Bioinformatics Challenges and Advances in RNA interference

Deepak Anand, Ph.D.*†
Prerna Pandey, Ph.D.¡

† Post Doctoral Fellow, Max Plank Institute for Terrestrial Microbiology, Karl-von-Frisch-Strasse 10- 35043, Marburg, Germany.
¡ Molecular Biologist, Science Writer and Editor, Maharashtra, India

* Corresponding Author
deeepak.anand@mpi-marburg.mpg.de

Abstract

RNA interference is a post-transcriptional gene regulatory mechanism to down-regulate the gene expression either by mRNA degradation or by mRNA translation inhibition. The mechanism involves a small partially complementary RNA against the target gene. To perform the action, it also requires a class of dedicated proteins to process these primary RNAs into mature microRNAs. The guide sequence determines the specificity of the miRNA. Therefore, the knowledge of guide sequence is crucial for predicting its targets and also exploiting the sequence to create a new regulatory circuit. In this short review, we will briefly discuss the role and challenges in miRNA research for unveiling the target prediction by bioinformatics and to foster our understanding and applications of RNA interference.

Introduction

Regulation of gene expression is the key component to any cell or organism. Each process, starting with birth to development and growth requires regulation of genes at a given time. The regulation of promoter activation by transcription factors, post-transcriptional regulation by regulatory RNAs [1], regulatory elements within the UTR [2], mRNA degradation by ribonucleases [3,
4] or by protein degradation [5] provides complete regulation of expression. In higher organisms, these mechanisms have evolved into complex and multi-tiered but well-structured systems to regulate gene expression.

RNA interference was discovered by Andrew Fire and Craig C. Mello in 1998 and jointly received the Nobel Prize for physiology and medicine in 2006 for the discovery of “RNA interference” [6]. RNA interference or RNAi is a post-transcriptional gene silencing (PTGS) mechanism in eukaryotes whose role is important in various stages of life or growth conditions. It can modulate the stability of mRNA or inhibit its translation. These effector molecules are named depending on their origin. While short interfering RNA (siRNA) are exogenously synthesized mature RNA or viral encoded RNA, microRNA (miRNAs) are endogenously encoded RNAs [7] and are required for the basic physiological function of the cell. These RNAs have unique secondary structure and exhibit regulatory functions exclusively. While the siRNAs can have perfect complementarity to their target, miRNAs share only partial complementarity with their target, therefore, allowing outputs different from that of siRNA. The identification new miRNAs and their targets are important as they not only control basic physiological function but are also associated with various human diseases. For example, miR-600 has been found to control the cell proliferation, migration, and invasion in human CRC cell lines [8]. Besides studies on human cell lines, miRNA investigations are also focused extensively on insects, and plants development [9-11].

**Synthesis and mechanism of action**

miRNAs are endogenous noncoding RNA, synthesized during various stages of growth. However, siRNA could be viral or synthetic in origin. Short hairpin RNA (shRNAs) are the class of regulatory RNA similar to miRNA in design, structure, and processing [6, 7, 12], and are expressed from a plasmid or through viral vectors. The miRNAs are transcribed as a primary RNA (pri-miRNA) with partially paired nucleotide overhangs at both 5' and 3' ends that form hairpin loops and stem structure in the middle. Specific ribo-endonucleases process these structures that serve as templates, resulting in a hairpin structure of 60-70 nucleotide known as precursor-miRNA (pre-miRNA). Pre-miRNA is transported to the cytoplasm with the help of a protein called Exportin-5. Here, a second endonuclease reaction cleaves off the hairpin loop from pre-miRNA and converts it into a short double stranded RNA product [6, 13] (Fig 1). The resulting miR-miR* duplex is partially complementary and contains 5' phosphate and 3' overhang of two nucleotides. The two strands of this duplex are called “Guide strand” and “Passenger-strand.” The strands are shortened, and the Argonaut protein retains the miR-guide strand, whereas miR*-passenger strand is released and degraded eventually (Fig1). The guide-miRNA and the Argonaut complex together form the RNA interference silencing complex (RISC), which can recognize complementary target sequence on 3' untranslated region (3'UTR) of the cognate mRNA. A *seed sequence* from 2-8 nucleotide of miRNA 5' end provides the specificity for target mRNA. The inhibition of target translation is achieved by association or recruitment of interacting partners to the RISC complex which either degrades the bound mRNA or blocks translation (Fig 1) [12, 14, 15].

**Identification of regulatory RNAs/miRNAs**

Initial studies of miRNA were discoveries were based on genetic screening [16]. Most frequent methods for detection and identification of novel miRNA include transposon RNA
miRNAs are transcribed as primary miRNA (Pri-miRNA) and processed at the 5′ and 3′ overhang by Drosha and DGCR8 to generate precursor-miRNA (pre-miRNA). These are exported to the cytoplasm with the help of Exportin 5. Pre-RNA binds to Dicer (RNase III nuclease) and TAR RNA-binding protein (TRBP). Dicer cleaves pre-miRNA into 20-22 nt mature miRNA duplex. Mature miRNA binds with Ago protein and TRBP to make RISC complex; eventually selected guide strand is directed to target mRNA while another strand is degraded. In the last step, repression can be achieved by induced degradation of target or translation block at initiation and elongation stages.

