ISCI, Volume *19*

Supplemental Information

A Learning-Based Method for LncRNA-Disease Association Identification Combing Similarity Information and Rotation Forest

Zhen-Hao Guo, Zhu-Hong You, Yan-Bin Wang, Hai-Cheng Yi, and Zhan-Heng Chen

Supplemental Figures

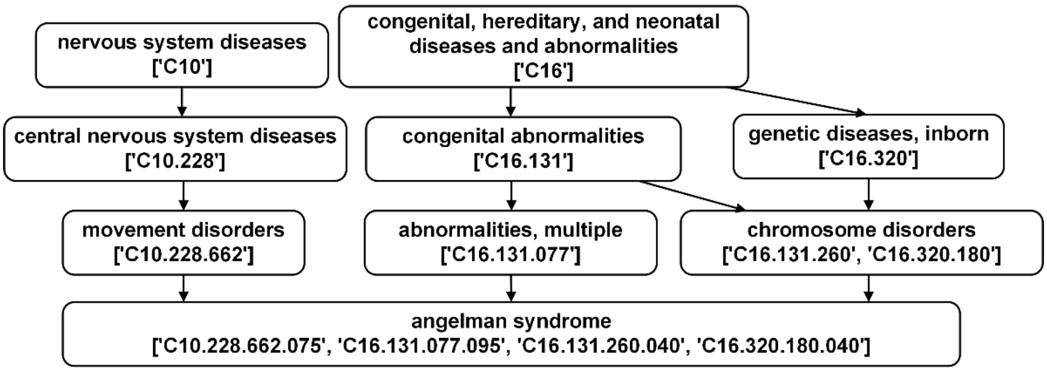
nervous system diseases ['C10']
congenital, hereditary, and neonatal diseases and abnormalities ['C16']
central nervous system diseases ['C10.228']
congenital abnormalities ['C16.131']
genetic diseases, inborn ['C16.320']
movement disorders ['C10.228.662']
abnormalities, multiple ['C16.131.077']
chromosome disorders ['C16.131.260', 'C16.320.180']
angelman syndrome ['C10.228.662.075', 'C16.131.077.095', 'C16.131.260.040', 'C16.320.180.040']

Figure S1. Construction of a disease's DAG. Related to Figure 1.

Transparent Methods

Data Collection

Known lncRNA-disease associations were downloaded from the LncRNADisease database (v2017) (Geng Chen *et al.*, 2012). which contained 2947 experimentally validated lncRNA–disease associations between 914 lncRNAs and 329 diseases. After deleting duplicate data caused by multiple experiment validations, we selected 1765 associations involving 881 lncRNAs and 328 diseases. The lncRNA–disease associations can be visualized as a network, the nodes represent specific lncRNA or disease, the edges connect a lncRNA to a disease. To extract positive and negative samples from this network, all experimentally validated lncRNA-disease pairs (i.e. 1765 lncRNA-disease pairs) constitute the golden standard positive dataset. The remaining edges of this network can be considered as nonassociation, and the corresponding lncRNA and disease can be collected as negative samples. In this paper, we followed previous method collect negative samples with the same size as positive samples using random selection (Ben-Hur and Noble, 2005). Although false negative samples may be included in the negative dataset, considering that the size of data collected only accounts for a small part of the whole network, the impact can be neglected. This can be treated as an issue with unbalanced data set processing, i.e. the process of down-sampling from negative sample (unlabeled sample). The picked negative samples are a very small percentage which only accounts for 0.61% (1765/ (881*328)-1765) and then a total of 3530 lncRNA–disease pairs were collected.

LncRNADisease v2017 and LncRNADisease v2012 are 2 different versions of the same database, of which v2012 is a true subset of v2017. The previous proposed by Chen *et al.* is to train and test based on lncRNADisease v2012, in order to ensure the fairness of the experiment, the 293 lncRNA–disease associations in version 2012 involving 118 lncRNAs and 167 diseases were also collected to constitute positive set. The negative set was constituted by the method mentioned above. As a result, the entire dataset consists of 586 lncRNA–disease pairs, of which half is from the positive samples and the other is from the negative samples.

