes for each method when the threshold r=100. It is worth mentioning that a smaller value of MPR is associated with a higher probability and a better effect. It can be seen that the MPR value of PUIMCHIF is the lowest and the recall rate reaches 50%, while the best method among other methods, IMC, is only 25%, that is, the recall is doubled. The precision rate was also twice that of Katz which is the best method of other methods, reaching 4.87%. The overall performance of PUIMCHIF has been further improved, confirming the superiority of our method.

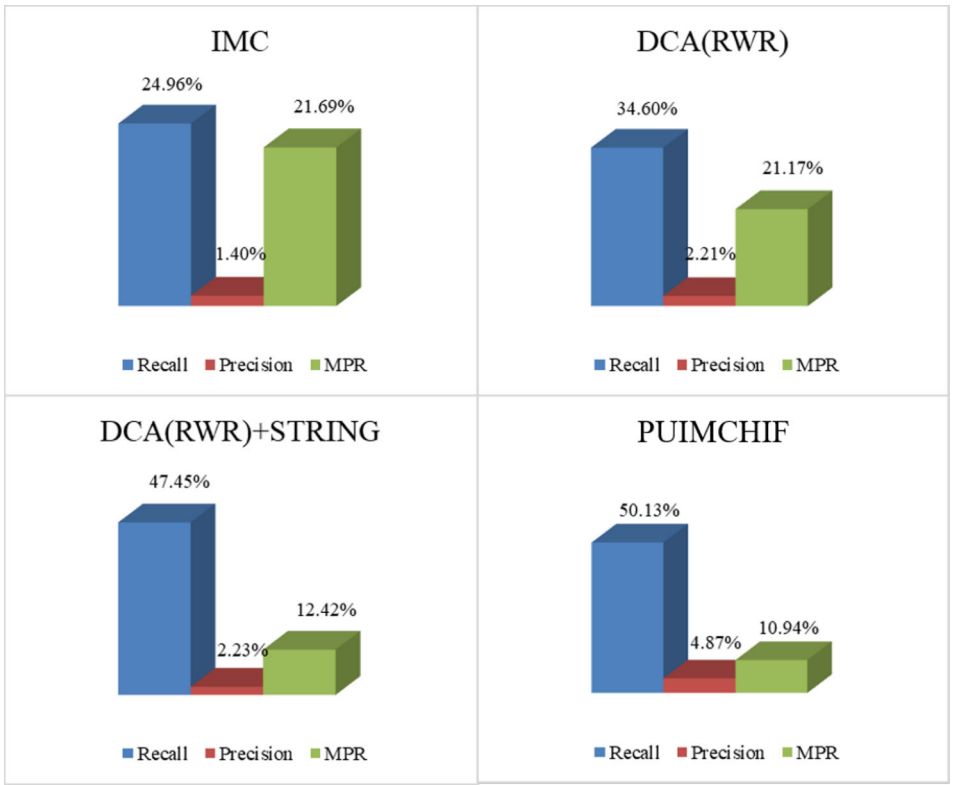

FIGURE 5 | Model Optimization Results.

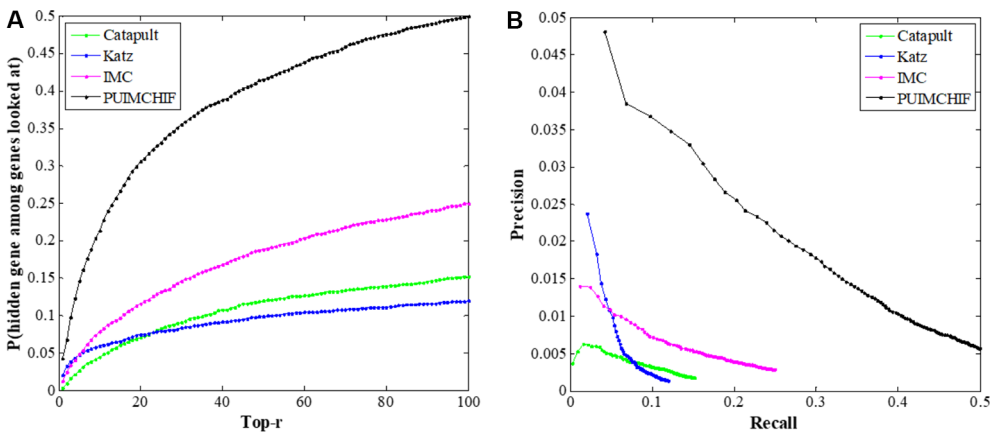

FIGURE 6 | Global Performance with Different Thresholds *r*. **(A)** Recall rate at different threshold *r*. **(B)** Precision-recall curve.

