

ABSTRACT

miRDiabetes: A microRNA-Diabetes Association Database Constructed With Data Mining on
Literature

By Hui Guo

June 2014

Director: Dr. Qin Ding

DEPARTMENT OF COMPUTER SCIENCE

MicroRNAs (miRNAs) are a growing class of non-coding RNAs that regulate gene expression by translational repression. A role for miRNA in diabetes was first established in 2004 and research in miRNA-diabetes association has been an increasing interest since then. However, no effort or computational tool has been put forward to retrieve and gather literature on this topic. In this research, we have designed and implemented a method of utilizing data mining techniques on textual data on this subject, which can automatically determine relevancy of new entries with high accuracy. With this method, we have constructed miRDiabetes, the first comprehensive database to collect information in publications from PubMed that profiles relations between miRNAs and diabetes. We have also developed an application to facilitate future updates and built a website for researchers to search and download the miRDiabetes database.

miRDiabetes: A microRNA-Diabetes Association Database Constructed With Data Mining on
Literature

A Thesis

Presented to the Faculty of the Department of Computer Science

East Carolina University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

By

Hui Guo

June 2014

©Copyright 2014

Hui Guo

miRDiabetes: A microRNA-Diabetes Association Database Constructed With Data Mining on
Literature

By

Hui Guo

APPROVED BY:

DIRECTOR OF THESIS: _____
Qin Ding, PhD

COMMITTEE MEMBER: _____
Junhua Ding, PhD

COMMITTEE MEMBER: _____
Nasseh Tabrizi, PhD

COMMITTEE MEMBER: _____
Qin Ding, PhD

CHAIR OF THE DEPARTMENT OF COMPUTER SCIENCE: _____
Karl Abrahamson, PhD

DEAN OF THE GRADUATE SCHOOL: _____
Paul J. Gemperline, PhD

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Qin Ding, for her patient guidance and valuable advice throughout my entire master degree study.

I would also like to give my sincere thanks to Dr. M.N.H. Tabrizi, who helped and encouraged me in the last two years, even when I had doubt in myself.

Finally, I would like to thank my parents for their love and support, and all those who have contributed to the success of this thesis one way or another.

TABLE OF CONTENTS

ABSTRACT.....	i
ACKNOWLEDGEMENTS.....	vi
TABLE OF CONTENTS.....	vii
LIST OF FIGURES	ix
LIST OF TABLES.....	x
CHAPTER 1: INTRODUCTION.....	1
CHAPTER 2: BACKGROUND AND RELATED WORK.....	7
2.1 PubMed and Queries.....	7
2.2 Databases on MicroRNA Publications	10
2.3 Biomedical Text Mining.....	13
2.3.1 Biomedical Text Mining Approaches.....	16
2.3.2 Manual Verification.....	18
CHAPTER 3: SYSTEM ARCHITECTURE AND METHODOLOGY	20
3.1 Retrieval from PubMed.....	22
3.1.1 Introduction to E-utilities.....	22
3.1.2 ESearch and EFetch.....	23
3.1.3 XML Data	27
3.2 Named Entity Recognition.....	28
3.2.1 NER of Diabetes	29
3.2.2 NER of microRNAs.....	30
3.3 Class Labeling.....	31
3.4 Classification on Relevancy.....	33
3.4.1 Classification in Data Mining	33
3.4.2 Classification on Textual Data.....	35
3.4.3 Evaluation of Classifiers.....	38
CHAPTER 4: IMPLEMENTATION AND RESULTS	39
4.1 Data Retrieving and Preprocessing.....	39

4.2 Data Investigation	41
4.3 Attributes Extraction.....	41
4.4 Classification.....	43
4.5 Results.....	44
4.6 Performing an Update.....	49
CHAPTER 5: DISCUSSIONS	53
5.1 Calculation of Attributes.....	53
5.2 Skewed Attributes.....	54
5.3 Classification.....	55
5.4 Correctness of Verification.....	56
CHAPTER 6: CONCLUSIONS AND FUTURE WORK.....	57
REFERENCES	59

LIST OF FIGURES

Figure 1. Number of papers on miRNA-diabetes association by year	4
Figure 2. Header of PubMed website	8
Figure 3. Number of papers on miRNAs by year	14
Figure 4. System architecture of this project	20
Figure 5. XML response from ESearch	24
Figure 6. XML response from EFetch	26
Figure 7. XML response for a retrieved article.....	28
Figure 8. Dictionaries used in miRDiabetes	42
Figure 9. miRDiabetes website homepage.....	47
Figure 10. Results for a search on search page.....	47
Figure 11. Detail page for an article	48
Figure 12. miRDiabetes website browse page.....	49
Figure 13. Snapshot of startup of the program	50
Figure 14. Snapshot of a to-do list.....	51
Figure 15. Snapshot of information of one article	51
Figure 16. Snapshot of an evaluation result.....	52

LIST OF TABLES

Table 1. miRNA suffixes	31
Table 2. Results of different classification techniques	44
Table 3. Distribution of classes in miRDiabetes as of June 2014.....	45
Table 4. Times used for classification	46

CHAPTER 1: INTRODUCTION

New scientific discoveries are based on the existing knowledge, which has to be accessible and therefore usable, by the scientific community [1]. The roles of microRNAs in the etiology, pathology, symptom, and therapeutics of diabetes did not receive much attention until recent years. This topic has been showing increasing potentials and starts to draw overwhelming interest among biomedical researchers. As publications on the association between microRNA and diabetes grow rapidly in number, researchers have found that retrieving them is a more difficult task than ever. A literature collection database on this topic is in need, as well as computational methods to perform 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 thesis focuses on the construction of this miRNA-diabetes association database, as well as the method of utilizing data mining techniques for literature retrieval, which we have designed and practiced in this process.

MicroRNAs (miRNAs) are a class of naturally occurring, small non-coding RNA molecules, about 22 nucleotides in length. They function via base pairing with complementary sequences within messenger RNA (mRNA) molecules, down-regulating gene expression in a variety of manners, including translational repression, mRNA cleavage, and deadenylation.

The miRNA genes are transcribed in the nucleus, yielding long primary transcripts of miRNA (pri-miRNAs). An enzyme (Nuclear RNase III Drosha) then cleaves pri-miRNAs into precursors of miRNA (pre-miRNAs). Pre-miRNAs, usually inactive, are then exported from the nucleus into the cytoplasm. With further processing, active mature miRNAs are integrated [2]. The transcription and maturation of miRNA is a precise controlled and collaborated process.

Although the first miRNAs were characterized in the early 1990s [3], it was not until the early 2000s that miRNAs were recognized as a distinct class of biological regulators with conserved functions and started to appeal to more researchers. Since then, studies and publications involving miRNAs have grown exponentially. During the last few years, advances have showed that aberrant expression of miRNA is associated with a broad spectrum of human diseases, such as cancer, diabetes, obesity, cardiovascular and psychological disorders [4].

The miRBase database (mirbase.org) [5] is a searchable database that provides miRNA location, sequence data, annotation, and target prediction information. This database was previously hosted and supported by the Wellcome Trust Sanger Institute, and now is hosted and maintained in the Faculty of Life Sciences at the University of Manchester. As of June 2013, it contains 24,521 entries representing precursor miRNAs, expressing 30,424 mature miRNA products, in 206 species. The entire data can be downloaded freely from its website.

Diabetes mellitus (DM), or simply diabetes, is a group of metabolic diseases in which a person has high blood glucose level. Diabetes is either due to the pancreas not producing enough insulin (Type 1 Diabetes, or T1D), or because cells of the body develop insulin resistance, that is, they become resistant to insulin and are unable to use it as effectively (Type 2 Diabetes, or T2D). Insulin is a peptide hormone, produced by beta cells in the pancreas, and is central to regulating carbohydrate and fat metabolism in the body. The third main form of diabetes is gestational diabetes (GDM), which occurs when pregnant women without a previous diagnosis of diabetes exhibit high blood glucose level. It happens when insulin receptors do not function properly [6]. Long-term diabetes affects many major organs, including heart, blood vessels, nerves, eyes, and kidneys. People with diabetes often develop complications such as heart and blood vessel disease, neuropathy, nephropathy (renal disease), retinopathy, foot damage, osteoporosis, hearing loss and so on.

In 2011, 25.8 million Americans, 8.3% of the population in the US had diabetes, and 26.9% of the population of people aged 65 or older had diabetes. Furthermore, estimated 79 million Americans had pre-diabetes [7]. This disease also resulted in 1.4 million deaths worldwide in the same year, making it the eighth leading cause of death.

Compelling evidence has proven that miRNAs contribute to the etiology of diabetes, especially Type 2 Diabetes. A role for miRNAs in T2D was first established 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 [8]. 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. More and more studies on the relation between miRNA and diabetes have led to a much greater understanding of the genetic basis of this disease and provided novel diagnostic, prognostic, and treatment alternatives.

Growing interest in miRNAs inspired swift increase in the number of miRNA-related publications, making it more difficult for researchers in biology or medicine to find papers related to their research areas. As of May 2014, more than 40,000 publications on PubMed involve miRNA [9]. Manual retrieval or classification of these papers is an increasingly demanding drudgery.

Figure 1 shows the growth of literature in the miRDiabetes (relevant papers on miRNA-diabetes association). It is reasonable to predict that the number of articles on this subject will keep growing rapidly in the future. As of May 2014, using merely keyword searching on PubMed (a searchable resource for biomedical literature), there are about 1,000 articles that mention both miRNAs and diabetes, and only one third of them are considered relevant by miRDiabetes. At this point, the task of gleaning relevant literature is still feasible through human reviewing and selection. With the exploding number of candidate papers, manual literature retrieval 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.

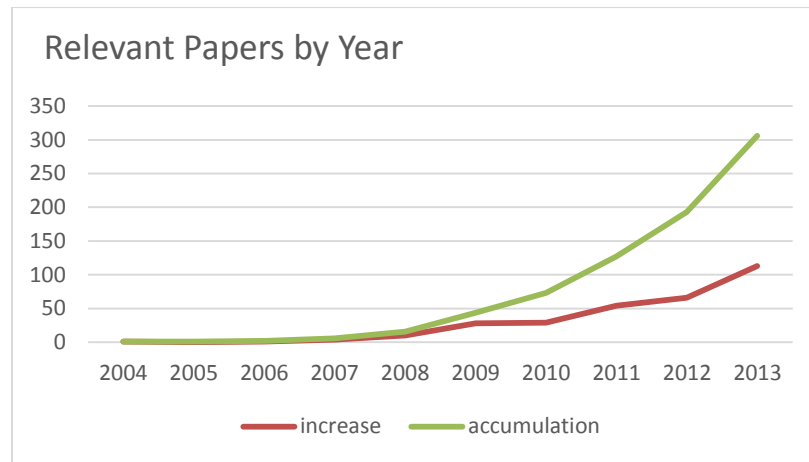


Figure 1. Number of papers on miRNA-diabetes association by year

Data mining, also referred to as knowledge discovery in databases (KDD), is a process of nontrivial extraction of implicit, previously unknown, and potentially useful information from data in large databases [10]. It involves analyzing data from different perspectives and summarizing it. Data mining is an important subject in computer science, and it has major practical applications. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years.

One basic technique of data mining is classification. Classification, sometimes referred to as prediction, is the process of assigning classes, or categories, to previously unseen records as accurately as possible, based on a training set of data containing observations whose classes are known. An algorithm implemented for this purpose is known as a classifier. A classifier also refers to the mathematical function implemented by a classification algorithm. One of the most important measures of a classifier is accuracy, which is the fraction of all testing items that it correctly classifies.

A variation of data mining is text mining, with the data being texts. Text mining, also known as text data mining, refers to the process of deriving high-quality information from text

[11]. However, text mining also involves other subjects of computer science, such as computational linguistics, also known as natural language processing (NLP), pattern recognition, information retrieval, etc.

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 a few 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 [12]. The accuracy of text mining on biomedical texts is debatable.

Labeling new entries with relevancy classes (relevant or irrelevant), or relevancy prediction, 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 made great contributions to computer science as well as biomedical research. We have designed and practiced a practical solution to relevancy prediction of texts, using basic classification techniques in data mining. This gives insight to future biomedical literature retrieval on a certain subject. We have constructed the miRDdiabetes database with this method, which can benefit biomedical researchers who are studying miRNA-diabetes associations. Each entry in the database has been manually verified by our developers. We have also implemented this method in an application to facilitate updates, which retrieves and classifies new entries automatically. This application can also break abstracts into sentences, and highlight key words, to make human verification much easier. We also built a website for all users to search and download this database. We are updating the miRDdiabetes database regularly to keep it up-to-date.

