Bioinformatics, 31(22), 2015, 3638–3644
doi: 10.1093/bioinformatics/btv417
Advance Access Publication Date: 20 July 2015
Original Paper

Systems biology

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

Yingli Lv[1], Shuyuan Wang[1], Fanlin Meng[1], Lei Yang[1], Zhifeng Wang[1], Jing Wang[1], Xiaowen Chen[1], Wei Jiang[1,*], Yixue Li[1,2,*] and Xia Li[1,*]

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China and [2]Bioinformatics 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), miR2Disease (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

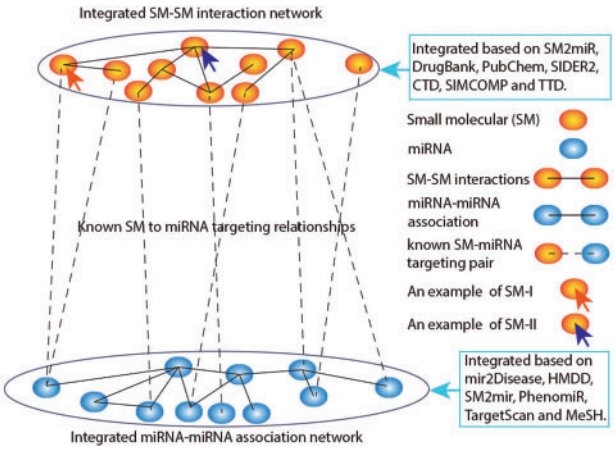

Fig. 1. The integrated multi-layer network for performing RWR method

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 weighed combination strategy to integrate the similarities. As shown in Equation (1), for each of the SM pairs, the integrated similarity S_S was defined as follows:

$$S_S = (\beta_1 S_S^D + \beta_2 S_S^T + \beta_3 S_S^C + \beta_4 S_S^S)/\sum_j \beta_j \, (j = 1, 2, 3, 4) \quad (1)$$

S_S^D, S_S^T, S_S^C and S_S^S indicate the similarity measurement based on indication phenotype, functional consistency, chemical structure and side effect, respectively. The default value $\beta_j = 1$ 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:

$$S_M = (\alpha_1 S_M^D + \alpha_2 S_M^T)/\sum_i \alpha_i \, (i = 1, 2) \quad (2)$$

S_M^D and S_M^T 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

matrix (Kohler *et al.*, 2008). Let p_0 be the initial probability vector and p_t be a vector in which the i-th element holds the probability of finding the random walker at node i at step t. The probability vector at step $t+1$ can be given by

$$p_{t+1} = (1-\gamma)A^T p_t + \gamma p_0$$

$\gamma \in (0,1)$ indicates the restart probability. We extend the traditional RWR algorithm and apply it into an integrated network with two types of nodes that are composed of SM-SM and miRNA-miRNA networks based on the described integrated SM and miRNA similarity measures (Fig. 1). The corresponding initial seed node probability in the integrated network is $p_0 = \begin{bmatrix} (1-\eta)u_0 \\ \eta v_0 \end{bmatrix}_{((m+n)\times 1)}$. Let $u_{0(m\times 1)}$ and $v_{0(n\times 1)}$ denote the initial probabilities of the miRNA network and SM network, respectively. m and n represent the number of miRNAs and SMs. With respect to the impact of the two types of nodes, the parameter $\eta \in (0,1)$ is used to weigh the importance of the miRNA and SM networks. A is a probability transition matrix of the integrated network. A_{ij} represents the probability of getting from one miRNA (SM) to another miRNA (SM) or from one miRNA (SM) to a SM (miRNA). $\lambda \in (0,1)$ is a jumping probability from the miRNA network to the SM network (and vice versa) which indicates the reinforcement between the two networks (Li and Patra, 2010). After some steps, the steady probability $p_\infty = \begin{pmatrix} (1-\eta)u_\infty \\ \eta v_\infty \end{pmatrix}_{((m+n)\times 1)}$ is obtained when the change between p_t and p_{t+1} is less than10^{-10}. miRNAs are ranked based on u_∞. miRNAs with maximum in u_∞ among all the non-seed nodes is considered as the most probable miRNA target of SM i (see Supplementary File for detailed description of improved RWR algorithm).

