Exploiting PERL for Micro-RNA Target Identification: A Potential Drug Site for COVID-19

Subhojyoti Chatterjee[1]| Jagriti Chatterjee[2]

[1]School of Biotechnology, Amrita Vishwa Vidhyapeetham, Kollam, Kerala, India.
[2]Department of Bioinformatics, Karunya Institute of Technology and Sciences, Coimbatore, Tamilnadu, India

To Cite this Article

ABSTRACT

The functioning of gene expression or ribonucleic acid (RNA) silencing is governed by microRNA also known as miRNA. It is a small non-coding RNA molecule which finds its existence in animals, viruses, or plants. The large part of miRNAs is found to be transcribed from DNA sequences to form primary miRNAs followed by the processing of the precursor miRNAs and mature miRNAs. The miRNAs interconnect with their target genes in an effective manner which is dependent on factors like sub-cellular location of miRNAs, the availability of miRNAs and target miRNAs along with their interaction affinities, as seen in case of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Together with its involvement in normal functioning of eukaryotic cells, microRNA deregulation is associated with cancer, i.e., chronic lymphocytic leukemia. Therefore, microRNA target identification becomes important to unwind the relationship between microRNA deregulation and human diseases, thereby paving a path for structure-baseddrug discovery. Towards that direction, we have attempted to use the platform of PERL programming (a user-friendly and dynamic language to easily process and manipulate long sequences), to detect the microRNA target sites in genomic sequences, thereby trying to suppress the expression level for prognosis.

KEYWORDS: *MicroRNA target identification, Gene expression, Drug discovery, Cancer, Coronavirus, Dynamic PERL programming*

I. INTRODUCTION

MicroRNAs (miRNA) are single-stranded RNA molecules of about 21-23 nucleotides in length (indicated in*Fig. 1*), which regulates gene expression (Ling et al., 2013). A huge part of miRNA is highly conserved across species. miRNAs were first discovered by Ambros and Ruvkun groups in 1993 when it was first identified in *Caenorhabditis elegans*(C.elegans) (Lee et al., 1993). This organism was further studied to observe the effect of developmental timing by heterochronic mutants. It

was found that*lin-4*(one of the genes in C.elegans) did not encode for a protein, although it possess a small part of homology to multiple motifs in the 3'-untranslated region (3'-UTR) of *lin-14*. This *lin-14* is a heterochronic gene whichalso does not encode for a protein(Lee et al., 1993).

miRNAs are encoded by genes that are transcribed from DNA but not translated into protein (non-coding RNA) (Lee et al., 1993); instead they are processed from primary transcripts known as *pri-mi*RNA to short stem-loop structures called

*pre-mi*RNA with the help of RNA-III type protein (MacFarlane et al., 2010) and finally to functional miRNA (O'Brien et al., 2018), as represented in *Fig. 1*. Mature miRNA molecules are partially complementary to one or more messenger RNA (mRNA) molecules (Wahid et al., 2010), and their main function is to down regulate gene expression (Bartel, 2009). There is no such strong ground that says RNA viruses is involved in the genesis of miRNAs that are responsible for targeting their own genes or modify the host gene interaction (Bartel, 2004). However, we can see instances where mi-RNA can play a crucial role in anti-viral defense mechanism (refer *Fig. 1*), in case of **Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2),** as can be seen in *Fig. 1 and 2*, respectively(Schmidt, 2014).
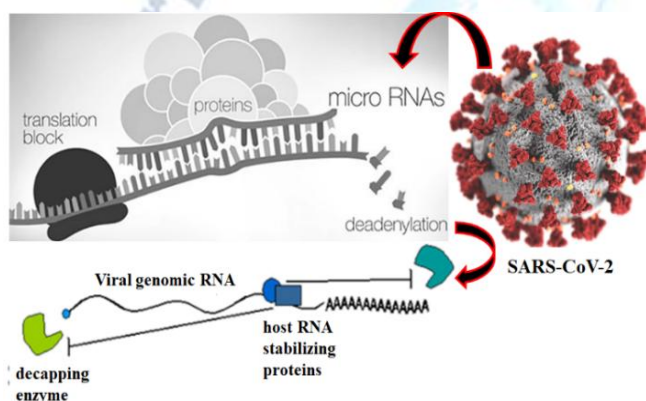
Fig. 1: Representation of MicroRNA expression along with coronavirus disguise (SARS-CoV-2) for mRNA.