Prediction of New Genes and New Diseases

Prediction of New Genes

One problem affecting prioritization assessments is that well-related genes and diseases tend to be more predictable and therefore tend to generate inflated recall rates. Here, we focued only on genes that are known to have a single association in the gene-disease association data set. In other words, we selected the gene corresponding to the row with only 1 non-zero element in the gene-disease association matrix as the validation set, and hided these known associations in the training process. After repeated three-fold cross validations, **Figure 7A** shows the predictive power of different methods within the threshold $r < 100$. The Y-axis represents the probability of a true known single gene association hidden during recovery training.

TABLE 2 | Experimental Results with Threshold r =100.

Methods	Recall	Precision	MPR
CATAPULT	0.152251	0.006289	0.319410
Katz	0.120132	0.023752	0.335564
IMC	0.249621	0.014036	0.216856
PUIMCHIF	0.501265	0.048681	0.109412

Table 3 shows the specific experimental results of each method when r = 100. For the prediction of new genes, although the precision rate was slightly lower than Katz, the recall rate of our PUIMCHIF was significantly higher than other methods, reaching 40.7% when the recall rate of IMC method was only 13.7%. At the same time, we found that using the MPR index to evaluate the results, the PUIMCHIF method was only 13.5%, much lower than Katz and CATAPULT. This also shows that our method is more reliable.

Prediction of New Diseases

Similar to the prediction of new genes, we only considered diseases with a single known association in the gene-disease association data set as the validation set, that is, diseases corresponding to the columns with only 1 non-zero element in the gene-disease association matrix, and hided these known associations during training. Similarly, a three-fold cross-validation analysis was used, and the results are shown in **Figure 7B**. The probability that the proposed method could recover the true association of new diseases reached 48%, which was a significant improvement compared with other methods. Moreover, the MPR value of our method was lower than that of other methods, and the precision rate was nearly 2.7 percentage points higher than that of IMC method. As can be seen from **Table 4**, PUIMCHIF method is superior to other methods in three evaluation indexes.

Newly Discovered Genes

Cross-validation of retrospective data can lead to overly optimistic performance estimates. For example, certain gene interactions may be found because of associations with specific diseases being evaluated. Although the association itself is hidden, other features are contaminated by this information. Therefore, the use of recently reported associations to assess gene prioritization tools is unbiased in this assessment.

We trained all methods using all the gene associations of the 3,209 OMIM diseases collected. We found 162 newly discovered associations, of which 83 genes had no known associations previously. Thus, the assessment of new associations also helps determine the ability of methods to recommend new genes. The ranking performance of each method in 162 new associations is shown in **Figure 8**. We can see that the IMC method is superior to other methods in the range of threshold $6 \leq r \leq 100$.

CONCLUSION

In this paper, a PU induction matrix completion algorithm based on heterogeneous information fusion, PUIMCHIF, was proposed to predict gene–disease associations. Based on the specific advantages of IMC method, PUIMCHIF can predict new genes and diseases, and has good predictive performance. In addition, because closely connected or functionally similar genes are more likely to cause the same or similar diseases, we

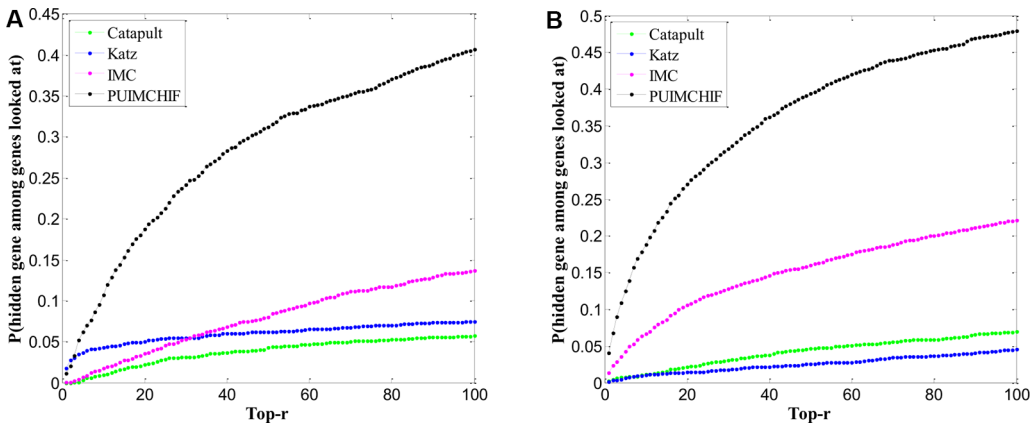

FIGURE 7 | Prediction of New Genes and New Diseases. **(A)** Prediction of New Genes. **(B)** Prediction of New Diseases.

