Bioinformation Volume 5 *open access*

www.bioinformation.net Issue 6 **Software**

Mi-DISCOVERER: A bioinformatics tool for the detection of mi-RNA in human genome

Saadia Arshad, Asia Mumtaz, Freed Ahmad, Sadia Liaquat, Shahid Nadeem, Shahid Mehboob, Muhammad Afzal*

Department of Bioinformatics, Government College University, Faisalabad, Pakistan; Muhammad Afzal- Email: afzalarsenal@yahoo.com; *Corresponding author.

Received July 26, 2010; Accepted August 26, 2010; Published November 27, 2010

Abstract:
MicroRNAs (miRNAs) are 22 nucleotides non-coding RNAs that play pivotal regulatory roles in diverse organisms including the humans and are difficult to be identified due to lack of either sequence features or robust algorithms to efficiently identify. Therefore, we made a tool that is Mi-Discoverer for the detection of miRNAs in human genome. The tools used for the development of software are Microsoft Office Access 2003, the JDK version 1.6.0, BioJava version 1.0, and the NetBeans IDE version 6.0. All already made miRNAs softwares were web based; so the advantage of our project was to make a desktop facility to the user for sequence alignment search with already identified miRNAs of human genome present in the database. The user can also insert and update the newly discovered human miRNA in the database. Mi-Discoverer, a bioinformatics tool successfully identifies human miRNAs based on multiple sequence alignment searches. It's a non redundant database containing a large collection of publicly available human miRNAs.

Keywords: miRNA prediction, Human miRNA detection, miRNA detection tool and database.

Background:
MicroRNAs (miRNAs) have lately gained much interest, as recent genome-wide studies have shown that they are widespread in a variety of organisms and are conserved in evolution. Hundreds of miRNAs have been identified by direct cloning and computational approaches in several species. However, there are still many miRNAs that remain to be identified because of their small size and sequence specificity make the detection of completely new miRNAs a challenging task. This cannot be based on sequence information alone, but requires structure information about the miRNA precursor **[1]**. Bioinformatics approaches have proved to be very useful toward this goal by guiding the experimental investigations **[13]**. Nearly 97% of the human genome is composed of noncoding DNA. Numerous genes in these non-protein-coding regions encode microRNAs, which are responsible for RNA-mediated gene silencing through RNA interference (RNAi)-like pathways. The miRNAs are initially expressed as part of an imperfect RNA hairpin of ~70 nucleotides in length that in turn forms part of a longer initial transcript termed a primary miRNA (pri-miRNA) **[9]**.The majority of long primary transcripts of the miRNA genes are transcribed by RNA polymerase II **[9,2]**. The 7-methylguanosine capped and poly (A) tailed transcripts are cleaved by the nuclear RNase III Drosha to release the precursors of miRNA (pre-miRNA) in the nucleus **[9]**. The precursors of miRNA that possess a thermodynamic stabile hairpin structure are exported into the cytoplasm by Exportin-5 or HASTY **[15]**, **(Figure 1)**. Once there, the pre-miRNA is processed by a second RNAse III family member called `Dicer' to give the mature ~22 nucleotides miRNA **[4,6,7]**. Mature miRNAs are generated by the RNase III-type enzyme 60 Dicer, producing a small double-stranded RNA from which one strand (called miRNA) is quickly degraded, releasing the small single-stranded miRNA **[3]**, **(Figure 1)**.

The miRNA is then incorporated into the RNA-induced silencing complex (RISC) **[5,11,14]** and guided to target sequences located at the 3'-terminal untranslated regions (3'-UTRs) of mRNAs by base pairing, resulting in the cleavage of target mRNAs or repression of their productive translation **[12]** as in **Figure 1**: The current release, version 11.0, contains over 6396 miRNAs from various organisms including 678 human and 472 mouse miRNAs. However, the function of each miRNA is mostly unknown except a few. All available miRNAs tools were web based; therefore aim of project was to provide the desktop access to scientist so that they can easily make alignment of their sequences with already discovered miRNAs without using internet. The desktop facility provides access to scientist to perform better task in less time. User can easily retrieve, update and insert new record about miRNA in database. The aim and objectives in developing Mi-Discoverer were: To discover microRNAs from the human genome; To provide desktop software tool facility to young scientist; To reduce cost to maintain the software; Software in which there is facility of finding out new miRNAs and saving, editing and searching the record of discovered miRNA. The new software should have user friendly environment that is every task should be perform on single click environment.