Computational research of miRNA, hence, has been extensively occupied in the field of drug discovery, especially as biomarkers in clinical medicine (Hanna et al.,2019). A single miRNA as a biomarker, is capable enough for identification of a health outcome. Studies show that patients with frontotemporal dementia expressed notable miRNA patterns between patients irrespective of the fact that they have mutations in gene for progranulin or not. There are many diseases like diabetes, breast cancer, coronary heart disease, for which miRNA has been selected as biomarker for noting the disease progression in patients, receiving drugs approved by Food and Drug Administration (FDA)(Abdel-Ghany el al., 2020).

This idea has been implemented in the international aim of curing the pandemic, i.e.,
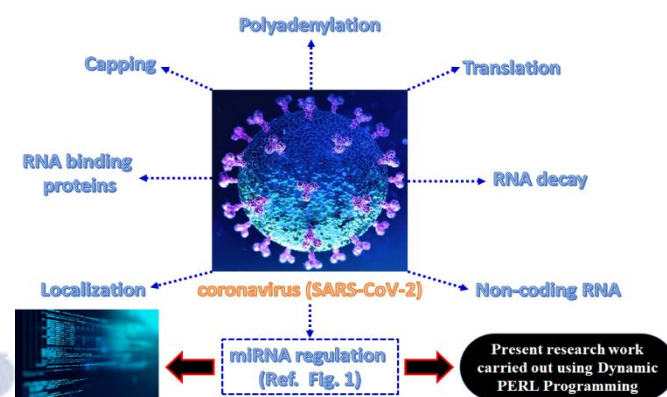
Fig. 2: The interface between coronavirus and miRNA solved using dynamic PERL programming.

SARS-CoV-2/COVID-19. Precisely, miRNA targets are in the process of being explored that is specific to SARS-CoV-2 RNA genome (as shown in *Fig. 2*) which oughts' to show its defense mechanism against the virus (Cavanagh, 2007). This paper gives one such perspective of the usage of miRNAs by topping it with PERL, a user-friendly computer language that makes easy analysis of high throughput data. This research aims at detecting the potential microRNA target sites in genomic sequences of Caenorhabditis elegans (as an example) and align the sequence based on complementarity.

The PERL program (see *Fig. 3*) reads micro RNA sequences from the user and genomic mRNA sequences from the flat files database of Caenorhabditis elegans (C.elegans) (Stein et al., 2001).Following which the dynamic PERL program runs local alignment, that involves one miRNA sequence (query sequence) been scanned against all sequences (reference sequences) in file2 to identify and report potential drug target sites. Nevertheless, it should be noted that the alignment scores are based on sequence complementarity and not on sequence identity.

