BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 5 2013, pages 638–644
doi:10.1093/bioinformatics/btt014

Data and text mining

Advance Access publication January 16, 2013

miRCancer: a microRNA–cancer association database constructed by text mining on literature

Boya Xie[1], Qin Ding[1,*], Hongjin Han[1] and Di Wu[2]

[1]Department of Computer Science, East Carolina University, Greenville, NC 27858 and [2]Department of Physiology, Brody School of Medicine, East Carolina University, Greenville, NC 27834, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Research interests in microRNAs have increased rapidly in the past decade. Many studies have showed that microRNAs have close relationships with various human cancers, and they potentially could be used as cancer indicators in diagnosis or as a suppressor for treatment purposes. There are several databases that contain microRNA–cancer associations predicted by computational methods but few from empirical results. Despite the fact that abundant experiments investigating microRNA expressions in cancer cells have been carried out, the results have remain scattered in the literature. We propose to extract microRNA–cancer associations by text mining and store them in a database called miRCancer.

Results: The text mining is based on 75 rules we have constructed, which represent the common sentence structures typically used to state microRNA expressions in cancers. The microRNA–cancer association database, miRCancer, is updated regularly by running the text mining algorithm against PubMed. All miRNA–cancer associations are confirmed manually after automatic extraction. miRCancer currently documents 878 relationships between 236 microRNAs and 79 human cancers through the processing of >26 000 published articles.

Availability: miRCancer is freely available on the web at http://mircancer.ecu.edu/

Contact: dingq@ecu.edu

Received on June 14, 2012; revised on December 18, 2012; accepted on January 8, 2013

1 INTRODUCTION

MicroRNAs (miRNAs) are small non-coding RNAs that regulate gene expressions by base pairing to messenger RNAs. Although the first miRNA, lin-4, was introduced in 1993 (Lee *et al.*, 1993), it did not draw much attention until 2000, when the second miRNA, let-7, was characterized. Studies found that let-7 regulates a couple of gene expressions in *Caenorhabditis elegans* development, and is conserved in many species (Pasquinelli *et al.*, 2000; Reinhart *et al.*, 2000). Since then, many techniques have been developed to identify miRNAs. As of November 2012, >30 000 mature miRNA sequences have been identified with a growth of >10 000 in the past year. Current findings suggest that miRNAs play an important role in crucial biological processes, such as cell proliferation (Hwang and Mendell, 2006), apoptosis (Jovanovic and Hengartner, 2006), development (Karp and Ambros, 2005), differentiation (Chen *et al.*, 2004; Shivdasani, 2006) and metabolism (Wienholds and Plasterk, 2005). Among these, miRNA association with disease, especially the association between miRNA and human cancers, has attracted a great deal of interest in both scientific and business fields.

Cancer is a leading cause of death worldwide. It is estimated that cancer caused 7.6 million deaths in 2008, about 13% of all deaths globally (http://turl.ca/xndn). By 2010, cancer surpassed heart disease as the top killer in USA for the first time. There are many possible carcinogens including tobacco, radiation, chemicals, environmental toxins, viruses and genetic problems. However, the causes of many cancers remain unknown. Cancer involves unregulated cell growth, thereby invading nearby parts of body during development. Early diagnosis before proliferation usually makes a difference in treatment and survival rate. The fact that miRNA expression levels vary significantly between normal cells and cancer cells suggests that miRNA might be associated with cancer development and potentially could be used for cancer diagnosis or even treatment. Even though it is uncertain whether cancer is a cause or consequence of deviant miRNA expression, miRNA fingerprints are found in all types of analysed cancers, such as lung cancer, breast cancer, cervical cancer and lymphoblastic leukemia (Iorio *et al.*, 2005; Lui *et al.*, 2007; Mi *et al.*, 2007; Takamizawa *et al.*, 2004).

Numerous databases have been created to document miRNA functionalities either from computational predictions or from experimental results. Although computational target prediction methods are fast, experimental validation of miRNA functionalities is also needed. The significant increase in validation experiments raises the need for having a database to store these results in some uniform way. However, compared with databases providing computationally predicted miRNA functions, databases storing experimental miRNA targets are rare. Most of these experimentally verified miRNA target databases are manually collected, such as miR2Disease (Jiang *et al.*, 2009), miRecords (Xiao *et al.*, 2009), miRTarBase (Hsu *et al.*, 2011) and TarBase (Sethupathy *et al.*, 2006) version 1 to 5. Rapid increase in the number of miRNA-related publications makes the manual collection more and more difficult. Three databases using text-mining strategies have been introduced in the past 2 years. miRSel (Naeem *et al.*, 2010) and miRWalk (Dweep *et al.*, 2011) contain miRNA interaction data, and miRNA targets were derived solely from text mining of MedLine abstracts. TarBase 6.0 (Vergoulis *et al.*, 2012) adds text mining to preprocess abstracts before human curation. The text mining

*To whom correspondence should be addressed.

preprocessing increased entries about 50-fold as compared with its previous version without text mining.

