Research Paper

FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model

Xing Chen[1,*], Yu-An Huang[2,*], Xue-Song Wang[1], Zhu-Hong You[3], Keith C.C. Chan[2]

[1]School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou, China

[2]Department of Computing, Hong Kong Polytechnic University, Hong Kong

[3]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China

[*]The first two authors should be regarded as joint First Authors

Correspondence to: *Xing Chen,* ***email:*** *xingchen@amss.ac.cn*
Zhu-Hong You, ***email:*** *zhuhongyou@gmail.com*

Keywords: *lncRNAs, functional similarity, disease, fuzzy measure, directed acyclic graph*

Received: *April 05, 2016* ***Accepted:*** *May 29, 2016* ***Published:*** *June 14, 2016*

ABSTRACT

Accumulating experimental studies have indicated the influence of lncRNAs on various critical biological processes as well as disease development and progression. Calculating lncRNA functional similarity is of high value in inferring lncRNA functions and identifying potential lncRNA-disease associations. However, little effort has been attempt to measure the functional similarity among lncRNAs on a large scale. In this study, we developed a Fuzzy Measure-based LNCRNA functional SIMilarity calculation model (FMLNCSIM) based on the assumption that functionally similar lncRNAs tend to be associated with similar diseases. The performance improvement of FMLNCSIM mainly comes from the combination of information content and the concept of fuzzy measure, which was applied to the directed acyclic graphs of disease MeSH descriptors. To evaluate the effectiveness of FMLNCSIM, we further combined it with the previously proposed model of Laplacian Regularized Least Squares for lncRNA-Disease Association (LRLSLDA). As a result, the integrated model, LRLSLDA-FMLNCSIM, achieve good performance in the frameworks of global LOOCV (AUCs of 0.8266 and 0.9338 based on LncRNADisease and MNDR database) and 5-fold cross validation (average AUCs of 0.7979 and 0.9237 based on LncRNADisease and MNDR database), which significantly improve the performance of previous classical models. It is anticipated that FMLNCSIM could be used for searching functionally similar lncRNAs and inferring lncRNA functions in the future researches.

INTRODUCTION

In recent years, the observation from the Next Generation Sequencing (NGS) project indicates that the number of non-coding sequences accounts for a large portion (more than 98%) of the complete human genome. A great number of non-coding RNAs (ncRNAs) are discovered which do not encode proteins, especially long noncoding RNAs (lncRNAs). LncRNA is the heterogeneous ncRNAs which consist of more than 200 nucleotides. According to the relative positions to the coding genes, there are five subgroups of lncRNAs (i.e. sense, antisense, bidirectional, intronic, and intergenic) [1–3]. As a traditional viewpoint from central dogma of molecular biology, the genetic information is mainly stored in the protein-coding genes. The special characters of lncRNAs, low expression level and high tissue-specific pattern, once leaded to a misconception that lncRNAs are purely "transcriptional noise" [4–6]. However, increasing evidences from biological experiments have shown that lncRNAs carry out various crucial functions, which clearly contradict to the traditional viewpoint. LncRNAs cover a wide range of functions of modulating gene expression at the epigenetic, transcriptional, and post-transcriptional levels [2]. Specifically, lncRNAs get involved in diverse biological processes, such as chromatin modification, cell differentiation and proliferation, RNA progressing, and cellular apoptosis [7–14]. For example, HOTAIR was verified as scaffold to bind histone modifiers, PRC2, and the LSD1 complex, carrying out functions of histone

modifications control and gene expression regulation [15]. Xist also proved to be a spliced and polyadenylated lncRNA which binds and recruits PRC2 to initiate X chromosome inactivation [16]. UCA1 is discovered to regulate the expression of several genes which are involved in tumorigenesis and embryonic development [17].

According to the new theory of competing endogenous RNA, lncRNAs interact with a wide range of RNA molecules and play a more important role in pathological conditions than previously expected [18]. Based on the existing experimental observations, lncRNAs emerge as important drivers of diverse diseases and modulate gene expression at several levels. The competing endogenous RNAs are considered to be involved in a large-scale regulatory network across the transcriptome, playing important roles in pathological conditions. Increasing evidence indicates that the lncRNA dysfunction is clearly associated with the development and progression of a wide range of diseases, such as diabetes [19, 20], HIV [21], breast cancer [22, 23], lung cancer [24, 25], colon cancer [26], prostate cancer [20], leukemia [27], and ovarian cancer [28]. Some lncRNAs are considered as biomarkers for specific diseases. For example, M41 was verified as a biomarker candidate for the prognosis of ER-associated breast cancers [29]. It was identified to be associated with preclinical cancer phenotype, tamoxifen resistance promotion, and poor outcomes in clinical samples. Except for M41, ANRIL was recently regarded as a potential prognostic biomarker in gastric cancer [30]. It recruits and binds to PRC2 and generally upregulates in human gastric cancer tissues. SNHG18 was also identified as a predictive biomarker for bladder cancer. The relevant study shows that the knockdown of SNHG18 can lead to decreased expression of several luminal PPARγ target genes including uroplakins and fibroblast growth factor receptor-3 (FGFR3), and further boost the development of muscle-invasive bladder cancer [31]. Even though the mechanisms of complex diseases are still unclear, the biological data collected from experimental discoveries is expected to shed light on the roles of lncRNAs in disease development and progression.

With the rapid development of experimental techniques and computational studies for lncRNA discovery, a large number of lncRNAs in various eukaryotic organisms have been discovered since H19 and XIST were first discovered in the early 1990s [32–34]. Many lncRNA-related biological datasets have been built and stored in some publicly available databases, such as NRED [35], NONCODE [36] and LncRNAdb [11]. However, the number of lncRNA-disease associations recorded in these databases is still limited. Even though more and more lncRNA functions have been identified by the disease-related studies, it is unrealistic to use experimental approaches to identify the functions of lncRNAs due to the high cost of time and money. In recent years, lncRNA-disease association identification and lncRNA function prediction have become hot research subjects attracting an increasing number of researchers. Based on the assumption that similar lncRNA functions are associated with the involvement in similar diseases, some computational models have been reported to calculate lncRNA functional similarity or identify lncRNA-disease associations [37–41]. These methods are mainly based on the recorded lncRNA-disease association networks. Due to the rapid computational process and the integration of various types of biological data, computational models can serve as a perfect complement for biological experiments by calculating lncRNA functional similarity or selecting the most probable candidate for further experimental validation. Therefore, computational methods for lncRNA-related studies are drawing increase attentions and expected to decrease the time and cost of experimental approaches [42–46]. In conclusion, developing computational models for calculating lncRNA functional similarity could not only boost the understanding on disease mechanism at lncRNA level, but also accelerates the process of new biomarker identification for drug discovery, disease diagnosis, treatment, prognosis, and prevention.