Materials and Methods:
Tools that are used for software development include MS-Access Version 2002(www.microsoft.com), (http://java.sun.com), JDK1.6.0 BioJava 1.0 (http://biojava.org) and NetBeans IDE 6.0 (http:// www.netbeans.org/ kb/60/java/gui-db.html). The software developed for the detection of miRNA in humans was based on the similarity searches of user sequence with already discovered miRNAs. The development of software involved following steps:

1. Data Mining:
Data about miRNA was collected from following mentioned online databases. miRNA data includes names of miRNA in human genome, their accession numbers, precursor sequences, precursor length, mature

sequences, mature length and different functions perform by different miRNAs. The following bioinformatics tools and databases were used to search the data related to miRNA human genome: miRanda, miRDB, MiRNAMap, MicroInspector, MiRGen, MicroRNAdb, Argonaute 2, MiRNAminer, PicTar, miRBase and Reference set (The miRNAs and their precursor sequences were downloaded from the MicroRNA Registry. This set contains 678 miRNAs and their precursors from Homo sapiens).

2. Use case:
A use case diagram is a type of behavioural diagram defined by the Unified Modeling Language (UML) and created from a Use-case analysis. Use case diagram is used here to show what system functions are performed for a user. Roles of the user in the system can be depicted. Here use case diagram of Mi-Discoverer is presented in **Figure 2**.

3. Database Design:
Mi-Discoverer database has a centralize data dictionary for the storage of information pertaining to data and its manipulation. The data of 678 human mature as well as precursor miRNAs (collected from Reference set) with information about their unique accession ID, chromosomal location, mature sequence, precursor sequence and function was added. The database represents complex relationship among MiRNA data. Our database keeps a tight control of data redundancy as primary key is defined in all the tables so there is no chance of data redundancy. The database is designed in Microsoft Access Version 2002.

4. ER Model:
An entity-relationship model (ERM) is an abstract conceptual representation of structured data. Entity-relationship modelling is a relational schema database modelling method, used in software engineering to produce a type of conceptual data model. ER model for Mi-Discoverer is represented in **Figure 3**.

5. Interface Design:
Mi-Discoverer has a cool interface design. It is also very important for an application as interface is used for the user interaction with the application. Interface is concerned with the layout of the screen which has to be presented to the user for various data entries and view purpose. The interface was designed in NetBeans IDE 6.0. The interface has the following features.
Same layout was used for the entire screen and was designed to avoid mistakes.
The screen suited the purpose for which it is design.
Buttons having meaningful labels was used so that an operator can easily realize what he is going to do.

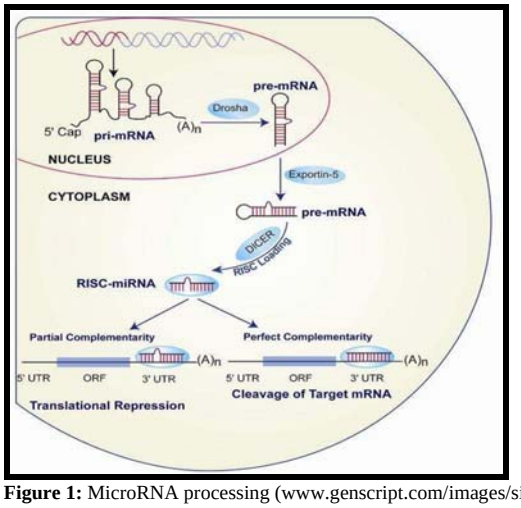

Figure 1: MicroRNA processing (www.genscript.com/images/sirna/miRNA.jpg)

Once the error has been detected then the user interface has to display the message for ease of user. These messages explain what is wrong and what the software is expecting.
Different forms are used to input data. There is a separate form for each table to insert, delete and update the record.