TABLE 3 | Prediction of New Genes with Threshold r =100.

Methods	Recall	Precision	MPR
CATAPULT	0.056943	0.001227	0.497410
Katz	0.074838	0.018446	0.466105
IMC	0.137195	0.001935	0.284610
PUIMCHIF	0.407281	0.013840	0.135043

TABLE 4 | Prediction of New Disease with Threshold r =100.

Methods	Recall	Precision	MPR
CATAPULT	0.070060	0.002392	0.346974
Katz	0.045454	0.001709	0.363452
IMC	0.221804	0.014012	0.226905
PUIMCHIF	0.479836	0.040671	0.112801

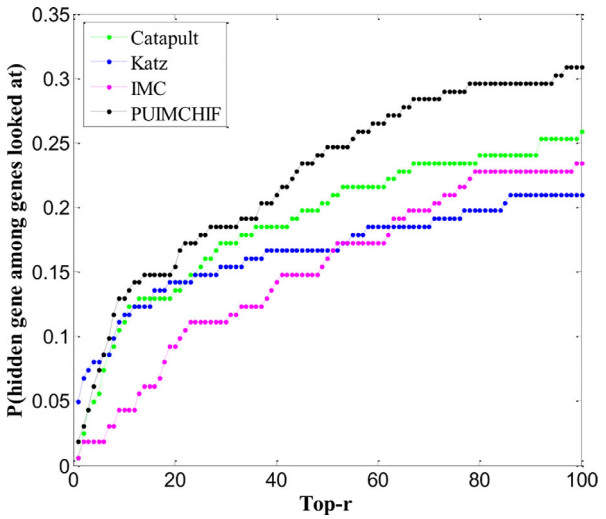

FIGURE 8 | Newly Discovered Prediction of Association.

constructed low-dimensional feature representations of genes and diseases from various data sources such as STRING using the compact feature learning method, which effectively alleviated the impact of data sparsity. Although there is no evidence that genes are unrelated to diseases in the data set, it is clear that most of the unknown associations are negative. PUIMCHIF conducts biased learning by treating unlabeled data as negative cases and constraining the penalty weight of known relationships to be greater than that of unknown relationships. Compared with the existing prediction methods, the PUIMCHIF method can significantly improve the prediction results regarding recall rate, precision rate, and MPR. According to the evaluation index of MPR, the experimental results of the PUIMCHIF method that we proposed are the lowest; that is to say, the candidate genes given by our algorithm have a higher priority for validation by biological experiments.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this article are not publicly available. Requests to access the datasets should be directed to chunyu@hit.edu.cn.

AUTHOR CONTRIBUTIONS

CW initiated the idea, conceived the whole process and drafted the manuscript. JZ and XW implemented the experiments and designed the figures. KH helped with data analysis and revised the manuscript. MG and finalized the paper. All authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

The work was supported Natural Science Foundation of China (61872114, 61571163, 61532014, and 61871020), and the National Key Research and Development Plan Task of China (No. 2016YFC0901902). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank Liwen Bianji, Edanz Group China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

REFERENCES

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43 (Database issue), D789–D798. doi: 10.1093/nar/gku1205

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Eppig JT *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25 (1), 25–29. doi: 10.1038/75556

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., et al. (2007). NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.* 35 (Database issue), D760–D765. doi: 10.1093/nar/gkl887

Becker, K. G., Barnes, K. C., Bright, T. J., and Wang, S. A. (2004). The genetic association database. *Nat. Genet.* 36 (5), 431–432. doi: 10.1038/ng0504-431

Cancer Genome Atlas Research, N., Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 45 (10), 1113–1120. doi: 10.1038/ng.2764

Cao, M., Pietras, C. M., Feng, X., Doroschak, K. J., Schaffner, T., Park, J., et al. (2014). New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics* 30 (12), i219–i227. doi: 10.1093/bioinformatics/btu263