Currently, measuring lncRNA functional similarity is still a challenging task. The difficulties lie in the undocumented structural features and weak conservation of lncRNA structures as well as the lack of reliable network model for uncovering the relationships between lncRNAs and other molecules [47]. However, accumulating studies provide the clues that mutations in the primary structure, secondary structure, expression levels and cognate RNA-binding proteins of lncRNAs underlie their biological functions [9, 48, 49]. Therefore, based on this fact, there are some computational models have been proposed for measuring lncRNA functional similarity by utilizing diverse expression profiles of lncRNAs. For example, Bellucci *et al.* proposed a method for measuring lncRNA functions by considering their potential associated proteins which were predicted based on sequence information [50]. In addition, Chen *et al.* reported a novel measurement of integrated lncRNA functional similarity by combining lncRNA expression similarity with lncRNA Gaussian interaction profile kernel similarity [37]. In this study, the lncRNA expression similarity was defined as the Spearman correlation coefficient between the expression profiles of each lincRNA pairs. Based on this measurement for lncRNA functional similarity, the first lncRNA-disease association prediction model of LRLSLDA was proposed. For lncRNA functional measurement, Liao *et al.* proposed a model based on coding-noncoding gene co-expression network which was constructed by re-annotating probes of the Affymetrix Mouse Genome 430 2.0 Array [51]. However, this kind of expression data suffers from the strong dependence on the design of the probes. Xiao *et al.* also recently proposed a model by mapping protein-

coding genes onto human PPI network based on Bayesian network [52]. In this work, lncRNA functional similarity could be measured by mining highly connected molecules in the network. In recent years, it's worth noting that the involvement of lncRNAs in a wide range of diseases could be far more prevalent than previously considered. Based on the assumption that lncRNAs which get involved in similar diseases are more likely to share the similar biological functions and vice versa, some computational models for calculating lncRNA functional similarity have been proposed. For example, Sun *et al.* proposed a calculation model for lncRNA functional similarity by adopting DOsim, an R package proposed by Wang *et al.*, to measure semantic similarity between disease directed acyclic graphs (DAGs) [53, 54]. Chen *et al.* developed the model of LFSCM for measuring lncRNA functional similarity [38]. This model was mainly based on the combination of microRNA (miRNA)-disease associations with lncRNA-miRNA interactions. Recently, Chen *et al.* further reported a computational method named LNCSIM for calculating lncRNA functional similarity by combining the experimentally confirmed lncRNA-disease associations and the information of DAGs constructed by disease Mesh descriptors [55]. The effectiveness of this kind of method mainly depends on the term-term similarity measure on the collocations derived from DAGs. In the model of LNCSIM, the traditional Jaccard similarity measure was adopted for calculating the semantic similarity between each disease pairs. With the continuous emergence of new clinical discoveries, this kind of method which uses know lncRNA-disease association can greatly benefit from the wealth of observation data. It is a current trend for network-based prediction models to consider additional biological knowledge [56–58].

In this study, we developed Fuzzy Measure-based LNCRNA functional SIMilarity calculation model (FMLNCSIM) based on the assumption that similar diseases tend to be involved with functionally similar lncRNAs and vice versa. The model of FMLNCSIM generally consists of two parts. In the first part, the terms of MeSH descriptors in a combined set describing two diseases would be considered as "information sources" which were used for calculating the similarities among diseases. Specifically, the semantic similarities of diseases were computed by combining the concepts of information content and fuzzy measure. In the second part, the functional similarity of two lncRNAs would be calculated based on the semantic similarities of their associated disease groups. To further evaluate the effectiveness of FMLNCSIM, we used the calculation results computed from FMLNCSIM to predict the lncRNA-disease associations based on the model of LRLSLDA which was proposed in the previous work. The performance of new integrated model would be directly influenced by FMLNCSIM and therefore reflected effectiveness of FMLNCSIM. We used two evaluation frameworks, namely global leave-one-out cross validation (LOOCV) and 5-fold cross validation, to evaluate the performance of FMLNCSIM. When exploring the LncRNADisease and MNDR databases by adopting the global LOOCV method, we obtained improved performance with AUCs of 0.8266 and 0.9338, respectively. By adopting 5-fold cross validation method based on LncRNADisease and MNDR databases, FMLNCSIM yielded average AUCs of 0.7979 and 0.9237, respectively. In addition, we further verified the top 10 prediction lists of acute myeloid leukemia and lung cancer by checking the updates of relevant databases and recent experimental literatures. As a result, six of them were confirmed. These reliable results demonstrate that FMLNCSIM is feasible and promising to quantify functional similarity of lncRNAs as well as to be combined with similarity-based computational models for lncRNA-disease association prediction.

RESULTS

Model design

FMLNCSIM is a computational model for calculating the functional similarity of lncRNAs by using the information of known lncRNA-disease associations and diseases DAGs (See Figure 1 and 2). It is mainly based on the assumption that functionally similar lncRNAs tend to be involved in similar disease and vice versa. The performance improvement of FMLNCSIM mainly comes from the combination of the concept of information content and the fuzzy set theory. Information content of disease terms in DAGs helps to retain their specificity and fuzzy measure is expected to lead to a more accurate similarity measurement based on the disease sets. The concept of information content has been adopted by some of previous researches related with gene ontology (GO) terms [59–61]. For example, Yu *et al.* have proposed an R package for semantic similarity among GO terms and gene products by introducing the concept of information content [59]. Fröhlich *et al.* have also used this concept and developed GOSim package for measuring functional similarity of gene products [60].

The similarity measure is one of useful tools for the degree of similarity between objects and is deeply studied in the fields of physical anthropology, numerical taxonomy, ecology, information retrieval, psychology, citation analysis, and automatic classification. Various term-term similarity measures have been proposed for expressing the degree of similarity of sets. For example, the Jaccard and Dice similarity measures, which were proposed in 1901 and 1945, respectively, have been widely used [62, 63]. However, it is shown that these functions have inherent limitations that they are only treated as discrete with loss of information and fail to measure the similarity between the trapezoidal intuitionistic fuzzy numbers (TIFNs) which should be

treated as continuous [62, 64]. Information available is sometimes vague, inexact or insufficient and fuzzy set theory proves to be ideally suited for solving these problems [65]. Since the scale of recorded MeSH descriptors associated with a specific disease greatly depend on the research degree on it, this bias in disease DAGs can cause partial, insufficient or redundant information for calculating lncRNA functional similarity