Disease MeSH Descriptors And Directed Acyclic Graph

Medical Subject Headings (MeSH) is an authoritative subject vocabulary compiled by the National Library of Medicine, which provide a hierarchically-organized terminology for indexing and cataloging of various diseases. Each disease can be represented as a Directed Acyclic Graph (DAG) by the information provided by MeSH, which is described as follows: *DAG(D)= (D,* N_D*,* E_D*)*. Here, *D* represents specific diseases, N_D is node set that contains all disease in *D*'s DAG. E_D represents the relationship between the nodes in *D*'s DAG. Specific examples are shown in Figure S1.

Disease Semantic Similarity Matrix 1

We computed disease semantic similarity based on DAG. the contribution of disease *t* to the semantic value of disease *D* is defined as:

$$\begin{cases} D1_D(t) = 1 & if\ t = D \\ D1_D(t) = \max\{\Delta * D1_D(t') | t' \in children\ of\ t\} & if\ t \neq D \end{cases} \tag{1}$$

Where Δ denotes the semantic contribution decay factor and equals to 0.5. In the DAG on disease *D*, disease *D* is at the top, and its contribution to its semantic value is defined as 1. The semantic contribution of the next layer to disease *D* is equal to the contribution of the layer disease to itself multiplied by the semantic contribution attenuation factor. Therefore, the semantic value of disease A can be defined as follows:

$$D1(D) = \Sigma_{t\in N_D} D1_D(t) \tag{2}$$

The measure of disease similarity can be derived from set theory. The similarity between two diseases is calculated by the following：

$$DS1(i,j) = \frac{\sum_{t\in N_i \cap N_j}(D1_i(t)+D1_j(t))}{DV1(i)+DV1(j)} \tag{3}$$

Disease Semantic Similarity Matrix 2

The above disease similarity measure only considers local information and the intersection between two sets. Some scholars considered that it was one-sided and incomplete. Another semantic similarity measure method is used to complement the previous one. Inspired by information theory, the method suggests that diseases that often occur in DAGs should have a higher status and contribute more to other diseases (Xing Chen *et al.*, 2015, Xuan *et al.*, 2013). The new disease contribution values are measured as follows:

$$D2_D(t) = -\log(\frac{the\ number\ of\ DAGs\ including\ t}{the\ number\ of\ disease}) \tag{4}$$

The sum of the contributions of all nodes in the DAG of disease *D* is as follows:

$$DV2(\mathrm{D}) = \Sigma_{t\in N_D} D2_D(t) \tag{5}$$

The semantic similarity value could be calculated just like *DS1*:

$$DS2(i,j) = \frac{\sum_{t\in N_i \cap N_j}(D2_i(t)+D2_j(t))}{DV2(i)+DV2(j)} \tag{6}$$

Gaussian Interaction Profile Kernel Similarity For Diseases And LncRNA

In order to overcome the gap caused by the lack of MeSH information, the idea of collaborative filtering is employed to construct the third similarity matrix. In this paper, we first construct an adjacency matrix using the association data of lncRNA and disease. The columns of the matrix represent lncRNA and the rows represent diseases. Then, the Radial Basis Function (RBF) Gaussian kernel function was applied to adjacency matrix to obtain similarity matrix of disease (van Laarhoven,Nabuurs and Marchiori, 2011, Xing Chen *et al.*, 2018). The similarity defined by the Gaussian interaction profile kernel is as follows:

$$DG(i,j) = exp(-\alpha_d \left\| d_i - d_j \right\|^2) \tag{7}$$

Where d_i and d_j are *i*-th row and the *j*-th row of the adjacency matrix, respectively. α_d that is the weight factor used to regulate the kernel bandwidth, can be defined as follows:

$$\alpha_d = \alpha'_d(\frac{1}{nd}\sum_{i=1}^{nd}\|d_i\|^2) \tag{8}$$

Here, *nd* is the number of the diseases, the parameter α'_d is set to 0.5 empirically.
Analogous to the Gaussian similarity calculation method of disease, the Gaussian similarity of RNA is calculated by the same method. Formula 7 is replaced by Formula 9:

$$RS(i,j) = RG(i,j) = exp(-\alpha_r \left\| r_i - r_j \right\|^2) \tag{9}$$

Where r_i and r_j are *i*-th column and the *j*-th column of the adjacency matrix, respectively. α_r is the weight factor used to regulate the kernel bandwidth, defined by Formula (10):

$$\alpha_r = \alpha'_r (\frac{1}{nr} \sum_{i=1}^{nr} \|r_i\|^2) \tag{10}$$

Here, *nr* is the number of the diseases, the parameter α'_r is set to 0.5 empirically. After constructing the similarity matrix based on adjacency matric *A*, the representation vector of each lncRNA or disease will not change with cross-validation. The impact of this on the results will be discussed in a follow-up work.