Text mining is the process in which useful information is extracted from text using computational approaches or tools. Unlike its application in other fields, accurate biomedical text mining remains an open problem as a result of specialized and complex vocabularies. There are three commonly used text-mining approaches in biomedical realm: (i) co-occurrence-based approach, which was adopted by miRSel and TarBase 6.0; (ii) rule-based approach, which keeps a set of rules that usually take significant amount of time to develop; and (iii) machine learning, where the required training data are usually expensive or even impossible to generate (Kohen and Hunter, 2008). Co-occurrence-based systems are normally easier to build while the other two provide better accuracy.

In this article, we proposed and developed rule-based text-mining approaches to extract miRNA and cancer association and store them in a database called miRCancer. All the discovered associations have been manually confirmed after automatic extraction. miRCancer currently documents 878 relationships between 236 microRNAs and 79 human cancers through the processing of >26 000 published articles from PubMed.

2 METHODS

All miRNA–cancer associations in our miRCancer database are extracted by a rule-based text-mining system and followed by manual confirmation. The rules used for text mining are hard-coded sentence structures. The system consists of literature collection, named entity and expression recognition, rule matching, voting, manual verification and recording (Fig. 1).

To collect all relevant publications, the query '*((((mir) OR mirna) OR microrna) OR micro-rna) OR micro rna*' is searched against the PubMed library. The PubMed provides search results in XML format with all literature details, including the abstract of the article. As of October 16, 2012, there have been 26 414 publications generated from the above query and those publications are the primary input to our text-mining system. The title and abstract of the articles are used for the following processes.

2.1 Named entity and miRNA expression recognition

The extraction ability of the system highly relies on the named entity recognition (NER), which is the miRNA and cancer name recognition. Compared with other biological sequences, such as genes and proteins, miRNA names have been formalized early enough so that they are quite unified and relatively simple. Most miRNAs are named with the 'miR' prefix, with the exception of a few miRNAs: bantam, let-7 family, lin-4 and lsy-6. We use regular expressions to identify miRNA names. The miRNA name prefix variations that our system handles include: miR, miR-, cases with species prefix, such as has- and mmu-, etc. Table 1 displays the complete list of miRNA prefix variations. A prefix is followed by a unique numerical number that is assigned sequentially. Because miRNAs are highly conserved between species and organisms, identical sequences have the same id number regardless of species (e.g., mmu-miR-21 in *Mus musculus*, hsa-miR-21 in *Homo sapiens*).

Additional suffixes sometimes are used in miRNA names for hairpin precursor and hairpin loci information. The miRNA name suffix variations that our system handles include the following: (i) single letter 'a' and 'b' for related miRNA sequences that only have small base changes (e.g. hsa-miR-216a and hsa-miR-216b); (ii) single digits such as '-1' and '-2' for identical miRNAs with different hairpin loci in a given organism (e.g. hsa-miR-16-1 and hsa-miR-16-2); and (iii) '-3p', '-5p' and '*' for miRNA sequences expressed from different arms of hairpin precursor (e.g. hsa-miR-339-5p) (Table 2). In addition to miRNA's naming variations, miRNA names are also written in different abbreviations, such as 'miR-123, -46, -23' (Table 3).

Cancer name recognition is performed by comparing the text with a cancer name dictionary. This dictionary is compiled from the international classification of diseases for oncology (ICD-O) (Fritz *et al.*, 2000). The ICD-O has been used in cancer registries for nearly 25 years. The classification has two axes: morphology, which describes the form and behaviour of the tumour, and topology, which describes the site of origin. Each description is assigned with a corresponding code. The morphology code starts with 'M' followed by four digits indicating the cell type, and one digit for the behaviour, such as M8010/3; the topology code begins with 'C' followed by a number ranging from 00 to 80, such as C15.0. We built our dictionary so that each entry contains a collection of common names and an ICD-O code that consists of morphology and topology codes. To illustrate, non-small cell lung cancer has the code 'C34.9, M8046/3', where 'C34.9' signifies that the cancer is located in lung, 'M8046' is the morphology code for non-small cell and the final digit, 3, is the behaviour code for malignant tumour. Additionally, the dictionary includes abbreviations for some cancer names, i.e. 'OSCC' stands for Oral Squamous Cell Carcinoma.