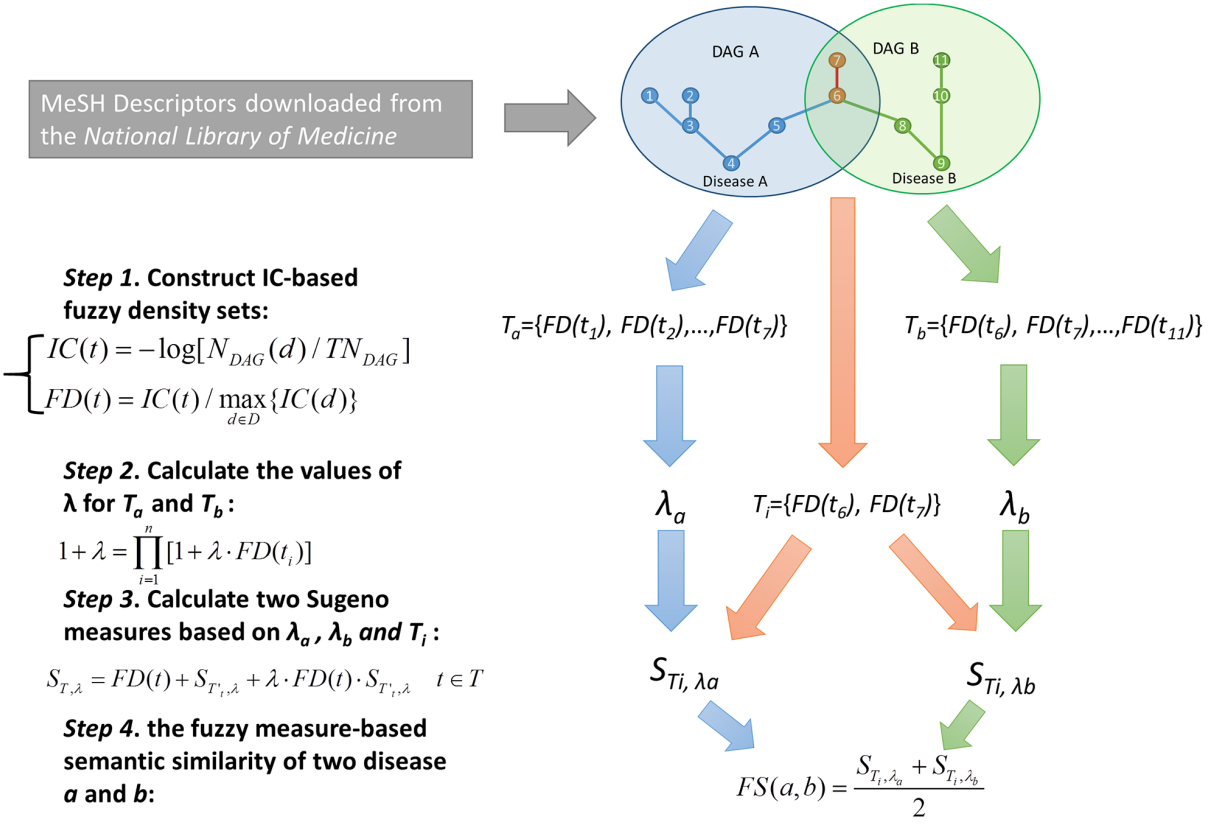

Figure 1: Flowchart of disease semantic similarity calculation in FMLNCSIM based on disease DAGs.

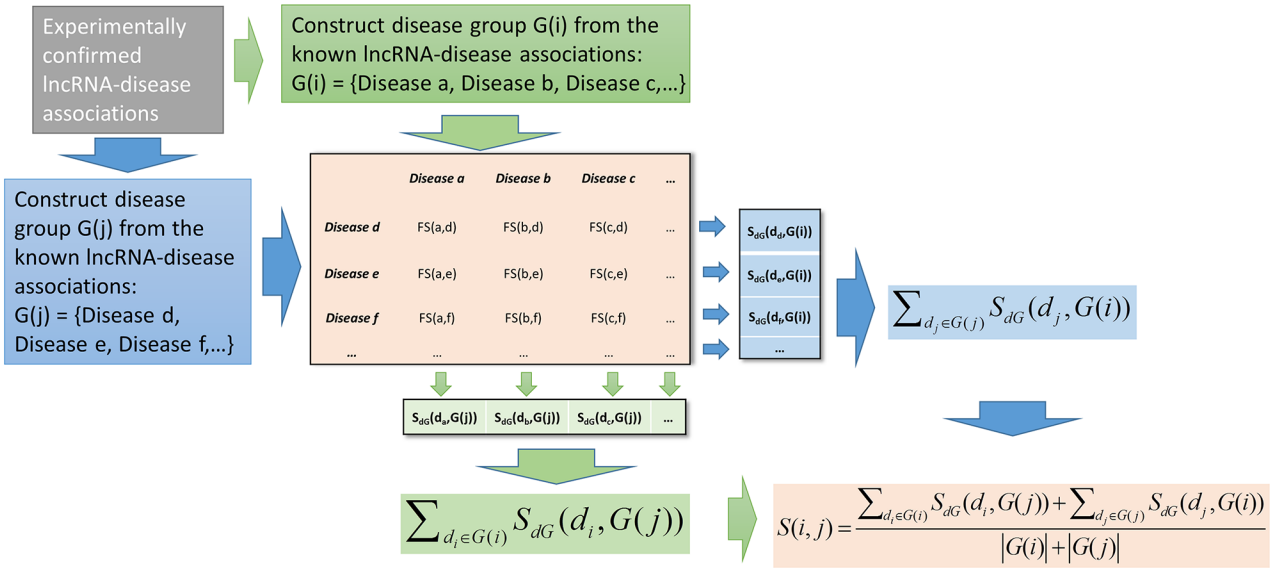

Figure 2: Flowchart of lncRNA functional similarity calculation based on disease semantic similarity.

and further greatly influence the effectiveness of disease-based computational models.

Performance evaluation

FMLNCSIM was developed to calculate the functional similarity scores of lncRNAs in LncRNADisease and MNDR databases, which are listed in Supplementary Table S1 and S2, respectively. To evaluate the performance of FMLNCSIM, we combined it with LRLSLDA and validated the effectiveness of the integrated model of FMLNCSIM-LRLSLDA. The performance of FMLNCSIM-LRLSLDA is directly influenced by the computed functional similarity of lncRNAs and therefore can reflect the effectiveness of FMLNCSIM. In the original version of LRLSLDA, Gaussian interaction profile kernel similarity and lncRNA expression similarity are integrated for constructing lncRNA similarity [37]. In this work, we applied a simple average operation to generate a new integrated disease similarity by combining calculated disease semantic similarity and disease Gaussian interaction profile kernel similarity. We further computed a new integrated lncRNA similarity by combining lncRNA functional similarity which was generated by ILNCSIM, lncRNA Gaussian interaction profile, and lncRNA expression similarity. By this means, we built a new model for quantifying the possibilities of potential lncRNA-disease associations, LRLSLDA-FMLNCSIM. It was mainly constructed by two parts – lncRNA functional similarity calculation model based on fuzzy measure and lncRNA-disease associations prediction model based on Laplacian regularized least squares.

