Systems biology

Identifying novel associations between small molecules and miRNAs based on integrated molecular networks

Yingli Lv¹, Shuyuan Wang¹, Fanlin Meng¹, Lei Yang¹, Zhifeng Wang¹, Jing Wang¹, Xiaowen Chen¹, Wei Jiang¹,⁎, Yixue Li¹,²,⁎ and Xia Li¹,⁎

¹College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China and ²Bioinformatics Center, Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

⁎To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on July 24, 2014; revised on July 1, 2015; accepted on July 14, 2015

Abstract

**Motivation:** miRNAs play crucial roles in human diseases and newly discovered could be targeted by small molecule (SM) drug compounds. Thus, the identification of small molecule drug compounds (SM) that target dysregulated miRNAs in cancers will provide new insight into cancer biology and accelerate drug discovery for cancer therapy.

**Results:** In this study, we aimed to develop a novel computational method to comprehensively identify associations between SMs and miRNAs. To this end, exploiting multiple molecular interaction databases, we first established an integrated SM-miRNA association network based on 690,561 SM to SM interactions, 291,600 miRNA to miRNA associations, as well as 664 known SM to miRNA targeting pairs. Then, by performing Random Walk with Restart algorithm on the integrated network, we prioritized the miRNAs associated to each of the SMs. By validating our results utilizing an independent dataset we obtained an area under the ROC curve greater than 0.7. Furthermore, comparisons indicated our integrated approach significantly improved the identification performance of those simple modeled methods. This computational framework as well as the prioritized SM-miRNA targeting relationships will promote the further developments of targeted cancer therapies.

**Contact:** yxli@sibs.ac.cn, lixia@hrbmu.edu.cn or jiangwei@hrbmu.edu.cn

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

1 Introduction

miRNAs are non-coding small RNAs with post transcriptional regulatory functions and are dysregulated in most of human cancers (Volinia *et al.*, 2006; Wu *et al.*, 2007). Approximately half of the known miRNAs are located in cancer-associated genome regions or fragile sites (Liu *et al.*, 2004; Lagana *et al.*, 2010). Cumulative studies demonstrated that the mature miRNAs as well as their precursors could be targeted by small molecular drugs (Liu *et al.*, 2008; Thomas and Hergenrother, 2008; Bose *et al.*, 2012; Srinivasan *et al.*, 2013; Hesse and Arenz, 2014). A recent clinical trial revealed SPC349, a newly developed drug, could successfully inhibit miR-122 which plays important roles in the duplication of hepatitis C viruses (Lanford *et al.*, 2010).

Several approaches then were developed to investigate the interactions between small molecules and miRNA. Structure-based approaches, such as molecular docking, is useful for identifying small molecular compounds that target miRNAs (Zhang *et al.*, 2010) with known 3D structures. Identifying associations between small molecules and miRNAs in 23 cancers types and Alzheimer have been established based on transcriptional responses of small
molecule and miRNA perturbation (Jiang et al., 2012; Meng et al., 2013). Jamal et al employed a chemical descriptors- and machine learning-based method, which is the first and most comprehensive computational analysis to predict small molecule modulators of miRNA (Jamal et al., 2012).

Meanwhile, based on the hypothesis that similar small molecules tend to target similar proteins, several computational systems biology approaches based on large scale molecular networks have been applied to identify molecular interactions related to small molecules (Van Laarhoven et al., 2011; Alaimo et al., 2013).

In this study, taking together the advantages of above structure- or experiment-based methods and the genome-wide exploration of biological network based approaches, we firstly apply network retrieval methods on integrated biological interactions to predict small molecule-miRNA associations for understanding miRNA binding activities of small molecules. This study provides researchers a practical method to identify the biological regulations of small molecules and will facilitate the further discovery of chemical drug on miRNA mediated human complex diseases.