The rest of the thesis is organized as follows. In Chapter 2, the major source of biomedical literature (PubMed) will be introduced, as well as previous collection databases of miRNA publications, and some of the relevant techniques. Chapter 3 will illustrate the methodology of this thesis, including the architecture and details of the project. Chapter 4 will present implementation and results. Discussions will be given regarding some details and shortcomings of this thesis, implying possible future work in Chapter 5. Conclusions of the thesis will be in Chapter 6.

CHAPTER 2: BACKGROUND AND 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 papers, researchers have realized the necessity of using computational tools, especially text mining techniques, to facilitate the collecting process. This chapter introduces PubMed and queries it uses, some collection databases, as well as techniques used for their construction.

2.1 PubMed and Queries

PubMed is the major source for most of the collection databases on miRNA publications. It is also the source of input textual data for this project.

PubMed (pubmed.gov) is a free resource developed and maintained by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM). It provides free access to MEDLINE (Medical Literature Analysis and Retrieval System Online, or MEDLARS Online), NLM's bibliographic database of citations and abstracts in the field of medicine, nursing, dentistry, veterinary medicine, health care systems, and preclinical sciences from more than 5,600 biomedical journals published in the United States and worldwide. As of 2014, more than 23 million citations for biomedical literature in the database can be accessed freely, 3 million of which have links to their full texts in PubMed Central (PMC), a free full-text

archive of journal literature at NLM [18]. Papers on PubMed are identified by their unique IDs, or PMIDs.

In biomedical research, new knowledge is primarily presented and disseminated in the form of peer-reviewed journal articles. Searching through literature to keep up with the state of the art is a task with increasing difficulty for many individual biomedical researchers. PubMed collects these articles and offers a powerful and user-friendly interface to simplify this process. Users can access information of thousands of articles with a proper query. Many biomedical researchers use PubMed for their studies.

A query is a request for information from a database, which must be in a stylized form set required by target database. To make such a request for PubMed, one can use the string query on its website, as shown in Figure 2. PubMed will return all qualified items according to its rules on retrieval.

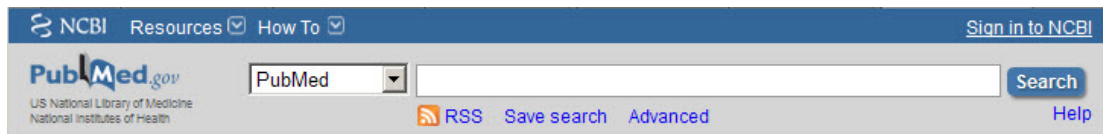


Figure 2. Header of PubMed website

The simplest query for PubMed comprises one term, or key concept. For example, one can simply search “cancer” on the PubMed website, and get information on about three million articles (as of May 2014). Furthermore, this term can be specified to appear in a certain field. In PubMed advanced search, there are 40 fields to choose from, including title, author, date, etc. For example, if one wants to search for articles with “cancer” in the title, the query should be:

cancer[Title]

or:

“cancer”[Title]

Using this query, PubMed will only return about six hundred thousand articles, much fewer than three million. Query in PubMed is case-insensitive.

One field that users can choose is MeSH terms. MeSH is the NLM's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. Using this field will provide much more specified results. Refer to <http://www.ncbi.nlm.nih.gov/mesh> for more information.

A complex query is a string consisting of legal queries and Boolean operators. Operators can be OR, AND, or NOT. A complex query usually contains more than two simple queries. The usage of operators is straightforward. Two queries connecting with AND will return an intersection of the articles retrieved separately. By the same fashion, OR will return a union set, while NOT will return a complementary set.

For example, if one wants to search for articles with both “cancer” and “microRNA” in the title, the following query can be used:

“cancer”[Title] AND “microRNA”[Title]

PubMed will return less than two thousand items this time. With a more complex query, fewer items are likely to be returned. PubMed offers a very versatile and convenient query builder interface for users to build very complicated queries in the advanced search page on its website. A few third-party developers have also developed more customized query builders, either website-based applications or local software, to help researchers save time on literature retrieval.

The query we used in this research is this:

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

This query to PubMed will return all articles with microRNA references. As of May 11th, 2014, this query returns 40,514 items. In this thesis, the number of miRNA-related articles refers to the count of returned entries from PubMed using this query. The term “micro RNA” will not return additional items. It is there for future-proof reasons. PubMed does more than keyword checking for these input terms. Each term will be translated to detailed queries and then queried against the database. However, this does not guarantee that returned articles are relevant.

E-utilities can be used to query PubMed in an application. Their usage and format of XML responses will be discussed in Chapter 3.

2.2 Databases on MicroRNA Publications

As researchers worldwide are undergoing an explosive growth of publications on miRNA, many have found the necessity of collecting them in a uniform format, either from experimental effort or from computational predictions. Therefore, numerous databases have been created in the past few years to collect, conclude, classify, and chalk marks on the published papers to carry forward the process of research.

The earliest ones of those databases were mostly created using experimental verification, namely with manual collection. The miR2Disease database is a very good example. miR2Disease (mir2disease.org) [13] is a manually curated database, aiming to provide a comprehensive resource of miRNA deregulation in various human diseases since 2009. It includes a comprehensive range of human diseases that are considered miRNA-related. Each entry in this database composes detailed information on a miRNA-disease association, including a miRNA ID, a disease name, and a detailed description of this association, often a literature reference along with other observations and attributes. As of 2011, this database contains 3,273 entries, representing relations between 349 miRNAs and 163 diseases. Each entry is supported by an

article, with a link to the corresponding item on PubMed. The process of collecting was presumably laborious, considering all the entries were verified and inserted manually.

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 far from comprehensive. There are only 10 entries on miRNA-diabetes association and 3 on miRNA-diabetes nephropathy, all of which were published before 2011, while our database shows that there were (at least) 29 relevant papers published in the year of 2010 alone. The last update of miR2Disease was made in March 2011, at which time less than 6,000 papers mentioned miRNA in PubMed. With a much larger and ever-growing number of papers on miRNA currently, to continue this project with human labor is unthinkable.

While the miR2Disease database focused on the relations between miRNAs and diseases, other databases accumulated microRNA-target interactions (MTIs). The miRecords database (<http://miRecords.umn.edu/miRecords>) [14] was created for curating high-quality experimentally validated 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.

The miRTarBase database (<http://miRTarBase.mbc.nctu.edu.tw/>) [15] is another database that collects experimentally validated MTIs. At the time of their publication of this database in 2011, there were 3,576 MTIs between 657 miRNAs and 2,277 target genes among 17 species, which were much more than those in miRecords in 2009. Hsu and colleagues first used data mining of text to filter articles related to functional studies of miRNAs, and then manually verified those literatures for MTIs. As of their latest update in November 2013, there are 51,460 entries of MTIs between 1,232 miRNAs and 17,520 target genes among 18 species. They collected 2,636 articles to support the interactions.

With the number of miRNA-related articles growing dramatically, researchers started to realize the power and importance of text mining techniques in helping with the collecting process. Application of text mining methods is one of the important features of more recent databases on miRNAs.

The miRSel database (<http://services.bio.ifi.lmu.de/mirsel>) [16] 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. Text mining enables the reliable extraction of miRNAs, genes and protein occurrences as well as their relationships from texts. It increased the number of human, mouse and rat MTIs by at least three-fold as compared to Tarbase. This database gives an excellent example for the practical uses of text mining on microRNAs in literature.

The miRWalk database (<http://mirwalk.uni-hd.de/>) [17] is another database solely based on text mining on PubMed abstracts. It provides comprehensive information on miRNAs from human, mouse, and rat on their predicted as well as validated binding sites on their target genes. The last updated to this database was made in March 2011.

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 less than 10 papers, describing the associations between miRNAs and diabetes mellitus, even though there are many more papers related to this topic. Furthermore, one of those papers (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. Developers 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/>) [12] is a collection database on association of microRNA and cancer. It was created at East Carolina University in 2012. The database was constructed using text mining on literature, and the selected papers were then manually verified to preserve precision. As of the first quarter of 2014, there are 1,363 papers presenting 2,120 miR-Cancer relations involving 161 different kinds of cancers. 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 for developers with the help of the filtering and prediction of text mining techniques. This database is updated by Xie quarterly hitherto.

2.3 Biomedical Text Mining

As mentioned in Chapter 1, text mining is the process of extracting useful information from textual data. This useful information often refers to some combination of relevance, novelty, and interestingness, in a uniform format that is readable to both human beings and machines. Biomedical text mining, as the name suggests, delimits the input data within texts and literature in the biomedical realm. This is an increasingly interesting subject due to the ballooning number of electronically available publications presented in databases such as PubMed. See Figure 3 for statistics of this growth [9]. Most of the databases mentioned in the

previous section use PubMed as their source of input, usually focusing on the abstracts of retrieved papers.

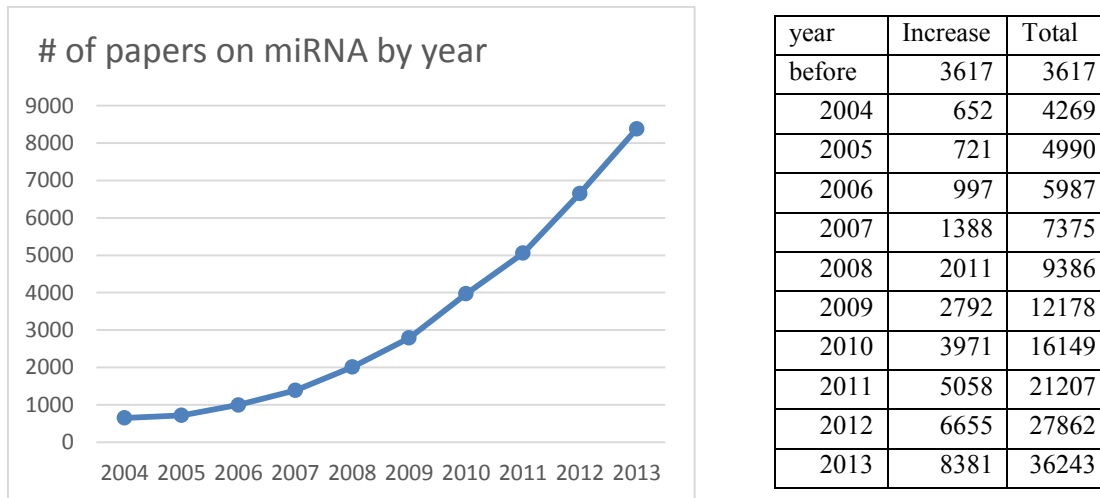


Figure 3. Number of papers on miRNAs by year

Classification is a typical task in text mining. Other tasks include clustering, concept extraction, summarization, relation modeling, etc. For example, databases on miRNA-target interactions (MTIs) aim at modeling relations between miRNAs and target genes. Text mining is sometimes used for sentiment analysis. However, in biomedical domain, most of text is factual statements.

Information retrieval is the activity of obtaining useful data, such as published papers, relevant to an information need from a collection of resources. Building a collection database on biomedical literatures is a process of information retrieval on full-text. It falls in the domain of text classification in text mining. Available texts are classified as relevant or irrelevant during this process, and only relevant documents are retrieved and collected.

Algorithms of literature retrieval must pay attention to two key measures: precision and recall. Precision is the fraction of the documents retrieved that are actually relevant to the topic

intended. Recall is the fraction of all relevant documents that are eventually retrieved by the algorithm.

$$precision = (|\{relevant\ documents\} \cap \{retrieved\ documents\}|) / (|\{retrieved\ documents\}|)$$

$$recall = (|\{relevant\ documents\} \cap \{retrieved\ documents\}|) / (|\{relevant\ documents\}|)$$

Both precision and recall are important to retrieval, but they sometimes are conflicting. Developers usually use F-measure to combine the two parameters. Traditional F-measure, or balanced F-score, is the harmonic mean of precision and recall.

$$F = 2 \cdot (precision \cdot recall) / (precision + recall)$$

In order to accomplish previously mentioned tasks in the biomedical realm, one necessary step is the identification of biomedical entities (named entity recognition, or NER), such as diseases, disorders, protein and gene names, or in this study, miRNA names. Biomedical domain differs from other domains with its very specialized, complicated, and fast-growing vocabularies. A collection of terms under the same category is usually called a dictionary. The performance of information extraction of biomedical text mining implementations depends on the completeness and uniqueness of the words in the dictionaries. Implementations of text mining often start with collecting related terms and storing them in separate dictionaries.