In this work, we explored the known lncRNA-disease associations stored in the manually crated diverse ncRNA-disease repository (MNDR) [66] and LncRNADisease [12] database by adopting two validation frameworks of global LOOCV and 5-fold cross validation. Specifically, for the framework of global LOOCV, each known lncRNA-disease association was left out in turn for testing and other samples were used for training, and all the lncRNA-diseases without recorded association evidence were considered as candidate samples. For 5-fold cross validation method, all known lncRNA-disease associations were randomly divided into five disjoint parts, of which four were used for training and the other one was used as testing samples. To visually evaluate the performance results of the proposed model, Receiver-operating characteristics (ROC) curves were drawn. For further evaluation, the value of AUC was also computed by measuring the area under ROC curves based on testing samples. The lncRNA-disease associations with higher ranks than the given threshold in the testing set were considered as successful predictions while predicted ranks lower than threshold lead to unsuccessful predictions. By setting different thresholds, we could obtain corresponding true positive rates (TPR, sensitivity) and false positive rates (FPR 1-specificity). Here, sensitivity was computed based on the percentage of samples which obtained higher ranks than the threshold, and specificity, on the other hand, denotes the percentage of negative samples with lower ranks than the threshold. ROC curves were further created by plotting the TPR against FPR at various threshold settings. The value of area under ROC curve (AUC) was computed for quantifying the performance results. In general, the value of AUC close to 0.5 means purely random performance while AUC close to 1 imply a promising prediction result.

In this work, we compared the performance of LRLSLDA-FMLNCSIM with three previously proposed computational methods (i.e. LRLSLDA [37], LRLSLDA-LNCSIM1 [55] and LRLSLDA-LNCSIM2 [55]). Figure 3 shows the comparison performance in the framework of global LOOCV. It can be observed that LRLSLDA-FMLNCSIM, LRLSLDA, LRLSLDA-LNCSIM1 and LRLSLDA-LNCSIM2 achieve AUCs of 0.8266, 0.7760, 0.8130 and 0.8198 on LncRNADisease dataset, and yielded AUCs of 0.9338, 0.8850, 0.9135 and 0.9169 on the MNDR dataset, respectively. We also adopted 5-fold cross validation method for further evaluation. To minimize the influence of random division, 5-fold cross validation was repeated 100 times and the average and standard deviation of AUCs yielded by the four models were computed. When we explored the LncRNADisease database, LRLSLDA-FMLNCSIM achieved the best performance with AUC of 0.7979+/-0.0098, significantly higher than those yielded by other methods (LRLSLDA: 0.7295+/-0.0089; LRLSLDA-LNCSIM1 0.7761+/-0.01; LRLSLDA-LNCSIM2 0.7872+/-0.0097). For the MNDR dataset, the yielded comparison results also demonstrated FMLNCSIM was superior to the other methods. LRLSLDA-FMLNCSIM achieved AUCs of 0.9237+/-0.0050 while LRLSLDA, LRLSLDA-LNCSIM1 and LRLSLDA-LNCSIM2 yielded poorer performance with AUCs of 0.8687+/-0.0053, 0.9012+/-0.0044 and 0.9050+/-0.0041, respectively. In conclusion, FMLNCS IM has proved to achieve greater effectiveness for calculating lncRNA functional similarity in the validation frameworks of global LOOCV and 5-fold cross validation.

Case studies

To further evaluate the performance of FMLNCSIM, we here applied LRLSLDA-FMLNCSIM to predict the most possible lncRNAs associated with two important diseases, acute myeloid leukemia and lung cancer, based on the known lncRNA-disease associations in the MNDR dataset. The lncRNA-disease association which obtained top 10 ranks were considered as the most potential candidates and further verified based on another existing databases about lncRNA-disease associations, Lnc2cancer [67], as well as recently published experimental literatures.

Acute myeloid leukemia is one of the high-mortality diseases with long-term overall survival (OS) rates of only 5–16% [68]. The older adults are considered as high-risk populations for acute myeloid leukemia due to the higher frequencies of secondary disease, adverse cytogenetics, comorbid conditions, and poor performance status [69]. An increasing number of novel genetic alteration including gene mutations and changes in gene expression are identified by recent works, which helps to improve the classification and risk stratification of acute myeloid leukemia patients [70]. We here applied LRLSLDA-FMLNCSIM to identify most potential lncRNAs associated with acute myeloid leukemia. As a result, lncRNA UCA1 and HOTAIR in the top 10 candidate list were verified by Lnc2cancer database.

Despite of the advances in clinical and experimental oncology, the prognosis of lung cancer is still unfavorable, with about 1.8 million new cases every year [71]. As one of the markedly leading causes of death worldwide, the 5-year survival rate of lung cancer is still dismal, only around 11% [72, 73]. In addition, lung cancer is usually hard to be diagnosed until advanced stage and therefore prognosis for lung cancer is important for the treatment. The participation of lncRNAs in the development of lung cancer has been intensely researched. Accumulating evidence link dysregulations of some lncRNAs to lung cancers and consider them as the biomarkers for lung cancer therapy. However, the number of detected lncRNAs associated with lung cancer is still limited. In this work, we applied LRLSLDA-FMLNCSIM to prioritize candidate lncRNAs based on known associations in the MNDR database. As a result, four potential lncRNAs with top 10 ranks (BC200, UCA1, HOTAIR, and XIST) were verified by Lnc2cancer database and relevant literatures [74]. Specially, UCA1 was predicted as the third candidate and confirmed by the recent observation that the overexpression of plasma UCA1 promoted the malignant progression of lung cancer [74].

The promising results obtained from global LOOCV, 5-fold cross validation and case studies have demonstrated the reliable performance of LRLSLDA-FMLNCSIM. Therefore, we further prioritize all the candidate lncRNAs for all the diseases recorded in MNDR database by utilizing the known experimentally confirmed lncRNA-disease associations stored in MNDR database and implementing the model of LRLSLDA-FMLNCSIM. The predicted