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 SM 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

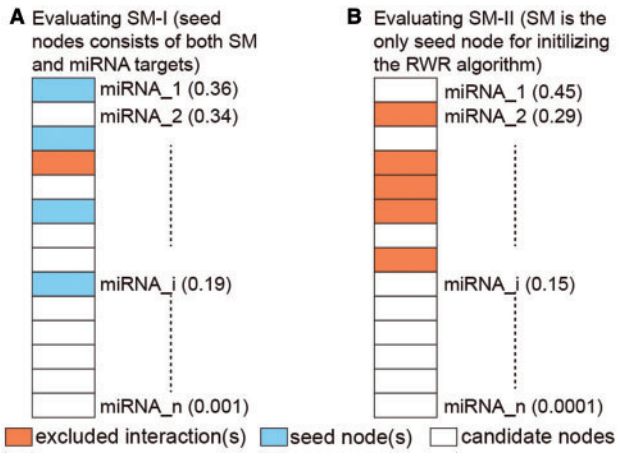

Fig. 2. Illustration of improved LOOCV procedures for validating our method. miRNAs were ranked according to the probability scores from RWR algorithm outputs (see Supplementary File)

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 empiric 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-Is (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 65% 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 similarities based on functional consistency, 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

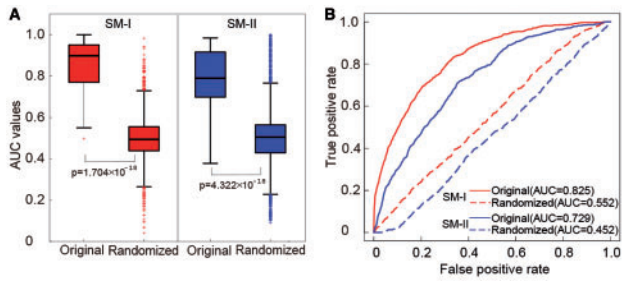

Fig. 3. (A) Box-plots of average AUC values reflecting original known and random model scenarios. The x-axes represent the original and random models for SM-I and SM-II, and the y-axes represent the average AUC values for each method. (B) The distribution of overall ROC curves between original and random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

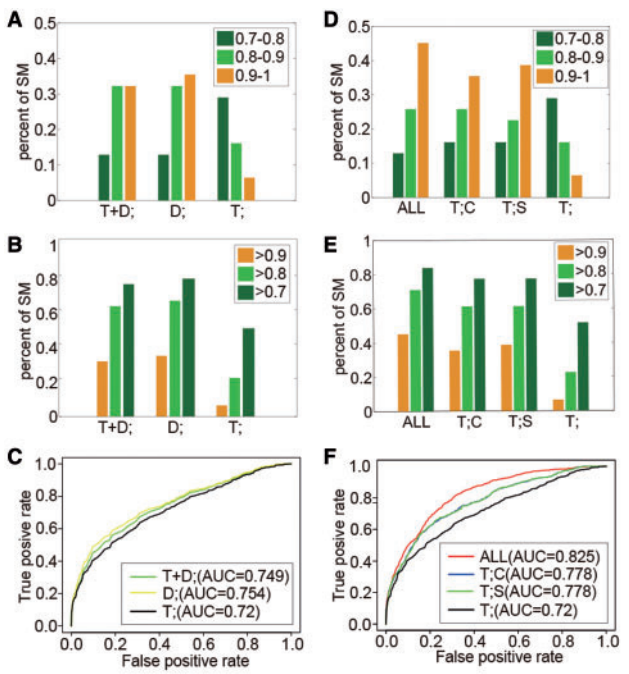

Fig. 4. (**A, D**). Bar graph of distribution of SMs in different ranges of AUCs. (**B, E**). Bar graph of cumulative distribution of SMs in different SM and its known target miRNAs or an SM alone, the known ranges of AUCs. The x-axes represent different models. while the y-axes represent the percent 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

'T;C' yielded AUCs greater than 0.8 for nearly 61% of the SMs and achieved an equal overall AUC (0.778) which demonstrated its usefulness as a biological information source. 'ALL' obtained the highest overall AUC. This indicates that adding various SM similarities to the integrated multi-level network improves prediction performance and is important for inferring new SM-miRNA relationships.

