frontiers in Genetics

ORIGINAL RESEARCH
published: 02 April 2019
doi: 10.3389/fgene.2019.00270

Identifying Disease-Gene Associations With Graph-Regularized Manifold Learning

Ping Luo[1], *Qianghua Xiao*[2], *Pi-Jing Wei*[1,3], *Bo Liao*[4] *and Fang-Xiang Wu*[1,4,5,6]*

[1] Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK, Canada, [2] School of Mathematics and Physics, University of South China, Hengyang, China, [3] College of Computer Science and Technology, Anhui University, Hefei, China, [4] School of Mathematics and Statistics, Hainan Normal University, Haikou, China, [5] Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SK, Canada, [6] Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

Complex diseases are known to be associated with disease genes. Uncovering disease-gene associations is critical for diagnosis, treatment, and prevention of diseases. Computational algorithms which effectively predict candidate disease-gene associations prior to experimental proof can greatly reduce the associated cost and time. Most existing methods are disease-specific which can only predict genes associated with a specific disease at a time. Similarities among diseases are not used during the prediction. Meanwhile, most methods predict new disease genes based on known associations, making them unable to predict disease genes for diseases without known associated genes.In this study, a manifold learning-based method is proposed for predicting disease-gene associations by assuming that the geodesic distance between any disease and its associated genes should be shorter than that of other non-associated disease-gene pairs. The model maps the diseases and genes into a lower dimensional manifold based on the known disease-gene associations, disease similarity and gene similarity to predict new associations in terms of the geodesic distance between disease-gene pairs. In the 3-fold cross-validation experiments, our method achieves scores of 0.882 and 0.854 in terms of the area under of the receiver operating characteristic (ROC) curve (AUC) for diseases with more than one known associated genes and diseases with only one known associated gene, respectively. Further *de novo* studies on Lung Cancer and Bladder Cancer also show that our model is capable of identifying new disease-gene associations.

Keywords: disease gene identification, manifold learning, disease module theory, gene ontology, multi-task learning

OPEN ACCESS

Edited by:
Chuan Lu,
Aberystwyth University,
United Kingdom

Reviewed by:
Ling-Yun Wu,
Academy of Mathematics and Systems Science (CAS), China
Min Chen,
Hunan Institute of Technology, China

***Correspondence:**
Fang-Xiang Wu
faw341@mail.usask.ca

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

Received: *21 December 2018*
Accepted: *12 March 2019*
Published: *02 April 2019*

Citation:
Luo P, Xiao Q, Wei P-J, Liao B and Wu F-X (2019) Identifying Disease-Gene Associations With Graph-Regularized Manifold Learning. Front. Genet. 10:270. doi: 10.3389/fgene.2019.00270

1. INTRODUCTION

Complex diseases are caused by a group of genes known as disease genes. Identifying disease-gene associations is of critical importance since it helps us unravel the mechanisms of diseases, which has many applications such as diagnosis, treatment and prevention of disease. With the advances in high-throughput experimental techniques, a large amount of data that indicate associations between diseases and their associated genes have been generated, which could accelerate the identification of disease-associated genes. However, it is expensive and time-consuming to

experimentally prove an association between a gene and a disease. Computational methods that translate the experimental data into legible disease-gene associations are necessary for in-depth experimental validation.

Currently, many algorithms have been developed to predict disease-gene associations, and they can be briefly divided into two categories: the machine learning-based methods and the network-based methods. The typical machine learning-based methods extract gene-related features and train models that can discriminate disease genes and passenger genes (Mordelet and Vert, 2011; Yang et al., 2012; Singh-Blom et al., 2013; Luo et al., 2019a,b). Since the features are extracted for genes, these algorithms are usually single-task algorithms which once can only predict disease genes for a specific disease. Thus, for diseases that have a few or no known associated genes, the number of the genes would be too small to train the model. In the meantime, the relationships among diseases are usually not used in the prediction since only one disease is considered at a time. Matrix completion methods, as a type of machine learning methods, can solve the above two issues by jointly predicting disease-gene associations and leveraging the similarities among diseases during the calculation (Natarajan and Dhillon, 2014; Zeng et al., 2017). However, matrix completion methods generally do not have the global optimal solutions and could take a very long time to converge to even a local optimal solution. Network-based methods are based on the assumption that genes close related in the network are associated with the same diseases. Centrality indices, random walk and network energy are used in many methods to predict disease-gene associations (Köhler et al., 2008; Vanunu et al., 2010; Chen et al., 2014a,b). Although most network-based methods are not affected by the above two issues, their performance is strongly affected by the quality of networks, and they usually perform worse than machine learning-based methods on diseases with many known associated genes (Chen et al., 2015, 2016).