Construction of Feature Vectors for Disease and lncRNA

Disease Semantic Similarity Matrix and Disease Gaussian Interaction Profile Kernel Similarity are two different types of information so neither is redundant. One of the above is often imperfect, to get a complete disease similarity matrix *DS*, we integrated disease semantic similarity matrix 1, disease semantic similarity matrix 2 and disease Gaussian interaction profile kernel similarity matrix by formula 11.

$$DS(i,j) = \begin{cases} \frac{DS1(i,j)+DS2(i,j)}{2} & if\ i\ and\ j\ have\ semantic\ similarity \\ DG(i,j) & otherwise \end{cases} \tag{11}$$

The row or column of matrix *DS* is regarded as the feature vectors of disease. Similarly, the row or column of matrix RS is regarded as the feature vectors of lncRNA. It's remarkable that all similarity matrices are symmetric matrices.
The *i*-th disease can be represented by the *i*-th row of the matrix *DS*:

$$DS_{i*} = (DS_{i1}, DS_{i2}, \dots, DS_{i328}) \tag{12}$$

The *j*-th disease can be represented by the *j*-th row of the similarity matrix RS:

$$RS_{j*} = (DS_{j1}, DS_{j2}, \dots, DS_{j881}) \tag{13}$$

The association consists of the *i*-th disease and the *j*-th lncRNA can be represented by the following vector:

$$Pair_{ij} = (DS_{i*}, RS_{j*}) = (DS_{i1}, DS_{i2}, \dots, DS_{i328}, RS_{j1}, RS_{j2}, \dots, RS_{j881}) \tag{14}$$

Each positive sample is given a label 1 and each negative sample is given a label 0.

AutoEncoder

Each association can be abstracted into a 1209-dimensional vector through the above step. Training set and test set consisting of thousands of such vectors take up a lot of storage space, which is not conducive to the training of classifiers. In order to reduce noise and improve feature quality, the autoencoder was used to obtain the optimal feature space from the original feature (Yi *et al.*, 2018). The autoencoder consists of an encoder and decoder. The coding part is responsible for compressing input data and the decoding part is responsible for restoring initial input. The main steps are as follows:
$f(x)$ is the activation function of the encoder, $g(h)$ is the activation function of the decoder. It will generally do this using a sigmoid function:

$$h = f(x) \coloneqq S_f(Wx + p) \tag{15}$$

$$y = g(h) \coloneqq S_g(W'x + q) \tag{16}$$

Here, we choose the sigmoid function as the activation function:

$$S_f(t) = S_g(t) = \frac{1}{1+e^{-t}} \tag{17}$$

The difference between *x* and *y* can be described by a reconstruction error function which is defined as follows:

$$L(x,y) = -\sum_{i=1}^{n}[x_i\, log(y_i) + (1-x_i)log(1-y_i)] \tag{18}$$

Through the above the loss function can be defined as follows:

$$Loss = \sum_{i=1}^{n} L(x_i, g(f(x_i))) \tag{19}$$

Therefore, the most suitable argument was obtained by minimizing the loss function. We can use *h* instead of *x* to represent the original vector. In this study, we used the keras library to implement the autoencoder and set the parameters batch size and epoch to 128 and 100, respectively.

