miRDiabetes: A microRNA-Diabetes Association Database Constructed by Classification on Literature

Hui Guo
Department of Computer Science
North Carolina State University
Raleigh, NC 27695, USA
hguo5@ncsu.edu

Qin Ding
Department of Computer Science
East Carolina University
Greenville, NC 27858, USA
dingq@ecu.edu

Abstract

MicroRNAs (miRNAs) are a growing class of non-coding RNAs that regulate gene expression by translational repression. miRNA plays a very important in many biological process. A role for miRNA in diabetes was first discovered in 2004 and miRNA-diabetes association has been an increasing interest since then. However, before this study, there is no computational tool available to be able to retrieve and gather literature on this topic. In this paper, we performed classification on the PubMed literature data to automatically retrieve associations between miRNAs and diabetes. After being verified, the retrieved associations are stored in a database called miRDiabetes, which is the first comprehensive database to collect articles that profile relations between miRNAs and diabetes. Performance study shows that our system can automatically determine relevancy of new entries with high accuracy. We also developed an application to facilitate regular updates on the database and built a website (http://mirdiabetes.ecu.edu) for researchers to search and download the miRDiabetes database.

Keywords: Text mining, miRNA, data mining, classification, diabetes, microRNA-Diabetes association

1 Introduction

New scientific discoveries are based on the existing knowledge, which has to be accessible and therefore by the scientific community [1]. The roles of microRNAs the etiology, pathology, symptom, and therapeutics of diabetes did not receive much attention until in recent This topic has been showing much potential and starts to draw overwhelming interest among biomedical As publications on the association between microRNA and diabetes grow rapidly in number, researchers have found that retrieving them from scatter literature is a more task. A literature collection database on this topic is in need, as well as computational methods to perform and update the collection. The miRDiabetes database was initiated by requests from some of these researchers. This database will provide a platform for them to feed on previous studies in order to conduct new ones. This paper focuses on the construction of a miRNA-diabetes association database called miRDiabetes, as well as the method of utilizing data mining techniques for literature retrieval, which we have designed and implemented.

MicroRNAs (miRNAs) are a class of naturally small non-coding RNA molecules, about 22 nucleotides in length. During the last few years, strong evidence showed that aberrant expression of miRNA is associated with a broad spectrum of human diseases, such as cancer, obesity, cardiovascular and psychological disorders [2].

Growing interest in miRNAs inspired swift increase in the number of miRNA-related publications, making it more time-consuming for researchers in biology or medicine to find articles related to their research areas. As of January 2016, almost 60,000 publications on PubMed involve miRNA [3]. Manual retrieval or classification of these articles has become a more demanding drudgery. Figure 1 shows the growth of miRNA-related articles in PubMed.

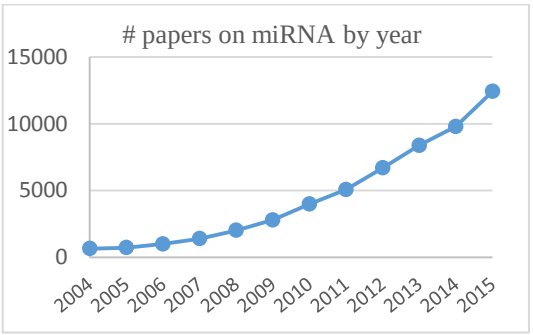

Figure 1. Number of papers on miRNAs by year

Diabetes mellitus (DM), or simply diabetes, is a group of metabolic diseases with which a person has high blood glucose level. There are three major types of diabetes,

Type 1 Diabetes (T1D), Type 2 Diabetes (T2D0) and gestational diabetes (GDM).

Compelling evidence has shown that miRNAs contribute to the etiology of diabetes, especially Type 2 Diabetes. A role for miRNAs in T2D was first discovered in 2004. Poy and colleagues showed that miR-375 was directly involved in the regulation of insulin secretion and might thereby constitute a novel pharmacological target for the treatment of diabetes [4]. As research advances, a link between miRNAs and diabetes now seems to be increasingly likely. Many miRNAs have been discovered to be associated with diabetes, as well as beta-cell biology, insulin resistance, and diabetic complications. Continuous studies on the relation between miRNAs and diabetes have led to a much greater understanding of the genetic basis of this disease and provided novel diagnostic, prognostic, and treatment alternatives.