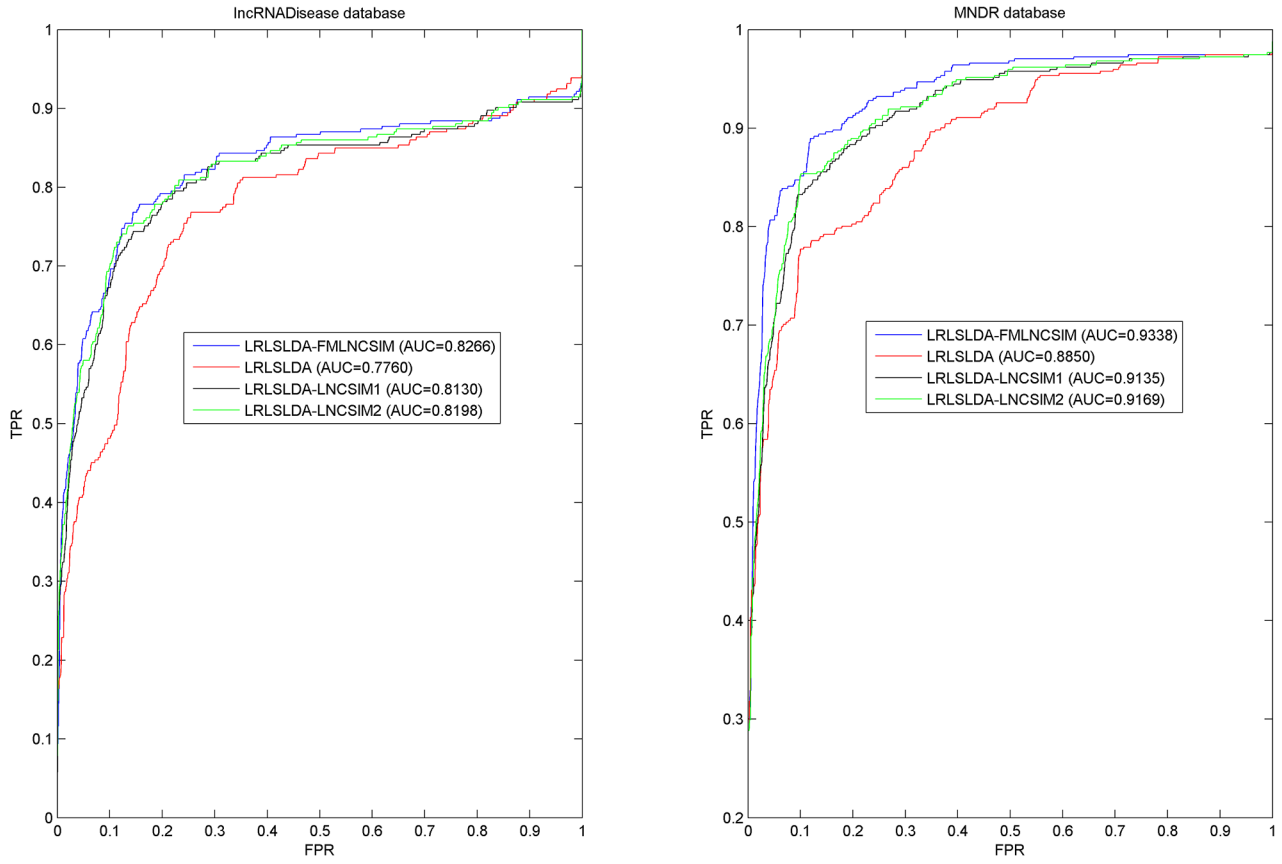

Figure 3: Performance comparisons between FMLNCSIM and three state-of-the-art disease-lncRNA association prediction models (LRLSLDA, LRLSLDA-LNCSIM1 and LRLSLDA-LNCSIM2) in terms of ROC curve and AUC based on global LOOCV. There are roughly 58 and 94 testing samples in LncRNADisease and MNDR databases respectively. As a result, FMLNCSIM achieved AUCs of 0.8266 and 0.9338 based on the LncRNADisease and MNDR databases, which significantly outperformed all the previous classical models and effectively demonstrated its reliable predictive ability.

ranks of lncRNAs for each disease were publicly released for further experimental validation (See Supplementary Table S3). The potential lncRNA-disease associations with high ranks are expected to be confirmed by biological experiments and clinical observation in the future.

DISCUSSION

Measuring lncRNA functional similarity has high value in inferring lncRNA functions as well as searching highly potential lncRNA-disease associations. Since the participation of lncRNAs has been confirmed to influence the development of diseases by increasing clinical observations, it is feasible to measure the functional similarity of lncRNAs based on lncRNA-disease associations. In this work, we proposed a novel computational model for calculating lncRNA functional similarity by combining the information of know lncRNA-disease associations and the disease semantic similarity. To our knowledge, fuzzy measure was first to be introduced for computational models associated with lncRNAs. For further evaluation, FMLNCSIM was integrated with previously proposed LRLSLDA model to quantify lncRNA-disease association probabilities. The reliable results yielded by LRLSLDA-FMLNCSIM in two evaluation frameworks (i.e. global LOOCV and 5-fold cross validation) demonstrated the high effectiveness of FMLNCSIM. Based on the model of FMLNCSIM, lncRNAs with the information of associated diseases could be efficiently searched for their functionally similar ones. The lncRNAs with predicted high ranks could be considered highly potential candidate for further biological experiment verification. Thus, we publicly released the potential lncRNA-disease pairs for all the diseases investigated in this work. We anticipate that there will be more predictions with high ranks confirmed by future biological experiments.

There are some limitations in the model of FMLNCSIM. Firstly, since the fuzzy measure needs to transform disease MeSH DAGs into fuzzy density sets for further computation, the hierarchical structure of DAGs would be failed to be retained in this transformations. Besides, considering the fact that the degrees of researches for different disease are imbalanced, the information bias in DAGs may influence the accuracy of FMLNCSIM. Finally, the proposed version of FMLNCSIM failed to integrated additional data from other types of biological datasets associated with lncRNAs.

MATERIALS AND METHODS

LncRNA-disease associations

In the previous work, we have constructed the first publicly available lncRNA-disease association database, LncRNADisease (http://cmbi.bjmu.edu.cn/lncrnadisease) [12], by manually collecting experimentally confirmed lncRNA-disease associations from accumulating experimental reports. We downloaded known lncRNA-disease associations and got rid of those duplicate samples which describe the same lncRNA-disease association based on different experimental evidences. As a result, there are 293 distinct high-quality lncRNA-disease associations, which include 118 lncRNAs and 167 diseases. For further performance evaluation, we downloaded known associations from another lncRNA-disease association database, the Mammalian ncRNA-disease repository (MNDR, http://www.rna-society.org/mndr/) [66], in March, 2015. The duplicate lncRNA-disease associations from different evidences were also removed. As a result, we obtained 471 high-quality experimentally verified samples of 127 diseases and 241 lncRNAs.

Disease MeSH descriptors

For the measurement of disease similarity, we downloaded MeSH descriptors from the National Library of Medicine (http://www.nlm.nih.gov) [75] to construct disease DAGs. Based on a strict system for disease classification, there are 16 categories included in MeSH descriptors (e.g. Category A: anatomic terms; Category B: organisms; Category C: diseases; Category D: drugs and chemicals). In this work, we downloaded the descriptors of Category C and constructed DAGs to depict the disease association. In the disease DAGs, the nodes represent disease MeSH descriptors and each edge denotes the connection from a more general term (parent node) to a more specific term (child node).

