Molecular BioSystems

PAPER

Cite this: *Mol. BioSyst.*, 2016, 12, 624

miREFRWR: a novel disease-related microRNA-environmental factor interactions prediction method†‡

Xing Chen*[ab]

Received 18th October 2015,
Accepted 6th December 2015

DOI: 10.1039/c5mb00697j

www.rsc.org/molecularbiosystems

Increasing evidence has indicated that microRNAs (miRNAs) can functionally interact with environmental factors (EFs) to affect and determine human diseases. Uncovering the potential associations between diseases and miRNA–EF interactions could benefit the understanding of the underlying disease mechanism at miRNA and EF levels, miRNA signatures identification, and drug repurposing. In this study, based on the assumption that similar miRNAs (EFs) tend to interact with similar EFs (miRNAs) in the context of a given disease and under the framework of random walk with restart (RWR), a novel method of miREFRWR was developed to uncover the hidden disease-related miRNA–EF interactions by implementing random walks on an miRNA similarity network and EF similarity network, respectively. miREFRWR was evaluated by leave-one-out cross-validation, which achieved an AUC of 0.9500. It has been demonstrated that miREFRWR can effectively identify potential interactions in all the test classes, even if these test samples only share either EFs or miRNAs with the training samples. Furthermore, many predictive results for acute promyelocytic leukemia and breast cancer (67 and 10 interactions out of the top 1% predictions, respectively) have been verified by independent experimental studies. It is anticipated that miREFRWR could be a useful and important biological resource for biomedical research.

Introduction

Phenotypes and diseases are often determined by the complex interactions between genetic factors (GFs) and environmental factors (EFs).[1–4] As one class of important and newly identified GFs and as one of the most important components of the cell, microRNAs (miRNAs) play critical roles in many important biological processes, including cell growth, proliferation, differentiation, apoptosis, signal transduction, and viral infection.[5–11] Evidence have indicated that miRNAs are associated with various diseases.[5,12–20] One typical example is insulin secretion, which can be regulated by mir-375.[21,22] Numerous miRNAs have been linked with the initiation and development of various cancers.[23] For example, Huang *et al.* (2008) confirmed the upregulation of miR-373 and miR-520c to be a significant player in tumour invasion and metastasis.[24] Therefore, the interactions between miRNAs and EFs may contribute to the development or treatment of many phenotypes and diseases.

Recently, increasing studies have indicated that miRNAs can functionally interact with plenty of EFs to affect and determine the phenotypes and diseases. Related EFs include drugs,[25] alcohol,[26] cigarette,[27] stress,[28] diet,[29] virus,[30] air pollution,[31] and radiation.[32] For example, cigarette smoke condensate (CSC) could lead to cancer by dramatically increasing the expression level of mir-31 and hence by activating LOC554202 in normal respiratory epithelia and lung cancer cells.[33] More importantly, the interactions between miRNA and EF also can benefit the disease treatment. For instance, during the clinical treatment of ovarian cancer, paclitaxel could significantly decrease the expression of mir-29c.[34] In breast cancer treatment, 3,3′-diindolylmethane (DIM) could inhibit the proliferation of breast cancer cell by increasing miR-21 expression and hence causing the downregulation of Cdc25A.[35] Therefore, identifying potential disease-related miRNA–EF interactions based on computational methods has become an important problem to solve in biomedical research and can play a critical role in understanding disease pathogenesis at the miRNA–EF interactions level, in miRNA signatures identification for given EFs, and for the new indication inference of approved drugs. Computational prediction has been an important complementary method for disease-related interactions identification, and could

[a] *National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 100190, China. E-mail: xingchen@amss.ac.cn*
[b] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. E-mail: xingchen@amss.ac.cn*

† XC conceived the prediction method, developed the prediction method, conceived, designed and implemented the experiments, analyzed the result, and wrote the paper. All authors read and approved the final manuscript.

‡ Electronic supplementary information (ESI) available. See DOI: 10.1039/c5mb00697j

be used to select promising disease-related interactions for further experimental validation, and hence it can decrease the time and cost of biological experiments.[36–42]

Yang *et al.* (2011) manually collected experimentally supported disease-related miRNA–EF interactions and further constructed an miREnvironment database, which included more than 2500 entries regarding ~800 miRNAs, ~260 EFs, ~180 phenotypes, and 17 species.[43] Qiu *et al.* (2012) analyzed disease-related human miRNA–EF interactions in the miREnvironment database and obtained some important conclusions about the association patterns of miRNA–EF interactions.[44] These conclusions indicated that miRNA–EF interactions had a significant correlation with characteristics such as miRNA expression level, tissue specificity, conservation, and disease spectrum width. They further developed several methods for EF relationship characterization, cancer treatment result prediction, and novel EF–disease interactions inference. Although these proposed methods cannot predict ternary relationships among miRNAs, EFs, and diseases simultaneously, they laid a theoretical foundation for disease-related miRNA–EF interactions prediction research. In my previous study, the similar nature of disease-related miRNA–EF interactions was proposed, *i.e.* similar miRNAs (EFs) tend to interact with similar EFs (miRNAs) in the context of a given disease.[45] Based on this assumption and under the framework of a semi-supervised classifier, a semi-supervised classifier-based method (miREFScan) to predict potential disease-related interactions between miRNAs and EFs has been developed. Reliable performance has been obtained in both cross-validation and in a case study about acute promyelocytic leukaemia (APL).[45] To the best of my knowledge, miREFScan is the first computational tool for the prediction of simultaneous ternary relationships among miRNAs, EFs, and diseases.

However, to date, little effort has been made to analyze and predict potential disease-related miRNA–EF interactions from a network perspective. Network medicine could effectively predict the potential interactions among biological molecules, investigate how cellular systems induce different biological phenotypes under different conditions, and provide a novel approach to understand the complicated mechanism of disease and drug treatment.[1] In particular, network-based computational models have been widely used to predict disease-related genes, miRNAs, long non-coding RNAs (lncRNAs), and drug–target interactions.[36,46–50] Therefore, it is fundamentally important to understand the mechanisms of complex diseases and to identify new indications of drugs in a network-centric perspective. In the present study, a novel method of miREFRWR (miRNA–EF interactions inference based on the random walk with restart) to infer potential disease-related miRNA–EF interactions is developed by making full use of the network tool for data integration to predict potential associations. The method comprises four steps: first, three networks (miRNA–miRNA similarity network, EF–EF similarity network, and a known miRNA–EF interaction network for a given disease) are constructed; second, random walks are implemented on the miRNA similarity network and EF similarity network; third, predictive results based on a random walk on a miRNA similarity network and EF similarity network are combined to obtain the final predictive results; finally, the most probable miRNA–EF pairs are selected according to the stable probability of the random walk. In the framework of ''leave-one-out'' cross-validation (LOOCV), miREFRWR achieved a comparable performance (AUC = 0.9500) with miREFScan in my previous work. It has also been demonstrated that miREFRWR has a reliable performance in all the test classes, even if the test samples only shared either EFs or miRNAs with the training samples. In the case studies about APL and breast cancer, miREFRWR further showed the advantages of making full use of network information to predict potential interactions. In particular, in the APL-related miRNA–EF interactions prediction, sixty-seven interactions out of the top 1% of predictions based on miREFRWR have been confirmed by experimental literature. I further applied miREFRWR to predict potential novel miRNA–EF interactions for all the investigated diseases in the dataset. The top 100 interactions for each disease have been publicly released for further biological experiment validation.