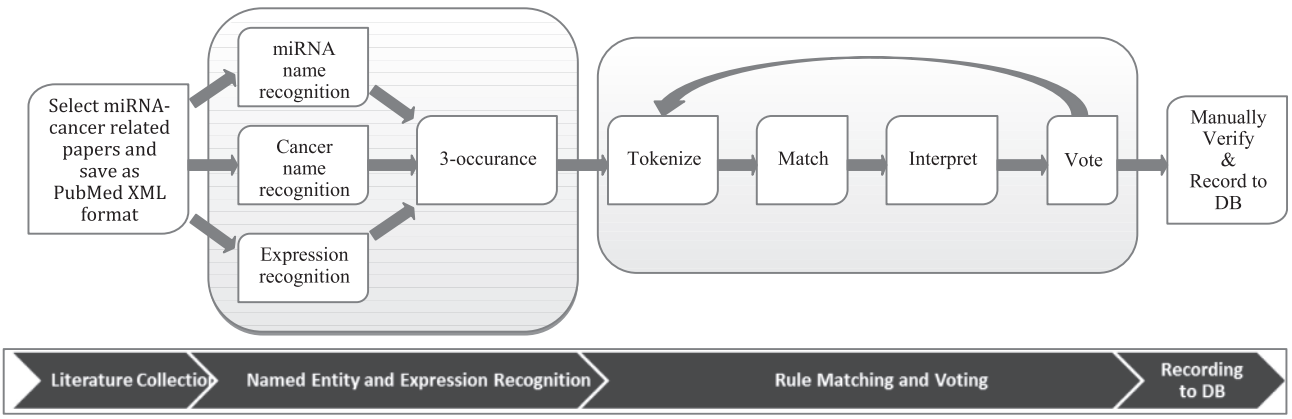


Fig. 1. Text mining workflow

Table 1. miRNA naming variations: miRNA prefix

miR, miR-, miRNA, miRNA-, microRNA, microRNA-, micro RNA, micro-RNA-, let-, miRNA-let-, miR-let-, has-, mmu-, and etc.

Table 2. miRNA naming variations: miRNA suffix

Situation	Example
Related sequence with base change	hsa-miR-216a and hsa-miR-216b
Identical sequence with different hairpin loci	hsa-miR-16-1 and hsa-miR-16-2
Sequence from 3′arms	hsa-miR-17-3p
Sequence from 5′arms	hsa-miR-17-5p
Minor sequence	has-miR-100*

Table 3. miRNA naming variations: miRNA variations

Short form	Formal form
miR-123, -46, -23	miR-123, miR-46, miR-23
miR-200a, b, c	miR-200a, miR-200b, miR-200c
miR-21/137	miR-21, miR-137
miR-99a/100	miR-99a, miR-100
miR-99a/b	miR-99a, miR-99b
miR-15a/16-1	miR-15a, miR-16-1
miR-221/-222	miR-221, miR-222
miR-23b/-27b	miR-23b, miR-27b
miR-let-7a/let-7b	let-7a, let-7b
miR-221&222	miR-221, miR-222

A list of 28 terms is further compiled to describe miRNA expression profiles in cancers. Fourteen of them indicate miRNA oncogenic function while the other 14 specify cancer suppressor miRNAs (Table 4). This expression dictionary is designed to maximize recall rather than precision, as the articles will be further examined by the rules and manual verification.

After an article title along with its abstract are processed with the NER and expression recognition module, the article is then defined as 3-occurrence if miRNA name, cancer name and expression terms are all present. A 3-occurrence article is then sent to the rule matching module to match with the pre-defined set of rules.

2.2 Rule construction

Though 3-occurrence likely indicates an miRNA–cancer association, it does not guarantee correctness. Alternatively we found that miRNA–cancer associations are usually stated with certain sentence structures. After reviewing a large pool of publications, we have collected 75 different sentence structures that researchers use to describe miRNA expression in cancers. Each sentence structure represents one rule, and all the structures are further categorized into four tiers depending upon their structure restrictiveness. There are 26 rules in tier 1, 5 rules in tier 2, 31 rules in tier 3 and 13 rules in tier 4.