6. Sequence Diagram:
A sequence diagram is a kind of interaction diagram in UML, which shows how processes operate with one another and in what order. It is a construct of Message Sequence Chart. Sequence Diagram of Mi-Discoverer is shown in **Figure 4**.

7. Testing the Software:
Testing was performed for Mi-Discoverer software. Basic objective of testing was to execute a program with intend to finding error. Sample data was entered to check whether system works properly or not. To test the performance of the software, Mi-Discoverer was used to predict pre-miRNAs from human genome using multiple sequence alignments. Among other things, Mi-Discoverer proposes a new criterion to select miRNAs from already entered miRNAs in the database that is based on the number of matching around their genome location. The program was tested on a PC with an Intel Pentium IV processor 2.8 GHz and 512MB RAM memory. The operation system was Windows XP 2003.The versions used were 1.6 JDK and 1.0 BioJAVA. The results and all additional information are saved in MS-Access database for each session. There are various techniques to test the software, here only two black box and unit testing is used.

8. Black Box Testing:
In this testing user interface is exercised over a full range of inputs and corresponding outputs observed for correctness. To perform the Black Box Test, open the mainframe form. Then open form to Search miRNA. Input the sequence and see output. Examine the output sequence in another form to view information about resulting miRNA. Save the results in save forms. MI-Discoverer qualified the Black Box Testing.

9. Unit Testing:
In this testing different modules of developed system are tested so that each module functions properly. The unit test of each individual unit was done. For Example, open data entry form and run it. Input the required data and save it. Its task to assure that a user can enter data and when user presses save button record must be save in database. Mi-Discoverer qualified the Unit Testing.