Methods

Disease-related miRNA–EF interactions

First, the whole dataset of known disease-related miRNA–EF interactions was downloaded from the miREnvironment database (http://cmbi.bjmu.edu.cn/miren, Version: September, 2011),[43] including more than 2500 entries, where each entry was composed of an miRNA name, an EF name, and their related phenotype/disease. This database is a very important and useful biological resource for research about the mutual relationships among miRNAs, EFs, and diseases and lays the data foundation for disease-related miRNA–EF interactions identification. Second, human disease-related miRNA–EF interactions were extracted from the abovementioned dataset and double-checked to eliminate entries with the phenotype named ''n/a''. Furthermore, the names of diseases, miRNAs, and EFs were normalized, and 862 distinct human disease-related miRNA–EF interactions were obtained, which contained information about 418 miRNAs, 138 EFs, and 97 diseases (see ESI,‡ Table S1). This dataset was regarded as the gold standard dataset in the cross-validation and case studies for performance evaluation. Finally, a disease-related miRNA–EF interaction adjacency matrix A was constructed for each given disease, where the entity $A(i,j)$ in row i and column j is 1 if EF j could interact with miRNA i and their interaction could contribute to the given disease, otherwise it takes the value 0.

Chemical structural similarity between EFs

In previous studies related to drugs,[45,46,51–58] chemical structure has been widely applied to effectively evaluate drug similarities. Considering the fact that many EFs are drugs in the known disease-related miRNA–EF interactions dataset, the chemical structural similarity matrix SCE was constructed (here, C denotes chemical structure and E denotes EF) for EFs based on the SIMCOMP tool[59] and the drug chemical structure

information derived from various databases, such as the KEGG database,[60] PubChem,[61] and ChemicalBook (http://www.chemicalbook.com/). The similarity score computed in this way is a global ratio between the size of common structures and the union structures of two drugs based on a graph alignment algorithm. The entity SCE(i,j) in the row i and column j is the chemical structure similarity score between EF i and j, if they are both drugs, otherwise it takes the value 0.

Functional similarity between miRNAs

Based on the assumption that miRNAs with similar functions are more likely to be related with similar diseases,[12] Wang *et al.* (2010) represented the relationships among different diseases with a directed acyclic graph (DAG) and further inferred miRNA functional similarity by calculating the similarity between their associated disease DAGs.[62] Herein, miRNA functional similarity scores were calculated by the MISIM tool in May 2011 (http://cmbi.bjmu.edu.cn/misim/).[62] Then, the miRNA functional similarity matrix SFM was constructed (M denotes miRNA and F denotes functional similarity), where the entity SFM(i,j) in row i column j is the functional similarity score between miRNA i and j. Previous studies have shown that functional similarity scores calculated in this way coincide well with prior knowledge about miRNA function annotations.[62] In addition, the miRNA functional similarity network plays a critical role in disease-related miRNAs and miRNA–EF interactions identification.[36,37,45]

Network-based similarity for miRNAs and EFs

Considering the fact that some EFs are not drugs and some miRNAs do not have any known associated diseases, the similarity scores for these EFs and miRNAs cannot be obtained based on the aforementioned chemical structure similarity and functional similarity. Herein, network-based similarity for miRNAs and EFs is proposed to improve the traditional similarity measure (see Fig. 1). A disease–miRNA, disease–EF, and EF–miRNA interactions network can be obtained from respective known disease-related miRNA–EF interactions. Based on the observation that EFs interacting with more common miRNAs or diseases tend to be more similar, the network-based EF similarity matrix SME and SDE (here, E denotes EF, M (D) indicates the network-based similarity obtained from the EF–miRNA (disease) interactions) are constructed by extracting the information from the disease–EF and EF–miRNA interactions networks, respectively, where the entity SME(i,j) SDE(i,j) in row i column j is the number of miRNAs (diseases) shared by EF i and j in EF–miRNA (disease) interactions network. Correspondingly, from the disease–miRNA and miRNA–EF interactions network, network-based miRNA similarity matrix SDM and SEM (here, M denotes miRNA, D (E) indicates the network-based similarity obtained from disease (EF)–miRNA interactions) can be obtained in a similar way, where the entity SDM(i,j) (SEM(i,j)) in row i column j is the number of diseases (EFs) shared by miRNA i and j. Network-based similarity must be normalized. Taking SME