Table 4. miRNA expression dictionary

Up regulate terms	Down regulate terms
over express[a]	under express[a]
overexpress[a]	underexpress[a]
over-express[a]	under-express[a]
highly express[a]	lower express[a]
high express[a]	lower-express[a]
up regulat[a]	down regula[a]
upregulat[a]	downregulat[a]
up-regulat[a]	down-regulat[a]
positive regulat[a]	negative regula
increase[a]	decrease[a]
forced expression	repress[a]
enhanced expression	suppress[a]
promote	inhibit[a]
oncogenic	delete[a]

[a]Could be any 1 or more characters, such as *s*, *es*, *ing*, *ion*, *ed*, *or*, *and* etc.

Tier 1 rules have miRNA names, cancer names and expression terms that form a complete sentence. The sentence could be simple or complex, but it should have a clear subject and predicate structure. The simplest rule in tier 1 is MIR-PHRASE EXP-V CANCER-PHRASE, where MIR-PHRASE could be a single miRNA name (e.g. miR-21), a group of miRNA names (e.g. miR-184, miR-200a, b, c and miR-205) or an miRNA name mixed with other words (e.g. expression of miR-21). EXP-V in the rule is an expression term in the verb form, including but not restricted to: inhibit, over-expressed, promotes and so on. CANCER-PHRASE is similar to MIR-PHRASE in that it could be a single cancer name (e.g. breast cancer), a group of cancer names (e.g. lung, liver and breast cancers) or a cancer name mixed with other words (e.g. highly invasive colorectal cancer cells). Tier 1 rules could also be as complex as EXP-ADJ MIR-PHRASE BE-PHRASE EXP-V 'in' CANCER-PHRASE COMPARE-PHRASE NORMAL-PHRASE. In this case, EXP-ADJ indicates an expression term as an adjective, while BE-PHRASE means a to-be phrase with a positive meaning (e.g. is confirmed to be, are found to be). COMPARE-PHRASE and NORMAL-PHRASE are not essential elements in this rule, but they make the sentence a comparison structure that strengthens the rule. A sample sentence matching this rule could be: increased expression of microRNA-21 is shown to be up-regulate in gastric carcinoma compared with corresponding non-cancerous tissues.

Unlike Tier 1 rules, miRNA, cancer and expression terms form a phrase rather than a complete sentence in tier 2 structures. There is neither a subject nor predicate, other than a phrase formed by the three essential elements of the miRNA–cancer association. There are two rules in this category: MIR-PHRASE 'as' EXP-N-PHRASE 'of' CANCER-PHRASE, and EXP-ADJ MIR-PHRASE 'in' CANCER-PHRASE. Examples of sentences matching these two rules could be: miR-34b and miR-129-3p as the known inhibitor of gastric cancers, increased miR-499-5p levels in highly invasive CRC cell lines.

In tier 3 rules, specific cancer names are absent from the structure; instead, there is a word or phrase referring to cancers in general, such as the word 'cancer' itself or 'cancer tissues'. Tier 3 rules are formed by inheriting all tier 1 and 2 rules and substituting CANCER-PHRASE with cancer-related words. As a result, tier 3 structures include both complete sentences and phrases. Because cancer names are not specified by tier 3 structures, the matching module infers the cancer name either from the current sentence, the abstract or the title of the article. When inference occurs, a tier 3 rule could be downgraded to tier 4 depending on the

cancer name source. The inference and rule downgrading will be described in the next section. Moreover, miRNA names and expression terms form tier 4 rules, which are all phrase structures. Similar to tier 3 rules, the cancer names are inferred, but there is no downgrading to this set of rules. Take tier 4 rule EXP-ADJ MIR-REF MIR-PHRASE as an example, it could match with the phrase: downregulated microRNAs miR-34b and miR-129-3p.

We have conducted a test on the accuracy of these rules by randomly choosing 70 matched sentences, and compared the results with human evaluations. Among the 70 sentences, 18, 6, 26 and 20 are matched with tier 1–4 rules, respectively. And 17, 5, 20 and 12 of them are correctly interpreted by our algorithm, which gives accuracy of 94%, 83%, 77% and 60%, respectively. Our test also showed that the results from tier 3 rules downgraded to tier 4 rules have similar accuracy as results from original tier 4 rules. Based on the test results, we assign voting weight of 0.94, 0.83, 0.77 and 0.60 to tier 1–4 rules respectively. These weights will be used in the voting module when rules are used to vote for different expression profiles related to one miRNA–cancer pair.

2.3 Rule matching and voting

The matching and voting module has four steps to translate a sentence into miRNA–cancer associations, namely: tokenization, matching, interpretation and voting.