Bioinformation **Volume 5** *open access*

www.bioinformation.net Issue 6 **Software**

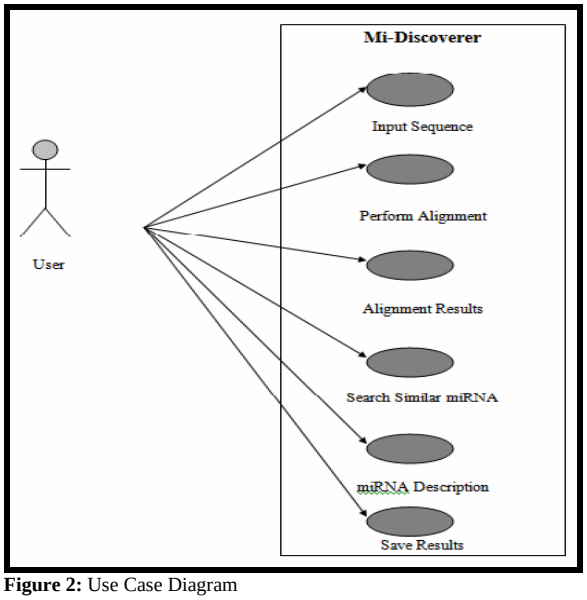

Figure 2: Use Case Diagram

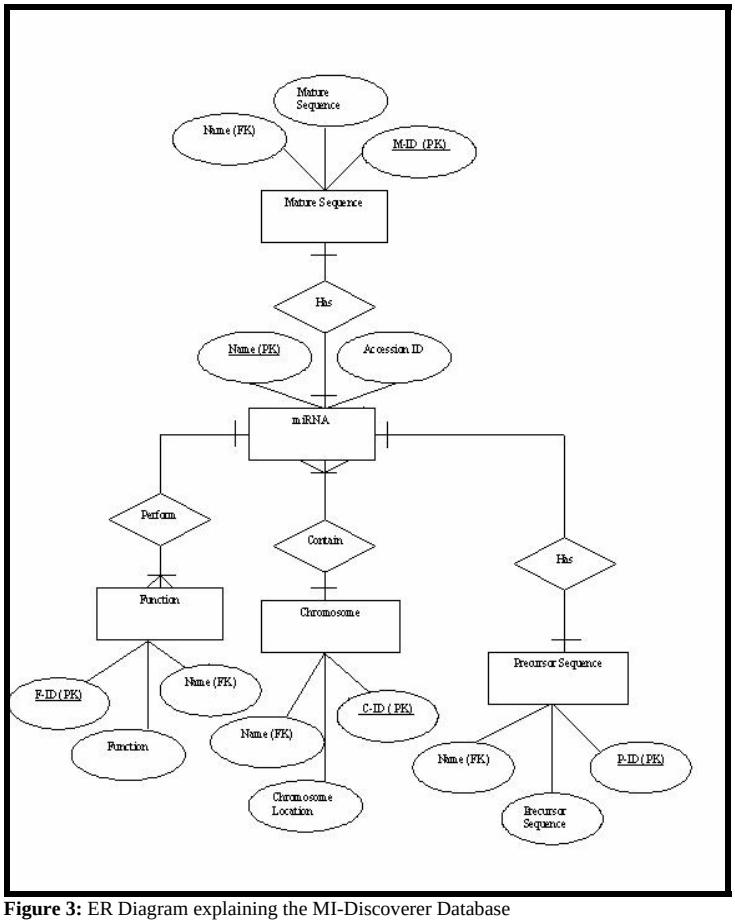

Figure 3: ER Diagram explaining the MI-Discoverer Database

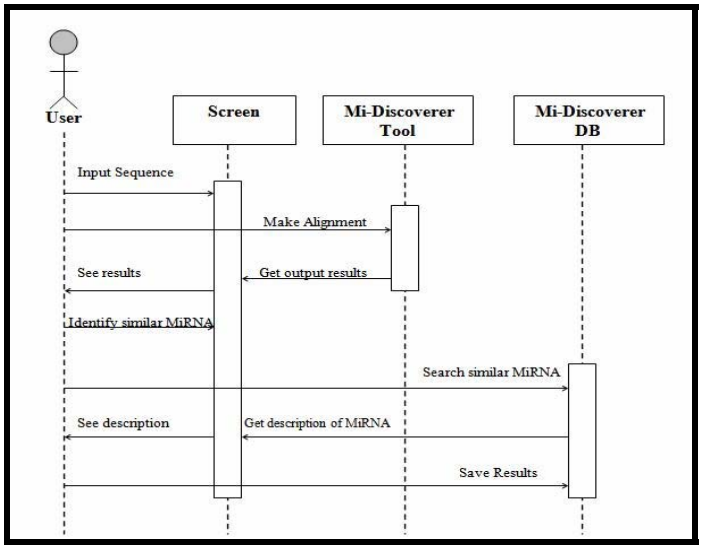

Figure 4: Sequence Diagram

The input RNA sequence:

caacuacagccacuacuacaggaccaucgaggaccugcgggacaagauucuuggugccaccauugagaacgccaggauuguccugcagau

caacaaugcucaacuggcugcagau

Resulting output

Query: caacuacagccacuacuacaggaccaucgaggaccugcgggacaagauucuuggugccaccauugagaacgccaggauuguccugcagaucaacaaugcucaacuggcugcagau

Subject1: ----------aggaccugcgggacaagauucuu----------

GUIDE TREE :((Subject1: 0.03409,Query:-0.03409)

Query: caacuacagccacuacuacaggaccaucgaggaccugcgggacaagauucuuggugccaccauugagaacgccaggauuguccugcagaucaacaaugcucaacuggcugcagau

Subject2: ----------auuccuagaaauuguucaua----------

GUIDE TREE:(Query: 0.35000,Subject2: 0.35000);

Query: caacuacagccacuacuacaggaccaucgaggaccugcgggacaagauucuuggugccaccauugagaacgccaggauuguccugcagaucaacaaugcucaacuggcugcagau

Subject3: ----------gaaguuguucguggguaauucg----------

GUIDE TREE:(Query :0.31818,Subject3 : 0.31818);

Figure 5: Result of output for the sequence based on the distance matrix.