Fig. 3: Pictorial depiction of microRNA target sites identified using dynamic PERL programming.
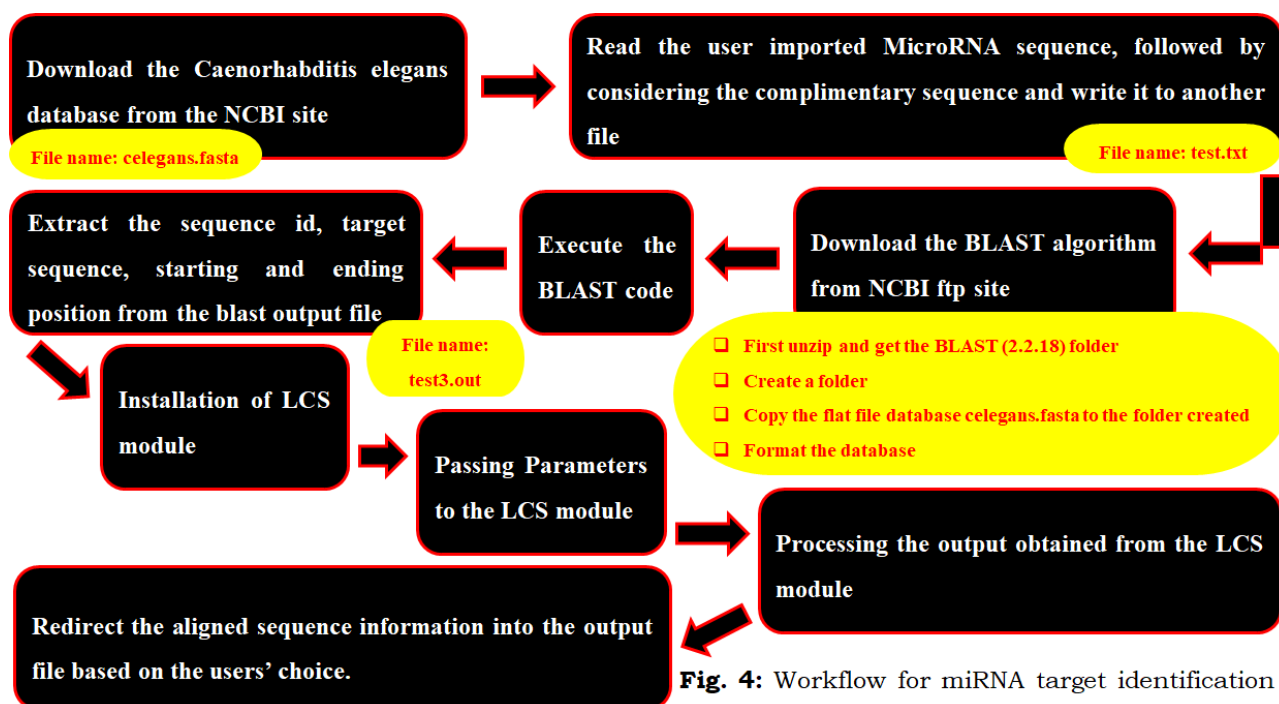
Fig. 4: Workflow for miRNA target identification using dynamic PERL programming.

II. REQUIREMENTS

To run the Micro-RNA target identification PERL program, the following minimum computer specification is required. To be precise, this work was done on a computer with **Windows 10** *(64-bit, x64-based processor)*. The *Random-Access Memory (RAM)* of the system was **8.00 GB**, with *processor* – **Intel® Core™ i5-6200U CPU @ 2.30GHz 2.40 GHz**. Apart from the mentioned system requirements, one need to install the PERL compiler (Strawberry Perl and ActiveState Perl (https://www.perl.org/get.html). It is to be notedthat, this program will work well with Linux and Mac-OS as well.

For the execution of the program, the following databases and algorithms was implemented: (a) **Database of Caenorhabditis elegans** which can bedownloaded from NCBI(https://www.ncbi.nlm.nih.gov), containing 28393 records,(b) **BLAST Algorithm**(Altschul et al.,1990),downloaded from the NCBI (https://www.ncbi.nlm.nih.gov/) ftp site, (c) **LCS (Longest common subsequence) Module**(Chvatal et al.,1975), downloaded from the CPAN site (https://www.cpan.org) and, (d) **Micro RNA**(Lee et al., 1993), accepted from the user and also downloaded from the microRNA database.