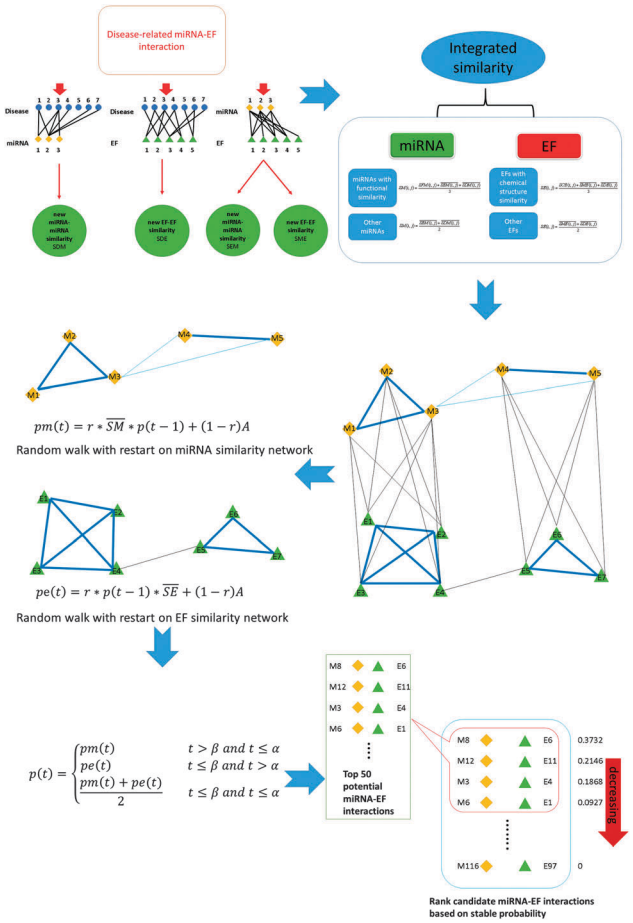

Fig. 1 Flowchart of miREFRWR. This flowchart provides a brief description of the new method developed in the present paper. Step 1: calculating network-based miRNA similarity and EF similarity. Step 2: calculating integrated miRNA similarity and EF similarity. Step 3: constructing a heterogeneous network. Here, a simple example is provided. The upper network is the miRNA similarity network and the lower network is the EF similarity network. They are connected into a heterogeneous network by known miRNA–EF interactions. Step 4: implementing random walks on the miRNA similarity network and EF similarity network and introducing two parameters to restrict the iteration steps of random walks on these two networks. Step 5: combining predictive results based on random walks on the miRNA similarity network and EF similarity network to obtain final predictive results. Step 6: ranking all the candidate miRNA–EF interactions based on stable probability and selecting potential disease-related miRNA–EF interactions for experimental validation.

as an example, the corresponding normalized matrix is defined as follows:

$$\overline{\mathrm{SME}} = (\mathrm{DME})^{-1/2}\mathrm{SME}(\mathrm{DME})^{-1/2}$$

where the diagonal matrix DME is defined such that DME(i,i) is the sum of the ith row of SME. The other three network-based similarity matrices were also normalized in a similar way. To avoid a circular design and optimistic prediction performance report in LOOCV, the network-based miRNA similarity and EF similarity were recalculated when each cross-validation run was implemented, *i.e.* the information of tested disease-related miRNA–EF interactions was discarded from the known

disease-related miRNA–EF interaction network and only the current training dataset was used to calculate the network-based similarity.

Integrated similarity for miRNAs and EFs

Based on the aforementioned drug chemical structure similarity and network-based EF similarity, an integrated similarity matrix SE for EFs can be constructed based on trivial combinatorial coefficients (see Fig. 1), where the entity $\mathrm{SE}(i,j)$ in row i column j is defined as follows:

$$\mathrm{SE}(i,j) = \begin{cases} \dfrac{\mathrm{SCE}(i,j) + \overline{\mathrm{SME}(i,j)} + \overline{\mathrm{SDE}(i,j)}}{3} & i,j \in \mathrm{IE} \\ \dfrac{\overline{\mathrm{SME}(i,j)} + \overline{\mathrm{SDE}(i,j)}}{2} & \text{otherwise} \end{cases}$$

where IE is the set of drugs among all the EFs investigated in the present study.

Moreover, an integrated similarity matrix SM for miRNAs can be defined in a similar way (see Fig. 1), where the entity $\mathrm{SM}(i,j)$ in row i column j is defined as follows:

$$\mathrm{SM}(i,j) = \begin{cases} \dfrac{\mathrm{SFM}(i,j) + \overline{\mathrm{SEM}(i,j)} + \overline{\mathrm{SDM}(i,j)}}{3} & i,j \in \mathrm{IM} \\ \dfrac{\overline{\mathrm{SEM}(i,j)} + \overline{\mathrm{SDM}(i,j)}}{2} & \text{otherwise} \end{cases}$$

where IM is the set of miRNAs that have known functional similarity with the other miRNAs investigated in this study. It must be pointed out that combinatorial coefficients could be better selected according to further cross-validation. For simplicity, the trivial combinatorial coefficients have been adopted here according to previous studies, where similar operations of combining different similarity measures into an integrated similarity have been adopted.[45,46] In these previous studies, reliable predictive performance has been obtained and the robustness of predictive accuracy towards combinatorial coefficients selection has been illustrated.[45,46]

miREFRWR

Herein, based on the assumption that similar miRNAs (EFs) tend to interact with similar EFs (miRNAs) in the context of a given disease and under the framework of random walk with restart (RWR),[36,46,48,50] I developed a novel method of miREFRWR to infer potential disease-related miRNA–EF interactions. As is well known, traditional RWR has been applied widely to plenty of biological problems such as disease genes prioritization,[48,50] drug–target interactions prediction,[46] and disease-related miRNAs inference.[36] However, traditional RWR has some critical limitations. miREFRWR is significantly different from traditional RWR. miREFRWR could make full use of the network tool for data integration to predict potential associations. The information of known disease-related miRNA–EF interactions, drug chemical structure similarity, and miRNA functional similarity was integrated in the framework of miREFRWR.

Based on the abovementioned basic ideas, miREFRWR was developed as follows (see Fig. 1, motivated by literature[63]). First, some matrices were normalized before random walk was implemented on the network. Taking the miRNA similarity matrix SM as an example, the corresponding normalized matrix is defined as follows:

$$\overline{\mathrm{SM}} = (\mathrm{DM})^{-1/2}\mathrm{SM}(\mathrm{DM})^{-1/2}$$

where the diagonal matrix DM is defined such that $\mathrm{DM}(i,i)$ is the sum of the ith row of SM. The EF similarity matrix SE is also normalized in a similar way. For the disease-related miRNA–EF interaction adjacency matrix A, the entity $A(i,j)$ in row i and column j is divided by the sum of elements in the matrix A. Thus, a normalized miRNA similarity matrix $\overline{\mathrm{SM}}$, normalized EF similarity network $\overline{\mathrm{SE}}$, and a normalized interaction adjacency matrix $\bar{A}$ (for a brief description of the following equation, I set $p(0) = \bar{A}$) were constructed, respectively. Second, considering the fact that there are different topologies and network structures in the miRNA similarity network and EF similarity network and hence the optimal iteration steps might be different with the two networks, two parameters were introduced to restrict the respective number of iteration steps of random walk on these two networks (motivated by literature[63]). Here, the parameters α and β are used to denote the number of maximum iterations in the miRNA similarity network and in the EF similarity network, respectively. Furthermore, the restart of random walk in every time step at source nodes can be allowed with the probability r $(0 < r < 1)$. Third, two random walks are implemented on the miRNA similarity network and EF similarity network, respectively. The random walks in these two networks finally converge to a unique solution after some steps of iterations. The predictive results from these two random walks are then combined to give the final prediction. miREFRWR is defined as follows (motivated by literature[63]):

$$\text{for } t = 1 \text{ to } \max(\alpha,\beta)$$

$$\text{if } t \le \alpha$$

$$pm(t) = r \times \overline{\mathrm{SM}} \times p(t-1) + (1-r)A$$

$$\text{if } t \le \beta$$

$$pe(t) = r \times p(t-1) \times \overline{\mathrm{SE}} + (1-r)A$$

$$p(t) = \begin{cases} pm(t) & t > \beta \text{ and } t \le \alpha \\ pe(t) & t \le \beta \text{ and } t > \alpha \\ \dfrac{pm(t) + pe(t)}{2} & t \le \beta \text{ and } t \le \alpha \end{cases}$$

where $p(t)$ is a matrix with the entity in row i and column j as the probability of arriving at the pair consisting of miRNA i and EF j at the time step t. The value of matrix p could be updated based on this iteration equation and the current value of matrix p. Finally, when the number of iterations exceeds the maximum of α and β, the random walk is terminated. Candidate miRNA–EF pairs are ranked according to corresponding values in the final probability

matrix p to select potential disease-related miRNA–EF interactions. The high-scoring interactions can be expected to have a high probability to be associated with the given disease and will have priority to be tested in the biological experiments.