In this study, we propose a manifold learning-based method (dgManifold) to predict disease-gene associations. In our dgManifold, genes and diseases are regarded as points in the same high-dimensional Euclidean space. Our assumption is that diseases and their associated genes should be consistent in some lower dimensional manifold, and the geodesic distance between a disease and its associated genes should be shorter than that of other non-associated disease-gene pairs. Although the Euclidean distance between diseases and genes in the high-dimensional space may not reflect their true geodesic distance, we can map the diseases and genes into a low-dimensional manifold based on the experimentally verified disease-gene associations (Tenenbaum et al., 2000; Ham et al., 2005). Then, the true geodesic distance between all the disease-gene pairs can be calculated. In the meantime, the mapping process is regularized by two affinity graphs, disease similarity network and gene similarity network, so that the learned representations with the similarity information can further increase the prediction accuracy. Additionally, since our dgManifold is a supervised method, and it is difficult (if possible) to learn valuable representations for diseases that only have a few or no known associated genes. A prior information vector calculated with the disease similarities and known disease-gene associations should be combined with the original association data to solve this issue. Similar strategies have been applied to calculate the initial probabilities used in the random walk, which have improved the accuracy of predicting miRNA-disease associations (Chen et al., 2016b, 2018a,b).

In the rest of the manuscript, section 2 describes our algorithm as well as the data sources and evaluation metrics used in the study. Section 3 discusses the evaluation results. Section 4 draws some conclusions.

2. MATERIALS AND METHODS

2.1. General Model

Given n diseases and m genes, the associations among them can be represented by a matrix $A \in R^{n\times m}$ in which $a_{ij} = 1$ if disease i is associated with gene j, and otherwise $a_{ij} = 0$. Intuitively, each disease can be represented by a binary m-dimensional row vector while each gene can be represented by a binary n-dimensional column vector. However, in these high-dimensional spaces, it is hard to calculate the actual distance between a disease and a gene.

If we map the diseases and genes into the same manifold with a lower dimensionality and assume that the distance between a disease and its associated genes should be as short as possible on this manifold, predicting disease-gene associations can be solved by computing this mapping based on known disease-gene associations, which can be mathematically formulated as: finding k-dimensional representatives of diseases $\mathbf{r}_1, \ldots, \mathbf{r}_n$ and k-dimensional representatives of genes $\mathbf{q}_1, \ldots, \mathbf{q}_m$ such that the following objective function is minimized

$$O_k = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}\|\mathbf{r}_i - \mathbf{q}_j\|^2. \tag{1}$$

However, without any constraints, the objective function (1) is not well defined. To illustrate this, if k-dimensional vectors $\mathbf{r}_i^+$ and $\mathbf{q}_j^+$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ minimize the objective function (1), then $\epsilon\mathbf{r}_i^+$ and $\epsilon\mathbf{q}_j^+$ can further minimize the objective function when $0 \leq \epsilon < 1$. Especially, when $\epsilon = 0$, any k-dimensional vectors $\mathbf{r}_i^+$ and $\mathbf{q}_j^+$ can minimize the objective function. Therefore, to make the optimization problem well defined, the following constraints are added

$$\sum_{i=1}^{n}\mathbf{r}_i\mathbf{r}_i^T = I_k \quad \text{and} \quad \sum_{j=1}^{m}\mathbf{q}_j\mathbf{q}_j^T = I_k. \tag{2}$$

where I_k is the $k \times k$ identity matrix. As a results, the learned representations are unique with these constraints.

To insure that the mapped representations of diseases and genes are in concert with their intrinsic properties, two affinity graphs, disease similarity network and gene similarity network are used to regularize the objective function (1), and the new objective function is as follows

$$O_k = \sum_{j=1}^{m}\sum_{i=1}^{n} a_{ij}\|\mathbf{r}_i - \mathbf{q}_j\|^2 + \frac{\alpha}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}^d\|\mathbf{r}_i - \mathbf{r}_j\|^2$$

$$+\frac{\beta}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}s_{ij}^{g}\|\mathbf{q}_i-\mathbf{q}_j\|^2 \quad (3)$$

where S^d and S^g are the adjacency matrices of the disease similarity network and the gene similarity network, respectively.