Chen, X., Pérez-Jiménez, M. J., Valencia-Cabrera, L., Wang, B., and Zeng, X. (2016). Computing with viruses. *Theor. Comput. Sci.* 623, 146–159. doi: 10.1016/j.tcs.2015.12.006

Chen, L., Zeng, T., Pan, X. Y., Zhang, Y. H., Huang, T., and Cai, Y. D. (2019). Identifying Methylation Pattern and Genes Associated with Breast Cancer Subtypes. *Int. J. Mol. Sci.* 20 (17), 20. doi: 10.3390/ijms20174269

Chen, L., Pan, X. Y., Zhang, Y. H., Kong, X. Y., Huang, T., and Cai, Y. D. (2019). Tissue differences revealed by gene expression profiles of various cell lines. *J. Cell Biochem.* 120 (5), 7068–7081. doi: 10.1002/jcb.27977

Cheng, L., Sun, J., Xu, W. Y., Dong, L. X., Hu, Y., and Zhou, M. (2016). OAHG: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.* 6, 1–9. doi: 10.1038/srep34820

Cheng, L., Jiang, Y., Ju, H., Sun, J., Peng, J., Zhou, M., et al. (2018). InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk. *BMC Genomics* 19 (Suppl 1), 919. doi: 10.1186/s12864-017-4338-6

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* 34 (11), 1953–1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* 47 (D1), D140–D144. doi: 10.1093/nar/gky1051

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2019). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* doi: 10.1093/nar/gkz843

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2019). MetSigDis: a manually curated resource for the metabolic signatures of diseases. *Brief Bioinform.* 20 (1), 203–209. doi: 10.1093/bib/bbx103

Deng, L., Li, W., and Zhang, J. (2019). LDAH2V: Exploring meta-paths across multiple networks for lncRNA-disease association prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* doi: 10.1109/TCBB.2019.2946257

Ding, Y., Tang, J., and Guo, F. (2017). Identification of drug-target interactions *via* multiple information integration. *Inf. Sci.* 418-419, 546–560. doi: 10.1016/j.ins.2017.08.045

Ding, Y., Tang, J., and Guo, F. (2019). Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing* 325, 211–224. doi: 10.1016/j.neucom.2018.10.028

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., et al. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* 8 (Suppl 2), I1. doi: 10.1186/1752-0509-8-S2-I1

Hao, M., Bryant, S. H., and Wang, Y. (2019). Open-source chemogenomic data-driven algorithms for predicting drug-target interactions. *Brief Bioinform.* 20 (4), 1465–1474. doi: 10.1093/bib/bby010

He, J., Fang, T., Zhang, Z., Huang, B., Zhu, X., and Xiong, Y. (2018). PseUI: Pseudouridine sites identification based on RNA sequence information. *BMC Bioinf.* 19 (1), 306. doi: 10.1186/s12859-018-2321-0

Hong, Z., Zeng, X., Wei, L., and Liu, X. J. B. (2019). Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics*. (Pisa, Italy: IEEE). doi: 10.1093/bioinformatics/btz694

Hu, Y., Koren, Y., and Volinsky, C. (2008). "Collaborative Filtering for Implicit Feedback Datasets," in *2008 Eighth IEEE International Conference on Data Mining: 15-19 Dec. 2008*, (Pisa, Italy: IEEE), 263–272. doi: 10.1109/ICDM.2008.22

Jiang, R. (2015). Walking on multiple disease-gene networks to prioritize candidate genes. *J. Mol. Cell Biol.* 7 (3), 214–230. doi: 10.1093/jmcb/mjv008

Johnson, C. (2014). "Logistic matrix factorization for implicit feedback data," in *Advances in Neural Information Processing Systems*. Montréal, Canada. vol. 27

Kim, M., and Leskovec, J. (2011). "The Network Completion Problem: Inferring Missing Nodes and Edges in Networks," in *Proceedings of the 2011 SIAM International Conference on Data Mining*. Mesa, Arizona, U. S. A. 47–58. doi: 10.1137/1.9781611972818.5

Kohler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* 82 (4), 949–958. doi: 10.1016/j.ajhg.2008.02.013

Krauthammer, M., Kaufmann, C. A., Gilliam, T. C., and Rzhetsky, A. (2004). Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl. Acad. Sci. U. S. A* 101 (42), 15148–15153. doi: 10.1073/pnas.0404315101