Figure 2 shows the growth of literature in the miRDiabetes (relevant articles on miRNA-diabetes association). It is reasonable to predict that the number of articles on this subject will keep growing rapidly in the future.

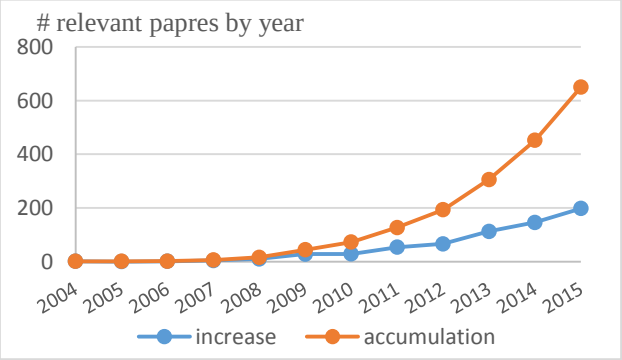

Figure 2. Papers on miRNA-diabetes association by year

With the exploding number of potential publications in this field, manual retrieval of literature will soon become too time-consuming and even infeasible. It is necessary to develop a computational tool to ease this process, using techniques from areas such as data mining.

One of our goals is to build a classifier that can decide relevancy of a new article based on its title and abstract, in the form of text. Intuitively, this falls in the realm of text mining, and there have been some studies that tried to analyze articles on miRNAs using text mining techniques. However, unlike the application of text mining in other fields, accurate biomedical text mining remains an open problem, because of very specialized, complicated, and fast-growing vocabularies [5]. The accuracy of text mining on biomedical text is debatable. Labeling new entries with relevancy classes (relevant or irrelevant), or relevancy prediction, using the model built from existing data, is a typical data mining classification problem. With proper manipulation of textual inputs, basic classification techniques in data mining can be used to determine their relevancy with high accuracy.

In this research, we have designed and implemented a practical solution to relevancy prediction of text, using classification techniques in data mining. This gives insight to future biomedical literature retrieval on a certain subject. We have constructed the miRDiabetes database with this method, which can benefit biomedical researchers who have interest in miRNA-diabetes associations. Once the relevancy of the entries has been predicted using classification, each entry in the database has also been manually verified to ensure the correctness of the content. We have also implemented this method in an application to facilitate updates, which retrieves and classifies new entries automatically. We also built a website (http://mirdiabetes.ecu.edu) for researchers to search and download this database. We update the miRDiabetes database regularly to keep it up-to-date.

2 Related work

A number of collection databases have been built to assist research on miRNA in the past decade. Researchers started this kind of collection by manually going through articles in PubMed, which is a major source of biomedical publications. In recent years, with the rapid growth of miRNA-related publications, researchers have realized the necessity of using computational tools, especially data mining techniques, to facilitate the collecting process.

The miR2Disease [6] is a manually curated database, aiming to provide a comprehensive resource of miRNA deregulation in various human diseases since 2009. The miR2Disease database offers an excellent framework on miRNA-disease relationships by providing detailed disease categories and a helpful description format. However, the entries currently in the database are outdated and far from comprehensive. The process of collecting was presumably laborious, considering that all the entries were collected manually. With the exploding number of publications on miRNAs, continuing this project with manual labor is unthinkable. The miRecords database [7] was created for curating high-quality experimentally validated microRNA-target interactions (MTIs) with systematic documentation of experimental support for each interaction. After the last update in 2009, the database included 1,135 records of validated MTIs between 301 miRNAs and 902 target genes in 7 animal species, each of which was manually validated. This database was considered large at that time, but it has now been discontinued.