Results

Leave-one-out cross-validation

Parameters $\alpha = 4$, $\beta = 4$ and $r = 0.8$ were chosen according to previous studies.[63] Actually, these parameters could be better selected based on further cross-validation, and the influence of parameter selection on the predictive results will be discussed in the following section.

LOOCV was implemented to evaluate the performance of miREFRWR. Considering the fact that miREFRWR cannot rank candidate miRNA–EF interactions for all the diseases simultaneously, LOOCV was instead implemented for each disease. In the known disease-related miRNA–EF interaction dataset, only about 8.89 miRNA–EF interactions have been associated with each disease on average, which means there is little difference between LOOCV and 10-fold cross-validation. Furthermore, 32, 17, 12, 9, and 3 out of all the 97 diseases have 1, 2, 3, 4, or 5 known related interactions, respectively, which means multi-fold cross-validation also cannot be implemented for most of the diseases. For these reasons, LOOCV was selected for performance validation.

In the LOOCV scheme, each known interaction associated with the given disease is taken in turn as the test sample and other known interactions associated with this disease are taken as the training samples. Therefore, if this disease has only one known miRNA–EF interaction, LOOCV cannot be implemented. The aforementioned, network-based miRNA similarity and EF similarity matrices were recalculated with each cross-validation run implemented to avoid a circular design and optimistic prediction performance of LOOCV. The performance of miREFRWR was evaluated based on the rank of this test sample in the candidate samples, which were composed of known left-out interactions and miRNA–EF pairs without the known associations with the given disease. Furthermore, a ROC curve (plotting true positive rate (TPR, sensitivity) *versus* the false positive rate (FPR, 1-specificity) at different cut-offs) was obtained and AUC was calculated (area under the ROC curve). AUC = 1 shows a perfect performance and 0.5 indicates a random performance.

miREFRWR was compared with miREFScan (see Fig. 2), which was the first disease-related miRNA–EF interaction prediction method. Consequently, miREFRWR achieved an AUC of 0.9500, which showed a comparable performance to miREFScan. However, as a network-based method for miRNA–EF interactions prediction, it could bring a novel network perspective to the current research and promote the progression of developing network-based methods for miRNA–EF interactions prediction in the future.

To evaluate whether the results of LOOCV by miREFRWR were likely to have been obtained by chance, 100 random

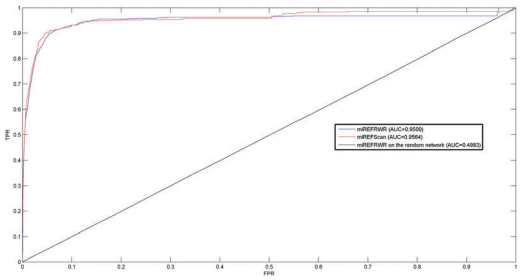

Fig. 2 Performance comparison. Comparison between miREFRWR and the first disease-related miRNA–EF interaction prediction method miREFScan in terms of ROC curve and AUC based on LOOCV. As a result, miREFRWR achieved an AUC of 0.9500, which showed its comparable performance with miREFScan.

disease-related miRNA–EF interaction networks were generated. LOOCV procedure was implemented over these random networks and the mean FPR and mean TPR were obtained to plot the ROC curve and to calculate AUC. As a result, the AUC of 0.4983 demonstrated that the observed excellent performance of miREFRWR could not be achieved by chance, and hence prediction results by miREFRWR would be of biological significance and could reflect some mechanisms of human complex diseases (see Fig. 2).

Furthermore, miREFRWR was compared with some similar versions of miREFRWR, which either ignored the use of network-based similarity or implemented miREFRWR only on a single network (see ESI,‡ Fig. S1). As a result, miREFRWR was significantly improved compared to other methods, demonstrating the reasonability of introducing a network-based similarity (details on the AUC comparison between miREFRWR and miREFRWR without introducing a network-based similarity are given in ESI,‡ Fig. S1) and implementing miREFRWR on both the miRNA similarity network and EF similarity network (AUC comparisons between miREFRWR and miREFRWR implemented only on the single network as shown in ESI,‡ Fig. S1).

Moreover, when LOOCV was implemented for each disease, the ROC curve and the corresponding AUC could be obtained to assess how well the known miRNA–EF interactions of this disease are ranked relative to the candidate pairs (see ESI,‡ Table S2). The performance of the miREFRWR was evaluated by counting how many diseases had an AUC greater than different cutoffs (see Fig. 3).

Parameter effects on the performance of miREFRWR

There are three parameters in miREFRWR, including the number of maximum iterations in the miRNA similarity network and EF similarity network, and the restart probability. To investigate the parameter effects on the performance of miREFRWR, various values were assigned to the three parameters (the number of maximum iterations was taken to be between 1 step and 5 steps and the restart probability was chosen from 0.1 to 0.9), and the corresponding AUC of miREFRWR was calculated under the framework of LOOCV (see ESI,‡ Table S3). The result demonstrated that miREFRWR could achieve an excellent performance with almost all the parameters selected (see Fig. 4 and 5).

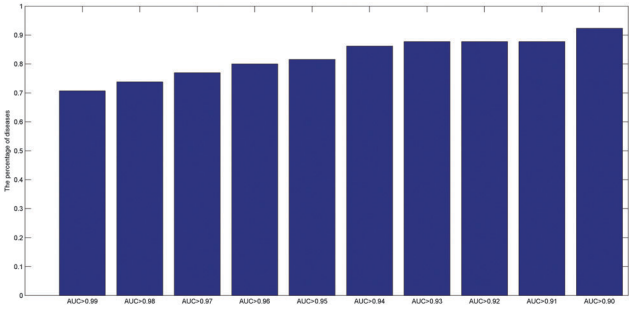

Fig. 3 Performance of miREFRWR based on AUC. The performance of miREFRWR was evaluated by counting how many diseases had an AUC greater than different cutoffs.