Note that

$$\begin{aligned}
O_k &= \sum_{i=1}^{n}(\sum_{j=1}^{m}a_{ij})\mathbf{r}_i^T\mathbf{r}_i+\sum_{j=1}^{m}(\sum_{i=1}^{n}a_{ij})\mathbf{q}_j^T\mathbf{q}_j-2\sum_{i=1}^{n}\sum_{j=1}^{m}a_{ij}\mathbf{r}_i^T\mathbf{q}_j\\
&+\alpha\sum_{i=1}^{n}(\sum_{j=1}^{n}s_{ij}^{d})\mathbf{r}_i^T\mathbf{r}_i-\alpha\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}^{d}\mathbf{r}_i^T\mathbf{r}_j\\
&+\beta\sum_{i=1}^{m}(\sum_{j=1}^{m}s_{ij}^{g})\mathbf{q}_i^T\mathbf{q}_i-\beta\sum_{i=1}^{m}\sum_{j=1}^{m}s_{ij}^{g}\mathbf{q}_i^T\mathbf{q}_j\\
&=\sum_{i=1}^{n}A_{ri}\mathbf{r}_i^T\mathbf{r}_i+\sum_{j=1}^{m}A_{cj}\mathbf{q}_j^T\mathbf{q}_j-2\sum_{i=1}^{n}\sum_{j=1}^{m}a_{ij}\mathbf{r}_i^T\mathbf{q}_j\\
&+\alpha\sum_{i=1}^{n}S_i^{d}\mathbf{r}_i^T\mathbf{r}_i-\alpha\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}^{d}\mathbf{r}_i^T\mathbf{r}_j\\
&+\beta\sum_{j=1}^{m}S_j^{g}\mathbf{q}_j^T\mathbf{q}_j-\beta\sum_{j=1}^{m}\sum_{i=1}^{m}\mathbf{s}_{ij}^{g}\mathbf{q}_i^T\mathbf{q}_j\\
&=\sum_{i=1}^{n}(A_{ri}+\alpha S_i^{d})\mathbf{r}_i^T\mathbf{r}_i+\sum_{j=1}^{m}(A_{cj}+\beta S_j^{d})\mathbf{q}_j^T\mathbf{q}_j\\
&-2\sum_{i=1}^{n}\sum_{j=1}^{m}a_{ij}\mathbf{r}_i^T\mathbf{q}_j-\alpha\sum_{i=1}^{n}\sum_{j=1}^{n}s_{ij}^{d}\mathbf{r}_i^T\mathbf{r}_j-\beta\sum_{j=1}^{m}\sum_{i=1}^{m}s_{ij}^{g}\mathbf{q}_i^T\mathbf{q}_j
\end{aligned} \quad (4)$$

where $S_i^d=\sum_{j=1}^{n}s_{ij}^d, S_i^g=\sum_{j=1}^{m}s_{ij}^g, A_{ri}=\sum_{j=1}^{m}a_{ij}, A_{cj}=\sum_{i=1}^{n}a_{ij}$. Let

$$\begin{aligned}
L^{11}&=diag[A_{r1}+\alpha S_1^d, A_{r2}+\alpha S_2^d,\ldots,A_{rn}+\alpha S_n^d]-\alpha S^d,\\
L^{22}&=diag[A_{c1}+\beta S_1^g, A_{c2}+\beta S_2^g,\ldots,A_{cm}+\beta S_m^g]-\beta S^g,
\end{aligned} \quad (5)$$

the objective function (3) can be simplified as

$$O_k=\sum_{i=1}^{n}\sum_{j=1}^{n}L^{11}\mathbf{r}_i^T\mathbf{r}_j+\sum_{i=1}^{m}\sum_{j=1}^{m}L^{22}\mathbf{q}_i^T\mathbf{q}_j-2\sum_{i=1}^{n}\sum_{j=1}^{m}a_{ij}\mathbf{r}_i^T\mathbf{q}_j \quad (6)$$

Furthermore, let

$$r_i=\begin{bmatrix}x_{i1}\\x_{i2}\\\vdots\\x_{ik}\end{bmatrix}, q_j=\begin{bmatrix}y_{j1}\\y_{j2}\\\vdots\\y_{jk}\end{bmatrix}, z_t=\begin{bmatrix}x_{1t}\\\vdots\\x_{nt}\\\hdashline y_{1t}\\\vdots\\y_{mt}\end{bmatrix}=\begin{bmatrix}x_t\\\hdashline y_t\end{bmatrix}, \quad (7)$$

$$\begin{aligned}
&A_r=diag[A_{r1},\ldots,A_{rn}],\quad A_c=diag[A_{c1},\ldots,A_{cm}],\\
&L^d=diag[S_1^d,\ldots,S_n^d]-S^d,\quad L^g=diag[S_1^g,\ldots,S_m^g]-S^g,
\end{aligned} \quad (8)$$

$$L=\begin{bmatrix}A_r+\alpha L^d & -A\\ -A^T & A_c+\beta L^g\end{bmatrix}, \quad (9)$$

objective function (6) can be simplified as

$$\begin{aligned}
O_k&=\sum_{t=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{n}L^{11}x_{it}x_{jt}+\sum_{t=1}^{k}\sum_{i=1}^{m}\sum_{j=1}^{m}L^{22}y_{it}y_{jt}\\
&-2\sum_{t=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m}a_{ij}x_{it}y_{jt}\\
&=\sum_{t=1}^{k}[\mathbf{x}_t^TL^{11}\mathbf{x}_t+\mathbf{y}_t^TL^{22}\mathbf{y}_t-2\mathbf{x}_t^TA\mathbf{y}_t]\\
&=\sum_{t=1}^{k}[\mathbf{x}_t^T\mathbf{y}_t^T]\begin{bmatrix}L^{11}&-A\\-A^T&L^{22}\end{bmatrix}\begin{bmatrix}\mathbf{x}_t\\\mathbf{y}_t\end{bmatrix}\\
&=Tr(Z^TLZ)
\end{aligned} \quad (10)$$