With the number of miRNA-related articles growing dramatically, text mining methods has been applied in

more recent databases on miRNAs. The miRTarBase database [8] is a database that collects experimentally validated MTIs. Hsu and colleagues first used data mining on text to filter articles related to functional studies of miRNAs, and then manually verified those articles for MTIs. The miRSel database [9] utilizes text mining techniques to automatically extract miRNA-gene associations. This database is updated daily with computational predictions based on text-mining results with existing databases. It increased the number of human, mouse and rat MTIs by at least three-fold as compared to Tarbase. The miRWalk database [10] is a database solely based on text mining on PubMed abstracts. One important feature of the miRWalk database is that it not only contains MTI information but also presents validated information on the association between miRNAs and pathways, diseases, organs, disorders, and so on. The miRWalk database contains more than 100 entries based on only less than 10 publications, describing the associations between miRNAs and diabetes mellitus, even though there exist many more publications related to this topic. Furthermore, one of those articles (PMID: 19896465) claims that the miRNA being discussed is not associated with diabetes. Both recall and precision of this database are yet to be improved.

From the aforementioned databases, we can see that both manual collection and computational prediction have their advantages and disadvantages. Manual collection can guarantee the precision of retrieved papers, but it requires too much human labor, especially with the publications on miRNAs increasing rapidly. Building a database purely based on manual selection is undesirable and impractical. Text mining techniques can improve the efficiency of literature collection. It enables databases to make updates daily. However, with the inadequacy of current text mining techniques on biomedical text, it is hard to build a reliable database with high precision. Some systems apply daily updates to keep the databases up-to-date and comprehensive, but it also causes the database to contain redundant information. More specified and precise databases are much more desirable to researchers on a certain topic.

The miRCancer database (http://mircancer.ecu.edu) [5] is a collection database on association of microRNA and cancer. It was first created in 2012. The database was constructed using text mining on literature, and the selected entries were then manually verified to preserve precision. This database specializes on the association of miRNAs and one category of human disease. The database is relatively small and easy to maintain, but it is also comprehensive on target subject. Verification is not too much work with the help of the filtering and prediction of text mining techniques. This database is updated quarterly hitherto.

There are three commonly used approaches in biomedical text mining: co-occurrence approach, rule-based approach, and machine learning-based approach. Since biomedical terms are usually long and unique, co-occurrence approach is usually the simplest and most effective method for information extraction. Co-occurrence approach gives text mining a great efficiency and works fairly well for applications with complete dictionaries in specialized domains. The miRSel database adopted this approach, which enables it to perform daily updates, keeping it comprehensive. However, co-occurrence does not guarantee relevancy or association, which means that systems based on this approach have relatively low precision.

Rule-based approach is a text mining method in which a set of rules are designed to determine certain characters of a text. Different rules are then united with either Boolean combination or voting, to determine the target character of this text. Rules and the mechanism of their combinations are usually empirical. Developers usually design more rules to construct algorithms that are more specific. This kind of text mining is suitable for classification or prediction of papers on a certain subject. For example, the miRCancer database implements rule-based text mining. The developers collected 75 different sentence structures that scholars used to describe miRNA expression in cancers, according to which they constructed 75 rules. For a new article, these 75 rules are calculated against each sentence of its abstract. The results are then combined by voting to decide relevancy of this article. The nature of this combination was tested and refined by experiments.

This research suggests that the mechanism of rule combination can also be determined by data mining techniques. The results of rule calculations against an article can be considered as its numerical or categorical attributes, which will be used to decide its target category.

Machine learning-based approach, usually combined with natural language processing (NLP), are known as text mining techniques, which accumulate known knowledge from text to predict new patterns or associations. Machine learning-based text mining requires training data that can be expensive or even impossible to generate. Meanwhile, building a reliable NLP engine for biomedical text may be costly and even counterproductive for a collection database. Building a processor that can process long sentences with professional words is hard and the result is prone to information loss.

Our goal of retrieval is to find relevant articles out of all retrieved articles. Text classification will suffice for our purpose. If we can somehow convert text into a well-selected list of representative attributes, basic data mining techniques can be applied properly to obtain promising results.

3 Methodology

The goals of this research are to build a collection database of literature on the subject of miRNA-diabetes associations, and to use classification techniques to perform future retrieval. When constructing the miRDiabetes database, we focused on obtaining training data and constructing a classifier. Updates to the database will involve information retrieval and the usage of the classifier, as well as human verification.