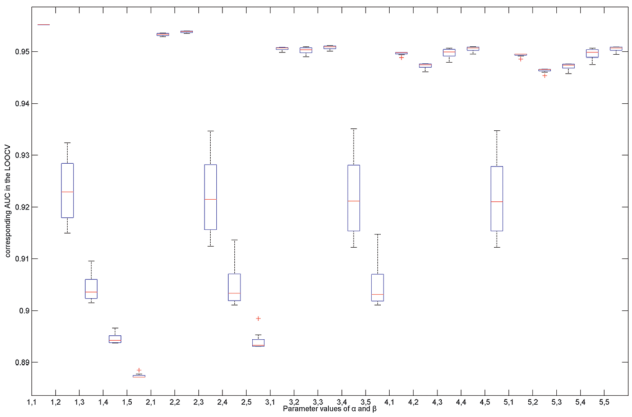

Fig. 4 Performance of miREFRWR based on different selections of the number of maximum iterations. To investigate the parameter effects on the performance of miREFRWR, various values were assigned to α and β, and the corresponding AUC of miREFRWR was calculated under the framework of LOOCV. The result demonstrates that miREFRWR can exhibit excellent performance with almost all the parameters selected. From this figure, it could be easily found that miREFRWR tends to show better performance when parameter α is greater than or equal to β, and a worse performance when α is less than β.

For the selection of parameters α and β, $\alpha = 4$ and $\beta = 4$ were reported to produce the best performance in the previous research about disease genes prioritization.[63] However, interesting conclusions can be obtained from the results in current disease-related miRNA–EF interactions prediction. A box plot for the AUCs under the framework of LOOCV and corresponding to different parameter values of α and β is shown in Fig. 4. From this figure, it could be easily found that miREFRWR tends to show better performance when parameter α is greater than or equal to β. Further confirmation can be obtained from the box plot for the AUCs under the framework of LOOCV when α is greater than or equal to β and α is less than β (see ESI,‡ Fig. S2). This observation may arise from the fact that most EFs show little similarity to other EFs and only local network information can be obtained for the random walk on the EF similarity network. Instead, the edges in the miRNA similarity network are denser than the edges in EF similarity network. Therefore, the number of random walk steps on the EF similarity network should be less than the steps on the miRNA similarity network.

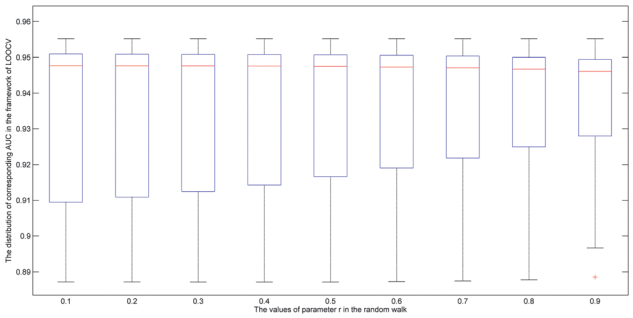

Fig. 5 Performance of miREFRWR based on different selections of restart probability. To investigate the performance of miREFRWR based on different selections of restart probability, various values of r ranging from 0.1 to 0.9 were set and AUC was calculated under the framework of LOOCV when different parameter values of α and β were chosen. A box plot for the AUCs corresponding to different parameter values of r is shown. It could be observed that the performance of miREFRWR is stable based on any selection of parameter values.

It has been demonstrated that the predictive results of random walk are robust to the restart probability in previous studies about disease-related genes identification and disease–miRNA association inference.[36,48,50] As mentioned before, to investigate the performance of miREFRWR based on different selections of restart probability, various values of r ranging from 0.1 to 0.9 were adopted, and the corresponding AUCs were calculated under the framework of LOOCV when different parameter values of α and β were chosen. A box plot for the AUCs corresponding to different parameter values of r is shown in Fig. 5. It can be observed that the performance of miREFRWR is stable based on any selection of parameter values.

LOOCV under the new framework

Flaws in the evaluation procedure for the pair-input computational prediction problems based on cross-validation have been pointed out in a recent study.[64] For instance, the paired nature of inputs causes a natural partitioning of test samples. Normally, pair-input computational methods achieve different predictive performances for distinct test classes.[64] Based on this new validation framework, the test pairs of disease-related miRNA–EF interactions are classified into four distinct classes: C1 is composed of the test samples sharing both EFs and miRNAs with the training samples; C2 is composed of the test samples sharing only miRNAs with the training samples; C3 is composed of the test samples sharing only EFs with the training samples; and C4 is composed of the test samples sharing neither EFs nor miRNAs with the training samples. LOOCV was implemented for these four test classes, and the corresponding performance of miREFRWR are shown in Fig. 6 (AUC of 0.9931 in C1, 0.7929 in C2, 0.9548 in C3, and 0.6803 in C4). The results demonstrate that miREFRWR has a reliable performance in all the test classes, even if the test samples only share either EFs or miRNAs with the training samples.

Case studies

APL, a subtype of acute myelogenous leukemia, is regarded as the most malignant form of acute leukemia, with a severe

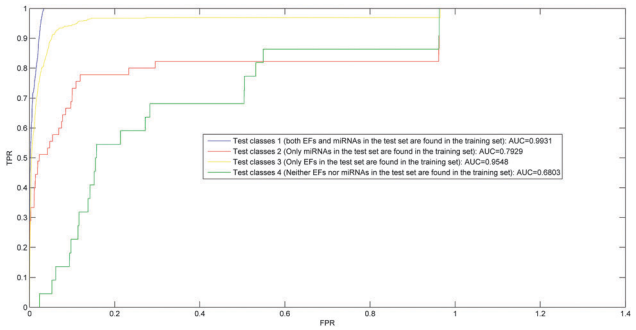

Fig. 6 Performance evaluation of miREFRWR based on the LOOCV under the new framework. To further investigate the performance of miREFRWR, LOOCV was implemented for four test classes, and the corresponding performance of miREFRWR is shown (AUC of 0.9931 in C1, 0.7929 in C2, 0.9548 in C3, and 0.6803 in C4). The results demonstrate that miREFRWR has a reliable performance in all the test classes, even if the test samples share neither EFs nor miRNAs with the training samples.