III. COMPUTATIONAL METHODOLOGY

The main steps (as seen in *Fig. 4*) that are used in the microRNA target detection involvesdownloading the Caenorhabditis elegans database from the NCBI site (**File name: celegans.fasta**). This file contains 28393 records. The next step is to readthe user imported MicroRNA sequence, followed by considering the complimentary sequence and write it to another file (**File name: test.txt).** Once the sequence is written to a file, we downloaded the BLAST algorithm to check the similarity of the sequences from NCBI ftp site. For that, we first unzipped the BLAST file and retrieved the folder (**blast – 2.2.18**). This folder contains sub folders like ***bin*** which contains executable files, and program is written in this folder, ***data*** containing sample database, ***doc*** where all installation details are available. Secondly we created a folder (**Folder name: bb)** where we copied the flat file database **celegans.fasta** to the folder. The next step involved formatting of the database using the code: **bin >. / formatdb -i / data / celegans.fasta -p F -o T / bb / celegans.fasta.** After which celegans.fasta file was copied to the local folder (**Folder name: bb).**

The next process involves executing the blast code by the command: **bin >./ blastall -p blastn -d ./ bb / celegans.fasta -i ./ bb / test.txt -o ./ bb / test3.out.** Where, **-p** indicates the type of blast being executed (blastn), **-d** indicates the database (celegans.fasta), **-i**indicates the input file (test.txt) and **-o** indicates the output file (**test3.out).** This process is followed byextraction of sequence id, target sequence, starting position and ending position from the blast output file – **test3.out.** In this step, we will get the array of target sequences,

starting position's array, ending position's array and array of the length of the target sequence.

Next, we installedthe LCS module, where we firstexecuted the Perl *pm* file, **LCS.pm.** Here the LCS module is installed in the folder **insmod** locally.For this, thefollowingcommand was used:

>**perl Makefile.pl PREFIX = / home / students / amm07bio19 / BLAST / insmod**
>**make**
>**make test**
>**make install**

Before executing the program, the following command was executed to export the local path to **@INC**,

>**export**
PERL5LIB = / home / students / amm07bio19 / BLAST / insmod / lib / perl5 / site_perl / 5.8.5 / i386-linux-thread-multi

Important Note:
"/home/students/amm07bio19" *in the commands/source-code,should be replaced by the root directory of your system.*

After the execution of the commands, we converted the microRNA sequence as well as each target sequence into arrays and passed them into the LCS module. The output from the LCS module contains the alignment position's index. We processed the output from the LCS module and converted the target sequence along with the microRNA sequence according to the alignment position's index, to get the alignment of the given input sequence. Finally, we redirected the aligned sequence information into the output file based on the users' choice.

IV. PROGRAM: THE SOURCE CODE

```
#! /usr/bin/perl
use Algorithm::LCS;
$alg = Algorithm::LCS−>new;
```

#Enter the micro RNA sequence
```
print "Enter the microRNA sequence:";
$miRNA=<stdin>;
chomp($miRNA);
$comRNA=$miRNA;
$miRNA=uc $miRNA;
$comRNA=uc $comRNA;
@miRNA=split("",$comRNA);
open(FLHD,">>DNASeq.txt"), or die("Cannot open this file");
```

#To find out the length
```
$len_mirna=length($miRNA);
```

#Find out the complementarity
```
$comRNA=~ tr/AUCG/TAGC/;
@comRNA=split("",$comRNA);
```

#Write the complementary sequence to a file
```
open(FILE,">/home/students/amm07bio19/BLAST/blast-2.2.18/bin/bb/test.txt") or ("Cannot open the file");
while(FILE)
{
    last if print FILE "$comRNA";
}
close(FILE);
```

#To execute the blast code
```
`./blastall -p blastn -d ./bb/celegans.fasta -i ./bb/test.txt -o ./bb/test3.out`;
open(FILE1,"/home/students/amm07bio19/BLAST/blast-2.2.18/bin/bb/test3.out") or ("Cannot open the file");
@data=<FILE1>;
close(FILE);
```

#For finding the potential mRNA sequence target
```
foreach $x(@data)
{
    if($x=~m/^Sbjct:/)
    {
        push(@mRNA,$x);
        $x=~m/^Sbjct:\s*(\d+)\D+(\d+)\s*$/;
        $start=$1;
        $end=$2;
        push(@start_pos,$start);
        push(@end_pos,$end);
    }
    elseif($x=~m/^>/)
    {
        $i=index($x,'|');
        $j=index(($x,'|',$i+1);
        $len=$j-$i;
        $sequen_id=substr($x,$i+1,$len-1);
        push(@seq_id,$sequen_id);
    }
}
```