2 Materials and methods

2.1 Datasets

2.1.1 Small molecule drugs
FDA-approved small molecule or drug compounds (SM) were obtained from SM2miR (Liu et al., 2013), DrugBank (Knox et al., 2011) and PubChem (Wang et al., 2009). After filtering redundant annotations across databases and many-to-one SM-miRNA relationships, we finally obtained a total of 1338 SMs including 101 SMs from SM2miR, 1291 FDA-approved SM drugs with DrugBank annotations and 1124 SM drugs from PubChem. 1077 SMs have both annotations in DrugBank and PubChem, 214 SMs unique annotated in DrugBank and 47 SMs identically annotated in PubChem (Supplementary Table S1).

2.1.2 miRNAs with phenotype annotations
miRNAs were compiled from the SM2miR, HMDD (Lu et al., 2008), mi2Disease (Jiang et al., 2009) and PhenomiR databases (Ruepp et al., 2010). Then, they were matched with their precursors using the miRBase database (release 20). After removing deprecated miRNAs, a total of 571 miRNAs were left for further analyses (Supplementary Table S2).

2.1.3 A compendium of known SM-miRNA targeting pairs
A collection of 664 SM-miRNA targeting pairs reported by SM2miR (Version 1) were used as seeds for Random Walk with Restart (RWR) algorithms. Another set of 78 newly updated SM-miRNA targeting pairs were used as an independent testing set for validation (Supplementary Table S3).

2.2 Similarity calculation and integration

2.2.1 Integrating SM-SM similarities to establish SM-SM interaction networks
Previous drug discovery studies demonstrated that similarities based on chemical structures, side effects, targeting functions and phenotypes are powerful computational tools to identify the associations among SMs for drug discoveries (Gottlieb et al., 2011; Chen et al., 2012a, b; Takarabe et al., 2012; Chen et al., 2013). In this study, we employed four commonly used similarity measurements which are based on side effect (Gottlieb et al., 2011), functional consistency (Lv et al., 2012), chemical structure (Hattori et al., 2003) and indication phenotype (Gottlieb et al., 2011), respectively (Fig. 1 and Supplementary File). Then, to reduce the bias of each similarity measurement and facilitate the discovery of novel interactions, we used a weighted combination strategy to integrate the similarities. As shown in Equation (1), for each of the SM pairs, the integrated similarity $S_i$ was defined as follows:

$$S_i = \left(\beta_1 S^a_{ST} + \beta_2 S^b_{ST} + \beta_3 S^c_{ST} + \beta_4 S^d_{ST}\right)/\sum_{j} \beta_j (j = 1, 2, 3, 4)$$ (1)

$S^a_{ST}$, $S^b_{ST}$, $S^c_{ST}$ and $S^d_{ST}$ indicate the similarity measurement based on indication phenotype, functional consistency, chemical structure and side effect, respectively. The default value $\beta_i = 1$ assigns the same weight to each separated similarities. After integration, we identified 690 561 SM-SM interactions (Supplementary Table S4), which increase our capability of finding novel SM-miRNA interactions. Also, the distribution of the integrated interactions shows a typical pattern of scale-free networks (Barabasi and Oltvai, 2004; Albert, 2005; Rai et al., 2014) (Supplementary Fig. S1).

2.2.2 Measuring similarities between miRNA and miRNA to establish miRNA-miRNA association networks
As shown in Figure 1, we employed two previous defined measurements, based on functional consistency (Lv et al., 2012) and phenotype of indications (Gottlieb et al., 2011), to determine raw miRNA-miRNA associations (Supplementary File). We integrated the two measurements to reduce the bias and extend the network for discovering novel miRNA-miRNA associations. As shown in Equation (2), for each of the miRNA pairs, the integrated associations $S_M$ was defined as:

$$S_M = (\alpha_1 S^a_M + \alpha_2 S^b_M) / \sum_{i} \alpha_i (i = 1, 2)$$ (2)

$S^a_M$ and $S^b_M$ indicate the indication phenotype- and functional consistency-based association respectively. For SM-SM interactions, $\alpha_i = 1$ was set as the default value. A total of 291 600 miRNA-miRNA interactions were identified after integrating (Supplementary Table S5).