For example, the miRSel database uses miRNA, gene, and protein name dictionaries in order to mine MTIs from literature. These dictionaries are compiled from several databases, including miRBase and gene-centered information at NCBI (Entrez Gene). The miRCancer database uses a cancer name dictionary to perform cancer name recognition. This dictionary was

compiled from the International Classification of Diseases for Oncology (ICD-O). The completeness of this dictionary will affect the recall of miRNA-cancer associations.

2.3.1 *Biomedical Text Mining Approaches*

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. If two unique words from separate dictionaries co-occur in one sentence, there is a high chance that these two subjects are related. The basic operation of this approach is keyword searching, i.e. searching the existence of words from given dictionaries. Besides building dictionaries on related subjects, developers of co-occurrence-based biomedical text mining applications usually spend time on building a predicate dictionary to determine the types of relations between those co-occurring subjects. Examples of those predicates are “regulate”, “express”, “repress”, and “increase” concerning miRNA expression.

Special care must be taken when keyword searching is used, especially when the keyword is short in length. For example, the string “MIR-1” can refer to a human microRNA miR-1 (or hsa-mir-1, miRNA-1), which regulates endothelin-1 in diabetes [19]. However, MIR-1 can also refer to a Russian deep seas submersible. This could result in error of keyword matching, even if this happens rarely in biomedical texts. Words for predicates are usually not as unique as biomedical terms, which also require additional attention. Anagrams are another challenge for keyword searching. These factors can affect the precision of literature retrieval.

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 developers design a set of rules to determine certain characters of a text. A rule can be the existence of a keyword, a co-occurrence, or a combination of a set of co-occurrences. Different rules are then united with either Boolean combination or voting, to construct a universal attribute, which decides the target character of this text. The results of natural language processing (NLP), such as sentence structures, can also be used as a rule.

Rules and the mechanism of their combinations are usually empirical. Developers will decide them according to their experience on this matter. They 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 decided 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 NLP, is a text mining technique, which accumulates known knowledge from text to predict new patterns or associations. A machine learning-based system usually takes advantage of all the available data to make predictions that are more accurate. This approach is not a new concept. PubMiner [20] is a machine learning based system for mining biological information from literature. It was

designed to imply new interactions between biological terms such as gene, protein, and enzymes, based on extracted interactions from existing papers. Some databases use machine learning for MTI predictions based on other collections of MTIs. TargetSpy (targetspy.gov) [21] is an example. TargetSpy is not a text mining system, but it shows how powerful machine learning is in miRNA-target predictions. It is believed that machine learning based text mining can give a good accuracy on predictions. In biomedical publications, many papers use multiple sentences to describe a relation between two subjects. Machine learning and reliable NLP engines are necessary to find these relations.

However, 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. Most of the databases mentioned above use titles and abstracts retrieved from PubMed as input data, in which sentences usually have complicated structures. In our database, each sentence has an average length of proximately 21 words, and a significant number of words are long, rare, and specialized. 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 mathematical attributes, basic data mining techniques can be applied properly to get satisfactory results.

2.3.2 Manual Verification

In information retrieval, precision is one the most important measures. For a collection database, the precision of literature retrieval is essential to its reliability. Reliability is important for a collection because users expect every entry in it to be on topic. Even though data mining techniques nowadays can offer a high accuracy, no one can guarantee a 100% precision,

especially not for textual data. Developers usually go through the effort of manually checking the classification results of all entries to make sure all retrieved papers are relevant.

The miR2Disease database was created entirely out of manual collection. If we assume that a well-trained scholar can make the correct decision on relevancy, this database has a 100% precision. The miRCancer database uses text mining techniques to extract associations, which are in turn validated by the developers through verification of the supporting texts. The miRCancer database also has a 100% precision.

Manual selection is only feasible when the source has limited number of entries. It is reasonable to assume that papers on miRNAs will keep growing exponentially in the future. Pure manual collection of them is impractical. However, papers that mention both miRNA and diabetes have a relatively small count at present. We can take this advantage and use the results of manual collection to train a classifier for future retrieval, using data mining techniques. With computational tools pruning the target collection with high accuracy, we can then verify the results on a relatively smaller candidate pool to make sure that the final collection is reliable.

CHAPTER 3: SYSTEM ARCHITECTURE AND METHODOLOGY

The goals of this thesis are to build a collection database of literature on the subject of miRNA-diabetes associations, and to use classification techniques in data mining to perform future retrieval. The architecture of this project is shown below in Figure 4. When constructing the miRDiabetes database, we focused on the obtaining of training data and the building of a classifier. Updates to the database will involve information retrieval and the usage of the classifier, as well as human verification.

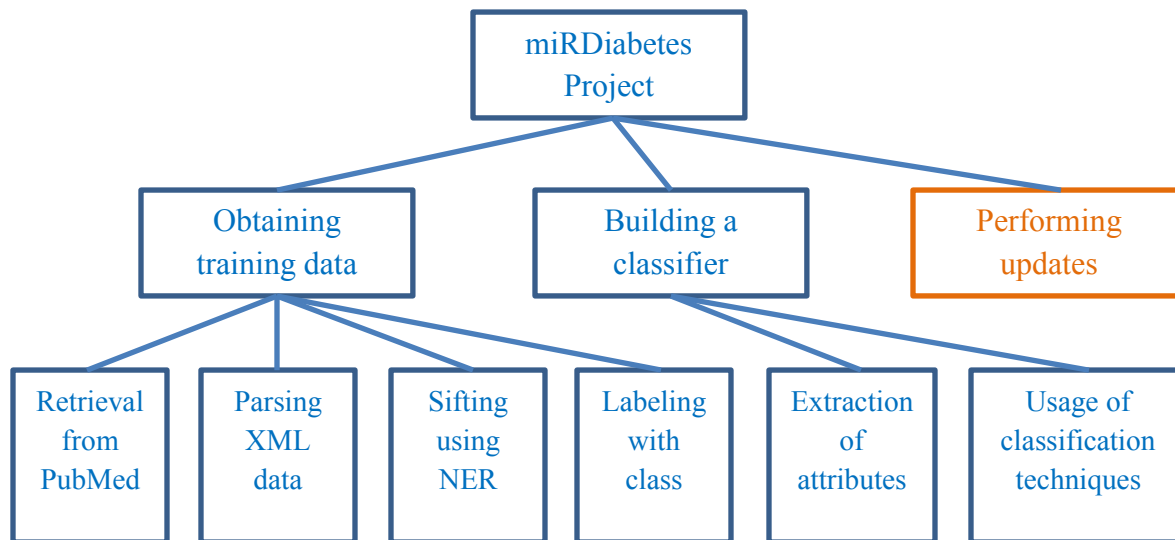


Figure 4. System architecture of this project

The first step of the project is obtaining training data for classification. We retrieved information of literature from PubMed and performed named-entity recognition (NER) of miRNAs and diabetes to leave only the papers that had references to both subjects. At the time of our construction, only a few hundred of articles were extracted after these two processes. We went through all of them and labeled them with relevancy, which is the target attribute, or class-label attribute, for the classification. Without loss of generality, we assume that a well-trained scholar can do this job correctly.

The second step is 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.

The third and final step is the usage of the classifier, in the process of 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 processes. Developers will verify the results in order to maintain reliability of the database.

To retrieve information from PubMed, the miRNA query mentioned in Chapter 2 is used. Information of papers can be downloaded from PubMed website manually. Since we have to keep updating our database regularly, we take advantage of E-utilities, and let the application do this job in order to minimize human labor. Two E-utilities are used, ESearch and EFetch.

The retrieved data is in XML format. XML responses from E-utilities are parsed, and important information of papers is extracted, including their titles and abstracts.

NER is used to guarantee that every paper in the database mentions both miRNA and diabetes. No papers will be deleted from database after sifting of NER, because both relevant and irrelevant papers are needed to train the classifier.

The application developed with our methodology retrieves and classifies papers automatically, which has made updates to the database much easier.

3.1 Retrieval from PubMed

After a query search on PubMed, the results can be downloaded manually. A downloadable file can be created for a search in a certain format on the website. This is very helpful if one wants to conduct a text mining research on the papers, or simply to read the results on a local machine. To make the downloaded file more machine-friendly, one can select XML format before creating a file. The output is a standard XML file, which can be easily interpreted by programming languages such as Perl, Python, JAVA, and C++. The XML tags that PubMed uses will be discussed in this section.

However, PubMed is updated daily from Tuesday through Saturday every week, and the number of citations and abstracts in the database grows rapidly. To build an up-to-date database based on PubMed, developers need to download packed files from PubMed website frequently, which can be tedious and quite slow. Luckily, we can take advantage of Entrez and E-utilities, and let machines do this job.

3.1.1 Introduction to E-utilities

Entrez (Entrez Global Query Cross-Database Search System) is a query and retrieval system developed by NCBI. The E-utilities (Entrez Programming Utilities) are a set of nine-server-side programs that provide a stable interface into the Entrez query and database system. The E-utilities can translate a strictly formatted URL string into the values necessary for various NCBI software components to search for and retrieve the requested data. Entrez and E-utilities can be used to access as many as 38 databases hosted by NCBI, including PubMed, Nucleotide, OMIM and Genome. Entrez system does not include all databases at NCBI, but the majority of data at NCBI are included. All E-utility requests should be made to URLs beginning with the following string:

<http://eutils.ncbi.nlm.nih.gov/entrez/eutils/>

Different E-utilities use different URLs by adding strings to this base string. For example, EInfo is an E-utility for database statistics (number of records, last update, etc.). Its base URL string is:

<http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi>

When using the Entrez system, developers are recommended to comply with NCBI's policies on frequency, timing, and registration of E-utility URL requests, in order not to overload the servers [22].

There are nine E-utilities available: EInfo (database statistics), ESearch (text searches), EPost (UID uploads), ESummary (document summary downloads), EFetch (data record downloads), ELink (Entrez links), EGQuery (global query), ESpell (spelling suggestions), and ECitMatch (batch citation searching in PubMed) [22]. Each of them has very special and much-needed functionalities.

In this project, we have only used ESearch and EFetch, and they have been very helpful. ESearch can return an ID list of desired articles, while EFetch can get comprehensive information on each individual article.

3.1.2 *ESearch and EFetch*

ESearch and EFetch are two of the most used E-utilities. They are powerful enough to conduct all retrievals from PubMed that we needed.

The base URL string for ESearch is:

<http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi>

Each Entrez database refers to a data record within it by an integer ID called a UID (unique identifier). The UIDs that PubMed uses are PMIDs. ESearch can search a database using a text query, and return a list of matching UIDs, along with the term translations of the query.

For example, the first step of this project was to retrieve all papers that had a miRNA reference. We already have a query for PubMed:

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

We can use ESearch to get a list of PMIDs, which are the unique identifiers of the papers we want to retrieve. The URL for this purpose is like this:

```
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=\(\(mir\)+or+mirna\)+or+microrna\)+or+micro-rna\)+or+micro+rna
```

The returned XML file contains information on search results, such as returned count, a UID list, and translation information. An example of such XML files is shown in Figure 5.

```
▼ <eSearchResult>
  <Count>40514</Count>
  <RetMax>20</RetMax>
  <RetStart>0</RetStart>
  ▼ <IdList>
    <Id>24812635</Id>
    <Id>24812632</Id>
    <Id>24812602</Id>
    <Id>24812525</Id>
    <Id>24812324</Id>
    <Id>24812123</Id>
    <Id>24812122</Id>
    <Id>24812086</Id>
    <Id>24812015</Id>
    <Id>24812010</Id>
    <Id>24811921</Id>
    <Id>24811707</Id>
    <Id>24811520</Id>
    <Id>24811488</Id>
    <Id>24811474</Id>
    <Id>24811402</Id>
    <Id>24811259</Id>
    <Id>24811246</Id>
    <Id>24811193</Id>
    <Id>24811064</Id>
  </IdList>
  ▶ <TranslationSet>...</TranslationSet>
  ▶ <TranslationStack>...</TranslationStack>
  ▼ <QueryTranslation>
    ((mir[All Fields] OR ("micrornas"[MeSH Terms] OR "micrornas"[All Fields] OR "mirna"[All Fields])) OR ("micrornas"[MeSH Terms] OR "micrornas"[All Fields] OR "microrna"[All Fields])) OR ("micrornas"[MeSH Terms] OR "micrornas"[All Fields] OR ("micro"[All Fields] AND "rna"[All Fields]) OR "micro rna"[All Fields]) OR ("micrornas"[MeSH Terms] OR "micrornas"[All Fields] OR ("micro"[All Fields] AND "rna"[All Fields]) OR "micro rna"[All Fields])
  </QueryTranslation>
</eSearchResult>
```

Figure 5. XML response from ESearch

The URL string is case-insensitive. There are two parameters in this string: “db” and “term”. The parameter “db” refers to the database name, which is PubMed here. Term refers to the query we wish to use.