#Validation of the sequence
```
foreach $y(@mRNA)
{
    $y=~s/Sbjct://;
```

```perl
        $y=~tr/atgc//cd;
    push(@mmRNA,$y);
}


#To find out the position of the target sequence
in the database
@start=map($_-10,@start_pos);
@end=map($_+10,@end_pos);
$size=@start_pos;


#To find out the length of the target sequence
for($i=0;$i<$size;$i++)
{
    $seq_length=$end[$i]-$start[$i];
    push(@seq_len,$seq_length);
}
#Taking the sequence from the c.elegans
database
open(FILE1,"/home/students/amm07bio19/BLA
ST/blast-2.2.18/bin/bb/celegans.fasta")      or
("Cannot open the file");
@data2=<FILE1>;
Close(FILE1);
#Join the array into a scalar variable
$dbinfo=join("",@data2);
@dbinfo2=split(">",$dbinfo);
#For extracting the target sequence from the
celegans database
$size=@seq_id;
for($i=0;$i<$size;$i++)
{
    foreach $z(@dbinfo2)
    {
        if($z=~m/$seq_id[$i]/)
        {
            $z=~s/.*mRNA//;
            $z=~s/.*alternatively spliced//;
            $z=~tr/[ATGC]//cd;
            $z=~tr/\s//d;
            push(@tar,$z);
        }
    }
}
$i=0;
    foreach $x(@tar)
    {
        $x=~s/.*alternatively spliced//;
        $x=~s/.*mRNA//;
        $x=~tr/[ATGC]//cd;
        $x=~tr/\s//d;

    push(@target,substr($x,$start[$i],$seq_len[$i]));
        $i++;
```

```perl
    }
$count=0;


#Output redirection
print "\nEnter the output file name:";
$outfile=<stdin>;
chomp($outfile);
open(FILE1,">>$outfile")  or  ("Cannot  open  the
file");
{
print FILE1 "\nMICRO-RNA TARGET
DETECTION\n***************************\n\nMicr
oRNA:$miRNA\nDatabase:
Caenorhabditis elegans\n";
}
close(FILE1);
#Each sequence in the array passed into the
LCS module
foreach $x(@target)
{
    @align=qw();
    @tar=split("",$x);
    $seq_len=@tar;
    $rna_len=@comDNA;
    @lc=$alg->LCS(\@comRNA,\@tar);
    foreach $y(@lc)
    {
        push(@arr,@{$y});
    }
    $lt=@arr;
$lt--;
    $a=0;
    for($i=0;$i<$lt;$i=$i+2)
{
    $z=$arr[$i];
        $w=$arr[$i+1];
            $align[$z]=$tar[$w];
            $z="";
            $w="";
}
    $len_alig=@align;
#For inserting the _ for alignment
for($i=0;$i<$len_alig;$i++)
{
if($align[$i])
{
    next;
}
else
{
    $align[$i]='-';
}
}
```

#Opening the file for appending each sequence alignment

```
open(FILE1,">>$outfile") or ("Cannot open the file");
{
    print FILE1 "\n\n$seq_id[$count]\tStarting
        position:$start[$count]\tEnd
        position:$end[$count]\nRNA:@miRNA\nS
EQ:@align\n";
}
close(FILE1);
$count++;
@arr=();
}
```

V. CONCLUSION AND FUTURE DEVELOPMENT

MicroRNA target identification is the first approach as a free accessible academic real time programming software for non-coding RNA / gene expression worldwide.The identification of the miRNA target sites from this program code, will serve as a platform for potential biomarkers for drug designing, in case of **SARS-CoV-2**. Also, the example C.elegans has served the purpose for demonstration and workability of the program code. However, users can change the genome database according to their choice and need for the diseases such as **SARS-CoV-1**, **SARS-CoV-2**, **Flu**, and **MERS-CoV**.