2.3 RWR method for the integrated network with two types of nodes

The RWR algorithm is derived from graph theory and simulates a random move from the seed node(s) to their immediate neighbors or stay at the current node(s) according to the probability transition...
3 Results

3.1 Performing improved RWR to prioritize SM targeting miRNAs

RWR was used to predict associations between SMs and their biological targets (Chen et al., 2012b). We constructed SM-SM and miRNA-miRNA association networks based on integrating their association measures. Using known SM-miRNA targeting relationships, we merged the two networks into one multiple layer network (Fig. 1). As shown in Figure 1, some of the SMs have known miRNA targets (illustrated with red arrow in Fig. 1) while others do not (illustrated with blue arrow in Fig. 1). Notably, 792 (out of 831, 95%) SMs have no validated targets according to SM2miR. In effort to exploit the RWR method on integrated networks for both SMs with/without validated targets, we classified the SMs into two groups and employed different strategies to predict their miRNA targets. The 39 SMs with known miRNA targets were grouped into the first type of SMs (SM-I), in which eight SMs have only one known miRNA target. 792 SMs lacking validated miRNA targets were grouped into the second type of SMs (SM-II). RWR requires a set of seed nodes to initiate the analysis and assign probability scores (scoring the relationship) to all nodes (including the seed nodes) according to the topological structure of the network. Accordingly, for each SM, the SM and its known miRNA targets were set as seed nodes in SM-I while only the SM was set as seed node in SM-II. RWR was then utilized to predict the miRNA targets of each SM based on assigned probability scores (see Supplementary File).

3.2 Performance evaluation and independent validation

Cross validation is essential for validating the performance of prediction methods. The receiver operating characteristic curve (ROC) plots the true (sensitivity) versus false positive rate (1-specificity) at different cutoffs (Yang et al., 2014) and area under curve (AUC) of ROC is commonly used to represent the results of cross validation. We utilized an improved leave-one-out cross validation (LOOCV) (Chen et al., 2012c) to validate our method on integrated networks. As described, different seed node(s) settings were utilized in predicting miRNA targets for SM-I and SM-II.

To validate SMs in SM-I, as illustrated in Figure 2A, one known SM-miRNA association was excluded and the SM to be validated along with remaining known targets of the SM were set as seed nodes. The rank of the excluded miRNA (colored in red) as determined by the RWR probability scores, and the AUC of ROC was then calculated as the index to score the recovery capability of the excluded known SM-miRNA association. The validation set comprised a set of 31 SMs in which each SM has at least two known miRNA targets (not including the excluded miRNA for which there should be at least one known target set as seed node(s)). These 31 SMs involved 656 known SM-miRNA associations. As a result, the AUC of 14 (45%) SMs is greater than 0.9, the AUC of 22 (71%) SMs is greater than 0.8, and the AUC of 26 (84%) SMs is greater than 0.7 (Supplementary Fig. S2). Due to insufficient known miRNA targets for SMs in SM-II, LOOCV cannot be performed directly validate the recovery capability. However, by using SMs from SM-I but excluding all the known miRNA targets, we can establish an RWR-defined initial seed node for SM-II. As illustrated in Figure 2B, all the known SM-miRNA associations for the validating SMs were excluded, and only the SM itself was set as the seed node. Then, by using the ranks of excluded miRNAs (colored in red) as determined by RWR, AUCs were calculated for each SM. A set of 39 SMs in which each SM has at least one known miRNA target comprised the validation dataset. The AUC of 11 (28%) SMs is greater than 0.9, the AUC of 18 (46%) SMs is greater than 0.8, and the AUC of 27 (69%) SMs is greater than 0.7 (Supplementary Fig. S2). The overall ROC predicting all of the SM-miRNA associations for the two types of SMs indicate that our method achieved satisfactory sensitivity and specificity (Supplementary Fig. S2).