Fuzzy measure-based disease semantic similarity

Fuzzy measures have recently proved to be useful and superior to additive probability measures for describing expert uncertainty. For example, k-additive fuzzy measure reduces the number of variables for definition by limiting the interaction between its subsets [76, 77]. In this work, Sugeno λ-measures, one of the most widely and successfully used class of fuzzy measures, were introduced to calculate disease semantic similarity based on disease MeSH DAGs [78, 79]. Specifically, the information content (IC) for each disease term were computed based on the corresponding MeSH DAG, and further used as fuzzy density values. The fuzzy measure-based disease similarity was then computed based on the IC-based fuzzy density values. The calculation process for disease semantic similarity mainly consists of four steps (See Figure 1).

Disease terms with higher specificity usually have a larger contribution to measuring disease similarity. In the first step, we introduced the concept of information content which can effectively depict how specific a term is. Specifically, we counted the number of occurrences in the DAGs of the term (say, disease d) and then converted it to information content by computing the negative log likelihood:

$$IC(t) = -\log[N_{DAG}(d) / TN_{DAG}] \quad (1)$$

where $N_{DAG}(d)$ denotes the number of DAGs including d; TN_{DAG} denotes the total number of diseases. We further proposed fuzzy density sets aiming at retaining the specificity information of disease group members for further set-set similarity measurements. To achieve this goal, we then defined the fuzzy density by using a normalization operation:

$$FD(t) = IC(t) / \max_{d \in D}\{IC(d)\} \quad (2)$$

where D denotes the whole disease set included in the dataset. In this way, diseases' DAGs were converted into real sets which were considered as fuzzy sets for further calculation.

In the second step, the values of λ for Sugeno measure were computed for each fuzzy density set. Given a fuzzy density set $T=\{FD(t_1), FD(t_2), \ldots, FD(t_n)\}$, the value of λ for T was computed based on the following equation:

$$1 + \lambda = \prod_{i=1}^{n}[1 + \lambda \cdot FD(t_i)] \quad (3)$$

For each fuzzy density set, λ has a unique value since equation (3) has a unique solution for $\lambda > -1$.

In the third step, the Sugeno measures for the intersection of two fuzzy density sets were computed. Given two fuzzy density sets, T_a and T_b, and their intersection T_i, two Sugeno measures for T_i were computed based on the values of λ form T_a and T_b. Given a fuzzy density set $T=\{FD(t_1), FD(t_2), \ldots, FD(t_n)\}$, we say the subset of T which excludes the element t as T'_t. Then, the Sugeno measure of T was calculated based on λ and the Sugeno measure of its subset T'_t, which could be defined as follow:

$$S_{T,\lambda} = FD(t) + S_{T'_t,\lambda} + \lambda \cdot FD(t) \cdot S_{T'_t,\lambda} \quad t \in T \quad (4)$$

In this recursive way, the Sugeno measure of T could be finally computed. Specially, for the fuzzy density set whose size equals one, its Sugeno measure was set to be the value of its only element. Assume that λ values of T_a and T_b are computed as λ_a and λ_b. Then, two Sugeno measures, S_{T_i,λ_a} and S_{T_i,λ_b} would be computed by the same way defined as equation 4. As a result, the value of $S_{T,\lambda}$ could be finally obtained in a recursive way.

In the final step, the fuzzy measure-based semantic similarity of two disease terms, a and b, were calculated based on the average of Sugeno measures:

$$FS(a,b) = \frac{S_{T_i,\lambda_a} + S_{T_i,\lambda_b}}{2} \quad (5)$$

In this way, the semantic similarity of each disease pair could be computed to constitute the disease similarity matrix FS, where the entity in row i column j represent the semantic similarity between *ith* disease and *jth* disease.

FMLNCSIM

Based on the fuzzy measure-based disease semantic similarity, FMLNCSIM was then developed to compute the lncRNA functional similarity by using the information of known lncRNA-diseases. Specifically, the lncRNA functional similarity of two lncRNAs was measured by the similarity between their associated disease groups. Given two disease groups, *G(i)* and *G(j)*, which are respectively associated with lncRNA *i* and lncRNA *j*, we calculated their similarity based on a group-based method (See Figure 2). The similarity between one of disease term (say d_i) in *G(i)* and *G(j)* was computed based on a maximum operation and could be defined as follow:

$$S_{dG}(d_i, G(j)) = \max_{d_j \in G(j)}(FS(d_i, d_j)) \quad (6)$$

The functional similarity between lncRNA *i* and lncRNA *j* was then computed based on the set-based similarity between *G(i)* and *G(j)*:

$$S(i,j) = \frac{\sum_{d_i \in G(i)} S_{dG}(d_i, G(j)) + \sum_{d_j \in G(j)} S_{dG}(d_j, G(i))}{|G(i)| + |G(j)|} \quad (7)$$

where $|G(i)|$ and $|G(j)|$ are the numbers of diseases in *G(i)* and *G(j)*, respectively.

Webserver

In order to provide convenience for applying our proposed model, we built a web server which implements the function of the proposed FMLNCSIM model. It is available at http://219.219.60.245/. This web server mainly carries out four functions. The function 1 and 2 enable visitors obtain functional similarities calculated by FMLNCSIM model based on two lncRNA-disease association databases (i.e. LncRNADisease and MNDR). The function 3 and 4 provide functional similarity calculation for new lncRNAs as long as users provided its associated diseases. When visitors provide a specific lncRNA with its associated diseases, function 3 and 4 could calculate the functional similarities between this query lncRNA and all lncRNAs in LncRNADisease and MNDR databases, and then list the results on the webpage.

ACKNOWLEDGMENTS

XC was supported by the National Natural Science Foundation of China under Grant No. 11301517. ZHY and YAH were supported by the National Natural Science Foundation of China under Grant No. 61572506.

CONFLICTS OF INTEREST

The authors declare no conflict(s) of interest.