3.1 Obtaining training data

The first step was to obtain training data for classification. We retrieved information of literature from PubMed using E-utilities [11], and performed named-entity recognition (NER) of miRNAs and diabetes to select only the publications that had references to both subjects.

The NER of diabetes is straightforward since diabetes has relatively small dictionary. We focused on the existence of the stem "diabet-", the recognition of the three types of diabetes, and occurrences of diabetes-related terms, such as insulin, pancreas, blood glucose level, etc.

The subject of miRNA can be referred to as microRNA, miRNA, micro-RNA, or micro RNA, which can be recognized using simple keyword searching. Thus, to retrieve information from PubMed, the following miRNA query is used.

(((mir) OR mirna) OR microrna) OR micro-rna) OR micro rna

While some relevant papers mention miRNA as one entire family, others investigate the exact types of miRNAs, by referring to their names. The annotation of miRNAs was uniformed quite early and their names are formalized and relatively simple to recognize [12]. The name of a miRNA is composed of a "mir-" prefix followed by a number, e.g., mir-121, with some variations. The popular prefixes and suffices are listed in Table 1. While conferring information in capitalization is highly discouraged, "miR-" is still one of the most used prefixes. In rare cases, papers refer to miRNAs with prefixes like "microRNA-", "miRNA-", "micro-RNA-", or even "micro ribonucleic acid". These are no standard miRNA names, but they happen from time to time.

Table 1: miRNA suffixes

Description	Example
Closely related sequences	hsa-mir-121a, hsa-miR-121b
Distinct precursor sequences and genomic loci that express identical mature sequences	hsa-mir-121-1, hsa-mir-121-2
Sequence from the 3' arm	miR-142-3p, formerly miR-142-as
Sequence from the 5' arm	miR-142-5p, formerly miR-142-s
Minor sequence	miR-56*

At the time of our construction, only a few hundred of articles were extracted after these two processes. We used those articles to build our training data. We went through all of them and labeled them with relevancy, which is the target attribute, or class-label attribute, for the classification. During this process, we categorized articles into the following six classes, regarding their relevancy on the subject of miRNA-diabetes association. Without loss of generality, we assume that a well-trained scholar can do this job correctly.

(1) Direct relation

This type of paper investigates the effect of a certain miRNA, or a group of miRNAs, on diabetes or certain aspect of diabetes, such as its pathology or treatments.

(2) Bridging relation

This type of paper focuses on the effect of miRNAs on a subject, which has deep relation with both miRNAs and diabetes. One example for this kind of subject is renal fibrosis.

(3) Potential relation

This type of paper may discuss the effect of miRNA on a subject, and suggests or claims that this bring novel ideas for understanding diabetes. It may otherwise discuss aspects of diabetes, and suggest miRNA as an intriguing idea, but it is not the main topic of the paper. This class is vague, but papers of this kind can be interesting for those who want to understand the relation between miRNA and diabetes.

(4) Unfocused relation

This type of paper claims that miRNAs have something to do with a range of diseases, which include diabetes. It either explains the effect of miRNAs on the similar symptoms of these diseases, or more frequently uses known fact to introduce miRNA.

(5) Introductive reference

This type of paper mentions relations between diabetes and miRNA, but the relation is not the main topic. It either refers to diabetes to state the importance of miRNA, or mentions miRNA in the discussion of diabetes as reference. Sometimes the paper discusses a topic that has little relation to either miRNA or diabetes. The two subjects are mentioned just as a reference.

(6) No relation

This type of paper mentions nothing about miRNA or diabetes. They appear in the search result when keyword searching is not sufficient or powerful enough.

In our database, papers in the first three classes are considered relevant, while other papers are considered irrelevant. However, the classification of papers in class III (potential relation) can be subjective, as some researchers may consider them as irrelevant. This database keeps them as relevant to guarantee the recall of this collection.

3.2 Construction of a classifier

The second step was the construction of a classifier. We experimented on possible set of quantifiable attributes extracted from text, as well as different types of classification techniques. We compared the classification results, based on their measures, including accuracy, F-measure, etc., and chose the best set of attributes and classification technique to build a classifier for future retrieval.