Therefore, minimizing the objective function (4) with constraints (2) is equivalent to minimize the following function

$$Q_k=Tr(Z^TLZ) \quad (11)$$

with constraints

$$Z^TZ=X^TX+Y^TY=2I_k \quad (12)$$

According to Bolla (2013), minimizing objective function (11) with constraints (12) can be solved by

$$Z^*=(\mathbf{u}_0,\mathbf{u}_1,\ldots,\mathbf{u}_{k-1}) \quad (13)$$

where $\mathbf{u}_0,\mathbf{u}_1,\ldots,\mathbf{u}_{k-1}$ are k eigenvectors correspond to the k smallest eigenvalues of L. Meanwhile, the smallest eigenvalue is 0, and the corresponding eigenvector $\mathbf{u}_0$ is a constant vector which does not contribute to the calculation of the geodesic distance. Thus, let $\hat{Z}$ denote the matrix by removing the fist column of Z^*. The first n rows of $\hat{Z}$ are the obtained $(k-1)$-dimensional representations of diseases, and the rest m rows of $\hat{Z}$ are the learned representations of genes. The geodesic distance between a disease i and gene j can be calculated by

$$gdist_{ij}=\|\hat{\mathbf{r}}_i-\hat{\mathbf{q}}_j\|^2. \quad (14)$$

2.2. Similarity Network

2.2.1. Gene Similarity

In this study, the learning process is regularized by similarity networks, and the similarities of genes are calculated based on the Gene Ontology (GO). GO database provides a set of vocabularies to describe the function of genes and gene products (Ashburner et al., 2000; Consortium, 2017). The GO terms and their relationships are manifested as a directed acyclic graph (DAG) where nodes represent terms while edges represent semantic relationships. Many algorithms have been proposed to calculate the similarities of genes using ontology data, and the approach proposed by Wang et al. (2007) is used in this study.

Let $DAG_h = (T_h, E_h)$ denote GO term h, where T_h contains all the successor GO terms of h in the DAG, and E_h contains the semantic relationships between h and other terms in T_h. Each term t in T_h has a τ-value related to h:

$$\begin{cases} \tau_h(t) = 1, if\ t = h \\ \tau_h(t) = max\{w_e * \tau_h(t^{'})|t^{'} \in children\ of\ t\}, otherwise \end{cases} \tag{15}$$

where w_e is the weight of the edge (semantic relationships) in the DAG. Two types of semantic relationships ("*is_a*" and "*part_of*") are used in the DAG, and the corresponding w_e is set to 0.8 and 0.6, respectively, as recommended in Wang et al. (2007).

Given $DAG_h = (T_h, E_h)$ and $DAG_b = (T_b, E_b)$ for GO terms h and b, their similarity can be computed by

$$sgo(h, b) = \frac{\sum_{t \in T_h \cap T_b}(\tau_h(t) + \tau_b(t))}{\sum_{t \in T_h} \tau_h(t) + \sum_{t \in T_b} \tau_b(t)} \tag{16}$$

Then, the similarity of one GO term t' and a set of GO terms $GO = \{t_1, t_2, \ldots, t_l\}$ is defined as

$$SGO(t', GO) = \max_{1 \le i \le l}(SGO(t', t_i)) \tag{17}$$

Finally, the functional similarity of two genes g_1 and g_2 is calculated by

$$s^{g}_{g_1,g_2} = \frac{\sum_{1 \le i \le n_1} SGO(t_{1i}, GO_2) + \sum_{1 \le j \le n_2} SGO(t_{2j}, GO_1)}{n_1 + n_2} \tag{18}$$

where $GO_1 = \{t_{11}, t_{12}, \ldots, t_{1n_1}\}$ and $GO_2 = \{t_{21}, t_{22}, \ldots, t_{2n_2}\}$ are two sets of GO terms that describe g_1 and g_2, respectively.

2.2.2. Disease Similarity

The similarities among diseases are also calculated with the ontology data. Instead of GO, the Human Phenotype Ontology (HPO) (Köhler et al., 2018) is used to characterize human diseases. The HPO provides a vocabulary of phenotypic terms related to human diseases. Each term represents a clinical abnormality, and all the terms are structured as a DAG, in which every term is related to their parent terms by "*is_a*" relationships. Although diseases are not directly described by the HPO, the annotation file provided by HPO contains terms associated with every disease, and thus Equations (17) and (18) can be used to compute the similarities of diseases. When we calculate the similarities of phenotypic terms based on the DAG, w_e in Equation (15) is set to 0.7 as recommended in Li et al. (2011).

2.3. Prior Information

For diseases with only a few associated genes, the limited information would affect the performance of any computational algorithms. This problem is especially serious for diseases with no known associated genes. To solve this problem, we add some prior information for diseases with no known associations.