Further, we applied the evaluation method to an independent testing set from recently updated SM-miRNA associations (6 SM targets).
were included which had 78 SM-miRNA interactions) to further evaluate the performance of this method (Supplementary Table S3). The results are shown in Supplementary Table S6. The AUC values of the newly found SMs CID: 5816 and CID: 2662 were 0.837 and 0.82, respectively. Moreover, four other SM AUCs were higher than or near 0.7. Thus, we propose that our prediction method is effective for finding potential miRNA targets of SMs.

3.3 Performance of identified SM-miRNA associations

We generated SM-miRNA interactions randomly in silico to evaluate whether or not the results of our prediction method were likely to be obtained by chance. To compare the recovery capability in SM-I we maintained the SM-SM and miRNA-miRNA similarity networks and randomly assigned a total of 656 miRNAs as the targets of 31 randomly selected SMs that have at least two miRNA targets. Our method and LOOCV procedures as described in Figure 2A were then performed to generate AUC values for scoring the recovery capability of the randomly generated SM-miRNA interactions. Randomization and subsequent validating procedures were executed 100 times. As shown in Figure 3A left, the average AUC of the randomized networks were significantly reduced (p-value = 1.7E-18) when comparing the original known SM-miRNA interactions based on multi-layer network and indicates the lower recovery capability of randomly generated networks. To compare the recovery capability in SM-II, a total of 664 miRNAs were assigned randomly as the targets of 39 SMs that have at least one target. As shown in Figure 3A right, networks based on randomized SM-miRNA interactions show significantly reduced average AUC values. We calculated empirical p-values for the AUC values individually and applied Benjamini-Hochberg (Benjamini and Hochberg, 1995) correction to adjust for multiple testing. The results showed that the AUCs of 27 SM-IIs (out of 31, 87%) and 30 SM-IIs (out of 39, 77%) were significantly higher (with FDR < 0.05) than the random AUCs (Supplementary Table S7).

Additionally, as shown in Figure 3B, we generated curves reflecting the overall ROC to predict all of the SM-miRNA interactions of SM-I and SM-II. The results demonstrate that AUC values reflecting random interactions were reduced compared with the true values of 0.825 and 0.729 for SM-I and SM-II, respectively, and reveal that results from our method cannot be achieved by chance and that the prediction results are significant.

3.4 Measuring the effects of the integrated similarity approach

We tested three methods ('T;', 'D;' and 'T+D;') to compare their performance on the miRNA similarity network. A semicolon was used to distinguish two single similarity networks, T and D represent miRNA similarities based on the functional consistency and phenotype of indications. 'T + D' represents the combination of similarities. Because the SM-IIs have no known miRNA targets, we can only compare the evaluation results of the three methods for SM-I implemented LOOCV. 'T + D;' successfully yielded AUC values for 23% of SMs and was in the range of 0.8 to 0.9 (Fig. 4A), and AUC values for more than 63% of the SMs was greater than 0.8 (Fig. 4B). The overall AUC of method ‘T + D;’ is 0.749 which is comparable with that of ‘D;’(0.754) and higher than that of ‘T;’(0.72) (Fig. 4C). We considered that miRNA related diseases were biologically relevant. If a miRNA network is constructed based on D alone it will miss some miRNAs having true relationships with diseases. Since, ‘D;’ and ‘T + D;’ produced a similar performance, it is convenient to use ‘T + D;’ to add biological association to the resulting SM-miRNA network.

We further compared ‘T + D;’ T + D + C + S’ (‘ALL’), ‘T;C’, ‘T;S’ and ‘T;’ in which T, D, C and S in SM networks represent SM similarity metrics in the integrated network and indicates indication phenotype, chemical structure and side effects. ALL, ‘T;C’ and ‘T;S’ integrate both miRNA and SM similarity networks using more SM similarities. As shown in Figure 4D, E and F, ‘ALL’ yielded AUCs greater than 0.8 for 72% of the SMs in which AUC values for 42% of SMs were in the range of 0.9 to 1. ‘T;’ only yielded AUCs greater than 0.8 for 35% of the SMs and 6% of SMs were in the range of 0.9 to 1. Although Wang et al. (Wang et al., 2013) showed that chemical structure similarity was least successful in predicting drug indications compared to functional consistency and side-effect, ‘T;S’ and

Fig. 3. (A) Box-plots of average AUC values reflecting original known and random model scenarios. The x-axes represent the original and random SM networks and the y-axes represent the average AUC values for each method. (B) The distribution of overall ROC curves between original and random for SM-I and SM-II.