Tokenization is the process that translates each article including the title and the abstract into generalized tokens. First, all of the miRNA and cancer names are tokenized as MIR-0 and CANCER-0, while the expression terms are turned into tokens such as EXP-*-0, where the '*' is N, V, ADJ and so on, depending on whether it is a noun, verb or adjective in the sentence. In addition to the three elements which composite an miRNA–cancer association, other words used to form rules are also tokenized. After the primary tokenization (Table 5), additional process encrypts the text with tokens into secondary tokens if possible. The secondary tokens (Table 6) are all phrases, including the phrases for the three essential elements of an association as well as other phrases necessary for the rules. During tokenization, the global cancer name for an article is saved if only one type of cancer is mentioned through the title and the abstract; otherwise the cancer name in the title is saved as the global cancer. In case there are multiple types of cancers occurring or no cancer name exists, the global cancer name is set to null value. The global cancer name helps tier 3 and 4 rules infer a cancer name to form a complete miRNA–cancer association.

Matching is the process to map the tokenized sentences with the rules, tier by tier, starting from tier 1. This means that each single sentence is compared with all 26 rules in tier 1, and any matches are interpreted and participate in the tier 1 vote. If there is a conclusive voting result, the matching and voting module completes its task for this sentence, and proceeds to the next sentence. On the other hand, if there is no match or the vote is a tie, tier 2 rules are used and the matches with tier 2 rules are translated and take part in the tier 2 vote. The results are recorded if tier 2 voting generates a conclusion; otherwise tier 3 rules will be processed and the same process continues on with tier 4 if necessary. The sentence to rule matching is a straightforward task, as all tokens and rules are formed as regular expressions. The situation is handled when a sentence, usually a compound one, is matched with multiple rules. Therefore, it is possible for one sentence to be matched with multiple rules multiple times, meaning that several miRNAs may have a many-to-many association with various cancers.

Interpretation is the process to decode the matching sentence–rule pairs into miRNA–cancer associations by tracing back the miRNA, cancer

Table 5. Primary tokens

Token	Meaning	Examples
MIR-\d+	An miRNA name	miR-21
CANCER-\d+	A cancer name	Breast cancer
EXP-V-\d+	A verb expression term	Suppress
EXP-N-\d+	A noun expression term	Overexpression
EXP-ADJ-\d+	An adjective expression term	Oncogenic
EXP-PASTVorADJ-\d+	An expression term could be both verb in past tense or adjective	Overexpressed
ART	Article	A, an, the
BE	Be verb	Is, are, was, were
NUM	Number	12, 3.4, sixteen

'-\d+': means the index number of the token in the text.

Table 6. Secondary tokens

Token	Meaning	Examples
MIR-PHRASE-\d+	A group of words for single or multiple miRNAs	miR-21 expression
CANCER-PHRASE-\d+	A group of words for single or multiple cancers	The breast cancer tissues
EXP-N-PHRASE-\d+	A group of words stating expression profile as noun	The known suppressor
EXP-PASSIVE-PHRASE-\d+	Words stating expression profile as verb in passive voice	Is significantly over-expressed
EXP-ADJ-PHRASE-\d+	A group of words stating expression profile as adjective	The six most over expressed
BE-PHRASE	A group of words equals to the function of a be verb	Are reported to be
NUM-PHRASE	Words for numbers	45% of
COMPARE-PHRASE	Compare	Compared with
NORMAL-PHRASE	A normal group which is compared with	Corresponding normal tissues

'-\d+': means the index number of the token in the text.

names and expression terms through index numbers attached to the tokens. Take the rule MIR-PHRASE EXP-V CANCER-PHRASE as an example: a sentence matching with this rule will have tokens arranged as MIR-PHRASE-1 EXP-V-0 CANCER-PHRASE-0. Each miRNA belonging to MIR-PHRASE-1 will pair up with every cancer name in CANCER-PHRASE-0, and the miRNAs will have an expression profile indicated by the term in EXP-V-0. The interpretation could be a little more challenging if there is more than one expression term in a rule. For example, the sentence 'suppression of miR-27a inhibits gastric cancer cell growth' matches with rule EXP-N of MIR EXP-V CANCER-PHRASE. Both the expression terms 'suppression' and 'inhibits' contribute to interpret that the miR-27a has an oncogenic function in gastric cancer. Another question in deciphering the rules is how to infer the implied cancer name when none is specified. Our algorithm is designed in such a way that it looks at the current sentence first. If there is a single cancer name present, it is assumed to be the implied cancer and forms the miRNA–cancer association. In case there is more than one or no cancer name, the global cancer name is used if there is any. While taking a global cancer name, the rule is downgraded from tier 3 to tier 4, and the voting weight is reduced accordingly. In the worst case scenario, when no cancer name can be inferred, the match is then discarded. When an interpretation is successful, it generates association sets in the form of miRNA–cancer–expression–weight. Each association has a single miRNA and a cancer name, while the expression is either 'up' or 'down' meaning oncogenic expression or suppressor functionality, and the weight is the voting weight of the matched rule.