3.5 Predicting novel SM-miRNA associations

Upon confirming the reliability of our method by cross-validation, an independent dataset, and statistical test we further predicted novel miRNA targets associated with the two types of SMs using integrated miRNA similarities, SM similarities, and known SM-miRNA interactions. For each of the 831 SMs, if we start from an

miRNA targets should be ranked at the top of the prioritized target list. As shown in Supplementary Table S8, an accumulative hypergeometric test found that known miRNA targets for every tested SM were significantly enriched in the top 50 of the predictions. For, known miRNA targets for 90% (28) of SMs in SM-I and 67% (26) of SMs in SM-II were significantly enriched in the top 50 pooled miRNAs. The P-values are smaller than 0.05. We speculated that miRNAs ranked in the top 50 are more likely to reveal hidden SM-miRNA associations (Supplementary Table S9). We confirmed this using the following example. Letrozole (letrozole tablets, CID: 3902) is an oral non-

steroidal aromatase inhibitor that has been introduced for the adjuvant treatment of hormone-responsive breast cancer. Letrozole decreases estrogen levels by inhibiting aromatase and thereby eliminating the effect of estrogen on tumor growth stimulation. CID: 3902 was taken as a seed for predicting the most likely targeted miRNAs in our method. We used the top 50 predictions as the potential miRNA targets for Letrozole. Of the top 50 predictions there were 36, 13 and 38 miRNAs in the HMDD, miR2Disease and PhenomiR databases, respectively. Each were further identified to be closely related to breast cancer in which 215, 56,and 272 breast cancer-related miRNAs in the respective databases were common to the candidate miRNAs, and a hypergeometric test found that predictions ranking in the top 50 were significantly enriched in breast cancer-related miRNAs. We hypothesize that these miRNAs may be potential targets of letrozole. Supplementary Table S10 shows the rank numbers and P-values of the top 50 ranked miRNAs.

The 47th ranked candidate miRNA is miR-206 which post-transcriptionally regulates ERalpha (estrogen receptor 1, Era) in breast cancer as verified by miRTarBase (Hsu *et al.*, 2014), TarBase (Vergoulis *et al.*, 2011) and miRecord (Xiao *et al.*, 2009). Adams et al. found that transfection of MCF-7 cells with miR-206 specifically decreased or increased the ERalpha mRNA levels (Adams *et al.*, 2007). In addition, over expression of miR-206 reduced ERalpha mRNA levels. miR-206 may be involved in ER expression in breast cancer and may have an estrogen-dependent growth regulatory role or be involved with malignant transformation. Jelovac et al. studied Ovariectomized mice bearing tumor xenografts grown from aromatase-transfected ER-positive human breast cancer cells (MCF-7Ca) that were injected s.c.with 10 μg/d of letrozole for up to 56 weeks (Jelovac *et al.*, 2005). A western blot analysis of the tumors revealed that ERalpha was increased at 4 weeks but decreased at 28 and 56 weeks. This suggests that letrozole may target ER indirectly by interacting with miR-206.

4 Discussions

miRNAs are a newly discovered SM drug targets that play crucial roles in multiple human diseases. The identification of SMs that target dysregulated miRNAs in cancer would be helpful for developing a novel effective miRNA-associated therapeutic strategy. miRNA therapeutics are attracting special attention from both academia and biotechnology companies and the development of miRNA-targeted strategies is challenging.