Spaces should be avoided in the URLs. We use a plus sign (+) instead of a space if it is required. Special characters should also be handled; for example, we use “%22” for a quotation mark (“), and “%23” for the “#” symbol.

Other parameters can be added to get more customizable results. “Retmax” defines the total number of UIDs in the retrieved ID list in an XML output. The default number is 20, i.e. in this example, 20 PMIDs will be returned, since no retmax value is specified. The maximum retmax value is 100,000. To retrieve more than 100,000 UIDs, one should submit multiple ESearch requests while incrementing the value of “retstart”.

“Retstart” defines the sequential index of the first UID in the retrieved set to be shown. Default value of retstart is zero. If user specifies this parameter, the XML output will only contain UIDs with the sequential numbers equal to or larger than it. Retmax and retstart can be used together to get an arbitrary sequential subset of the entire set. This is very useful for retrieval, especially for searches with large returned set.

After we get the UIDs of retrieved papers, the next step is to get detailed information for each individual entry. EFetch can be used for this purpose. The base URL for EFetch is:

<http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi>

This utility can return formatted data records for a list of input UIDs. For example, if we want to retrieve data for the paper with this PMID: 24741571, we can post a request to Entrez using this URL string:

<http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=24741571>

The parameter “id” is set to be the requested UID or UID list. To retrieve data for two entries or more, use commas to separate those UIDs. For example:

<http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=24741571,24752729>

The returned information includes titles, authors, abstracts, dates, and so on. The returned data is well formatted, but it is a self-defined format, which is not easy to interpret for beginners.

An XML file can be requested by setting “rettype” parameter to “xml”:

<http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&rettype=xml&id=24741571,24752729>

An example of returned XML file is shown in Figure 6 (for PMID: 24741571). The structure of tags will be discussed in the following section.

```
▼ <PubmedArticleSet>
  ▼ <PubmedArticle>
    ▼ <MedlineCitation Owner="NLM" Status="In-Process">
      <PMID Version="1">24741571</PMID>
      <DateCreated>...</DateCreated>
      ▼ <Article PubModel="Print-Electronic">
        ▶ <Journal>...</Journal>
        ▶ <ArticleTitle>...</ArticleTitle>
        ▶ <Pageitation>...</Pageitation>
        <ELocationID EldType="doi" ValidYN="Y">10.1155/2014/761938</ELocationID>
        ▼ <Abstract>
          ▶ <AbstractText>...</AbstractText>
          </Abstract>
        ▼ <AuthorList CompleteYN="Y">
          ▼ <Author ValidYN="Y">
            <LastName>Chang</LastName>
            <ForeName>Xiangyun</ForeName>
            <Initials>X</Initials>
            ▶ <Affiliation>...</Affiliation>
            </Author>
          ▶ <Author ValidYN="Y">...</Author>
          ▶ <Author ValidYN="Y">...</Author>
          ▶ <Author ValidYN="Y">...</Author>
          ▶ <Author ValidYN="Y">...</Author>
          ▶ <Author ValidYN="Y">...</Author>
          ▶ <Author ValidYN="Y">...</Author>
          </AuthorList>
          <Language>eng</Language>
          ▶ <PublicationTypeList>...</PublicationTypeList>
          ▶ <ArticleDate DateType="Electronic">...</ArticleDate>
          </Article>
          ▶ <MedlineJournalInfo>...</MedlineJournalInfo>
          <CitationSubset>IM</CitationSubset>
          ▶ <CommentsCorrectionsList>...</CommentsCorrectionsList>
          <OtherID Source="NLM">PMC3972833</OtherID>
          </MedlineCitation>
        ▶ <PubmedData>...</PubmedData>
      </PubmedArticle>
    </PubmedArticleSet>
```

Figure 6. XML response from EFetch

3.1.3 XML Data

XML (eXtensible Markup Language) is a markup language designed to carry data and represent arbitrary data structures, in a format that is both human-readable and machine-readable. Data structures are defined by tags. Tag structure for all retrieved PubMed data is uniform. Different entries differ only in data carried within tags. In the returned XML of ESearch, the count of the search results is stored in “<Count>” tag. The count of returned UIDs in one response is stored in “<RetMax>” tag, and the sequential number of the first UID is stored in “<RetStart>” tag. The UID list is stored in “<IdList>” tag, which contains a list of UIDs, each of which is casted in an “<Id>” tag. Query translation information is stored in “<TranslationSet>”, “<TranslationStack>”, and “<QueryTranslation>” tags. The entire data is stored in a tag called “<eSearchResult>”.

In the example showed above, 40,514 entries were found for the input query, 20 of which have PMIDs returned. The translation of the query is:

```
((mir[All Fields] OR ("micrornas"[MeSH Terms] OR "micrornas"[All Fields] OR "mirna"[All Fields])) OR ("micrornas"[MeSH Terms] OR "micrornas"[All Fields] OR "microrna"[All Fields])) OR ("micrornas"[MeSH Terms] OR "micrornas"[All Fields] OR ("micro"[All Fields] AND "rna"[All Fields]) OR "micro rna"[All Fields]) OR ("micrornas"[MeSH Terms] OR "micrornas"[All Fields] OR ("micro"[All Fields] AND "rna"[All Fields]) OR "micro rna"[All Fields])
```

We can extract the PMIDs from the ID list, in order to retrieve detailed information using EFetch utility.

The format for the returned XML file from EFetch is more complicated, but also very straightforward. The entire information is casted in a tag called “<PubmedArticleSet>”, with information of each article casted in a “<PubmedArticle>” tag. Each article has a “<MedlineCitation>” tag, containing information of PMID, date of creation, and article

information in “<Article>” tag. Within the “<Article>” tag, there are tags like “<Journal>”, “<ArticleTitle>”, “<Abstract>”, “<AuthorList>”, etc., which are all self-explanatory. This research only focuses on the titles and abstracts of articles. Therefore, we can get the following information from an XML file as in Figure 7.

```

▼ <Article PubModel= "Print-Electronic" >
  ▶ <Journal> ... </Journal>
  ▼ <ArticleTitle>
    Ethnic differences in microRNA-375 expression level and DNA methylation status in type 2 diabetes of Han and Kazak populations.
  </ArticleTitle>
  ▶ <Pageination> ... </Pageination>
  <ELocationID EldType= "doi" ValidYN= "Y" >10.1155/2014/761938 </ELocationID>
  ▼ <Abstract>
    ▼ <AbstractText>
      Han population is six times as likely as Kazak population to present with type 2 diabetes mellitus (T2DM) in China. We hypothesize that differential expression and CpG methylation of miR-375 may be an ethnic-related factor that influences the incidence of T2DM. The expression level of miR-375 was examined using real-time PCR on Kazak and Han T2DM plasma samples. Furthermore, the methylation levels of CpG sites of miR-375 promoter were determined by MassARRAY Spectrometry in these samples. The relative expression levels of plasma miR-375 in Kazak T2DM samples are 1, and the relative expression levels of plasma miR-375 in Han T2DM samples are 3. The mean level of miR-375 methylation, calculated from the methylation levels of the CpG sites, was 8.47% for the Kazak T2DM group and 10.38% for the Han T2DM group. Further, five CpG units showed a statistically significant difference between Kazak and Han T2DM samples, and, among them, four were hypomethylated and only one CpG unit showed hypermethylation in Kazak T2DM samples. These findings indicate that the expression levels of plasma miR-375 and its CpG methylation in the promoter region are ethnically different, which may contribute to the different incidence of diabetes observed in Kazak and Han populations.
    </AbstractText>
  </Abstract>
  ▶ <AuthorList CompleteYN= "Y" > ... </AuthorList>
  <Language> eng </Language>
  ▶ <PublicationTypeList> ... </PublicationTypeList>
  ▶ <ArticleDate DateType= "Electronic" > ... </ArticleDate>
</Article>

```

Figure 7. XML response for a retrieved article

3.2 Named Entity Recognition

Named Entity Recognition (NER, sometimes spelt named-entity recognition), also known as entity identification, or entity extraction, is a subtask of text mining and information retrieval. The performance of an information retrieval system relies highly on the correctness of NER. In this project, the main task of NER is the recognition of miRNAs and diabetes. NER of miRNAs

and diabetes can sift candidate papers from all raw entries, leaving only ones that mention both subjects. NER is also useful for attribute extraction, since most of the attributes that we use involve existence of miRNAs or diabetes in sentences.

3.2.1 NER of Diabetes

Diabetes has a relatively small dictionary. For most of the time, this disease is referred to as “diabetes mellitus” or “diabetic mellitus”, or simply “diabetes”. In rare cases, this disease is referred to as “glycuresis”, which focuses more on the fact of having excess sugar in the urine (glycosuria), as in diabetes. However, this term has never been used in a miRNA-related paper. Some papers may refer to this disease as “dibetes”, but none on miRNA-diabetes association.

There are three types of diabetes, Diabetes Mellitus Type I (T1DM, or T1D, or DM1), Diabetes Mellitus Type II (T2DM, or T2D, or DM2), and gestational diabetes (GDM). Type 2 Diabetes was formerly called Noninsulin-Dependent Diabetes Mellitus (NIDDM). DM1 and DM2 are rarely used for diabetes, rather for Dystrophia Myotonica. Abbreviations are usually used to substitute the exact types of diabetes. However, more often than not, the word “diabetes” is mentioned at least once throughout a text.

NER of most diabetic complications can simply be performed by checking the existence of the word “diabetic”. Even though papers on those complications do not always mention diabetes, there are still high chances that they are relevant to this disease. Papers on diabetic objects (diabetic patients, rats, mice, etc.) can also be relevant to this disease, but it varies from case to case.

Thus, the stem “diabet-” is essential for NER of diabetes. There are other terms with high interest on the subject of diabetes, such as “insulin”, but they do not refer to the disease itself. These words will be discussed in Chapter 4.

3.2.2 *NER of microRNAs*

The subject of miRNA can be referred to as microRNA, miRNA, micro-RNA, or micro RNA, which can be recognized using simple keyword searching. It is worth mentioning that mRNA stands for messenger RNA, instead of miRNA.

While some relevant papers mention miRNA as a whole 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 [23].

The name of a miRNA is composed of a “mir-” prefix followed by a number, e.g., mir-121, with some variations. Sometimes a name can start with a “hsa-mir-” prefix, with the first three letters signify the organism of this miRNA, which in this case is human. Different organisms have slightly different naming conventions. For example, in plants, the prefix “MIR” (with capitalization and no hyphen) is used, such as MIR121. In much of literature, the prefix “miR-” (with capitalization) designates mature miRNAs, e.g. miR-121, while “mir-” often refers to a miRNA precursor. 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.

The numbering of miRNAs is simply sequential, with identical miRNAs having the same number, regardless of organism. For instance, if the last published miRNA was mouse mir-352, the next novel published miRNA would get the number 353, even though it might be a human miRNA. Nearly identical orthologs can also be given the same number, at the discretion of the researcher. Identical or very similar miRNA sequences within a species can also get the same number, with their genes distinguished by letter and/or numeral suffixes, according to the convention of the organism. See Table 1 for these variations.

Table 1. miRNA suffixes

Description	Example
Closely related sequences	hsa-mir-121a and hsa-mir-121b hsa-miR-121a and hsa-miR-121b
Distinct precursor sequences and genomic loci that express identical mature sequences	hsa-mir-121-1 and 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 (not predominant)	miR-56*

There are exceptions to the naming scheme mentioned above. Examples are let-7 and lin-4. These names remain for historical reasons. A few articles that refer to these names also mention diabetes, so they must be taken into account of miRNA NER as well.

3.3 Class Labeling

We reviewed all the papers retrieved from PubMed, by reading their titles and abstracts. During this process, we categorized these articles into the following six classes, regarding their relevancy on the subject of miRNA-diabetes association. This was done to construct the training set for later classification process. This label can be used as class-label attribute for classification.

I. 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.

II. 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. We know that elevated blood glucose is the major character of diabetes. Over time, especially in genetically susceptible individuals, such chronic hyperglycemia can cause tissue injury. One pathological response to tissue injury is the development of fibrosis. Renal fibrosis is one of the major diabetic

complications. There have been papers investigating the effect of certain miRNAs, such as miR-21, on renal fibrosis, in order to propose novel targets for diabetic nephropathy.