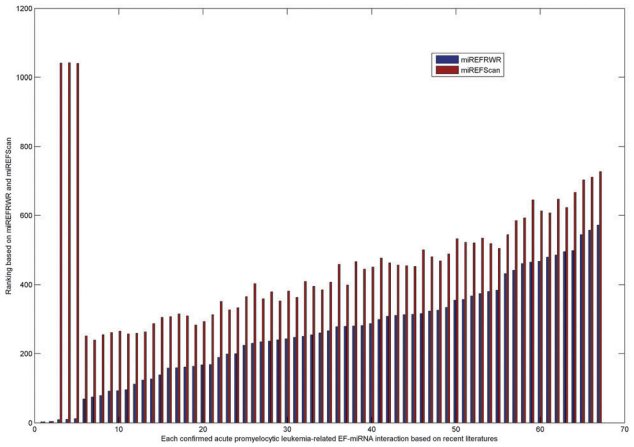

Fig. 7 Case study of APL. For APL-related miRNA–EF interactions prediction, 67 out of the top 1% candidate interactions have been confirmed by latest experimental literature. The previous method, miREFScan, only found 53 confirmed interactions. Moreover, all the confirmed interactions always obtained better ranking in the predictive list of miREFRWR over miREFScan.

bleeding tendency and a highly fatal course of only weeks.[65,66] Many studies have shown that the combined action of miRNAs and EFs could contribute to the development of effective therapy ways for APL. For example, four known APL-related miRNA–EF interactions have been provided in the training dataset. All-trans retinoic acid (ATRA) can benefit the treatment of APL by suppressing the regulation of let-7a, mir-15a, and mir-16.[67] The interaction between mir-21 and arsenic trioxide (ATO) may have a great curative effect on APL by regulating ATO-induced cell death.[68] Therefore, identifying disease-related miRNA–EF interactions could play a great role during clinical treatment.

Herein, potential APL-related miRNA–EF interactions were predicted based on miREFRWR. As a result, 67 out of the top 1% candidate interactions have been confirmed by latest experimental literature[65,69,70] (see ESI,‡ Table S4). The previous method, miREFScan, only found 53 confirmed interactions. Fourteen confirmed interactions predicted by miREFRWR cannot be obtained by miREFScan, while all the confirmed interactions predicted by miREFScan can be obtained by miREFRWR. Moreover, all the confirmed interactions always obtain better ranking in the predictive list of miREFRWR than with miREFScan (see Fig. 7). For the top 0.5% and 0.1% of the predictive list, 40 and 5 interactions based on miREFRWR have been confirmed, respectively. However, miREFScan only found 12 and 2 confirmed interactions. The abovementioned comparisons between miREFRWR and miREFScan fully demonstrate the superior performance of the new proposed methods and their potential value for disease diagnosis and treatment.

In a recent study,[70] authors found the upregulation of miR-15a, miR-16, and let-7a in APL patients and cell lines treated by ATRA based on a miRNA microarrays platform and quantitative real time-polymerase chain reaction (qRT-PCR). The interactions between ATRA and these three miRNAs were ranked 9th, 10th, and 12th in the predictive list based on miREFRWR, respectively. These interactions were ranked 1041st, 1042nd, and 1040th by miREFScan, respectively. In another experimental investigation,[65] authors demonstrated that mir-16 and let-7a were significantly differentially expressed after ATO treatment in APL cell NB4. These two APL-related miRNA–EF interactions were ranked 3rd and 4th in the predictive list based on miREFRWR among more than 50 000 candidate interactions.

To further evaluate the performance of miREFRWR on an independent dataset, a case study concerning breast cancer was implemented. Breast cancer is one of the most commonly occurring female cancers and makes up about 22% of all cancers in women. Recent biological experiments confirmed that miRNA let-7g was affected by Trastuzumab treatment in BT474 human breast cancer cells based on miRNA microarray profiling.[71] The interaction between let-7g and Trastuzumab was ranked 7th in the predictive list for breast cancer based on miREFRWR. In contrast, in the potential breast cancer-related miRNA–EF interactions list predicted by miREFScan, this interaction was only ranked 121st. Ichikawa *et al.* (2012) confirmed that miR-30b and miR-26a were upregulated in breast cancer cells after Trastuzumab treatment.[72] In the predictive list based on miREFRWR, these two interactions were ranked 75th and 88th. They are ranked 143rd and 165th by the previous method miREFScan. In the top 1% predictive list based on miREFRWR, 10 interactions associated with breast cancer were confirmed[23,73–75] (see ESI,‡ Table S5), while only 8 interactions can be found in the predictive list produced by miREFScan. Similar to the results in the case study of APL, two confirmed interactions predicted by miREFRWR cannot be obtained by miREFScan, while all the confirmed interactions predicted by miREFScan can be obtained by miREFRWR. Moreover, all the confirmed interactions always obtain better ranking in the predictive list of miREFRWR than miREFScan.

Predicting novel human disease-related miRNA–EF interactions

After confirming reliable predictive accuracy of miREFRWR based on LOOCV and the case studies about APL and breast

cancer, miREFRWR was further applied to predict novel disease-related miRNA–EF interactions for all the 97 diseases investigated in this article. The top 100 potential miRNA–EF interactions associated with each disease were publicly released to facilitate further experimental validation from biologists (see ESI,‡ Table S6). Reliable performance demonstrated in previous LOOCV and case studies leads us to believe that these predicted novel relationships among miRNAs, EFs, and human diseases could benefit the diagnosis and treatment of diseases.

Discussion

The reliable performance of miREFRWR could be mainly attributed to the combination of two factors: one is that potential disease-related miRNA–EF interactions were analyzed and predicted from a network perspective. Network-based methods can effectively identify biological properties at a network level and predict potential interactions among biological molecules. More importantly, global network information was adopted here, whose advantages over local network information methods have been demonstrated in many previous studies. The other is that known experimentally verified disease-related miRNA–EF interactions were used as the seed dataset to capture the potential associations between diseases and miRNA–EF interactions. Furthermore, drug chemical structure similarity, miRNA functional similarity, and networked-based similarity were also integrated into miREFRWR. These two factors also constitute the novelties of miREFRWR. In conclusion, miREFRWR could be a novel, important and effective biomedical tool in computational biology research.

Although excellent performances have been obtained in both the cross-validation and case studies, it should be noted that some limitations still exist in the current version of miREFRWR. First, although miREFRWR can obtain excellent performance in almost all the parameters selections, how to decide the parameter values is still not solved well. Second, introduction of more reliable similarity measures into this computational model, such as disease phenotypical similarity, drug side-effect similarity, and miRNA functional similarity, based on miRNA–target interactions has been planned. Moreover, how to integrate different similarity measures is an interesting and important problem in computational biology. Furthermore, the current version of miREFRWR cannot be applied to diseases without any known related miRNA–EF interactions. The performance of miREFRWR could be further improved when more experimentally confirmed human disease-related miRNA–EF interactions have been obtained in the future. Finally, the relationship between miRNA–EF interactions and cancer hallmark would be a very important problem to address in future studies. In particular, a cancer hallmark network could be constructed at the miRNA and EF levels to effectively evaluate cancer risks.[76]

Conclusions

Disease-related miRNA–EF interactions prediction is an important goal of biomedical research and plays a critical role in understanding the disease pathogenesis at the miRNA and EF levels and in the design of specific molecular tools for the prognosis, diagnosis, treatment and prevention of human disease. In this study, a novel method of miREFRWR was developed to predict potential disease-related miRNA–EF interactions. LOOCV and case studies about APL and breast cancer demonstrated that miREFRWR can effectively identify potential disease-related miRNA–EF interactions on a large scale by integrating the information of known disease-related miRNA–EF interactions, drug chemical structure, and miRNA functional similarity. In particular, miREFRWR has a reliable predictive accuracy in different test datasets according to the evaluation methods proposed in a recent article. Furthermore, the top 100 interactions associated with each disease have been publicly released to guide future biological experiments. It is anticipated that miREFRWR could be an effective and important biological tool for the research of non-coding RNAs, complex diseases, and drug design in the future.