Specifically, given a disease i', $\mathbf{p}_{i'}$ is added to the i'-th row of the matrix A as prior information so that the shortage of known information can be alleviated. The j-th entry of $\mathbf{p}_{i'}$ is calculated by

$$p_{i'j} = \left(\sum_{i=1, i \ne i'}^{n} s^{d}_{ii'} a_{ij}\right) / \left(\sum_{i=1, i \ne i'}^{n} a_{ij}\right) \tag{19}$$

In our experiments, when cross-validation is used to evaluate the algorithm, the prior information is added to the i-th row of matrix A as long as one of the associated genes of disease i is left to test the model. Meanwhile, in the *de novo* study, prior information is also added to the diseases used for evaluation.

2.4. Data Sources

The disease-gene association data are downloaded from the Online Mendelian Inheritance in Man (OMIM) database (Amberger et al., 2014) in August 2018. The Morbid Map at OMIM contains nearly seventy-five hundred entries sorted alphabetically by disorder names. Each entry represents an association between a gene and a disease. Different entries are labeled with different tags ("(3)," "[]," and "?") which indicate their reliabilities. To obtain a reliable association dataset, based on (Goh et al., 2007), three steps were performed to preprocess the originally downloaded data. First, entries with the tag "(3)" are selected while others are abandoned. We adopt this strategy because diseases with tag "(3)" indicate that the molecular basis of these diseases is known and the associations are reliable, while entries with "[]" represent abnormal laboratory test values, and entries with "?" represent provisional disease-gene associations. Second, disease entries are classified into distinct diseases by merging disease subtypes based on their given disorder names. For instance, 17 entries of "Leigh syndrome" are merged into disease "Leigh syndrome," and the 19 complementary terms of "Lung cancer somatic" are merged into "Lung Cancer." Third, 74 diseases are removed because they are not annotated by any HPO terms. During the classification, string match was used to classify adjacent entries, followed by a manual verification. Finally, we obtain a dataset consisting of 4,770 associations between 1,537 diseases and 3,320 genes. Among the 1,537 diseases, 917 have only one associated gene (single-gene disease), while the rest diseases have at least two associated genes (multiple-gene disease).

The ontology data of genes and phenotypes are downloaded from the GO database (Ashburner et al., 2000; Consortium, 2017), and the HPO database (Köhler et al., 2018), respectively. The PPI network used in the competing algorithms is downloaded from the InWeb_InBioMap database (version 2016_09_12) (Li et al., 2016).

2.5. Evaluation Metrics

In this study, the algorithm is evaluated in two steps. In the first step, our dgManifold is compared with two competing algorithms: PCFM (Zeng et al., 2017) and Katz (Singh-Blom et al., 2013). PCFM is a matrix completion method which integrates disease similarities and gene similarities to predict disease-gene associations. Katz is a classic network-based method which uses Katz centrality to rank the disease-gene associations.

We choose these two algorithms because they are all multi-task algorithms which can predict all disease-gene associations as our dgManifold does. The AUC (area under of the receiver operating characteristic (ROC) curve) scores calculated from 3-fold cross-validation are used to compare these three algorithms.

ROC curve plots the true positive rate [TP/(TP+FN)] verses the false positive rate [FP/(FP+TN)] at different thresholds, and a larger AUC represents better overall performance. In this study, a true positive (TP) is a known disease-gene association (positive sample) predicted as a disease-gene association, while a false positive (FP) is a non-disease-gene association (negative sample) predicted as a disease-gene association. A false negative (FN) is a positive sample predicted as negative while a true negative (TN) is a negative sample predicted as negative. Since negative samples are not included in existing databases, we randomly select a set of unknown disease-gene pairs as negative samples. The number of negative samples is equal to that of positive samples. Considering that the selected negative samples may have small possibilities to be a real disease-gene association, the random selection was run for five times to generate 5 sets of negative samples. The final AUC score is the average score obtained from the 5 sets of samples.

During the cross-validation, the known disease-gene associations are split into 3 groups, and the algorithm is run for 3 rounds. In each round, one group of associations is regarded as unknown ($a_{ij} = 0$), while the rest two groups of associations are used to train the model. The prior information is recomputed during every round of the cross-validation. Considering that single-gene diseases would have no known associated genes if they are left for testing the model during the cross-validation, predicting disease genes for these diseases is similar to predict disease genes for a completely new disease. Thus, the three algorithms are compared on multiple-gene diseases and single-gene diseases separately. Additionally, to show the effect of the prior information, the AUC scores of our method without prior information are also calculated.

In the second step, the model is trained with all the known associations, and the geodesic distance between every unknown disease-gene pairs is calculated. To find out whether our new predictions are in concert with existing experimental studies, the top-10 predictions of two diseases, Lung Cancer and Bladder Cancer, are searched from the existing literature. In our dataset, Lung Cancer has 16 associated genes, and Bladder Cancer has 4 associated genes. We choose these two types of cancer because they are experimentally well studied which could better prove our results.