Kuang, L., Yu, L., Huang, L., Wang, Y., Ma, P., Li, C., et al. (2018). A Personalized QoS Prediction Approach for CPS Service Recommendation Based on Reputation and Location-Aware Collaborative Filtering. *Sensors* 18 (5), 1556. doi: 10.3390/s18051556

Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., and Marcotte, E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* 21 (7), 1109–1121. doi: 10.1101/gr.118992.110

Li, Y., and Patra, J. C. (2010). Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics* 26 (9), 1219–1224. doi: 10.1093/bioinformatics/btq108

Li, W., Yu, J., Lian, B., Sun, H., Li, J., Zhang, M., et al. (2015). Identifying prognostic features by bottom-up approach and correlating to drug repositioning. *PloS One* 10 (3), e0118672. doi: 10.1371/journal.pone.0118672

Li, J. R., Lu, L., Zhang, Y. H., Liu, M., Chen, L., Huang, T., et al. (2019). Identification of synthetic lethality based on a functional network by using machine learning algorithms. *J. Cell Biochem.* 120 (1), 405–416. doi: 10.1002/jcb.27395

Liao, Z. J., Li, D. P., Wang, X. R., Li, L. S., and Zou, Q. (2018). Cancer diagnosis through isomiR expression with machine learning method. *Curr. Bioinform.* 13 (1), 57–63. doi: 10.2174/1574893611666160609081155

Liu, H., Zhang, W., Zou, B., Wang, J., Deng, Y., and Deng, L. (2020). DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Res.* 48(D1):D871–D881. doi: 10.1093/nar/gkz1007

Liu, B. (2019) BioSeq-Analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches. *Briefings In Bioinf.* 20 (4),1280–1294. doi: 10.1093/bib/bbx165

Liu, B., Gao, X., and Zhang, H. (2019a). BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* 47 (20), e127. doi: 10.1093/analys/anz032

Liu, B., Li, C., and Yan, K. (2019b). DeepSVM-fold: Protein fold recognition by combining Support Vector Machines and pairwise sequence similarity scores generated by deep learning networks. *Briefings Bioinf.* doi: 10.1093/bib/bbz098

Liu, B., Zhu, Y., and Yan, K. (2019c). Fold-LTR-TCP: protein fold recognition based on triadic closure principle. *Briefings Bioinf.* doi: 10.1093/bib/bbz139

Mordelet, F., and Vert, J. P. (2011). ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinf.* 12, 389. doi: 10.1186/1471-2105-12-389

Natarajan, N., and Dhillon, I. S. (2014). Inductive matrix completion for predicting gene-disease associations. *Bioinformatics* 30 (12), i60–i68. doi: 10.1093/bioinformatics/btu269

Navlakha, S., and Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. *Bioinformatics* 26 (8), 1057–1063. doi: 10.1093/bioinformatics/btq076

Pan, X. Y., Chen, L., Feng, K. Y., Hu, X. H., Zhang, Y. H., Kong, X. Y., et al. (2019). Analysis of Expression Pattern of snoRNAs in Different Cancer Types with Machine Learning Algorithms. *Int. J. Mol. Sci.* 20 (9), 16. doi: 10.3390/ijms20092185

Pan, X. Y., Hu, X. H., Zhang, Y. H., Chen, L., Zhu, L. C., Wan, S. B., et al. (2019). Identification of the copy number variant biomarkers for breast cancer subtypes. *Mol. Genet. Genomics* 294 (1), 95–110. doi: 10.1007/s00438-018-1488-4

Peng, L., Peng, M. M., Liao, B., Huang, G. H., Li, W. B., and Xie, D. F. (2018). The advances and challenges of deep learning application in biological big data processing. *Curr. Bioinform.* 13 (4), 352–359. doi: 10.2174/1574893612666170707095707

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y. W., Mazaitis, M., Felix, V., et al. (2012). Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res.* 40 (Database issue), D940–D946. doi: 10.1093/nar/gkr972

Singh-Blom, U. M., Natarajan, N., Tewari, A., Woods, J. O., Dhillon, I. S., and Marcotte, E. M. (2013). Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PloS One* 8 (5), e58977. doi: 10.1371/journal.pone.0058977

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). Bork P *et al*: The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45 (D1), D362–D368. doi: 10.1093/nar/gkw937

van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G., and Leunissen, J. A. (2006). A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.* 14 (5), 535–542. doi: 10.1038/sj.ejhg.5201585

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PloS Comput. Biol.* 6 (1), e1000641. doi: 10.1371/journal.pcbi.1000641