REFERENCES

1. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J and Hofacker IL. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science. 2007; 316:1484-1488.
2. Mercer TR, Dinger ME and Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet. 2009; 10:155-159.
3. Guttman M, Russell P, Ingolia NT, Weissman JS and Lander ES. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell. 2013; 154:240-251.
4. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A and Nusbaum C. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol. 2010; 28:503-510.
5. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A and Searle S. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012; 22:1760-1774.
6. Ponting CP, Oliver PL and Reik W. Evolution and functions of long noncoding RNAs. Cell. 2009; 136:629-641.
7. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M and FitzHugh W. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860-921.
8. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP and Cabili MN. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009; 458:223-227.
9. Wapinski O and Chang HY. Long noncoding RNAs and human disease. Trends Cell Biol. 2011; 21:354-361.
10. Wilusz JE, Sunwoo H and Spector DL. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. 2009; 23:1494-1504.
11. Amaral PP, Clark MB, Gascoigne DK, Dinger ME and Mattick JS. lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res. 2011; 39:D146-D151.
12. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G and Cui Q. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res. 2013; 41:D983-D986.
13. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Morales DR, Thomas K, Presser A, Bernstein BE and Van Oudenaarden A. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. P Natl Acad Sci USA. 2009; 106:11667-11672.
14. Yu F, Zheng J, Mao Y, Dong P, Li G, Lu Z, Guo C, Liu Z and Fan X. Long non-coding RNA APTR promotes the activation of hepatic stellate cells and the progression of liver fibrosis. Biochem Biophys Res Commun. 2015; 463:679-685.
15. Tsai M-C, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E and Chang HY. Long noncoding RNA as modular scaffold of histone modification complexes. Science. 2010; 329:689-693.
16. Penny GD, Kay GF, Sheardown SA, Rastan S and Brockdorff N. Requirement for Xist in X chromosome inactivation. Nature. 1996; 379:131-137.
17. Fang Z, Wu L, Wang L, Yang Y, Meng Y and Yang H. Increased expression of the long non-coding RNA UCA1 in tongue squamous cell carcinomas: a possible correlation with cancer metastasis. Oral surgery, oral medicine, oral pathology and oral radiology. 2014; 117:89-95.
18. Salmena L, Poliseno L, Tay Y, Kats L and Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell. 2011; 146:353-358.
19. Alvarez ML and DiStefano JK. Functional characterization of the plasmacytoma variant translocation 1 gene (PVT1) in diabetic nephropathy. PLoS One. 2011; 6:e18671.
20. Pasmant E, Sabbagh A, Vidaud M and Bièche I. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. FASEB J. 2011; 25:444-448.
21. Zhang Q, Chen C-Y, Yedavalli VS and Jeang K-T. NEAT1 long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression. MBio. 2013; 4:e00596-00512.
22. Guffanti A, Iacono M, Pelucchi P, Kim N, Soldà G, Croft LJ, Taft RJ, Rizzi E, Askarian-Amiri M and Bonnal RJ. A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. BMC Genomics. 2009; 10:163.
23. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai M-C, Hung T, Argani P and Rinn JL. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010; 464:1071-1076.
24. Zhang X, Zhou Y, Mehta KR, Danila DC, Scolavino S, Johnson SR and Klibanski A. A pituitary-derived MEG3 isoform functions as a growth suppressor in tumor cells. J Clin Endocrinol Metab. 2003; 88:5119-5126.
25. Ji P, Diederichs S, Wang W, Böing S, Metzger R, Schneider PM, Tidow N, Brandt B, Buerger H and Bulk E. MALAT-1, a novel noncoding RNA, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene. 2003; 22:8031-8041.
26. Pibouin L, Villaudy J, Ferbus D, Muleris M, Prospéri M-T, Remvikos Y and Goubin G. Cloning of the mRNA of overexpression in colon carcinoma-1: a sequence overexpressed in a subset of colon carcinomas. Cancer Genet Cytogenet. 2002; 133:55-60.
27. Zhang X, Lian Z, Padden C, Gerstein MB, Rozowsky J, Snyder M, Gingeras TR, Kapranov P, Weissman SM and Newburger PE. A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. Blood. 2009; 113:2526-2534.

28. Guan Y, Kuo W-L, Stilwell JL, Takano H, Lapuk AV, Fridlyand J, Mao J-H, Yu M, Miller MA and Santos JL. Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer. Clin Cancer Res. 2007; 13:5745-5755.

29. Feng FY, Ma T, Speers C, Iyer MK, Zhao S, Prensner JR, Rae JM, Pierce LJ and Chinnaiyan AM. Abstract PD6-1: The long noncoding RNA M41 promotes aggressiveness and tamoxifen resistance in ER-positive breast cancers. Cancer Res. 2015; 75:PD6-1-PD6-1.

30. Xu Z, Chen J, Luk JM and De W. LncRNA ANRIL indicates a potential prognostic biomarker in gastric cancer and promotes tumor growth by silencing of miR-99a/miR-449a. Cancer Res. 2015; 75:157-157.

31. Ochoa AE, Zhang J, Choi W, Malouf GG, Thompson EJ, Weinstein JN, Tannir NM, Dinney C, McConkey DJ and Su X. Abstract A1-68: The long noncoding RNA SNHG18 promotes PPARγ function and luminal gene expression in muscle-invasive bladder cancer. Cancer Res. 2015; 75:A1-68-A61-68.

32. Borsani G, Tonlorenzi R, Simmler MC, Dandolo L, Arnaud D, Capra V, Grompe M, Pizzuti A, Muzny D and Lawrence C. Characterization of a murine gene expressed from the inactive X chromosome. Nature. 1991; 351:325-329.

33. Brannan CI, Dees EC, Ingram RS and Tilghman SM. The product of the H19 gene may function as an RNA. Mol Cell Biol. 1990; 10:28-36.

34. Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S and Rastan S. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell. 1992; 71:515-526.

35. Dinger ME, Pang KC, Mercer TR, Crowe ML, Grimmond SM and Mattick JS. NRED: a database of long noncoding RNA expression. Nucleic Acids Res. 2009; 37:D122-D126.

36. Bu D, Yu K, Sun S, Xie C, Skogerbø G, Miao R, Xiao H, Liao Q, Luo H and Zhao G. NONCODE v3.0: integrative annotation of long noncoding RNAs. Nucleic Acids Res. 2011; 40:D210-D215.

37. Chen X and Yan G-Y. Novel human lncRNA–disease association inference based on lncRNA expression profiles. Bioinformatics. 2013; 29:2617-2624.

38. Chen X. Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. Sci Rep. 2015; 5:13186.

39. Chen X. KATZLDA: KATZ measure for the lncRNA-disease association prediction. Sci Rep. 2015; 5:16840.

40. Liu M-X, Chen X, Chen G, Cui Q-H and Yan G-Y. A computational framework to infer human disease-associated long noncoding RNAs. PLoS One. 2014; 9:e84408.

41. Huang Y, Chen X, You Z, Huang D and Chan K. ILNCSIM: improved lncRNA functional similarity calculation model. Oncotarget. 2016; 7:25902-25914. doi: 10.18632/oncotarget.8296.

42. Chen X, Yan CC, Zhang X, You Z-H, Deng L, Liu Y, Zhang Y and Dai Q. WBSMDA: Within and Between Score for MiRNA-Disease Association prediction. Scientific reports. 2016; 6:21106

43. Chen X. miREFRWR: a novel disease-related microRNA-environmental factor interactions prediction method. Mol Biosyst. 2016; 12:624-633.

44. Chen X, Yan CC, Zhang X, Li Z, Deng L, Zhang Y and Dai Q. RBMMMDA: predicting multiple types of disease-microRNA associations. Sci Rep. 2015; 5:13877.

45. Chen X and Yan G-Y. Semi-supervised learning for potential human microRNA-disease associations inference. Sci Rep. 2014; 4:5501.