During the investigation of retrieved articles, we focused on sentences where miRNAs and diabetes were mentioned. We considered 10 properties that an article might have, regarding relevancy of the text. For example, miRNAs and diabetes might be referred to in an enumeration or a quotation, indicating that they were not the main topic of the paper. Diabetes could also be brought up as a risk factor, or in phrases like "diabetes drugs" or "diabetes patients". With proper calculations, these 10 properties can be extracted from each sentence.

The following 10 properties are calculated for each sentence (the title is considered as a separate sentence). Keyword searching is the main method we used to perform this calculation. The values of these properties are Boolean type. These 10 properties are all self-explanatory.

a. Mentioning terms relevant to miRNA;
b. Mentioning miRNA in an enumeration;
c. Mentioning miRNA, but not in situation b;
d. Seeming like a quotation;
e. Mentioning terms related to diabetes;
f. Mentioning diabetes in diabetes drugs;
g. Mentioning diabetes in an enumeration;
h. Mentioning diabetes as a risk factor;
i. Mentioning diabetic objects, i.e. patients, mice;
j. Mentioning diabetes, but not in situations f, g, h or i.

Properties from all sentences can be combined into attributes for the entire text. For example, one attribute of a text can be true, if one sentence has a certain property, while none of other sentences has another property. Not all combinations are helpful. For example, even though titles are treated as sentences, they usually do not contain quotations. With thorough discussion and evaluation, we eventually chose 22 attributes that were most likely to be useful for classification, as shown below.

A01: Title mentions miRNA;
A02: Title mentions diabetes;
A03: Title mentions diabetic object;
A04: Title contains words related to diabetes;
B00: Possibility for bridging;
B01: miRNA appears and only appears in enumerations;
B02: miRNA is mentioned;
B03: miRNA is mentioned more than once;
B04: miRNA only appears in quotations;
B05: diabetes only appears in enumerations;
B06: diabetes only appears as a risk factor;
B07: only diabetic objects are mentioned;
B08: diabetes only appears in quotations;
B09: miRNA and diabetes coexist in an enumeration;
B10: miRNA coexists with diabetes as a risk;
B11: miRNA coexists with diabetic object;
B12: diabetes is mentioned;
B13: miRNA and diabetes coexist;
B14: only diabetes drug is mentioned;
B15: miRNA coexists with relevant words;
B16: diabetes/relevant terms happen more than once;
B17: diabetes/relevant terms coexist with miRNA related words.

Attributes A01 – A04 are properties of the title sentence. The title is extremely important in deciding the topic of a paper. If a title mentions both miRNAs and diabetes, it is probable that this paper is relevant to our subject. A few papers do not have abstract available on PubMed (e.g. PMID: 19145005).

Attributes B01 – B17 are attributes derived from properties of sentences in abstracts. All attributes are desirable features for a paper to be relevant or irrelevant. However, some of them are not very helpful in determining relevancy. For example, B02 being true is necessary for a paper to discuss miRNA-diabetes association. However, this attribute is true for most of the retrieved papers (506 out of 520), and may not be helpful in classification. On the contrary, if B01 (miRNA only appears in enumerations) is true for a paper, it is most likely to be irrelevant, but for most of the retrieved papers (506 out of 520), this attribute is false.

All the attributes are calculated based on properties of sentences, except B00. B00 is true if one relevant term coexists with miRNA in one sentence and with diabetes in another sentence. This attribute involves both the title and the abstract.

These 22 attributes can be calculated against all papers in a dataset, therefore constituting 22 quantifiable predicting attributes for the classification. A classifier can be built based on values of these 22 attributes for all papers in the training dataset, along with their relevancy as the class-label attribute.

We tested multiple classification methods in Weka [13] against our dataset multiple times during the construction

as preliminary analysis. The results of one test are shown in Table 2. This test was done based on the 520 entries as of April 2014.