3. RESULTS

3.1. Model Parameters

In our study, several parameters affect the performance of the model. To obtain the optimal parameters, the grid search is conducted by searching k from $\{20, 30, 50, 100, 500, 800, 1,000, 1,200, 1,500\}$ and α from $\{0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$. β is set to be equal to α. The AUC score is used to determine whether the selected parameters are optimal. Finally, for multiple-gene diseases, the model performs best when $k = 30, \alpha = \beta = 0.2$, and for single-gene diseases, the optimal parameters are $k = 30, \alpha = \beta = 0.1$.

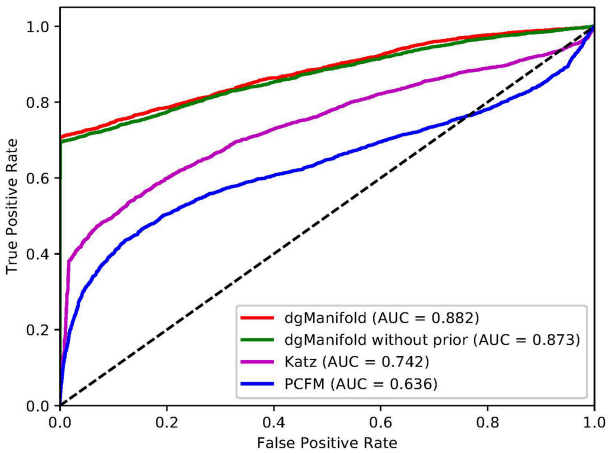

FIGURE 1 | ROC curves of the three competing algorithms on multiple-gene diseases.

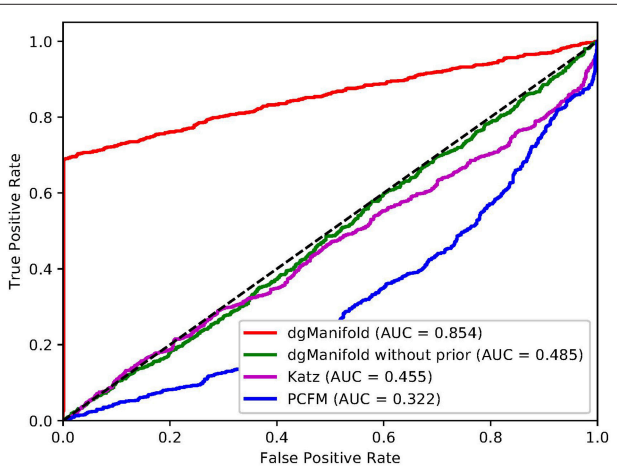

FIGURE 2 | ROC curves of the three competing algorithms on single-gene diseases.

3.2. Cross-Validation

Figures 1, **2** show the resulted ROC curves and AUC scores of the three competing algorithms on multiple-gene diseases and single-gene diseases, respectively. For multiple-gene diseases, our dgManifold achieves AUC score of 0.882 with prior information and 0.873 without prior information, while the AUC scores of Katz and PCFM are 0.742 and 0.636, respectively. For single-gene diseases, the AUC score of our dgManifold is 0.854 when prior information is used and 0.485 with no prior information, while the AUC scores of Katz and PCFM are 0.455 and 0.322, respectively. These results show that our method is superior to the competing methods in terms of the AUC scores.

It is worth noting that the AUC scores of all three algorithms are less than 0.5 when they are applied to single-gene diseases.

TABLE 1 | Top 10 predictions for lung cancer and bladder cancer.

Gene symbol	References
LUNG CANCER	
SEMA4A	
KCNK9	Sun et al., 2016
MYL2	Che et al., 2013
DENND5A	
HTRA1	Esposito et al., 2006
GABRA1	
ATP6AP1	Sabrkhany et al., 2018
KCTD17	
HFE	McLarty et al., 2008
BCS1L	
BLADDER CANCER	
PDYN	
DKC1	
SMAD3	Tong et al., 2018
MCC	
DMP1	Peng et al., 2015
MGP	
CALR	Kageyama et al., 2004
CASQ2	
SOX18	
GATM	

This is mainly because that single-gene diseases have no known associated genes during the cross-validation, and algorithms can only use disease similarities and association data of other diseases to perform the prediction. These data are not enough to generate accurate results, especially for supervised algorithms. Thus, prior information is necessary for the algorithm. In fact, the results of our experiments have shown that the prior information is beneficial to the prediction of disease-gene associations, especially when the diseases have no known associated genes.

3.3. *De novo* Study

In addition to AUC scores, we evaluate the performance of our dgManifold in predicting new disease-gene associations. Specifically, Lung Cancer and Bladder Cancer are selected, and prior information corresponded to these two diseases is added to matrix A. Then, all known disease-gene associations are used to train the model ($k = 30, \alpha = \beta = 0.2$), and the geodesic distance between all the unknown disease-gene pairs is calculated. For each of the two selected diseases, the unknown disease-gene pairs are ranked based on the geodesic distance in ascending order, and the top-10 predictions are searched from existing literature.