III. 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.

IV. 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.

V. 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.

VI. 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. For example, some papers use DM2 to denote Diabetes Mellitus Type 2. However, DM2 is more likely to be the abbreviation of Myotonic Dystrophy Type 2, which has nothing to do with diabetes. In addition, “miR” is case-sensitive for microRNA, since MIR can mean something different.

In this 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 researcher may consider them as irrelevant. This database keeps them as relevant to guarantee the recall of this collection.

3.4 Classification on Relevancy

If all papers in a collection database are labeled with classes, it is actually ready to use. However, maintaining this database will involve endless human labor, since new publications will keep coming and this database will soon fall out of date if it is not updated regularly. To alleviate the pain of updating, we need computational tools to retrieve and classify new entries. Our method of text classification that we have designed will be presented in this section. We have developed an application using this method, which can extract features of a text and determine its relevancy. We can then verify this result and update the database. The implementation of this application will be presented in Chapter 4.

3.4.1 *Classification in Data Mining*

In data mining, classification is a technique used to predict group membership for data instances. In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. Supervised learning fits to the need of this project, since a researcher can label a paper with relevancy on the subject of miRNA-diabetes association.

The individual observations, or instances, are analyzed into a set of quantifiable properties, known as various explanatory variables, or attributes, which may be categorical, ordinal, or numerical. The correct identification of an instance can be considered as a special attribute called class-label attribute. Classification is the process of building a model (or

classifier) that describes the relationship between a set of observable attributes (or predicting attributes) and the class-label attribute.

A classifier is built based on training data. Training data is correctly identified observations, with both predicting attributes and the class-label attribute ready. A classifier learns upon these classified instances (model construction), and then evaluates the value of the class-label attribute of new data (model usage), with only predicting attributes available. After a classifier is built, it is usually tested on a set of testing data. The results of this classifier are compared against the correct classes. The purpose of this step is to determine the accuracy of this classifier, in order to make sure its output is reliable.

Popular classification techniques include decision tree, linear regression, Bayesian classifications (Naïve Bayesian classification, Bayesian belief networks, etc.), neural networks, and so on. Numerous algorithms have been developed based on these techniques.

One example is the C4.5 algorithm developed by Ross Quinlan [24]. C4.5 is one of the best-known algorithms used to generate a decision tree. It uses the concept of information entropy to decide splitting attributes in each step of the tree building.

In this research, after comparative study, we eventually chose logistic model trees (LMT) algorithm to perform the classification because it gave satisfactory results on average. For predicting numeric quantities, model trees have been developed, combining both linear models and trees. Model trees are decision trees that contain linear regression functions at the leaves. In 2005, Landwehr, Hall, and Frank developed LMT, combining trees with logistic regression, instead of linear regression, to achieve better performance [25]. Later, Sumner, Frank, and Hall applied different validation and a weight trimming heuristic to LMT, and produced a significant speedup with limited accuracy loss [26].

3.4.2 Classification on Textual Data

To apply above-mentioned classification techniques, we need to get quantifiable attributes from data first. In this research, relevancy of an article is considered as a class-label attribute, which is a Boolean one (true or false). As mentioned in Chapter 2, rule-based text mining approach can be applied to predict relevancy, which utilizes rules and their combinations. The results of rule checking can be treated as Boolean attributes. We can then use data mining techniques to determine the way of combining these attributes to perform the classification. Extracting these attributes from text, i.e. titles and abstracts, is the major problem that we must conquer.

During the investigation of retrieved articles, we paid attention to sentences where miRNAs and diabetes were mentioned. We concluded 10 properties that they might have, regarding relevancy of the whole 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 (title is considered as a separate sentence). The values of these properties are Boolean, i.e., they can be either true or false. 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 (patients, mice, etc.);
- j. Mentioning diabetes, but not in situations f, g, h or i.

Properties from all sentences can be combined into attributes for the whole 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. Title is extremely important in deciding the topic of the paper. If a title mentions both miRNAs and diabetes, it is probable that this paper is relevant to our subject. It is also worth mentioning that 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 abstract. 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 is true if the abstract has miRNA references. It 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. These attributes will be further discussed in Chapter 5.

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. To predict relevancy of a new entry, the 22 predicting attributes are calculated for its text, and the classifier outputs the prediction result according to these 22 values.

The performance of a classifier can be evaluated using techniques mentioned in the following section.

3.4.3 Evaluation of Classifiers

After a classifier is built, it must be properly evaluated to estimate how accurately the predictive model will perform in practice. Evaluation on different classifiers can also help us choose better models. Usually a dataset is divided into two parts: training data and testing data. Classifiers are built based on the training data. They are then tested against the testing data to determine their accuracy, by comparing the predicted classes to actual ones.

However, for relatively small datasets like ours, using only half (or a relatively larger part) of the data for training is unwise. In this case, cross-validation, sometimes called rotation estimation, gives better estimation. K-fold cross-validation first divides the dataset into k subsets. In one round of cross-validation, k-1 subsets are used as training data and one subset is used as testing data. To reduce variability, multiple rounds are performed using different partitions, and the validation results are averaged over the rounds [27].

Cross-validation is also used in LMT algorithm for stopping criterion. Additional iterations are only performed if they actually improve predictive accuracy. However, cross-validation usually uses multiple rounds, and this affects the speed of this algorithm. In the speedup, AIC criterion (Akaike Information Criterion) is used instead of cross-validation.

The most popular cross-validations are 2-fold and 10-fold cross validations. This research uses 10-fold cross validation for accuracy of the classification. We compared the cross validation results of different classification techniques, and chose the one with best result to build the classifier of our application.

It is worth mentioning that, since cross validation uses random partitions of the dataset, the exact numbers of accuracy can vary.

CHAPTER 4: IMPLEMENTATION AND RESULTS

We have practiced our methodology, and created the miRDiabetes database, as well as an application that can perform update to it and help verification. This database will benefit biomedical researchers on miRNA-diabetes association. The application contains a tool to classify new entries based on the current entries in the database. It not only makes the updates to database much easier for us, but also gives insight on literature retrieval for computer science researchers. We have also built a website for all users to search and access miRDiabetes.

4.1 Data Retrieving and Preprocessing

The first task of building the database was to retrieve information on articles from PubMed. Retrieval is also needed to update the database. Then NER of miRNAs and diabetes is performed to keep in database only the ones that mention both subjects. The application will take the following steps in this process.

I. Retrieve Papers from PubMed

The application uses the query in Chapter 2 to retrieve papers from PubMed. A URL string is generated according to this query. A response, or a list of PMIDs, is requested from Entrez through ESearch utility using this URL. Finally, it requests an XML of paper information using each PMID through EFetch utility. The XML responses, with special characters converted, are parsed into update queries, which are used to insert new entries to the database.

In order to minimize the number of requests to Entrez, we encourage future developers to retrieve paper information using more than one PMID at a time. In fact, our default number of PMIDs in one request is 100.

The exact URLs and XML format can be found in Section 1 of Chapter 3.

If we wish to make an update, the application can find the date of the most recent entry, and only retrieve articles published after that. This date information can be reflected in the query. The “[Date - Create]” field is used in the query to specify the time span desired.

During the construction of the database, we improved our codes several times to make sure that it got information of all papers that PubMed returned. For example, most of the papers on PubMed have one section or paragraph in their abstracts. However, some papers have multiple sections. Without considering this, we could suffer information loss for them.

II. Perform NER of miRNAs and diabetes

NER of miRNAs and diabetes is discussed in Section 2 of Chapter 3. This step will eliminate most of the irrelevant papers. Only the papers having both miRNA and diabetes references are further examined.

Texts are first tokenized into words, and then keyword checking or name recognition is performed. Names of miRNAs, if any, are also extracted and saved for future reference.

III. Break abstracts into sentences

Attributes are extracted from properties of sentences. Thus, this is necessary, not only in computational predication, but also in helping the verification process. Abstract that is broken into sentences are relatively easier to read.

This step involves splitting the abstract string with the period signs, or dots. Special care must be taken in this step. Numbers with decimals, such as “3.14”, and words with dots, such as “i.e.” or “U.S.”, are exceptions of dots representing ends of sentences. We paid special attention to these exceptions, and merged suspiciously short sentences into neighboring sentences. This function has been proven to work as expected.

4.2 Data Investigation

As of April 2014, we have 520 papers in the database after the first step. Each paper has been investigated and classified into six classes: direct relation, bridging relation, potential relation, unfocused relation, introductory reference, and no relation. Every classified entry in the database will be later used in computational classification of new ones.

Relevancy can be a subjective property. In this study, we do not argue the correctness of manual classification. For example, obesity has many similarities to diabetes. Many researchers investigate obesity and diabetes at the same time. Nevertheless, in miRDiabetes database, association of miRNA and obesity is considered irrelevant, unless diabetes is also one of the focuses of the paper. Diabetic nephropathy is a major diabetic complication. In this research, papers on diabetic nephropathy will be considered as directly relevant.

The correctness of manual classification does matter in practical use. This will be further discussed in Chapter 5.

4.3 Attributes Extraction

To calculate the 10 properties for each sentence, keyword searching is performed. We use a few dictionaries to execute this process. If a sentence contains a word from a dictionary, then the keyword searching will return true for this dictionary.

Figure 8 lists the words in each dictionary. We will keep updating them if necessary.

The dictionary of bridges is used for detecting bridging between sentences, which means a term coexists with miRNA in one sentence and with diabetes in another. The dictionary of disease is used for detecting enumerations. Most of the time, an enumeration starts with a phrase like “such as”, “including”, or “like”. The application searches these phrases first. This is done recursively, since a sentence may contain multiple enumerations. Meanwhile, existence of such

phrases does not guarantee an enumeration. This dictionary contains terms that are most likely to be enumerated along with diabetes. The application then examines the existences of such words to determine enumerations.

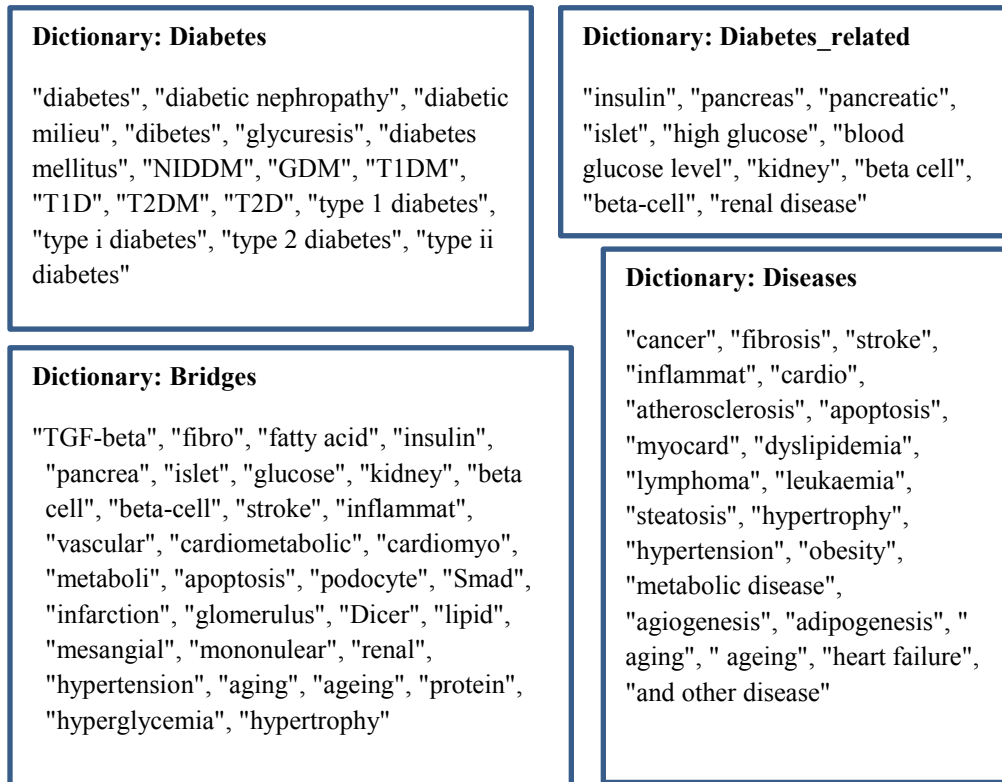


Figure 8. Dictionaries used in miRDiabetes

Smaller dictionaries are usually hard-coded. For example, keywords for miRNA are small in number, and the naming of miRNAs is unified. NER of miRNAs is executed by a hard-coded function. Furthermore, terms for diabetes drugs, terms related to miRNAs, and terms to determine miRNA enumerations are relatively fixed, and their counts are rather small. For the time being, they are all hard-coded as well.