Overall, the classifications gave very good results. The exact numbers can vary based on the nature of the data set. Adding or removing an entry can cause a difference in ranking. We found that C4.5 [14] and logistic model trees (LMT) [15] [16] had the desirable accuracy and F-measure in most occasions. We eventually selected LMT for classification, because it gave better results on average for all tests that we performed during the construction. The classification has a current accuracy of 90.4%. Results from C4.5 are also available on our website.

Table 2: Results of different classification techniques

Algorithm	Accuracy	Recall	Precision	F-measure
J48	0.9173	0.9222	0.9477	0.9347
J48 Graft	0.9135	0.9222	0.9419	0.9319
LMT (speeded)	0.9019	0.9132	0.9327	0.9228
Logistic Function	0.9000	0.9132	0.9299	0.9215
Naïve Bayes	0.9000	0.9132	0.9299	0.9215
Simple Cart	0.8981	0.9251	0.9169	0.9210
LAD Tree	0.8981	0.9222	0.9194	0.9208
NB Tree	0.8981	0.9222	0.9194	0.9208
AD Tree	0.8942	0.9222	0.9139	0.9180
BF Tree	0.8923	0.9222	0.9112	0.9167
SPegasos	0.8904	0.9192	0.9110	0.9151
REP Tree	0.8904	0.9132	0.9159	0.9145
Bayesian Logistic Regression	0.8885	0.9222	0.9059	0.9139
Functional Tree	0.8865	0.9102	0.9129	0.9115
ID3	0.8865	0.9042	0.9179	0.9110
Decorate	0.8865	0.9012	0.9205	0.9107
K Star	0.8846	0.8952	0.9228	0.9088
Random Forest	0.8827	0.9042	0.9124	0.9083
Random Tree	0.8788	0.9042	0.9069	0.9055

3.3 Performing updates

The third and final step was to apply the classifier, in the process of making updates to the database. To perform an update, we need to retrieve newly published papers from PubMed, apply NER on them, train a classifier with classified data in database, and classify new entries with their relevancy. We have developed an application to carry out all of these steps. We have also verified the results in order to maintain the reliability of the database.

4 Results

We have implemented our method which has shown promising results. Now we have a database that contains all retrieved papers that mention both microRNA and diabetes. Each entry is labeled with a verified class to indicate relevancy.

As of January 2016, there are 1055 papers in the database, with 804 relevant papers (Class I, II, and III). Table 3 shows distributions of the six classes. This database is also accessible on our website. The URL of our miRDiabetes website is: http://mirdiabetes.ecu.edu.

Table 3: Distribution of classes in miRDiabetes as of Jan. 2016

Class	Count
1. Direct Relation	252
2. Bridging Relation	293
3. Potential Relation	122
4. Unfocused Relation	145
5. Introductive reference	218
6. No relation	25

LMT is used for the classification. As of January 2016, this algorithm has accuracy of 90.4%, recall of 90.4%, precision of 90.4%, and F-measure of 90.4%. During a recent update, 71 out of 80 new entries were correctly classified. The accuracy of predicting new entries was 88.8%, which was very promising.

Not all attributes are used in the current classification process. Attributes with skewed distributions are less likely to be used by a classification algorithm, especially after pruning. These attributes are designed to handle special cases. For example, B06 is only true for 21 papers, but 18 (85.7%) of them are irrelevant to our subject. We think they can be put into better use with careful investigation and better application of data mining techniques.

The execution of the classification process is fast. The times in Table 4 were recorded for a classification based on the 520 entries in miRDiabetes as of April 2014. The classification time was the total time of evaluating all 520 entries.

Table 4: Times used for classification

	Test 1	Test 2	Test 3	Average
Attribute extraction (s)	0.925	0.854	0.867	0.882
Classifier building (s)	0.463	0.445	0.456	0.455
Cross-validation (s)	2.058	2.066	2.068	2.064
Classification (ms)	1.274	1.250	1.650	1.391

5 Conclusion

In this research, we have designed a new solution to literature retrieval on miRNA-diabetes association. By extracting Boolean attributes from text, we can utilize classification techniques to determine relevancy of new text with high accuracy. The idea has been implemented with good performance. Our approach achieves classification accuracy of 90.4% and F-measure of 90.4%. This framework alleviates the workload of future retrieval. It may give some insight on biomedical literature retrieval on other subjects.