In this study, we integrated an array of SM similarity metrics and miRNA similarity metrics, combining them with known associations between SMs and miRNAs, and constructed an integrated network. After applying a similarity-based RWR on this integrated network, we successfully predicted potential miRNA targets for 831 SMs on a large scale. We utilized LOOCVs, an independent dataset evaluations and Wilcoxon rank-sum test to estimate the stable of our method. Next, we analyzed the effect of merging similarity metrics on multiple networks and known targets of SMs on the performance of the method. As a result, it produced better results than single network similarity metrics or single networks alone. In addition, known targets of SMs can be used to improve the accuracy of the predictions as well. Taken together, the good performance of the method showed that our method can be used to discover novel SM-miRNA associations. Ultimately, for all the 831 SMs, we speculated that those miRNAs ranked in the top 50 are more likely to lead to SM-miRNA associations. This should be verified by further experiments and may provide future guidance for clinical treatments. In addition, known targets of SMs can be used to improve the accuracy of the predictions as well. To illustrate whether the known target factors of SMs influenced the results, we compared the prediction methods of SM-Is and SM-IIs using the same 31 SMs. Supplementary Figure S3 shows that the results of the SM-Is are better than those of the SM-IIs. In this case, potential miRNA targets of the predicted SM can be surmised according to the miRNA information of special SMs that are similar to the predicted SM. However, if the SM has known miRNAs, we can obtain predictions using miRNA similarity networks directly. The number of known targets of SMs can improve prediction performance.

In this work, the SM2miR dataset we developed provides reliable network seeds for predicting SM-miRNA associations. In addition, the creation of integrated heterogeneous network is valuable for inferring new SM-miRNA relationships. We integrated a variety of similarity metrics for SMs and miRNAs separately, which improved the identification of SMs targeting miRNAs. Importantly, this methodology can predict the potential targets of SMs even without the SMs having known target information.

The RWR method involves three parameters: the restart probability γ, jumping probability λ, controlling the impact of two kinds of seed nodes η, 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: α_1,α_2 β_1, β_2, β_3, β_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-Is and 45 for SM-IIs. We can't evaluate the performance of 'T;' 'D;' and 'T + D;' for SM-IIs, because the SM-IIs 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 inpart 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 2014 CB910504], 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

Adams,B.D. *et al.* (2007) The micro-ribonucleic acid (miRNA) miR-206 targets the human estrogen receptor-alpha (ERalpha) and represses ERalpha messenger RNA and protein expression in breast cancer cell lines. *Mol. Endocrinol.*, **21**, 1132–1147.

Alaimo,S. *et al.* (2013) Drug-target interaction prediction through domain-tuned network-based inference. *Bioinformatics*, **29**, 2004–2008.

Albert,R. (2005) Scale-free networks in cell biology. *J. Cell Sci.*, **118**, 4947–4957.

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser.*, **57**, 289–300.

Bose,D. *et al.* (2012) The tuberculosis drug streptomycin as a potential cancer therapeutic: inhibition of miR-21 function by directly targeting its precursor. *Angew. Chem. Int. Ed. Engl.*, **51**, 1019–1023.

Chen,L. *et al.* (2013) Prediction of drug target groups based on chemical-chemical similarities and chemical-chemical/protein connections. *Biochim. Biophys. Acta*, **1844**, 207–213.

Chen,X. *et al.* (2012a) Prediction of disease-related interactions between microRNAs and environmental factors based on a semi-supervised classifier. *PLoS One*, **7**, e43425.

Chen,X. *et al.* (2012b) Drug–target interaction prediction by random walk on the heterogeneous network. *Mol. BioSyst.*, **8**, 1970–1978.

Chen,X. *et al.* (2012c) RWRMDA: predicting novel human microRNA–disease associations, *Mol. BioSyst.*, **8**, 2792–2798.

Gottlieb,A. *et al.* (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.*, **7**, 496.

Hattori,M. *et al.* (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.*, **125**, 11853–11865.