Voting for a single sentence only takes place when multiple associations have the same miRNA and cancer but different expressions. Associations vote by adding their weights together, and the associations with higher total weight win and the voting iteration is terminated. If there is still a tie after the tier 4 votes, both associations with opposite expressions are saved for later use, where associations from other sentences could help break the tie. Upon completion of processing the article title and every sentence of the abstract, the collection of associations will perform a final vote to decide which associations should be inserted into the database. We set the cut-off point at 0.8 so that any voting result with a score >0.8 is recorded for manual validation, otherwise discarded. Recall the weights assigned to each tier rules, tier 1 and 2 rules have weights >0.8, therefore an association matched with these rules is guaranteed to be recorded for further manual verification, while the other two tier rules need more supportive votes to qualify for human analysis. The even votes are also documented for human interaction.

While recording associations, every sentence where an miRNA–cancer relationship emanates is known. As a result, we therefore only need to manually review that single sentence to verify the corresponding relationship; once it is confirmed, it will be added to the miRCancer database. Besides the three essential elements, i.e. miRNA–cancer–expression, additional information is also documented to facilitate building the mirCancer website content. Each miRNA name is saved with an id and URL link to the central online miRNA repository, miRBase (Jones-Griffiths *et al.*, 2008), if it is reported in miRBase. Every relation is noted with the article title and PubMed Id from which the association is extracted.

3 RESULTS AND DISCUSSIONS

By processing 26 414 publications, miRCancer currently records 878 pairs of miRNA–cancer associations from 573 publications. Among these relationships, there are 236 miRNAs and 79 human cancers involved. miRCancer provides a web interface to retrieve these associations by miRNA and/or cancer names. The results are showed with miRNA (linking to miRBase entry), cancer, expression profile and the literature (linking to PubMed) (Fig. 2). miRCancer also provides some miRNA sequence analysis

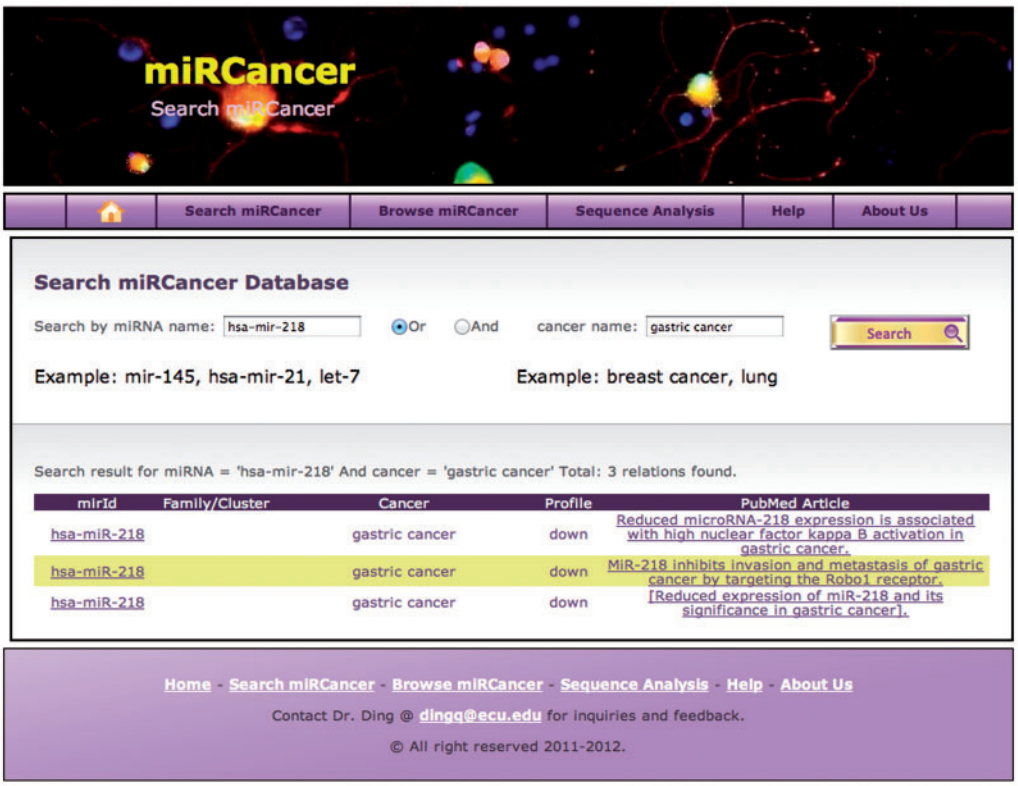
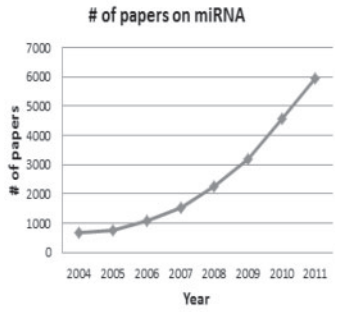