Name	Accession_ID	Chromosome_Location	Mature_Sequence	Precursor_Sequence	Function
hsa-miR-492	MI0003131	12	AGGACCUGCGGGACAAGAUUC...	CAACUACAGCCACUACUACAG...	----------

Figure 6: Result of sequence information.

Results:
In the present study, Mi-Discoverer, a bioinformatics strategy was developed that relies on a multiple sequence alignment to predict human miRNAs and successfully applied the program to identify miRNAs. Java language was used to write the program. This program aligned an input human gene sequence to a large collection of publicly available human MiRNAs. Database Mi-Discoverer was created that comprised on almost all miRNAs sequences which have been discovered yet from human genome. It contains 678 miRNA for humans.

When a human RNA sequence was submitted to find the miRNA the CLUSTLW program gives the output in two files, one is .aln and other is .dnd file. The .aln file is the alignment and the .dnd file is a guide tree. A pairwise score was calculated for every pair of sequences that are to be aligned. These scores are presented in a table in the results. Pairwise scores were calculated as the number of identities in the best alignment divided by the number of residues compared (gap positions are excluded). Both of these scores were initially calculated as percent identity scores and are converted to distances by dividing by 100 and subtracting from 1.0 to give number of differences per site. A guide tree was calculated based on the distance matrix that is generated from the pair wise scores **(Figure 5)**. It was found that the query sequence was best aligned with subject1 sequence as its guide tree score is lowest. Now we can find out the details of this sequence by the mature sequence table. The mature sequence table that gives the resulting sequence information is presented here **(Figure 6)**.

Discussion:
Experimental cloning efforts have successfully identified highly expressed miRNAs from various tissues. However, cloning methods are highly biased towards miRNAs that are abundantly and/or ubiquitously expressed. On the other hand, computational prediction of miRNAs could become a robust approach for tissue-specific or lowly expressed miRNAs. Several computational methods have been developed to find close homologs among related miRNAs **[17, 8]**. Both their small size and sequence specificity make the detection of completely new miRNAs a challenging task. This cannot be based on sequence information alone, but requires structure information about the miRNA precursor **[1]**. Mi-Discoverer was accurately detecting miRNAs in human genome as it could be a better alternative of other miRNAs discovery softwares and its main features were:
It is universally useful software for prediction of new pre-miRNAs in unique specie that is human genome.
It automatically process large amounts of genomic sequences.
It finds as many as possible miRNA from pre miRNA candidates
It can easily used by people who have little knowledge about the computational theory and processes used behind the scenes.

Two programs have already been reported in literature for miRNA detection. The MiRscan program successfully predicted close homologs of *Caenorhabditis briggsae* with statistically conserved patterns of *Caenorhabditis elegans* miRNAs (Lim et al., 2003) while the program mirAlign was based on animal data from 12 species, *C. elegans*, *Caenorhabditis briggsae*, *Drosophila melanogaster*, *Drosophila pseudoobscura*, *Danio rerio*, *Gallus gallus*, *Homosapiens*, *Mus musculus*, *Rattus norvegicus*, *Epstein Barr virus*, *Arabidopsis thaliana* and *Oryza sativa* **[16]**. But Mi-Discoverer is based on data from human genome only. Both of the software miRAlign and miRscan are online while Mi-Discoverer is a desktop application that facilitates the user to make identification of miRNA from human genome even without using internet. It allows sequence based search that used to find potential miRNAs in the query sequence based on established CLUSTALW alignment algorithms. The accurate prediction of a comprehensive set of messenger RNAs (targets) regulated by animal microRNAs (miRNAs) remains an open problem. In particular, the prediction of targets that did not possess evolutionarily conserved complementarity to their miRNA regulators were not adequately addressed by available tools.