Detection of quotations is harder to do. For simplicity, we only performed keyword searching on phrases like “known to be”, “it has been shown”, “previous/recent studies”, “been described to”, and so on. Of course, if one sentence also contains phrases like “this research”, it is less likely to be a quotation. This detection is not a thorough examination of possible references.

In our database, it works fine for most of the situations. The usage of attributes involving this function will be investigated in Chapter 5.

4.4 Classification

With the results of attribute calculations and class labeling, we created a relative small, well-formatted data file. This file can be used for C4.5 program and other statistics or data mining software. Our application can generate such a data file for the current entries on miRDiabetes. This file considers papers with direct relation, bridging relation, and potential relation as relevant ones. Different formats and versions of this file are available on our website.

Weka (Waikato Environment for Knowledge Analysis) [28] is a popular suite of machine learning software written in JAVA. Its development started at the University of Waikato, New Zealand, intended to aid in the application of machine learning techniques to solve real-world problems. Currently it is widely used for educational purpose and research. Weka is a free application under the GNU General Public License. By integrating algorithms from Weka, we can perform our classification with advanced data mining techniques.

The data file that Weka uses is in the format of ARFF (Attribute-Relation File Format), which can be generated by our application and is downloadable on our website.

We tested multiple classification methods against such data files several times during the construction. The results of one test are shown in Table 2. This test was done based on the 520 entries as of April 2014. The names of algorithms follow the naming in Weka. For example, J48 is an implementation of C4.5 in Weka; hence, the name J48 is used instead of C4.5.

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 and LMT 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 current accuracy of 90.0% (the exact number may vary). Files for 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/Simple Logistic Function	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

4.5 Results

We have practiced our method and got a good result. Now we have a database that contains all the papers that mention both microRNA and diabetes. Each entry is labeled with a verified class to indicate relevancy. We are using MySQL database for the project.

As of June 2014, there are 617 papers in the database, with 381 relevant papers (Class I, II, and III). Table 3 shows distributions of the six classes. This database is also accessible on our website. The miRDdiabetes website can be visited from <http://miridiabetes.ecu.edu>.

Table 3. Distribution of classes in miRDiabetes as of June 2014

Class	Count
1. Direct Relation	183
2. Bridging Relation	114
3. Potential Relation	84
4. Unfocused Relation	93
5. Introductive reference	131
6. No relation	12

The miRDiabetes database has four tables. All tables use PMID as their primary keys. The SQL file for these tables is available on our website.

I. Table Papers

This table contains all information of retrieved papers except for journal information. This information includes PMID, date of creation, title, abstract, author, affiliation, language, keywords, pagination, and so on.

II. Table Journal

This table contains journal information. This information includes PMID, ISSN, volume, issue, year, month, title, and so on. Pagination is stored in Papers.

III. Table SimpleLit

This table is a simplified table of papers for the purpose of classification. This table contains PMID, title, and abstract of all papers. This table also includes class, comment, and attributes of every paper, which are the results of either classification or manual verification. This table is built so that Table Papers will not be modified in the processes after initial retrieval.

IV. Table Indexes

This table contains NER results, including names of miRNAs and diabetes that each paper mentions. We include this table so that users can search papers by miRNA and diabetes on our website.

We have implemented an application that can be used to perform updates. The application will retrieve possibly relevant papers from PubMed, apply classification based on classified entries in the database, and allow developers to conduct verification and save the results into database. This application also breaks abstracts into sentences, highlights key words, and shows values of attributes. It has made verification much easier for updaters.

LMT is used for the classification. As of June 2014, this algorithm has accuracy of 90.0%, recall of 92.4%, precision of 91.4%, and F-measure of 91.9%. During the most recent update, 50 out of 56 new entries were correctly classified. The accuracy of predicting new entries was 89.3%, which was very satisfactory.

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
10-fold cross-validation (s)	2.058	2.066	2.068	2.064
Classification (ms)	1.274	1.250	1.650	1.391

The following figures are snapshots of this website.

Figure 9 is the home page (retrieved on June 8, 2014), which shows the current statistics of the database. There are links to search page, browse page, download page, as well as information page.

In the search page, one can specify the target miRNA and diabetes to initiate a search. Figure 10 shows a returned set for associations between miR-29 and Type 2 Diabetes. As of June 2014, the miRDiabetes database contains papers on 263 different miRNAs.

miRDiabetes
Literature Collection Database on microRNA and diabetes association

Home Search miRDiabetes Browse miRDiabetes Downloads About Us

First Things First

Welcome. miRDiabetes is a literature collection database on the subject of microRNA-diabetes associations. This database was built in 2014 as a part of my thesis. Papers in this database were extracted from [PubMed](#).

— Hui Guo

Database stats

Updated on: 06/07/2014

Relevant papers: 381

All papers: 617

New entries: 97

Why Is This Important?

MicroRNAs (miRNAs) are a class of naturally occurring, small non-coding RNA molecules, about 22 nucleotides in length. They function via base-pairing with complementary sequences within messenger RNA (mRNA) molecules, down-regulating gene expression in a variety of manners, including translational repression, mRNA cleavage, and deadenylation.

Figure 9. miRDiabetes website homepage

miRNA name: e.g. miR-192, mir-375, etc. Click [HERE](#) for miRNA list.

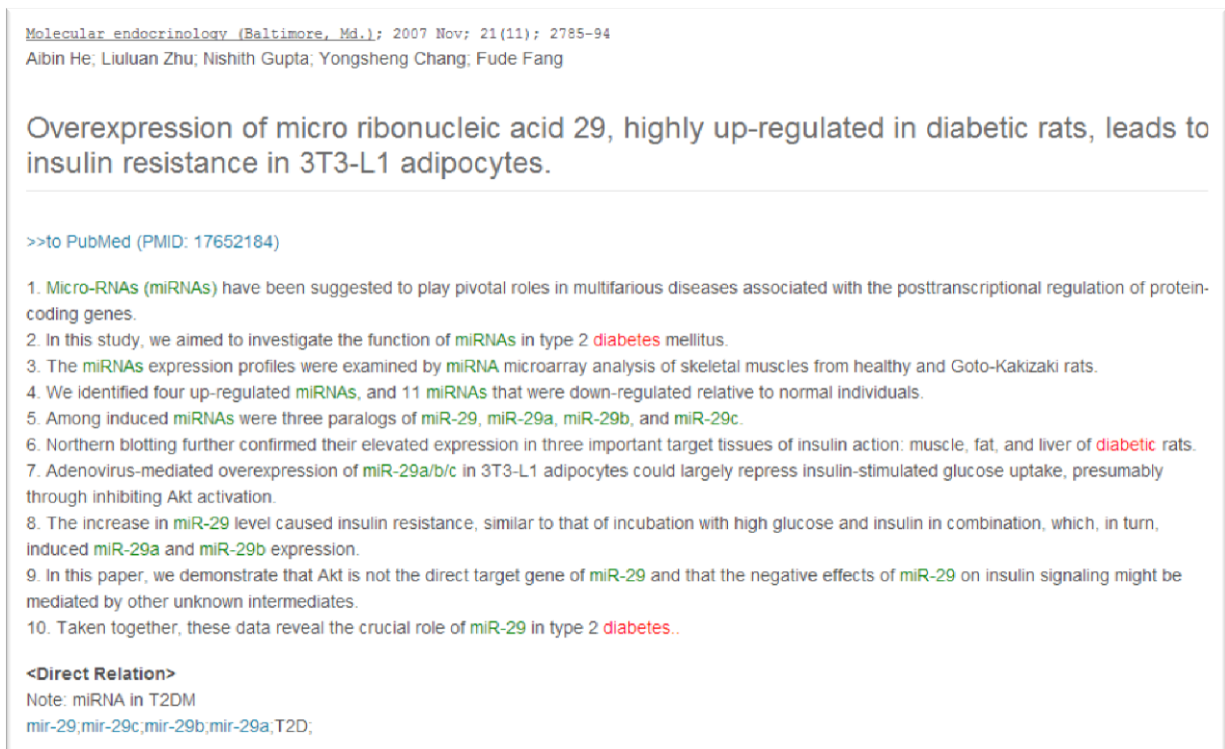
AND

5 entries are returned.

PMID	Relation	Year	Author	Title
17652184	Direct	2007	He <i>et al.</i>	Overexpression of micro ribonucleic acid 29, highly up-regulated in diabetic rats, leads to insulin resistance in 3T3-L1 adipocytes.
23834149	Direct	2013	Chakraborty <i>et al.</i>	miRNAs in insulin resistance and diabetes-associated pancreatic cancer: the 'minute and miracle' molecule moving as a monitor in the 'genomic galaxy'.
24039891	Direct	2013	Baran-Gale <i>et al.</i>	Beta cell 5'-shifted isomiRs are candidate regulatory hubs in type 2 diabetes.
24349318	Direct	2013	Peng <i>et al.</i>	Urinary miR-29 correlates with albuminuria and carotid intima-media thickness in type 2 diabetes patients.
24722248	Direct	2014	Kurtz <i>et al.</i>	microRNA-29 fine-tunes the expression of key FOXA2-activated lipid metabolism genes and is dysregulated in animal models of insulin resistance and diabetes.

Figure 10. Results for a search on search page

PMIDs in the search results contain links to corresponding detail pages. One example is shown in Figure 11. In the detail page, there is a link to the same paper on PubMed, and links to all mentioned miRNAs on miRBase. On this page, abstract of a paper is broken into sentences, and all miRNA and diabetes references are colored.



Molecular endocrinology (Baltimore, Md.); 2007 Nov; 21(11); 2785-94
Aibin He; Liuluan Zhu; Nishith Gupta; Yongsheng Chang; Fude Fang

Overexpression of micro ribonucleic acid 29, highly up-regulated in diabetic rats, leads to insulin resistance in 3T3-L1 adipocytes.

>>to PubMed (PMID: 17652184)

1. Micro-RNAs (miRNAs) have been suggested to play pivotal roles in multifarious diseases associated with the posttranscriptional regulation of protein-coding genes.
2. In this study, we aimed to investigate the function of miRNAs in type 2 diabetes mellitus.
3. The miRNAs expression profiles were examined by miRNA microarray analysis of skeletal muscles from healthy and Goto-Kakizaki rats.
4. We identified four up-regulated miRNAs, and 11 miRNAs that were down-regulated relative to normal individuals.
5. Among induced miRNAs were three paralogs of miR-29, miR-29a, miR-29b, and miR-29c.
6. Northern blotting further confirmed their elevated expression in three important target tissues of insulin action: muscle, fat, and liver of diabetic rats.
7. Adenovirus-mediated overexpression of miR-29a/b/c in 3T3-L1 adipocytes could largely repress insulin-stimulated glucose uptake, presumably through inhibiting Akt activation.
8. The increase in miR-29 level caused insulin resistance, similar to that of incubation with high glucose and insulin in combination, which, in turn, induced miR-29a and miR-29b expression.
9. In this paper, we demonstrate that Akt is not the direct target gene of miR-29 and that the negative effects of miR-29 on insulin signaling might be mediated by other unknown intermediates.
10. Taken together, these data reveal the crucial role of miR-29 in type 2 diabetes..

<Direct Relation>
Note: miRNA in T2DM
mir-29,mir-29c,mir-29b,mir-29a,T2D;

Figure 11. Detail page for an article

The browse page shows the distribution of papers in miRDiabetes, as shown in Figure 12. One can browse all papers in a certain category. For example, one can browse all papers published in the year of 2014 from this page.

Downloads page provides links to all files related to this project, for biomedical researchers and for data miners. The entire database is downloadable in textual format, and files for data mining software are available, such as ARFF files for Weka, as well as C4.5 files.