RotationForest

Building an integrated learning algorithm by merging multiple models helps to achieve better prediction effects (Wang *et al.*, 2017, Li *et al.*, 2017). The idea of ensemble learning is to solve the defects and limitations inherent in the model of a single model by integrating more models. In 1990, Schapire analyzed and proved the equivalence between the weak learning algorithm and the strong learning algorithm based on the PAC (Probably Approximately Correct) learning model (Schapire, 1990). Since then, it has gradually attracted the focus of a wide range of scholars and shown outstanding effects on many classification or regression tasks. Assemble learning classifiers have stronger generalization capabilities and simpler parameter adjustments than traditional single models. Rotation Forest here was chosen to carry out the prediction. The Rotation Forest algorithm is based on the idea of feature transformation and focuses on improving the variability and accuracy of the base classifier (Rodriguez,Kuncheva and Alonso, 2006). Suppose $x = [x_1, x_2, \dots, x_n]^T$ represents the sample with *n* features. A matrix *X* of *N*n* to represent a training sample set with *N* data records. $y = [y_1, y_2, \dots, y_n]^T$ represents the corresponding sample class label in the training sample set *X. F* represents the attribute set, and *D1*, *D2*, …, *DL* represent *L* base classifiers. The main steps are as follows:

(1) The attribute set *F* is randomly divided into *K* sub-attribute sets, and each sub-attribute set contains about *M* = *n/K* attributes.

(2) Denote by $F_{i,j}$ the *j*-th subset of features for the training set of classifier *Di*. Then a bootstrap subset of objects is drawn with the size of 75% of the dataset to form a new training set, which is denoted by X'_{ij}. Using the selected subset of samples to transform the sub-attribute set in $F_{i,j}$, the principal component analysis (PCA) is used to obtain *Mj* principal components: $a_{ij}^1, a_{ij}^2, \dots, a_{ij}^{M_j}$.

(3) Repeat step 2 to store the obtained *K* principal component coefficients into a coefficient matrix *Ri.* According to the order of the original data attribute set, rearrange the matrix *Ri* to obtain R'_i, then the training set will be transformed into XR'_i. The base classifier *Di* will be trained on the new training set.

$$R_i = \begin{bmatrix} [a_{ij}^1, a_{ij}^2, \dots, a_{ij}^{M_1}] & [0] & \dots & [0] \\ [0] & [a_{ij}^1, a_{ij}^2, \dots, a_{ij}^{M_2}] & \dots & [0] \\ \vdots & \vdots & \vdots & \vdots \\ [0] & [0] & \dots [a_{ij}^1, a_{ij}^2, \dots, a_{ij}^{M_K}] \end{bmatrix} \tag{20}$$

(4) After the above steps, *L* base classifiers can be obtained. The final prediction category is determined with maximum confidence.

Supplemental References

Ben-Hur, A. and Noble, W.S. (2005) 'Kernel methods for predicting protein–protein interactions'. *Bioinformatics,* 21 (suppl_1), pp. i38-i46.

Chen, G. *et al.* (2012) 'LncRNADisease: a database for long-non-coding RNA-associated diseases'. *Nucleic acids research,* 41 (D1), pp. D983-D986.

Chen, X. *et al.* (2018) 'Novel Human miRNA-Disease Association Inference Based on Random Forest'. *Molecular Therapy-Nucleic Acids,* 13 568-579.

Chen, X. *et al.* (2015) 'Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity'. *Scientific reports,* 5 11338.

Li, J.-Q. *et al.* (2017) 'PSPEL: in silico prediction of self-interacting proteins from amino acids sequences using ensemble learning'. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB),* 14 (5), pp. 1165-1172.

Rodriguez, J.J., Kuncheva, L.I. and Alonso, C.J. (2006) 'Rotation forest: A new classifier ensemble method'. *IEEE transactions on pattern analysis and machine intelligence,* 28 (10), pp. 1619-1630.

Schapire, R.E. (1990) 'The strength of weak learnability'. *Machine learning,* 5 (2), pp. 197-227.

van Laarhoven, T., Nabuurs, S.B. and Marchiori, E. (2011) 'Gaussian interaction profile kernels for predicting drug–target interaction'. *Bioinformatics,* 27 (21), pp. 3036-3043.

Wang, L. *et al.* (2017) 'An ensemble approach for large-scale identification of protein-protein interactions using the alignments of multiple sequences'. *Oncotarget,* 8 (3), pp. 5149.

Xuan, P. *et al.* (2013) 'Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors'. *PloS one,* 8 (8), pp. e70204.

Yi, H.-C. *et al.* (2018) 'A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information'. *Molecular Therapy-Nucleic Acids,* 11 337-344.