Table 1 shows the results of *de novo* studies. 5 out of 10 predicted genes have been experimentally confirmed as associated with Lung Cancer. Among these genes, KCNK9 is a potential therapeutic target (Sun et al., 2016). HTRA1 contributes to the tumor formation by inhibiting the TGF-beta pathway (Esposito et al., 2006). ATP6AP1 and MYL2 are two potential biomarkers (Che et al., 2013; Sabrkhany et al., 2018). Mutation of C282Y allele in HFE is associated with Lung Cancer (McLarty et al., 2008). Although SEMA4A is not proved to be associated with Lung Cancer yet, it is related to Lung Inflammation and Colorectal Cancer, and its role in Lung Cancer genesis might be discovered in the future (Iyer and Chapoval, 2019). For Bladder Cancer, 3 out of 10 genes have been experimentally verified. Among them, SMAD3 mediates epithelial-mesenchymal transition which affects the invasion and migration of Bladder Cancer (Tong et al., 2018). DMP1 is a tumor suppressor gene of Bladder Cancer (Peng et al., 2015). CALR is potential biomarker (Kageyama et al., 2004). These results show that our predictions are in concert with existing reports, and thus our dgManifold is valuable for predicting new disease-gene associations.

4. CONCLUSION

In this study, we have proposed dgManifold to predict disease-gene associations with manifold learning. Our dgManifold assumes that the distance between diseases and their associated genes should be shorter than that of other non-associated disease-gene pairs and maps the diseases and genes into a lower dimensional manifold based on known disease-gene associations, disease similarity and gene similarity. The prediction of new associations can be achieved by sorting the geodesic distance between unknown disease-gene pairs. The cross-validation results show that our model outperforms the competing algorithms in terms of AUC scores for both multiple-gene diseases and single-gene diseases. The further *de novo* studies also demonstrate that our dgManifold is valuable in predicting new disease-gene associations.

Note that dgManifold is only regularized by disease similarities and gene similarities at the current version, and the prior information is also obtained from the disease similarities. In the future, we can improve our method by regularizing the objective function with more types of data and computing the prior information with clinical evidences.

DATA AVAILABILITY

The datasets generated for this study and a reference implementation of the algorithm can be found in the GitHub repository of the study.

AUTHOR CONTRIBUTIONS

F-XW conceived this study. F-XW, PL, QX, P-JW, and BL discussed about the methods. PL implemented the algorithm, designed and performed the experiments. PL and F-XW wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work is supported in part by Natural Science and Engineering Research Council of Canada (NSERC) and China Scholarship Council (CSC).

REFERENCES

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2014). Omim. org: online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. *Nucl. Acids Res.* 43, 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). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25:25. doi: 10.1038/75556

Bolla, M. (2013). *Spectral Clustering and Biclustering: Learning Large Graphs and Contingency Tables*. Chichester: John Wiley & Sons.

Che, C.-L., Zhang, Y.-M., Zhang, H.-H., Sang, Y.-L., Lu, B., Dong, F.-S., et al. (2013). DNA microarray reveals different pathways responding to paclitaxel and docetaxel in non-small cell lung cancer cell line. *Int. J. Clin. Exp. Pathol.* 6:1538.

Chen, B., Li, M., Wang, J., and Wu, F.-X. (2014a). Disease gene identification by using graph kernels and markov random fields. *Sci. China Life Sci.* 57, 1054–1063. doi: 10.1007/s11427-014-4745-8

Chen, B., Shang, X., Li, M., Wang, J., and Wu, F.-X. (2015). "A two-step logistic regression algorithm for identifying individual-cancer-related genes," in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on (IEEE)* (Washington, DC), 195–200.

Chen, B., Shang, X., Li, M., Wang, J., and Wu, F.-X. (2016). Identifying individual-cancer-related genes by rebalancing the training samples. *IEEE Trans. Nanobiosci.* 15, 309–315. doi: 10.1109/TNB.2016.2553119

Chen, B., Wang, J., Li, M., and Wu, F.-X. (2014b). Identifying disease genes by integrating multiple data sources. *BMC Med. Genom.* 7:S2. doi: 10.1186/1755-8794-7-S2-S2

Chen, M., Liao, B., and Li, Z. (2018a). Global similarity method based on a two-tier random walk for the prediction of microrna'disease association. *Sci. Rep.* 8:6481. doi: 10.1038/s41598-018-24532-7

Chen, M., Lu, X., Liao, B., Li, Z., Cai, L., and Gu, C. (2016b). Uncover mirna-disease association by exploiting global network similarity. *PLoS ONE* 11:e0166509. doi: 10.1371/journal.pone.0166509

Chen, M., Peng, Y., Li, A., Li, Z., Deng, Y., Liu, W., et al. (2018b). A novel information diffusion method based on network consistency for identifying disease related micrornas. RSC Adv. 8, 36675–36690. doi: 10.1039/C8RA07519K

Consortium, G. O. (2017). Expansion of the gene ontology knowledgebase and resources. *Nucl. Acids Res.* 45, D331–D338. doi: 10.1093/nar/gkw1108

Esposito, V., Campioni, M., De Luca, A., Spugnini, E. P., Baldi, F., Cassandro, R., et al. (2006). Analysis of htra1 serine protease expression in human lung cancer. *Anticancer Res.* 26, 3455–3459.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8685–8690. doi: 10.1073/pnas.0701361104

Ham, J., Lee, D. D., and Saul, L. K. (2005). "Semisupervised alignment of manifolds," in *AISTATS*, eds R. G. Cowell and Z. Ghahramani (Society for Artificial Intelligence and Statistics), 120–127.