Fig. 2. miRCancer user interface



year	# of papers on miRNA	# of test papers with association
2004	671	5
2005	765	6
2006	1089	9
2007	1536	12
2008	2273	18
2009	3174	25
2010	4573	36
2011	5944	46
2012	5581	44
Total	25606	200

Fig. 3. Number of articles on miRNAs by year

tools that we developed, such as a sequence clustering tool, which performs clustering on mirRNAs based on sequence similarity. More information about the tools can be found in our previous article (Xie *et al.*, 2010).

3.1 Evaluation and comparison

To estimate the system's performance, measurements such as recall, precision and f-measures are estimated by comparing the system output with a test pool of 400 articles. Among them, 200 articles contain miRNA–cancer associations, while the other articles contain miRNA, cancer and expression terms but no associations. All the results are manually verified and the precision is 100%, which is the same as miR2Disease. Furthermore, to facilitate the comparison with the miR2Disease database, the number of articles in the test pool for each year is proportional to the total published miRNA-related articles in that year. The total published articles on miRNAs and selected number of test articles from each year are shown in Figure 3.

For the 200 articles with the associations present, associations were correctly extracted from 157 articles, while nothing was found in 43 articles, which gives a recall rate of 78.5%. With the other 200 articles containing no association, the results were all correctly identified as true negatives. There are three causes for not discovering any association in those 43 uncalled articles: (i) The miRNA–cancer relationship is not stated within one sentence. For example, an article may introduce the miRNA target gene in one sentence, and describe the gene–cancer relationship in another sentence; (ii) Associations captured by tier 3 or 4 rules have a low voting score. Including the associations detected by tier 3 and 4 rules to manual verification will definitely increase the recall rate. Nonetheless, there are >1000 such associations, and most of them are false positives. With limited time and resources, we do not manually evaluate those associations. We aim to improve the recall rate by using other methods; and (iii) There are odd spellings or typos in some articles. There are instances where a cancer name is spelled in an unusual way, which was not captured when our cancer dictionary was constructed, such as spelling 'nasopharayngeal' instead of 'nasopharyngeal'. Among these three causes, the first type contributes the most to the missing hits. We plan to include miRNA targets for consideration and allow miRNA–cancer relation to be formed from multiple sentences in our future studies. With a comprehensive miRNA targets dictionary and careful construction of rules, the first type of problems may be solved.

Table 7. System performance

Database	Recall	Precision	F-measure
miRCancer	78.50%	100.00%	88.00%
Mir2Disease	28.50%	100.00%	44.40%
Mir2Disease (Before 2010)	77.00%	100.00%	87.00%

As the precision is 100%, by combining recall rate with precision, the F-measure for miRCancer is estimated to be 88.0%, as shown in Table 7.

Currently there are few databases recording experimental results on miRNA–cancer associations. To our knowledge, miR2Disease in the only database collecting experimentally verified miRNA deregulations in human diseases, including cancers. It was first constructed in 2008. All data in miR2Disease are manually curated from published articles. We evaluated miR2Disease against the same pool of test articles, and then compared the results with miRCancer. miR2Disease has the same perfect precision, but it has a much lower recall rate (Table 7). Although the latest update shown on miR2Disease website is listed as March 2011, we are not convinced that it contains any data after 2009 because none of the associations presented in test articles after 2009 could be found in the database. To have a fair evaluation, we also measured miR2Disease performance on test articles published before 2010. The recall rate increased from 28.5 to 77.0% as shown in Table 7. The low overall recall in miR2Disease indicates the difficulty in keeping the database up to date while the literature data is increasing exponentially. As shown in Figure 3, there were >5000 articles related to miRNAs published in both 2011 and 2012. It is much more difficult and more time consuming than in previous years for human to manually identify associations directly from large volume of articles and update the database. When compared with miR2Disease, miRCancer has the advantage in reducing human work to a minimum. The text-mining approach automatically extracts the candidate associations; therefore, for the manual verification, it narrows down to a much smaller set of articles where associations were extracted. Also it is much easier to verify a single sentence where the association is automatically extracted than reading the entire abstract or text of an article. More importantly, it still maintains a reasonably high recall rate. As a result, miRCancer is more effective and feasible to update while maintaining the same or higher recall rate.