Acknowledgements

This study was supported by the National Natural Science of Foundation of China under Grant No. 11301517 and National Center for Mathematics and Interdisciplinary Sciences, CAS.

Notes and references

1 A. L. Barabási, N. Gulbahce and J. Loscalzo, *Nat. Rev. Genet.*, 2011, **12**, 56–68.
2 W.-H. Chow, L. M. Dong and S. S. Devesa, *Nat. Rev. Urol.*, 2010, **7**, 245–257.
3 U. N. Das, *Nutrition*, 2010, **26**, 459–473.
4 A. M. Soto and C. Sonnenschein, *Nat. Rev. Endocrinol.*, 2010, **6**, 363–370.
5 A. Esquela-Kerscher and F. J. Slack, *Nat. Rev. Cancer*, 2006, **6**, 259–269.
6 X. Karp and V. Ambros, *Science*, 2005, **310**, 1288.
7 A. M. Cheng, M. W. Byrom, J. Shelton and L. P. Ford, *Nucleic Acids Res.*, 2005, **33**, 1290–1297.
8 E. A. Miska, *Curr. Opin. Genet. Dev.*, 2005, **15**, 563–568.
9 P. Xu, M. Guo and B. A. Hay, *Trends Genet.*, 2004, **20**, 617–624.
10 Q. Cui, Z. Yu, E. O. Purisima and E. Wang, *Mol. Syst. Biol.*, 2006, **2**, 46.
11 D. P. Bartel, *Cell*, 2004, **116**, 281–297.
12 M. Lu, Q. Zhang, M. Deng, J. Miao, Y. Guo, W. Gao and Q. Cui, *PLoS One*, 2008, **3**, e3420.
13 M. V. G. Latronico, D. Catalucci and G. Condorelli, *Circ. Res.*, 2007, **101**, 1225–1236.
14 Q. Jiang, Y. Hao, G. Wang, L. Juan, T. Zhang, M. Teng, Y. Liu and Y. Wang, *BMC Syst. Biol.*, 2010, **4**, S2.
15 G. A. Calin and C. M. Croce, *Nat. Rev. Cancer*, 2006, **6**, 857–866.
16 R. F. Duisters, A. J. Tijsen, B. Schroen, J. J. Leenders, V. Lentink, I. van der Made, V. Herias, R. E. van Leeuwen, M. W. Schellings and P. Barenbrug, *Circ. Res.*, 2009, **104**, 170–178.