Wang, S., Cho, H., Zhai, C., Berger, B., and Peng, J. (2015). Exploiting ontology graph for predicting sparsely annotated gene function. *Bioinformatics* 31 (12), i357–i364. doi: 10.1093/bioinformatics/btv260

Wang, L., Ping, P. Y., Kuang, L. N., Ye, S. T., Lqbal, F. M. B., and Pei, T. R. (2018). A novel approach based on bipartite network to predict human microbe-disease associations. *Curr. Bioinform.* 13 (2), 141–148. doi: 10.2174/1574893612666170911143601

Wu, X., Jiang, R., Zhang, M. Q., and Li, S. (2008). Network-based global inference of human disease genes. *Mol. Syst. Biol.* 4, 189. doi: 10.1038/msb.2008.27

Xiong, Y., Liu, J., Zhang, W., and Zeng, T. (2012). Prediction of heme binding residues from protein sequences with integrative sequence profiles. *Proteome Sci.* 10 (Suppl 1), S20. doi: 10.1186/1477-5956-10-S1-S20

Xiong, Y., Wang, Q., Yang, J., Zhu, X., and Wei, D. Q. (2018). PredT4SE-Stack: Prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. *Front. Microbiol.* 9, 2571. doi: 10.3389/fmicb.2018.02571

Xu, Q., Xiong, Y., Dai, H., Kumari, K. M., Xu, Q., Ou, H. Y., et al. (2017). PDC-SGB: Prediction of effective drug combinations using a stochastic gradient boosting algorithm. *J. Theor. Biol.* 417, 1–7. doi: 10.1016/j.jtbi.2017.01.019

Xu, H., Zeng, W., and Zhang, D. (2019). Zeng XJIToC: MOEA/HD: A Multiobjective Evolutionary Algorithm Based on Hierarchical Decomposition. *IEEE Trans. Cybernetics* 49 (2), 517–526. doi: 10.1109/TCYB.2017.2779450

Zeng, X., Zhang, X., and Liao, Y. (2016). Pan LJBeBA-GS: Prediction and validation of association between microRNAs and diseases by multipath methods. *Biochim. Biophys. Acta -General Subj.* 1860 (11), 2735–2739. doi: 10.1016/j.bbagen.2016.03.016

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017). Prediction and Validation of Disease Genes Using HeteSim Scores. *Ieee-Acm Trans. Comput. Biol. And Bioinf.* 14 (3), 687–695. doi: 10.1109/TCBB.2016.2520947

Zeng, X., Liu, L., Lü, L., and Zou, Q. (2018). Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* 34 (14), 2425–2432. doi: 10.1093/bioinformatics/bty112

Zeng, X. X., Wang, W., Deng, G. S., Bing, J. X., and Zou, Q. (2019). Prediction of potential disease-associated microRNAs by using neural networks. *Mol. Ther.-Nucl. Acids* 16, 566–575. doi: 10.1016/j.omtn.2019.04.010

Zeng, X., Zhong, Y., Lin, W., and Zou, Q. (2019). Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Briefings In Bioinf.* doi: 10.1093/bib/bbz080

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. (2019). deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*. 35 (24), 5191–5198. doi: 10.1093/bioinformatics/btz418

Zhang, J., Zhang, Z., Wang, Z., Liu, Y., and Deng, L. (2018). Ontological function annotation of long non-coding RNAs through hierarchical multi-label classification. *Bioinformatics* 34 (10), 1750–1757. doi: 10.1093/bioinformatics/btx833

Zhang, X., Zou, Q., Rodriguez-Paton, A., and Zeng, X. X. (2019). Meta-path methods for prioritizing candidate disease miRNAs. *Ieee-Acm Trans. Comput. Biol. Bioinf.* 16 (1), 283–291. doi: 10.1109/TCBB.2017.2776280

Zhang, Z., Zhang, J., Fan, C., Tang, Y., and Deng, L. (2019). KATZLGO: Large-scale prediction of LncRNA functions by using the KATZ measure based on multiple networks. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 16 (2), 407–416. doi: 10.1109/TCBB.2017.2704587

Zhu, X., He, J., Zhao, S., Tao, W., Xiong, Y., and Bi, S. (2019). A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of Saccharomyces cerevisiae. *Brief Funct. Genomics*. 18 (6), 367–376. doi: 10.1093/bfgp/elz018

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., and Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Front. In Genet.* 9, 515. doi: 10.3389/fgene.2018.00515

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.