The completeness and accuracy of miRCancer highly relies on the number and quality of the rules. The more comprehensive the rule set is, the better recall rate the system can generate. Nevertheless, some rules, such as tier 3 and tier 4 rules, could be too weak. They might introduce errors to be able to increase recall. We have been frequently reviewing the rules with newly published articles. We would add new rules and modify existing rules if necessary in order to correctly extract more associations. In addition to the rules, the cancer name dictionary also affects the results. Currently there are 149 cancer entries in the dictionary. We update the cancer dictionary quarterly to include possible new names.

4 CONCLUSIONS

miRCancer provides a comprehensive collection of miRNA expressions in human cancers based on results extracted from literature. Currently it has 236 miRNA expression profiles in 79 human cancers, totaling 878 associations extracted from 573 publications. All profiles are manually confirmed. The extraction process is automated by applying text mining and rule matching on article titles and abstracts. The miRCancer system has a reasonable good recall rate, but there is still potential for improvement by adding new rules. The miRCancer database can be queried via the free access web interface. MiRNA–cancer associations could be searched by miRNA name and cancer name independently or a combination of miRNA and cancer names.

The proposed text-mining approach can be extended to find experimentally verified target genes of microRNA. We plan to include this expansion in the future.

REFERENCES

Chen,C. *et al*. (2004) MicroRNAs modulate hematopoietic lineage differentiation. *Science*, **303**, 83–86.

Dweep,H. *et al*. (2011, May) miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *J. Biomed. Inform*., **44**, 839–847.

Fritz,A. *et al*. (2000) *International Classificaiton of Diseases for Oncology*, 3rd edn. World Health Organization, Geneva.

Hsu,S. *et al*. (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res*., **39**, D163–D169.

Hwang,H.-W. and Mendell,J.T. (2006) MicroRNAs in cell proliferation, cell death, and tumorigenesis. *Br. J. Cancer*, **94**, 776–780.

Iorio,M.V. *et al*. (2005) MicroRNA gene expression deregulation in human breast cancer. *Cancer Res*., **65**, 7065–7070.

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

Jones-Griffiths,S. *et al*. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res*., **36**, D154–D158.

Jovanovic,M. and Hengartner,M.O. (2006) miRNAs and apoptosis: RNAs to die for. *Oncogene*, **25**, 6176–6187.

Karp,X. and Ambros,V. (2005) Encountering MicroRNAs in cell fate signaling. *Science*, **310**, 1288–1289.

Kohen,K. and Hunter,L. (2008) Getting started in text mining. *PLoS Comput. Biol*., **4**, e20.

Lee,R.C. *et al*. (1993) The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.

Lui,W. *et al*. (2007) Patterns of known and novel small RNAs in human cervical cancer. *Cancer Res*., **67**, 6031–6043.

Mi,S. *et al*. (2007) MicroRNA expression signatures accurately discriminate acute lymphoblastic leukemia from acute myeloid leukemia. *Proc. Natl Acad. Sci. USA*, **104**, 19971–19976.

Naeem,H. *et al*. (2010) miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature. *BMC Bioinformatics*, **11**, 135.

Pasquinelli,A.E. *et al*. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86–89.

Reinhart,B.J. *et al*. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, **403**, 901–906.

Sethupathy,P. *et al*. (2006) TarBase: a comprehensive database of experimentally supported animal microRNA targets. *RNA*, **12**, 192–197.

Shivdasani,R.A. (2006) MicroRNAs: regulators of gene expression and cell differentiation. *Blood*, **108**, 3646–3653.

Takamizawa,J. *et al*. (2004) Reduced expression of the let-7 MicroRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res*., **64**, 3753–3756.

Vergoulis,T. *et al*. (2012) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res*., **40**, 222–229.

Wienholds,E. and Plasterk,R.H. (2005) MicroRNA function in animal development. *FEBS Lett*., **579**, 5911–5922.

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

Xie,B. *et al*. (2010) MIRSAT and MIRCDB: an Integrated microRNA Sequence Analysis Tool and a Cancer-associated microRNA Database. In: *Proceedings of International Conference on Bioinformatics and Computational Biology*. Honolulu, Hawaii, pp. 159–164.