46. Chen X, Liu MX and Yan G. RWRMDA: predicting novel human microRNA-disease associations. Mol Biosyst. 2012; 8:2792-2798.

47. Sun L, Luo H, Liao Q, Bu D, Zhao G, Liu C, Liu Y and Zhao Y. Systematic study of human long intergenic non-coding RNAs and their impact on cancer. Sci China Life Sci. 2013; 56:324-334.

48. Shi X, Sun M, Liu H, Yao Y and Song Y. Long non-coding RNAs: a new frontier in the study of human diseases. Cancer letters. 2013; 339:159-166.

49. Batista PJ and Chang HY. Long noncoding RNAs: cellular address codes in development and disease. Cell. 2013; 152:1298-1307.

50. Bellucci M, Agostini F, Masin M and Tartaglia GG. Predicting protein associations with long noncoding RNAs. Nature Methods. 2011; 8:444-445.

51. Liao Q, Liu C, Yuan X, Kang S, Miao R, Xiao H, Zhao G, Luo H, Bu D and Zhao H. Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network. Nucleic Acids Res. 2011; 39:3864-3878.

52. Xiao Y, Lv Y, Zhao H, Gong Y, Hu J, Li F, Xu J, Bai J, Yu F and Li X. Predicting the Functions of Long Noncoding RNAs Using RNA-Seq Based on Bayesian Network. Biomed Res Int. 2015; 2015.

53. Wang JZ, Du Z, Payattakool R, Philip SY and Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007; 23:1274-1281.

54. Sun J, Shi H, Wang Z, Zhang C, Liu L, Wang L, He W, Hao D, Liu S and Zhou M. Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network. Mol Biosyst. 2014; 10:2074-2081.

55. Chen X, Yan CC, Luo C, Ji W, Zhang Y and Dai Q. Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. Sci Rep. 2015; 5:11338.

56. Wang E, Zaman N, Mcgee S, Milanese J-S, Masoudi-Nejad A and O'Connor-McCourt M. Predictive genomics: A cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. Semin Cancer Biol. 2015; 30:4-12.

57. Li J, Lenferink AE, Deng Y, Collins C, Cui Q, Purisima EO, O'Connor-McCourt MD and Wang E. Identification

of high-quality cancer prognostic markers and metastasis network modules. Nature communications. 2010; 1:34.

58. Masoudi-Nejad A and Wang E. Cancer modeling and network biology: Accelerating toward personalized medicine. Seminars in cancer biology. 2015; pp. 1-3.
59. Yu G, Li F, Qin Y, Bo X, Wu Y and Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics. 2010; 26:976-978.
60. Fröhlich H, Speer N, Poustka A and Beißbarth T. GOSim–an R-package for computation of information theoretic GO similarities between terms and gene products. BMC bioinformatics. 2007; 8:166.
61. Luiten G, Froese T, Björk B, Cooper G, Junge R, Karstila K and Oxman R. An information reference model for architecture, engineering, and construction. Management of information technology for construction, World scientific & global publication services, Singapore. 1993; 1993:391-406.
62. Ye J. Multicriteria group decision-making method using vector similarity measures for trapezoidal intuitionistic fuzzy numbers. Group Decis Negot. 2012; 21:519-530.
63. Parvathi R and Malathi C. Arithmetic operations on symmetric trapezoidal intuitionistic fuzzy numbers. International Journal of Soft Computing and Engineering. 2012; 2.
64. Murofushi T and Sugeno M. An interpretation of fuzzy measures and the Choquet integral as an integral with respect to a fuzzy measure. Fuzzy Set Syst. 1989; 29:201-227.
65. Nguyen HT, Kreinovich V, Lorkowski J and Abu S. Why Sugeno lambda-Measures. 2015.
66. Wang Y, Chen L, Chen B, Li X, Kang J, Fan K, Hu Y, Xu J, Yi L and Yang J. Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network. Cell Death Dis. 2013; 4:e765.
67. Ning S, Zhang J, Wang P, Zhi H, Wang J, Liu Y, Gao Y, Guo M, Yue M and Wang L. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. Nucleic Acids Res. 2015; 44:D980-D985.
68. Farag SS, Archer KJ, Mrózek K, Ruppert AS, Carroll AJ, Vardiman JW, Pettenati MJ, Baer MR, Qumsiyeh MB and Koduru PR. Pretreatment cytogenetics add to other prognostic factors predicting complete remission and long-term outcome in patients 60 years of age or older with acute myeloid leukemia: results from Cancer and Leukemia Group B 8461. Blood. 2006; 108:63-73.
69. Kantarjian H, O'Brien S, Cortes J, Giles F, Faderl S, Jabbour E, Garcia-Manero G, Wierda W, Pierce S and Shan J. Results of intensive chemotherapy in 998 patients age 65 years or older with acute myeloid leukemia or high-risk myelodysplastic syndrome. cancer. 2006; 106:1090-1098.
70. Schlenk RF, Döhner K, Krauter J, Fröhling S, Corbacioglu A, Bullinger L, Habdank M, Späth D, Morgan M and Benner A. Mutations and treatment outcome in cytogenetically normal acute myeloid leukemia. N Engl J Med. 2008; 358:1909-1918.
71. Gutschner T, Hämmerle M, Eißmann M, Hsu J, Kim Y, Hung G, Revenko A, Arun G, Stentrup M and Groß M. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. Cancer Res. 2013; 73:1180-1189.
72. Liu X-h, Liu Z-l, Sun M, Liu J, Wang Z-x and De W. The long non-coding RNA HOTAIR indicates a poor prognosis and promotes metastasis in non-small cell lung cancer. BMC Cancer. 2013; 13:464.
73. Enfield KS, Pikor LA, Martinez VD and Lam WL. Mechanistic roles of noncoding RNAs in lung cancer biology and their clinical implications. Genet Res Int 2012; 2012.
74. Wang H-M, Lu J-H, Chen W-Y and Gu A-Q. Upregulated lncRNA-UCA1 contributes to progression of lung cancer and is closely related to clinical diagnosis as a predictive biomarker in plasma. Int J Clin Exp Med. 2015; 8:11824-11830.
75. Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc. 2000; 88:265-266.
76. Kruse R. A note on λ-additive fuzzy measures. Fuzzy Sets and Systems. 1982; 8:219-222.
77. Miranda P and Grabisch M. Caracterizing k-Additive Fuzzy Measures. Technologies for Constructing Intelligent Systems 2: Springer. 2002; pp. 209-222.
78. Al Boni M, Anderson DT and King RL. Hybrid Measure of Agreement and Expertise for Ontology Matching in Lieu of a Reference Ontology. International Journal of Intelligent Systems. 2015.
79. Štefka D and Holeňa M. Dynamic classifier aggregation using interaction-sensitive fuzzy measures. Fuzzy Sets and Systems. 2015; 270:25-52.