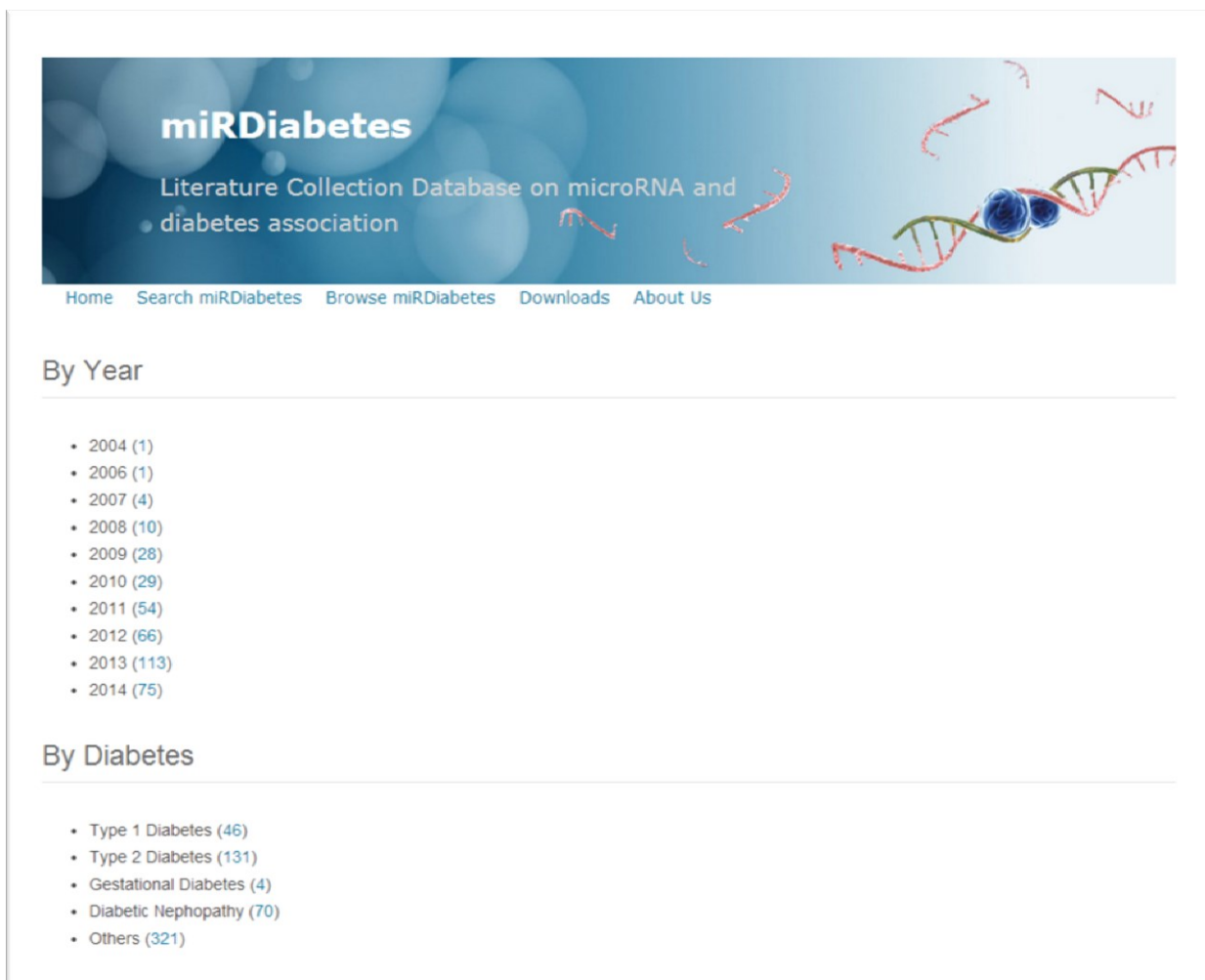


Figure 12. miRDiabetes website browse page

4.6 Performing an Update

Upon startup, the update application will find the date of the latest paper, as shown in Figure 13.

To make an update to the database, one should follow the steps below.

1) Press the “Update” button and the application will retrieve papers from PubMed since this date. Preliminary keyword checking and NER will also be performed. Only papers that mention both miRNA and diabetes will be saved into the database.

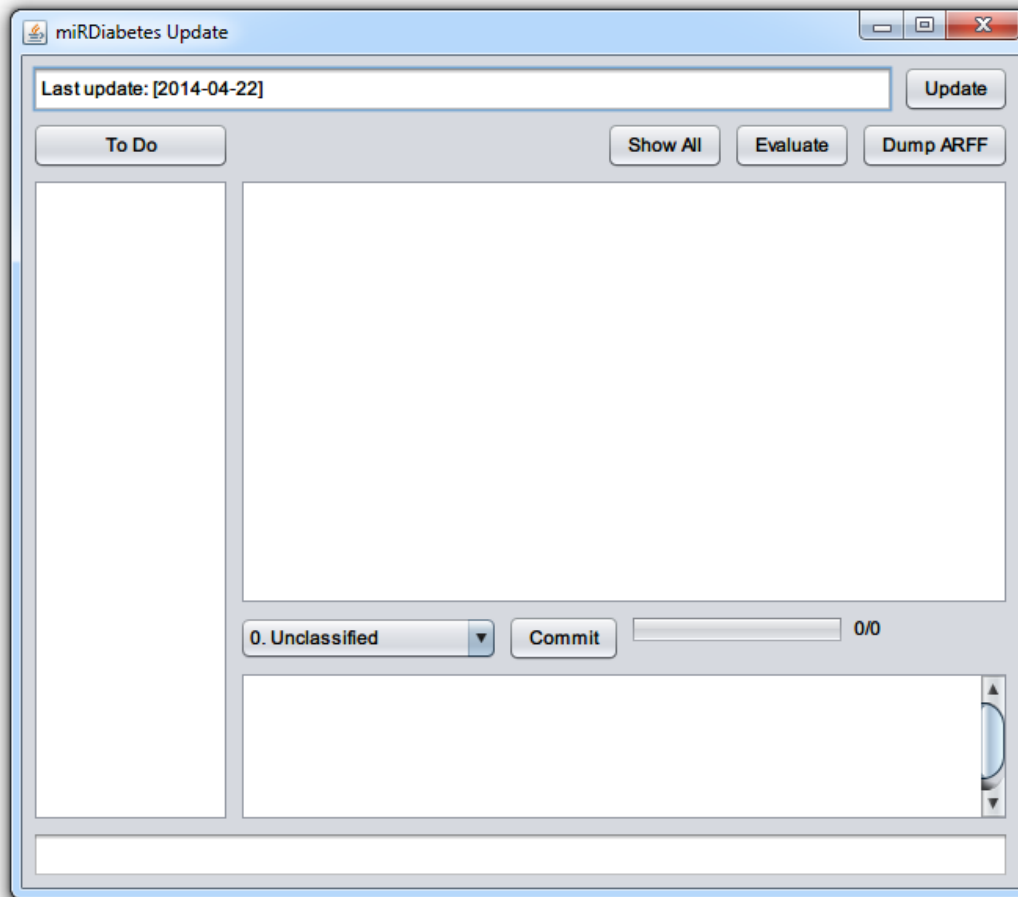


Figure 13. Snapshot of startup of the program

2) Press the “To do” button and the application will show all unclassified or unverified entries, as shown in Figure 14. If the classifier has not classified them, it will retrieve all classified item and classify new entries based on them.

3) Click the PMID from the left panel. The title and abstract of the corresponding paper will be presented on the right, along with values of its attributes, as shown in Figure 15. The classification result will be shown in the bottom right box.

4) Verify the classification results. Choose a class and press the “Commit” button. This result will be saved onto the database. A note can be added in the bottom right box and it will be saved as well.

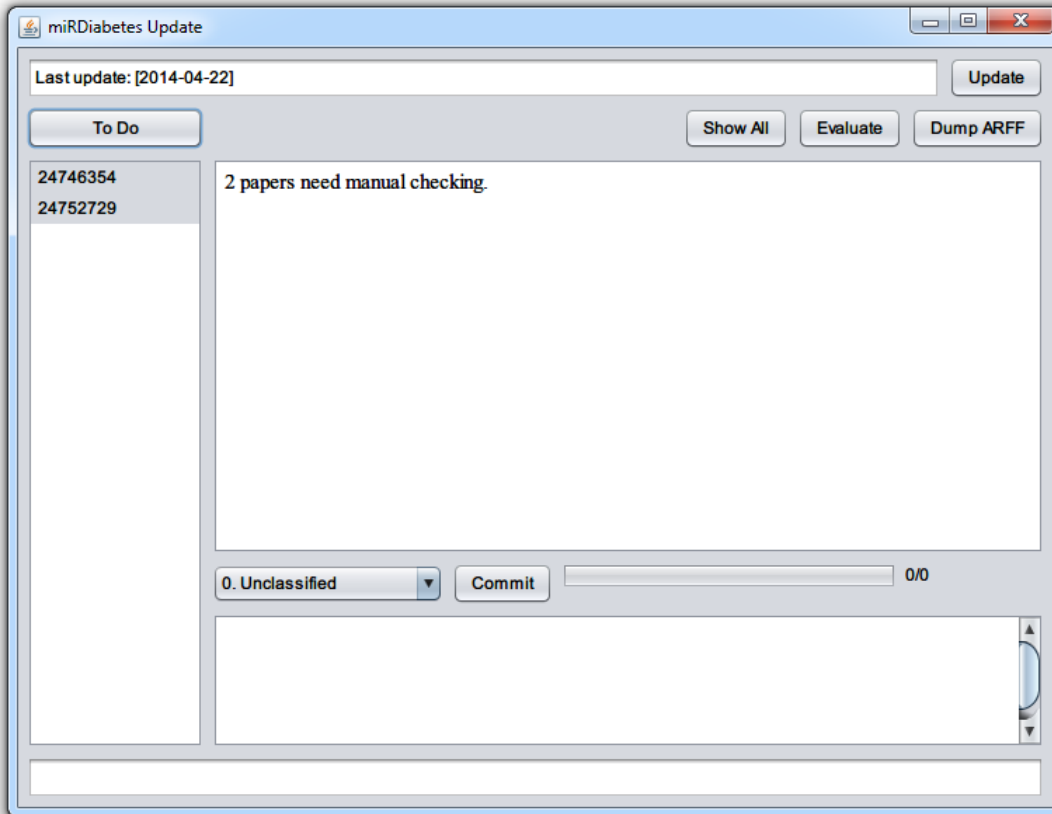


Figure 14. Snapshot of a to-do list

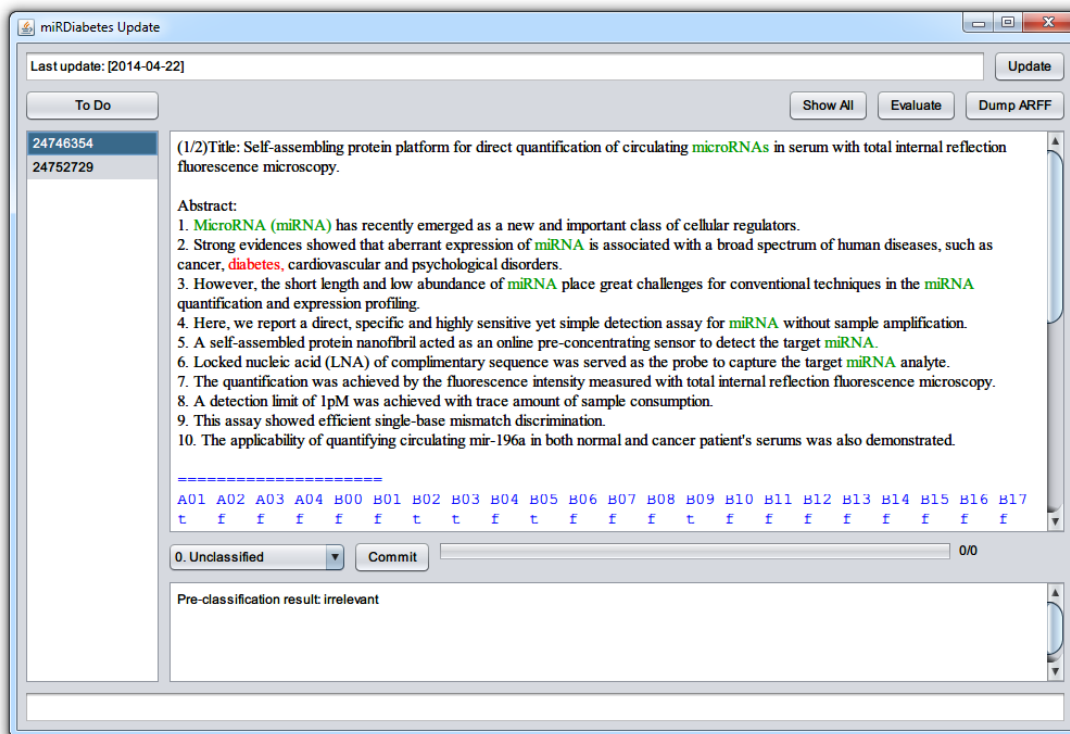


Figure 15. Snapshot of information of one article

“Evaluation” button can be used to check the accuracy of the classification algorithm, as shown in Figure 16. The “Dump ARFF” button can be used to generate an ARFF file for Weka.