With this method, we have created the miRDiabetes database, the first collection database on the subject of miRNA-diabetes associations, as well as an application to perform updates using data mining techniques, which can retrieve and classify new publications. We have also built a website for users to access miRDiabetes database.

We have carefully verified all papers in miRDiabetes database to preserve the reliability of the database. The database is being updated regularly.

Besides keeping the database up-to-date regularly, we plan to explore better ways to select and calculate the attributes. We may consider adding more attributes that are relevant. We also aim to further improve classification accuracy by combining multiple classification techniques. In addition, we will continue to consult experts on miRNA-diabetes association to offer classification that is more accurate with better verification process.

References

[1] M. Andrade and P. Borka, "Automated extraction of information in molecular biology," *FEBS Letters*, no. 476, pp. 12-17, 2000.

[2] Y. Gusev and D. J. Brackett, "MicroRNA expression profiling in cancer from a bioinformatics prospective," *Expert Review of Molecular Diagnostics*, vol. 7, no. 6, pp. 787-792, 2007.

[3] US National Library of Medicine, "PubMed," [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/. [Accessed 20 5 2014].

[4] M. N. Poy, L. Eliasson, J. Krutzfeldt, S. Kuwajima, X. Ma, P. E. Macdonald, S. Pfeffer, T. Tuschl, N. Rajewsky, P. Rorsman and M. Stoffel, "A pancreatic islet-specific microRNA regulates insulin secretion," *Nature*, vol. 432, no. 7014, pp. 226-230, 2004.

[5] B. Xie, Q. Ding, H. Han and D. Wu, "miRCancer: a microRNA-cancer association database constructed by text mining on literature," *Bioinformatics*, vol. 29, no. 5, pp. 638-644, 2013.

[6] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng and X. Zhang, "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Research*, vol. 37, pp. D98-D104, 2009.

[7] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao and T. Li, "miRecords: an integrated resource for microRNA-targe interations," *Nucleic Acids Research*, vol. 37, pp. D105-D110, 2009.

[8] S. D. Hsu, F. M. Lin, W. Y. Wu, C. Liang, W. C. Huang, W. L. Chan, W. T. Tsai, G. Z. Chen, C. J. Lee, C. M. Chiu, C. H. Chien, M. C. Wu, C. Y. Huang, A. P. Tsou and H. D. Huang, "miRTarBase: a database curates experimentally validated microRNA-target interactions," *Nucleic Acids Research*, vol. 39, pp. D163-D169, 2011.

[9] H. Naeem, R. Küffner, G. Csaba and R. Zimmer, "miRSel: automated extraction of association between microRNAs and genes from the biomedical literature," *BMC Bioinformatics*, vol. 11, p. 135, 2010.

[10] H. Dweep, C. Sticht, P. Pandey and N. Gretz, "miRWalk--database: prediction of possible miRNA binding sites by "walking" the genes of three genomes," *Journal of Biomedical Informatics*, vol. 44, no. 5, pp. 839-847, 2011.

[11] US National Library of Medicine, "Entrez programming utilites help," 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK25501/. [Accessed 20 5 2014].

[12] V. Ambros, B. Bartel, D. P. Bartel, C. B. Burge, J. C. Carrington, X. Chen, G. Dreyfuss, S. R. Eddy, S. Griffiths-Jones, M. Marshall, M. Matzke, G. Ruvkun and T. Tuschl, "A uniform system for microRNA annotation," *RNA*, vol. 9, no. 3, pp. 277-279, 2003.

[13] G. Holmes, A. Donkin and I. H. Witten, "Weka: a machine learning workbench," in *Second Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia, 1994.

[14] J. Quinlan, C4.5: programs for machine learning, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[15] N. Landwehr, M. Hall and E. Frank, "Logistic Model Trees," *Machine Learning*, vol. 59, pp. 161-205, 2005.

[16] M. Summer, E. Frank and M. Hall, "Speeding up logistic model tree induction," *Lecture Notes in Computer Science*, vol. 3721, pp. 675-683, 2005.