Iyer, A., and Chapoval, S. (2019). Neuroimmune semaphorin 4a in cancer angiogenesis and inammation: a promoter or a suppressor? *Int. J. Mol. Sci.* 20:124. doi: 10.3390/ijms20010124

Kageyama, S., Isono, T., Iwaki, H., Wakabayashi, Y., Okada, Y., Kontani, K., et al. (2004). Identification by proteomic analysis of calreticulin as a marker for bladder cancer and evaluation of the diagnostic accuracy of its detection in urine. *Clin. Chem.* 50, 857–866. doi: 10.1373/clinchem.2003.027425

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

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J.-P., et al. (2018). Expansion of the human phenotype ontology (hpo) knowledge base and resources. *Nucl. Acids Res.* 47, D1018–D1027. doi: 10.1093/nar/gky1105

Li, J., Gong, B., Chen, X., Liu, T., Wu, C., Zhang, F., et al. (2011). Dosim: an r package for similarity between diseases based on disease ontology. *BMC Bioinformatics* 12:266. doi: 10.1186/1471-2105-12-266

Li, T., Wernersson, R., Hansen, R. B., Horn, H., Mercer, J., Slodkowicz, G., et al. (2016). A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat. Methods* 14, 61–64. doi: 10.1038/nmeth.4083

Luo, P., Ding, Y., Lei, X., and Wu, F.-X. (2019a). deepdriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Front. Genet.* 10:13. doi: 10.3389/fgene.2019.00013

Luo, P., Tian, L.-P., Ruan, J., and Wu, F.-X. (2019b). Disease gene prediction by integrating ppi networks, clinical rna-seq data and omim data. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*. 16, 222–232. doi: 10.1109/TCBB.2017.2770120

McLarty, J., Ma, Y., Smith, M., and Glass, J. (2008). "Iron metabolism and the risk of lung and head and neck cancers," in *AACR Annual Meeting*, Vol. 68 (AACR) (San Diego, CA), 3923–3923.

Mordelet, F., and Vert, J.-P. (2011). Prodige: prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinformatics* 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, i60–i68. doi: 10.1093/bioinformatics/btu269

Peng, Y., Dong, W., Lin, T.-X., Zhong, G.-Z., Liao, B., Wang, B., et al. (2015). Microrna-155 promotes bladder cancer growth by repressing the tumor suppressor dmtf1. *Oncotarget* 6:16043. doi: 10.18632/oncotarget.3755

Sabrkhany, S., Kuijpers, M. J., Knol, J. C., Damink, S. W. O., Dingemans, A.-M. C., Verheul, H. M., et al. (2018). Exploration of the platelet proteome in patients with early-stage cancer. *J. Proteomics* 177, 65–74. doi: 10.1016/j.jprot.2018.02.011

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:e58977. doi: 10.1371/journal.pone.0058977

Sun, H., Luo, L., Lal, B., Ma, X., Chen, L., Hann, C. L., et al. (2016). A monoclonal antibody against kcnk9 k+ channel extracellular domain inhibits tumour growth and metastasis. *Nat. Commun.* 7:10339. doi: 10.1038/ncomms10339

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323. doi: 10.1126/science.290.5500.2319

Tong, H., Yin, H., Hossain, M. A., Wang, Y., Wu, F., Dong, X., et al. (2018). Starvation-induced autophagy promotes the invasion and migration of human bladder cancer cells via tgf-I1/smad3-mediated epithelial-mesenchymal transition activation. *J. Cell. Biochem.* 120, 5118–5127. doi: 10.1002/jcb.27788

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:e1000641. doi: 10.1371/journal.pcbi.1000641

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C.-F. (2007). A new method to measure the semantic similarity of go terms. *Bioinformatics* 23, 1274–1281. doi: 10.1093/bioinformatics/btm087

Yang, P., Li, X.-L., Mei, J.-P., Kwoh, C.-K., and Ng, S.-K. (2012). Positive-unlabeled learning for disease gene identification. *Bioinformatics* 28, 2640–2647. doi: 10.1093/bioinformatics/bts504

Zeng, X., Ding, N., Rodríguez-Patón, A., and Zou, Q. (2017). Probability-based collaborative filtering model for predicting gene–disease associations. *BMC Med. Genomics* 10:76. doi: 10.1186/s12920-017-0313-y

Conflict of Interest Statement: 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.

The reviewer MC declared a past co-authorship with one of the authors, BL, to the handling editor.