We developed a new computational approach, called Mi-Discoverer that can accurately detect miRNAs in human genome. Compared to similar methods, our method had been a better performance as its main advantage is that our application is purely desktop. Mi-Discoverer does not relay on different genomes and is able to address the unique human genome. A down side might be that the species specific i.e. Homo sapiens could be predicted since these miRNAs would be left out in the sequence alignment step before starting the prediction. The test results showed that this program met the goals relatively well.
Mi-Discoverer has also a comprehensive database for human genome miRNAs, including all available latest information about pre-miRNA sequence and length of the stem loop region and function. The major diagnostic features of present software include, that the software is so easy and convenient to user and it has easy graphical user interface. This software is so fast that it works in no time. All Modules of Mi-Discoverer works properly and gives effective output. It is so easy that anybody can operate it. It also checks input and avoids duplication of data and gives steady information to user. Overall the software is error free. In short, Mi-Discoverer achieves better sensitivity than previously reported softwares.

References:
[1] M Brameier & C Wiuf. *BMC Bioinformatics* **8**: 478 (2007) [PMID: 18088431]
[2] X Cai *et al. RNA* **10**: 1957 (2004) [PMID: 15525708]
[3] BR Cullen. *Molecular Cell* **16**: 861 (2004) [PMID: 15610730]
[4] A Grishok *et al. Cell* **106**(1): 23 (2001) [PMID: 11461699]
[5] SM Hammond *et al. Nature* **404**: 293 (2000) [PMID: 10749213]
[6] G Hutvagner *et al. Science* **293**(5531): 834 (2001) [PMID: 11452083]
[7] RF Ketting *et al. Genes Development* **15**: 2654 (2001) [PMID: 11641272]
[8] EC Lai *et al. Genome Biology* **4**(7) (2003) [PMID: 12844358]
[9] Y Lee *et al. Nature* **425** (6956): 415 (2003) [PMID: 14508493]
[10] SL Lin *et al. Journal of Biomedicine and Biotechnology* **2006**: 26818 (2006) [PMID: 17057362]
[11] J Martinez *et al. Cell* **110**(5): 563 (2002) [PMID: 12230974]
[12] G Meister & T Tuschl. *Nature* **431**(7006): 343 (2004) [PMID: 15372041]
[13] A Sewer *et al. BMC Bioinformatics* **6**: 267 (2005) [PMID: 16274478]
[14] DS Schwarz *et al. Molecular Cell* **10**(3): 537 (2002) [PMID: 12408822]
[15] R Yi *et al. Genes Development* **17**: 3011 (2003) [PMID: 14681208]
[16] X Wang *et al. Bioinformatics* **21**(18): 3610 (2005) [PMID: 15994192]
[17] MJ Weber. *Journal of FEBS* **272**(1): 59 (2005) [PMID: 15634332]
[18] www.genscript.com/images/sirn miRNA.jpg

Edited by P Kangueane
Citation: **Afzal *et al.*** Bioinformation 5(6): 271-276 (2010)

Supplementary material:

Table 1: Name and accession ID of miRNAs

Field name	Datatype	Size	Attributes
Name	Text	50	PK
Accession ID	Text	50	

Table 2: Chromosome Location of miRNA in human genome.

Field name	Data type	Size	Attributes
C_ID	Auto number	Long integer	PK
Name	Text	50	FK
Chromosome Location	Memo		

Table 3: Mature sequence of miRNAs.

Field name	Data type	Size	Attributes
M_ID	Auto number	Long integer	PK
Name	Text	50	FK
Mature sequence	Memo		

Table 4: Precursor sequence of miRNAs.

Field name	Data type	Size	Attributes
P_ID	Auto number	Long integer	PK
Name	Text	50	FK
Precursor sequence	Memo		

Table 5: Function of miRNAs.

Field name	Data type	Size	Attributes
F_ID	Auto number	Long integer	PK
Name	Text	50	FK
Function	Memo		