Hesse,M. and Arenz,C. (2014) miRNAs as novel therapeutic targets and diagnostic biomarkers for Parkinson's disease: a patent evaluation of WO2014018650. *Expert. Opin. Ther. Pat.*, **24**, 1271–1276.

Hsu,S.D. *et al.* (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.*, **42**, D78–D85.

Jamal,S. *et al.* (2012) Computational analysis and predictive modeling of small molecule modulators of microRNA. *J. Cheminform.*, **4**, 16.

Jelovac,D. *et al.* (2005) Activation of mitogen-activated protein kinase in xenografts and cells during prolonged treatment with aromatase inhibitor letrozole. *Cancer Res.*, **65**, 5380–5389.

Jiang,Q. *et al.* (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.*, **37**, D98–D104.

Jiang,W. *et al.* (2012) Identification of links between small molecules and miRNAs in human cancers based on transcriptional responses. *Sci. Rep.*, **2**, 282.

Knox,C. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.

Kohler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.

Lagana,A. *et al.* (2010) Variability in the incidence of miRNAs and genes in fragile sites and the role of repeats and CpG islands in the distribution of genetic material. *PLoS One*, **5**, e11166.

Lanford,R.E. *et al.* (2010) Therapeutic silencing of microRNA-122 in primates with chronic hepatitis C virus infection. *Science*, **327**, 198–201.

Li,X. *et al.* (2004) Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling. *Nucleic Acids Res.*, **32**, 2685–2694.

Li,Y. and Patra,J.C. (2010) Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, **26**, 1219–1224.

Liu,C.G. *et al.* (2004) An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues. *Proc. Natl. Acad. Sci. USA*, **101**, 9740–9744.

Liu,X. *et al.* (2013) SM2miR: a database of the experimentally validated small molecules' effects on microRNA expression. *Bioinformatics*, **29**, 409–411.

Liu,Z. *et al.* (2008) MicroRNA: An emerging therapeutic target and intervention tool. *Int. J. Mol. Sci.*, **9**, 978–999.

Lu,M. *et al.* (2008) An analysis of human microRNA and disease associations. *PLoS One*, **3**, e3420.

Lv,S. *et al.* (2012) A novel method to quantify gene set functional association based on gene ontology. *J. R. Soc. Interface*, **9**, 1063–1072.

Meng,F. *et al.* (2013) Constructing and characterizing a bioactive small molecule and microRNA association network for Alzheimer's disease. *J. R. Soc. Interface*, **11**, 20131057.

Rai,A. *et al.* (2014) Randomness and preserved patterns in cancer network. *Sci. Rep.*, **4**, 6368.

Ruepp,A. *et al.* (2010) PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. *Genome Biol.*, **11**, R6.

Srinivasan,S. *et al.* (2013) MicroRNAs -the next generation therapeutic targets in human diseases. *Theranostics*, **3**, 930–942.

Takarabe,M. *et al*. (2012) Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics*, **28**, i611–i618.

Thomas,J.R. and Hergenrother,P.J. (2008) Targeting RNA with small molecules. *Chem. Rev.*, **108**, 1171–1224.

Van Laarhoven,T. *et al*. (2011) Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics*, **27**, 3036–3043.

Vergoulis,T. *et al*. (2011) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res*., **40**, D222–D229.

Volinia,S. *et al*. (2006) A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc. Natl. Acad. Sci. USA*, **103**, 2257–2261.

Wang,Y. *et al*. (2013) Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PLoS One*, **8**, e78518.

Wang,Y. *et al*. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.

Wu,W. *et al*. (2007) MicroRNA and cancer: current status and prospective. *Int. J. Cancer*, **120**, 953–960.

Xiao,F. *et al*. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res*., **37**, D105–D110.

Yang,L. *et al*. (2014) Human proteins characterization with subcellular localizations. *J. Theor. Biol.*, **358**, 61–73.

Zhang,S. *et al*. (2010) Targeting microRNAs with small molecules: from dream to reality. *Clin. Pharmacol. Ther*, **87**, 754–758.