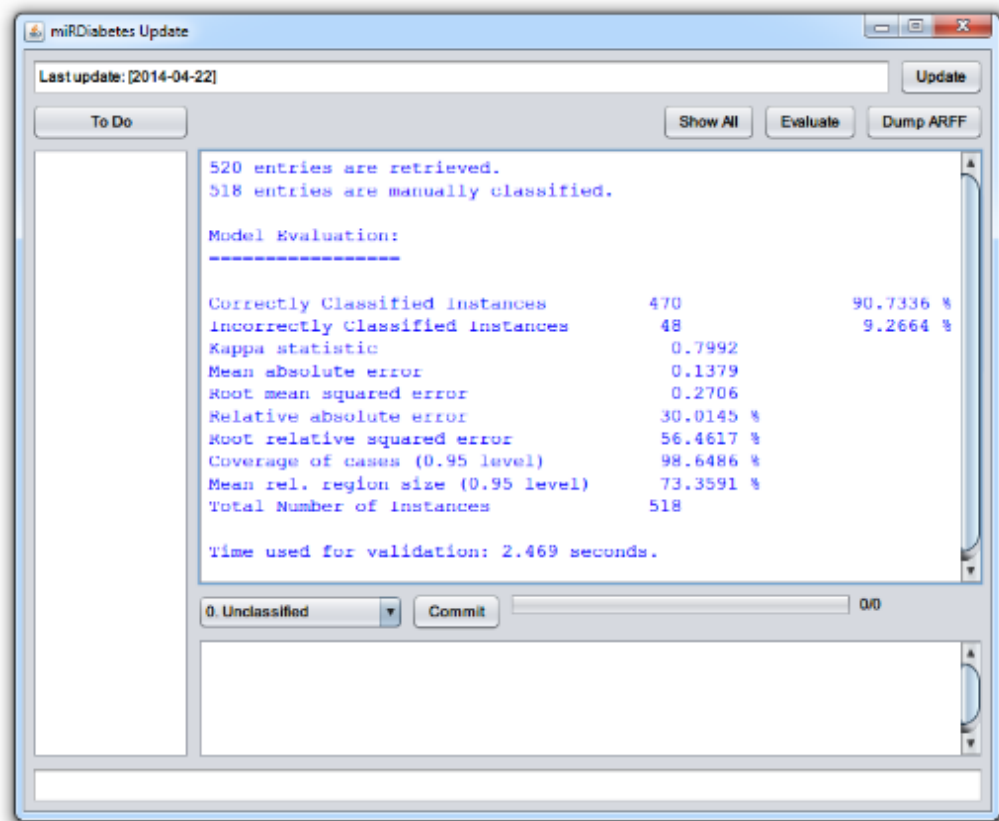


Figure 16. Snapshot of an evaluation result

CHAPTER 5: DISCUSSIONS

In this research, we have designed and experimented on a method of utilizing data mining techniques on biomedical texts for literature retrieval on a target subject. As a result, a well-classified database has been built, and an application has been implemented with this method, making the database easy to maintain. A number of classification algorithms have shown to achieve very good results. This confirms the capability of the attributes we selected. The classification used in the application also has high accuracy. This chapter will discuss possible improvements to our methodology and its implementation.

5.1 Calculation of Attributes

Most of the attributes are calculated based on keyword searching. Some attributes are designed to represent existences of relevant terms, while some attributes are designed to represent coexistences of terms on two subjects. The correctness of such attributes relies on the completeness of the dictionaries. This means that the dictionaries must be maintained in database and updated regularly.

Other attributes are based on the nature of sentences, such as enumeration, risk factor, or quotation. Currently we simply search the existences of possible indicative words or phrases, such as “risk” and “including”. To make the calculations more accurate, closer analysis of text should be implemented. This may involve natural language processing, and may be time-consuming and even counterproductive.

However, attributes to reflect special situations are very likely to be skewed. It must be carefully discussed whether these attributes should be removed in future versions, or how they should be handled.

5.2 Skewed Attributes

Not all attributes are used in the current classification process. Some have more deciding ability than others do. Different data mining algorithms may use different subsets of attributes for different datasets.

During the test in Chapter 4, J48 Graft pruned tree only used A01, A02, A03, A04, B03, B08, B11, B12, B13, and B16. J48 pruned tree only used A01, A02, A03, B08, B12, and B16. Yet the two algorithms had similar accuracy and F-measure. Attributes B00, B05, B10 were not used in these two trees, but they were used in LMT classification.

Attributes with skewed distributions are less likely to be used by a classification algorithm, especially after pruning. Take B06 as an example. B06 is true if diabetes is only mentioned as a risk factor. This kind of paper is more likely to be irrelevant. There are only 21 papers whose B06 is true, and for the rest 499 papers, B06 is false. Therefore, it was not used in the three algorithms mentioned above. However, of the 21 papers with true B06, 18 (85.7%) are irrelevant and only three are relevant. This attribute can be helpful in some cases. These attributes is better considered as asymmetric attributes.

There are altogether seven skewed attributes. Attributes B01, B04, B06, B08, B10, and B14 are designed for special cases. This means that they are false for a majority of articles. Attributes like this are not used in most of the classification techniques. We have kept these attributes because we believe that we can find ways to use them to improve accuracy of classification, after comprehensive investigation and experiments. For example, rule-based approach may put them into good use.

B02, on the other hand, is true for most of papers after preliminary keyword checking. It is a desired feature for relevant papers, but it does not have classification abilities. We can add additional irrelevant papers to make it less skewed. However, this increases workload, and there

is no guarantee that classification results will be better. It can be removed in future implementations, or it can be used as criteria of the preliminary screening.

5.3 Classification

As shown in previous sections, many classification techniques have high accuracy. During the construction of this database, LMT gave better results than others did. However, J48 and J48 graft had better accuracy and F-measure in the test in Chapter 4. It may be wise to use multiple classification algorithms to perform the classification, and use the combination of their results as the final one.

There are six classes and currently only binary classification is executed. These six classes are neither ordinal nor categorical. Class I and II indicate close relevance, while classes IV to VI indicate irrelevance, and class III is a class between them. There have been discussions about classifying class III as irrelevant. Classification algorithms also gave satisfying results for this method. However, to build a collection database, we need to increase recall as much as possible. This is because we have to keep the database comprehensive, and irrelevant articles can be removed during the verification process. This is the reason why class II is considered relevant in our classification process.

We can also perform classification over three classes. Currently we have a relatively small database, with only 13.3% of the papers in class three. It is unwise to further divide the dataset and cause unreliable results. This could be a possible improvement in the future, as our database grows much larger to support this classification.

Fuzzy classification can be a solution to articles with a vague relevancy. Fuzzy classification is the process of grouping elements into a fuzzy set whose membership function is defined by the truth-value of a fuzzy propositional function. Whether it is wise to invite fuzzy classification into our project needs further discussion and analysis.

5.4 Correctness of Verification

This research was conducted based on the assumption that a well-trained researcher could decide the relevancy of a biomedical paper correctly. Our developers that performed verification might lack biological and medical background, and interests of biomedical researchers might not be met. However, for papers in Class III, relevancy is vague, and a researcher from biomedical background may be biased as well, with his/her personal area of interest.

Nevertheless, it is wise to benefit from mutual discussion. We will consult biomedical researchers and collect feedbacks from our users to offer better classification results.

CHAPTER 6: CONCLUSIONS AND FUTURE WORK

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 basic classification techniques in data mining to determine relevancy of new texts with high accuracy. This solution has been implemented and practiced in this research, with satisfactory results. The result of classification has accuracy of 90.0% and F-measure of 91.9%. This classification alleviates the workload of future retrieval. This solution 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 literature retrieval using data mining techniques, which can retrieve and classify new publications automatically. Relevancy of a new entry can be determined based on the classified items in the database with high accuracy. We have also built a website for users to access miRDiabetes database.

Our developers have carefully verified all papers in miRDiabetes database to preserve the reliability of the database. With the help of the update application, updates to the database have been greatly accelerated, because it not only retrieves and classifies new entries automatically, but also breaks abstracts into sentences and highlights key words and attributes. The database is being updated regularly.

Future work of this thesis includes:

- 1) Keep updating the database constantly. We plan to update this database quarterly for the time being.

- 2) Find better ways to select and calculate the attributes. We will discuss and analyze all attributes and consider other possible ones.
- 3) Try to implement better classification techniques, or combine them, in the following versions. We will investigate and make better use of skewed attributes.
- 4) Consult experts on miRNA-diabetes association to offer classification that is more reliable, better verification, and comments that are more professional.

REFERENCES

- [1] M. Andrade and P. Borka, "Automated extraction of information in molecular biology," *FEBS Letters*, no. 476, pp. 12-17, 2000.
- [2] T. D. Schmittgen, E. J. Lee, J. Jiang, A. Sarkar, L. Yang, T. S. Elton and C. Chen, "Real-time PCR quantification of precursor and mature microRNA," *Methods*, vol. 44, no. 1, pp. 31-39, 2008.
- [3] R. C. Lee, R. L. Feinbaum and V. Ambros, "The *C.elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*," *Cell*, vol. 75, no. 5, pp. 843-854, 1993.
- [4] Y. Gusev and D. J. Brackett, "MicroRNA expression profiling in cancer from a bioinformatics perspective," *Expert Review of Molecular Diagnostics*, vol. 7, no. 6, pp. 787-792, 2007.
- [5] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman and A. J. Enright, "miRBase: microRNA sequences, targets and gene nomenclature," *Nucleic Acids Research*, no. 34, pp. D140-D144, 2006.
- [6] WHO Media Center, "Diabetes fact sheet N°312," 10 2013. [Online]. Available: <http://www.who.int/mediacentre/factsheets/fs312/en/>. [Accessed 20 5 2014].
- [7] Centers for Disease Control and Prevention, "2011 National Diabetes Fact Sheet," 2011. [Online]. Available: <http://www.cdc.gov/DIABETES/pubs/factsheet11.htm>. [Accessed 20 5 2014].
- [8] 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.
- [9] US National Library of Medicine, "PubMed," [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/>. [Accessed 20 5 2014].
- [10] H. Sahu, S. Shma and S. Gondhalakar, "A brief overview of data mining survey," *International Journal of Computer Technology and Electronics Engineering*, vol. 1, no. 3, pp. 114-121, 2011.

- [11] V. Rao, "An introduction to text analytics," 2012. [Online]. Available: <http://www.datasciencecentral.com/profiles/blogs/an-introduction-to-text-analytics>. [Accessed 20 5 2014].
- [12] 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.
- [13] 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.
- [14] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao and T. Li, "miRecords: an integrated resource for microRNA-target interactions," *Nucleic Acids Research*, vol. 37, pp. D105-D110, 2009.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] US National Library of Medicine, "PubMed: MEDLINE retrieval on the world wide web," 2002. [Online]. Available: <http://www.nlm.nih.gov/pubs/factsheets/pubmed.html>. [Accessed 20 5 2014].
- [19] B. Feng, Y. Cao, S. Chen, M. Ruiz and S. Chakrabarti, "miRNA-1 regulates endothelin-1 in diabetes," *Life Science*, vol. 98, no. 1, pp. 18-23, 2014.
- [20] J. H. Eom and B. T. Zhang, "PubMiner: Machine learning-based text mining system for biomedical information mining," *Artificial Intelligence: Methodology, Systems, and Applications Lecture Notes in Computer Science*, vol. 3192, pp. 216-225, 2004.
- [21] M. Sturm, M. Hackenberg, D. Langenberger and D. Frishman, "TargetSpy: a supervised machine learning approach for microRNA target prediction," *BMC Bioinformatics*, vol. 11, p. 292, 2010.

- [22] US National Library of Medicine, "Entrez programming utilities help," 2010. [Online]. Available: <http://www.ncbi.nlm.nih.gov/books/NBK25501/>. [Accessed 20 5 2014].
- [23] 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.
- [24] J. Quinlan, *C4.5: programs for machine learning*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.
- [25] N. Landwehr, M. Hall and E. Frank, "Logistic Model Trees," *Machine Learning*, vol. 59, pp. 161-205, 2005.
- [26] M. Summer, E. Frank and M. Hall, "Speeding up logistic model tree induction," *Lecture Notes in Computer Science*, vol. 3721, pp. 675-683, 2005.
- [27] J. Schneider, "Cross Validation," [Online]. Available: <http://www.cs.cmu.edu/~schneide/tut5/node42.html>. [Accessed 20 5 2014].
- [28] 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.