Empirical evidences of efficacy for: (a) spotting the probable microRNA target sites in genomic sequence, (b) aligning the sequence based on the genomic mRNA sequences (dynamic programming local alignment), (c) percentage sequence identity between the query miRNA sequences, in the program MicroRNA target identification and (d) artificial development of sequence annotations would be helpful to identify areas for future research in providing a quick review on the down regulation of gene expression, that are transcribed from DNA for the benefit of scientific community, mainly in treating various diseases like **COVID-19**.

We are going to incorporate more scientific functionalities regarding **primary transcripts** and **gene functionalities** as well as the complications arising due to short stem loop structures. Periodically continuous updates shall be released to include other biological functionalities like, *t*RNA synthesis, the transcription and translation process in proteins, to name a few. We plan to incorporate a provision to avail the required information using **graphical user interface (GUI)** in near future.

REFERENCES

[1] Hui, Ling; Muller Fabbri, George A., Calin (2013). MicroRNAs and other non-coding RNAs as targets for anticancer drug development.Nature Reviews Drug Discovery, Volume 12, Issue 11, pp. 847 – 865.

[2] Rosalind, C. Lee; Rhonda; L. Feinbaum; Victor,Ambros (1993). The C. elegans Heterochronic Gene *lin-4* Encodes Small RNAs with Antisense Complementarity to *lin-14*". Cell, Volume 75, Issue 5, pp. 843 – 854.

[3] Leigh-Ann, MacFarlane;Paul, R. Murphy (2010).MicroRNA: Biogenesis, Function and Role in Cancer. Current Genomics, Volume 11, Issue 7, pp. 537 – 561.

[4] Jacob, O'Brien; Heyam, Hayder; Yara, Zayed; Chun, Peng (2018). Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Frontiers in Endocrinology, Volume 9, Issue 402, pp. 1 - 12.

[5] Fazli, Wahid;Adeeb, Shehzad;Taous, Khan;You, Young Kim (2010). MicroRNAs: Synthesis, mechanism, function, and recent clinical trials.Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, Volume 1803, Issue 11, pp. 1231 – 1243.

[6] David, P.Bartel (2009). MicroRNAs: target recognition and regulatory functions. Cell, Volume 136, Issue 2, pp. 215 – 233.

[7] David, P.Bartel (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, Volume 116, Issue 2, pp. 281 – 297.

[8] Marco, F. Schmidt (2014). Drug target miRNAs: chances and challenges. Trends in Biotechnology, Volume 32, Issue 11,pp. 578 – 585.

[9] Johora, Hanna; Gazi, S. Hossain;Jannet, Kocerha (2019). The Potential for microRNA Therapeutics and Clinical Research. Frontiers in Genetics, Volume 10, Issue 478, pp. 1-6.

[10] Shaimaa, Abdel-Ghany; Hussein, Sabit (2020). microRNA-Based Vaccination and Treatment for COVID-19. Current Trends in Vaccine and Vaccinology, Volume 3, Issue 1, pp. 109-110.

[11] Dave, Cavanagh (2007). Coronavirus avian infectious bronchitis virus. Veterinary Research, Volume 38, Issue 2, pp. 281 – 297.

[12] Lincoln, Stein; Paul, Sternberg; Richard, Durbin; Jean, Thierry-Mieg; John, Spieth (2001). WormBase: network access to the genome and biology of *Caenorhabditis elegans*. Nucleic Acids Research, Volume 29, Issue 1, pp. 82 – 86.

[13] Stephen, Altschul; Warren, Gish; Webb, Miller; Eugene, Myers; David, J. Lipman (1990). Basic local alignment search tool. Journal of Molecular Biology, Volume 215, Issue 3, pp. 403 – 410.

[14] Václáv, Chvatal; David,Sankoff (1975). Longest common subsequences of two random sequences. Journal of Applied Probability, Volume 12, Issue 2, pp. 306 – 315.