17 A. Markou, E. G. Tsaroucha, L. Kaklamanis, M. Fotinou, V. Georgoulias and E. S. Lianidou, *Clin. Chem.*, 2008, **54**, 1696–1704.
18 T. E. Miller, K. Ghoshal, B. Ramaswamy, S. Roy, J. Datta, C. L. Shapiro, S. Jacob and S. Majumder, *J. Biol. Chem.*, 2008, **283**, 29897–29903.
19 F. J. Slack and J. B. Weidhaas, *N. Engl. J. Med.*, 2008, **359**, 2720–2722.
20 M. S. Weinberg and M. J. A. Wood, *Hum. Mol. Genet.*, 2009, **18**, R27–R39.
21 M. N. Poy, L. Eliasson, J. Krutzfeldt, S. Kuwajima, X. Ma, P. E. MacDonald, S. Pfeffer, T. Tuschl, N. Rajewsky and P. Rorsman, *Nature*, 2004, **432**, 226–230.
22 H. H. G. van Es and G. J. Arts, *Drug Discovery Today*, 2005, **10**, 1385–1391.
23 F. Xin, M. Li, C. Balch, M. Thomson, M. Fan, Y. Liu, S. M. Hammond, S. Kim and K. P. Nephew, *Bioinformatics*, 2009, **25**, 430–434.
24 Q. Huang, K. Gumireddy, M. Schrier, C. Le Sage, R. Nagel, S. Nair, D. A. Egan, A. Li, G. Huang and A. J. Klein-Szanto, *Nat. Cell Biol.*, 2008, **10**, 202–210.
25 R. T. Lima, S. Busacca, G. M. Almeida, G. Gaudino, D. A. Fennell and M. H. Vasconcelos, *Eur. J. Cancer*, 2011, **47**, 163–174.
26 Y. Ladeiro, G. Couchy, C. Balabaud, P. Bioulac-Sage, L. Pelletier, S. Rebouissou and J. Zucman-Rossi, *Hepatology*, 2008, **47**, 1955–1963.
27 A. Izzotti, P. Larghero, C. Cartiglia, M. Longobardi, U. Pfeffer, V. E. Steele and S. De Flora, *Carcinogenesis*, 2010, **31**, 894–901.
28 Y. Gidron, M. De Zwaan, K. Quint and M. Ocker, *Mol. Med. Rep.*, 2010, **3**, 455.
29 A. Alisi, L. Da Sacco, G. Bruscalupi, F. Piemonte, N. Panera, R. De Vito, S. Leoni, G. F. Bottazzo, A. Masotti and V. Nobili, *Lab. Invest.*, 2010, **91**, 283–293.
30 Z. Lin and E. K. Flemington, *Cancer Lett.*, 2011, **305**, 186–199.
31 M. J. Jardim, *Mutat. Res.*, 2011, **717**, 38–45.
32 O. M. Niemoeller, M. Niyazi, S. Corradini, F. Zehentmayr, M. Li, K. Lauber and C. Belka, *Radiat. Oncol.*, 2011, **6**, 29.
33 S. Xi, M. Yang, Y. Tao, H. Xu, J. Shan, S. Inchauste, M. Zhang, L. Mercedes, J. A. Hong and M. Rao, *PLoS One*, 2010, **5**, e13764.
34 T. Boren, Y. Xiong, A. Hakam, R. Wenham, S. Apte, G. Chan, S. G. Kamath, D.-T. Chen, H. Dressman and J. M. Lancaster, *Gynecol. Oncol.*, 2009, **113**, 249–255.
35 Y. Jin, X. Zou and X. Feng, *Anticancer Drugs*, 2010, **21**, 814–822.
36 X. Chen, M. X. Liu and G. Yan, *Mol. BioSyst.*, 2012, **8**, 2792–2798.
37 X. Chen and G.-Y. Yan, *Sci. Rep.*, 2014, **4**, 5501.
38 X. Chen, *Sci. Rep.*, 2015, **5**, 13186.
39 X. Chen, C. C. Yan, X. Zhang, Z. Li, L. Deng, Y. Zhang and Q. Dai, *Sci. Rep.*, 2015, **5**, 13877.
40 X. Chen, C. C. Yan, C. Luo, W. Ji, Y. Zhang and Q. Dai, *Sci. Rep.*, 2015, **5**, 11338.
41 X. Chen and G.-Y. Yan, *Bioinformatics*, 2013, **29**, 2617–2624.
42 X. Chen, *Sci. Rep.*, 2015, **5**, 16840.
43 Q. Yang, C. Qiu, J. Yang, Q. Wu and Q. Cui, *Bioinformatics*, 2011, **27**, 3329–3330.
44 C. Qiu, G. Chen and Q. Cui, *Sci. Rep.*, 2012, **2**, 318.
45 X. Chen, M. X. Liu, Q. H. Cui and G. Y. Yan, *PLoS One*, 2012, **7**, e43425.
46 X. Chen, M. X. Liu and G. Yan, *Mol. BioSyst.*, 2012, **8**, 1970–1978.
47 J. Sun, H. Shi, Z. Wang, C. Zhang, L. Liu, L. Wang, W. He, D. Hao, S. Liu and M. Zhou, *Mol. BioSyst.*, 2014, **10**, 2074–2081.
48 S. Köhler, S. Bauer, D. Horn and P. N. Robinson, *Am. J. Hum. Genet.*, 2008, **82**, 949.
49 M. Zhou, X. Wang, J. Li, D. Hao, Z. Wang, H. Shi, L. Han, H. Zhou and J. Sun, *Mol. BioSyst.*, 2015, **11**, 760–769.
50 X. Chen, G. Y. Yan and X. P. Liao, *OMICS*, 2010, **14**, 337–356.
51 T. van Laarhoven, S. B. Nabuurs and E. Marchiori, *Bioinformatics*, 2011, **27**, 3036–3043.
52 Y. Yamanishi, M. Kotera, M. Kanehisa and S. Goto, *Bioinformatics*, 2010, **26**, i246–i254.
53 K. Bleakley and Y. Yamanishi, *Bioinformatics*, 2009, **25**, 2397–2403.
54 Y. C. Wang, Z. X. Yang, Y. Wang and N. Y. Deng, *Lett. Drug Des. Discovery*, 2010, **7**, 370–378.
55 Z. Xia, L. Y. Wu, X. Zhou and S. T. C. Wong, *BMC Syst. Biol.*, 2010, **4**, S6.
56 Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda and M. Kanehisa, *Bioinformatics*, 2008, **24**, i232–i240.
57 W. Yu, X. Cheng, Z. Li and Z. Jiang, *Drug Dev. Res.*, 2011, **72**, 219–224.
58 A. Gottlieb, G. Y. Stein, Y. Oron, E. Ruppin and R. Sharan, *Mol. Syst. Biol.*, 2012, **8**, 592.
59 M. Hattori, Y. Okuno, S. Goto and M. Kanehisa, *J. Am. Chem. Soc.*, 2003, **125**, 11853–11865.
60 M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki and M. Hirakawa, *Nucleic Acids Res.*, 2006, **34**, D354–D357.
61 E. E. Bolton, Y. Wang, P. A. Thiessen and S. H. Bryant, *Annu. Rep. Comput. Chem.*, 2008, **4**, 217–241.
62 D. Wang, J. Wang, M. Lu, F. Song and Q. Cui, *Bioinformatics*, 2010, **26**, 1644–1650.
63 M. Xie, T. Hwang and R. Kuang, *Advances in Knowledge Discovery and Data Mining*, Springer, 2012, pp. 292–303.
64 Y. Park and E. M. Marcotte, *Nat. Methods*, 2012, **9**, 1134–1136.
65 S. H. Ghaffari, D. Bashash, A. Ghavamzadeh and K. Alimoghaddam, *Tumor Biol.*, 2012, **33**, 157–172.
66 J. H. Martens, A. B. Brinkman, F. Simmer, K.-J. Francoijs, A. Nebbioso, F. Ferrara, L. Altucci and H. G. Stunnenberg, *Cancer Cell*, 2010, **17**, 173–185.
67 C. D. Davis and S. A. Ross, *Nutr. Rev.*, 2008, **66**, 477–482.
68 J. Gu, X. Zhu, Y. Li, D. Dong, J. Yao, C. Lin, K. Huang, H. Hu and J. Fei, *Med. Oncol.*, 2011, **28**, 211–218.
69 Y. Wu, X. Li, J. Yang, X. Liao and Y. Chen, *Zhonghua Xueyexue Zazhi*, 2012, **33**, 546.
70 R. Garzon, F. Pichiorri, T. Palumbo, M. Visentini, R. Aqeilan, A. Cimmino, H. Wang, H. Sun, S. Volinia and H. Alder, *Oncogene*, 2007, **26**, 4148–4157.

71 X.-F. Le, M. I. Almeida, W. Mao, R. Spizzo, S. Rossi, M. S. Nicoloso, S. Zhang, Y. Wu, G. A. Calin and R. C. Bast Jr, *PLoS One*, 2012, **7**, e41170.
72 T. Ichikawa, F. Sato, K. Terasawa, S. Tsuchiya, M. Toi, G. Tsujimoto and K. Shimizu, *PLoS One*, 2012, **7**, e31422.
73 S. L. Tilghman, M. R. Bratton, H. C. Segar, E. C. Martin, L. V. Rhodes, M. Li, J. A. McLachlan, T. E. Wiese, K. P. Nephew and M. E. Burow, *PLoS One*, 2012, **7**, e32754.
74 M. Jansen, E. Reijm, A. Sieuwerts, K. Ruigrok-Ritstier, M. Look, F. Rodríguez-González, A. Heine, J. Martens, S. Sleijfer and J. Foekens, *Breast Cancer Res. Treat.*, 2012, **133**, 937–947.
75 E. J. Jung, L. Santarpia, J. Kim, F. J. Esteva, E. Moretti, A. U. Buzdar, A. Di Leo, X. F. Le, R. C. Bast and S. T. Park, *Cancer*, 2012, **118**, 2603–2614.
76 E. Wang, N. Zaman, S. Mcgee, J.-S. Milanese, A. Masoudi-Nejad and M. O'Connor-McCourt, *Semin. Cancer Biol.*, 2015, **30**, 4–12.