Slide -2 -Selected Points

* lncRNAs-Long non-coding RNAs (lncRNAs) are RNA molecules longer than 200 nucleotides that do not encode proteins. They play diverse roles in cellular processes, including epigenetic regulation, transcriptional control, and post-transcriptional regulation
* potential association between lncRNA and diseases-LncRNAs play an important role in various biological processes. The identification of disease-related lncRNAs is of great signifi cance for understanding the pathogenesis of diseases at the lncRNA level and contributes to disease prevention and treatment.
* characteristics of the associated data were seldom explored--Long non-coding RNAs (lncRNAs) have been shown to play a regulatory role in various processes of human diseases. However, lncRNA experiments are inefficient, time-consuming and highly subjective, so that the number of experimentally verified associations between lncRNA and dis eases is limited. In the era of big data, numerous machine learning methods have been proposed to predict the potential association between lncRNA and diseases, but the characteristics of the associated data were seldom explored. In these methods, negative samples are randomly selected for model training and the model is prone to learn the potential positive association error, thus affecting the prediction accuracy.
* PU positive and unlebeled
* COPTLDA-The COPTLDA model uses the following strategies: (1) The two-step strategy is used to select more likely negative samples. The selected negative samples and some known positive samples are used to train the first sub-model and record the evaluation indicators of the sub-model. (2) Without putting back the negative samples selected in the last round, the process of selecting negative samples and training sub-models is repeated constantly, and the evaluation index of each sub-model is recorded. (3) Count the number of negative samples that have been selected, sort by the number of selection times, and select a certain number of negative samples with higher ranking. The selected negative samples are trained together with the positive samples to obtain the final model, and then all unknown associations are predicted.
* **AUC-** In this paper, the AUC value of the COPTLDA model reached the maximum value of 0.9348, indicating that COPTLDA performed better than the other five models and had the higher confidence in predicting the potential as sociation between lncRNA and diseases.
* **What others did(3 calculations examples)-** traditional biological experiments are time-consuming and costly and it is difficult to make achievements from numerous data, so the number of experimentally verified lncRNAs associated with diseases is limited. Therefore, it is necessary to predict the potential associations between lncRNAs and diseases with computational methods. Currently, the calculation methods of predicting the po tential associations between lncRNAs and diseases can be roughly divided into the following three categories. Firstly, based on traditional calculation methods, Chen et al. constructed a semi-supervised learning framework named LRLSLDA to predict potential disease-related lncRNAs [15]. Lu et al. used the induction matrix completion model SIMCLDA to predict the associations between lncRNA and diseases [16]. Secondly, based on deep learning methods, Zeng et al. used a neural network model based on deep matrix factorization to predict potential lncRNA-disease associations [17]. Thirdly, with the calculation results from other biological infor mation, the correlations between lncRNA and diseases were predicted. Chen et al. integrated the miRNA-disease correlation with lncRNA-miRNA interaction to predict the potential associations between lncRNA and diseases [18].
* **Datasets(**lncRNADisease Database V2.0)- wo datasets were downloaded from lncRNADisease Database V2.0 [26]. One dataset contained the experimentally verified as sociations between lncRNAs and diseases. After removing the information of other species and repeated entries, 581 lncRNAs and 215 diseases were obtained as the training model. The statistics are shown in Table 1. From the lncRNA-disease association matrix, 1477 positive samples were obtained. In the obtained samples, 60% of positive samples were used as the training set and 30% of positive samples were used as the test set to determine the accuracy of the model. The remaining 10% of positive samples were used as the test set to determine the recall rate of the model. The other dataset contained the association information between lncRNA and diseases predicted by other models and was used as the verification cases for case study. The two datasets were also downloaded from the RNADisease Database V4.0 [27]. One dataset was verified by experiments and the other was predicted by other models (Table 2). The two datasets were also used to evaluate the case study results.

Slide-3-The Methodology

2-step-Step 1: Train a model on the positive examples and a set of unlabeled examples.

Step 2: Use the trained model to predict labels for the remaining unlabeled examples.

COPTLDA- In the basic model of COPTLDA, a fully connected neural network is used to train the data. The row vectors and column vectors of the association matrix are respectively used as the inputs of a three-layer neural network and a four-layer neural network. For the intersection of the row and column, the value is masked and replaced with 0. The basic model diagram is shown in Fig. 1.

Slide-4-strategy of COPTLDA model

in COPTLDA, each lncRNA is represented as the row L i of the association matrix and each disease is represented as the column D of the association matrix, so that the association matrix R is obtained. Each lncRNA has a one-to-one association with all kinds of diseases. When the i-th lncRNA is associated with the j-th disease, Rij = 1 is a positive sample; otherwise, it is an unknown sample. In ssociation matrix, 1477 positive samples and 123,438 unlabeled samples are obtained. Then, 60% of the positive samples, 886 cases, are taken as the training set and 10% of positive samples, 147 cases, are doped into the unknown sample set after changing their labels. Finally, 30% of the samples, 444 cases, are taken as the test set. he quantity of the data is small and negative samples are randomly selected, so the obtained model is rather rough. However, it is still considered to have a certain predictive ability. At this time, the untrained unknown samples are predicted with the first model and then the most possible negative samples, the highest ranked top 886 samples, are selected as the negative samples of the second model. Then, the 886 negative samples are combined with the 886 positive samples to train the second model. Then, the negative samples in the previous round of training are added into the unknown sample pool to repeat the training process of the second model. In the process, the most possible negative samples are continuously selected and combined with the fixed positive samples to train the models. Each obtained model is verified with the test set to evaluate its predictive ability and then the remaining unlabeled samples are predicted to obtain the predicted value. The negative samples selected in all the training models are statistically analyzed, and the highest ranked top 886 samples are selected as negative samples for training according to the descending order of selection times. In this way, the final model was obtained.

Slide-5-conclusion

In this paper, the two-step strategy in PU learning was adopted to select the training data and train the model. Unknown positive samples mixed with unlabeled samples for training affected the predictive performance of the model, indicating that PU learning method performed better in predicting the associations between lncRNA and diseases. Then, COPTLDA model was compared with the single model. The results of multiple parameters indicated that the two-step strategy in training data could greatly improve the predictive performance of lncRNA-disease association model. Based on the statistics of the first 60 sub-models, the top 886 samples with the largest probability of being selected were obtained and the higher accuracy and recall rate were finally realized, thus further confirming the effectiveness and superiority of the two-step sampling strategy.

In COPTLDA, only the association between lncRNAs and diseases is used to construct the association matrix without using other biological data. However, other kinds of biological data are also significant for the prediction of association between lncRNAs and diseases. In future studies, it is necessary to integrate various biological data for joint prediction and develop more effective algorithms to adapt to complex data relationships.