Fig. 4. (A, D). Bar graph of distribution of SMs in different ranges of AUCs. (B, E). Bar graph of cumulative distribution of SMs in different SM and its known target miRNAs or an SM alone, the known ranges of AUCs. The x-axes represent different models, while the y-axes represent the percent of SM in different ranges of the AUC. (C, F). The overall ROC curves of the different models. T represents miRNA similarity based on the functional consistency and D indicates indication phenotype in the miRNA similarity network; C and S represent SM similarity based on chemical structure and side effect in the SM similarity network. ALL represents the two miRNA similarity metrics and four SM similarity metrics in the integrated network.
The RWR method involves three parameters: the restart probability $\gamma$, jumping probability $\lambda$, controlling the impact of two kinds of seed nodes $\eta$, seed miRNAs and seed SMs. We provide some analysis on the choice of the parameters for the algorithm in Supplementary File. In addition, the integrated similarity involves six parameters: $\alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3, \beta_4$. The alpha parameters represent the weights of different similarity evaluations in the integrated miRNA similarities and the beta parameters represent that in the integrated SMs similarities. The six similarity measures reflect the miRNA similarities and the SM similarities in terms of different biology. SM's chemical structure provides information by the 'structure determines function' paradigm and side effect hints the unwanted effect at phenotype level. All of the measures are important in terms of biology. Thus, we select equal weigh for the six weight parameters (Li et al., 2004).

Considering the existing methods predicting SM-miRNA interactions are rarely and these methods implemented on different level datasets (Jamal et al., 2012; Jiang et al., 2012; Meng et al., 2013). We analyzed the result of method ALL and other methods that implemented on all integrated network combined different miRNA
similarity metrics with different SM similarity metrics. Every type of similarity combination of miRNA and SM was firstly applied to construct the integrated network and was then used to predict SM-miRNA associations up to now. We traverse the different combination of small molecular similarity and miRNA similarity to construct the integrated network and calculate the overall AUC of performance for this algorithm. There are 48 combination strategies for SM-I$s$ and 45 for SM-II$s$. We can’t evaluate the performance of ‘T’, ‘D’ and ‘T + D’ for SM-II$s$, because the SM-II$s$ have no known target miRNAs. We found the AUCs of different combination were fluctuant with increasing of similarity metrics, which indicate differential combination method are competitive (Supplementary Table S11). Although the AUC of method ALL is not the best, the method based on miRNA network constructed with T and D reduce the missing of candidate of miRNAs to some extent. In assuming the six similarity metrics for all miRNAs and SMs, the method ALL still obtained comparative performance on a perfect candidate miRNAs set of SMs.

In addition, Supplementary Table S11 enlighten us when a SM or miRNA only has one or few metrics, we can select alternative combination method with better performance result to infer SM-miRNA associations. For instance, when a miRNA has information of T and D, a SM only has information of C and S, it is also to get a prediction result based on the integrated network consisted of ‘T + D + C + S’. But this method may be not the best with development of more biologically relevant information defining miRNA-miRNA similarity and SM-SM similarity. For all this, Supplementary Table S11 is also able to provide some guidance for our prediction method.

However, known SM-miRNA links are quite scarce, which impacts the evaluation of the method. In the future, with increasing reports of SM-miRNA links, we will further improve the performance of the SM-miRNA associations prediction method.

Funding
This work was supported in part by the National High Technology Research and Development Program of China [863Program, Grant 2014AA021102], the National Program on Key Basic Research Project [973 Program, Grant 2014CB910504], the National Natural Science Foundation of China [Grant 91129710 and 61170154] and the Science Foundation of Hei Long Jiang Province Health Department [Grant 2013125].

Conflict of Interest: none declared.

References