**Fig. 1 Biogenesis of microRNA and target repression.**

miRNAs are transcribed as primary miRNA (Pri-miRNA) and processed at the 5′ and 3′ overhang by Drosha and DGCR8 to generate precursor-miRNA (pre-miRNA). These are exported to the cytoplasm with the help of Exportin 5. Pre-RNA binds to Dicer (RNase III nuclease) and TAR RNA-binding protein (TRBP). Dicer cleaves pre-miRNA into 20-22 nt mature miRNA duplex. Mature miRNA binds with Ago protein and TRBP to make RISC complex; eventually selected guide strand is directed to target mRNA while another strand is degraded. In the last step, repression can be achieved by induced degradation of target or translation block at initiation and elongation stages.

sequencing (RNA-seq), mutagenesis [17], RNA sequencing, Watson-Crick base pairing with probes (northern hybridization, in-Situ hybridization, and RT-PCR), cloning and splinted ligation. Sequencing methods use robust and high throughput technology. While RNA sequencing provides direct identification of novel transcripts, other methods of hybridization need prior knowledge of miRNA sequence. For experimental detection, use of locked nucleic acid (LNA) probes improves hybridization efficiency and specificity [18] and helps in the validation of the predicted transcripts. Different methods have their advantages and disadvantages. For example, besides detecting the miRNA using northern hybridization, quantitation can also be performed by direct detection. Additionally, PCR-based detection is favored for low copy templates, which is enabled by many log amplifications.

**miRNA identification and target prediction: the role of bioinformatics**

Although predicting novel RNAi genes through computational approach presents a challenge. However, the methods are good for shortlisting the putative candidates that can be verified through experimental methods. The criteria used for candidate miRNA prediction are based on the comparative genomics with related organisms, genetic loci, the size of the transcript, the structure of primary RNA and free energy of the structure. Report of various strategies used for prediction of candidate micro-RNA-like micro-conserved elements (MCE) could be one of the many approaches [19]. The efficiency of prediction can be achieved through several modifications in the algorithms, as well as using the experimental evidence. A list of such tools has been reviewed by Salim and Chandra [20]. Some of these tools are stand alone and can function independently whereas others are based on web services or require other platforms. While all these tools used for prediction and validation of targets bring fewer publications, a huge number of research publications comes from experimental data (Fig 2) suggesting exponential growth in RNA interference research.
Fig 2. Increase in the publications on miRNA target prediction and miRNA research. After the discovery of miRNA in 1998 (red line on the x-axis), there has been slow increase in the research on prediction of miRNAs compared to the logarithmic increase in other miRNA related research work have been noted. The pink line corresponds to predicted miRNA publications, and blue denotes the research works carried out on miRNAs. Data on publication were collected from PubMed database search.

Some bioinformatics analysis includes unraveling counterparts and critical components of the RNAi pathway in various organisms. The genes involved in miRNA pathway of many organisms have been experimentally identified, and the information can be accessed from a registry that was constructed by sequencing small RNAs (~22 nt) and mapping them to the genome [13]. The miRNA sequence can occur in intergenic regions, in introns of protein coding regions or in exons and introns of noncoding genes. Through bioinformatics tools, the miRNA genes can be identified by using the known biases in targeting strands. This reduces the unwieldy work of analyzing the whole genome of the organism, although, it could be difficult to achieve easily in the case of miRNA, as there are significant mismatches between the miRNA single strand and the cognate mRNA. For this purpose, there are programs like the MiRscan57 and MiRseeker [21] that use the conserved regions of the stem-loop structures to identify novel miRNA genes across species. Since miRNAs work jointly with other members of pathways, the study of tissue specificity of genes can help in reviewing miRNA and their targets. Genome sequences also provide ways to predict miRNAs, but verification of expression need to be done by sequencing of RNA species or by other methods of choice.

For RNA interference, a strong binding of the 5' end (the first eight base pairs) of the mature miRNA to the target 3' UTR sequence is essential. Silencing efficiency is affected and reduced by G:U wobble pairing in the interacting region. Algorithms and models have been developed for miRNA target interactions, based on experimental evidence. Since the microarray experiments show that miRNAs exhibit tissue-specific expression patterns, correlation of mRNA with miRNA expression can help narrow down the predicted targets. Moreover, based on the patterns of expression, the miRNAs and mRNAs can be clustered. Several online analysis tools for miRNA target prediction have been documented in recent past [13, 20, 22-24]. Together with the miRNA identification, the prediction of their promoters is an ever-present challenge, although many underlying assumptions like homologies, UTR location of target sites, can help in identifying good targets. Therefore, caution must be taken for false positive predictions [25] as the coverage of all possible sites, or complete sequence may not be possible, some of which may be unique to certain species and may not follow the rules [22, 23].

siRNA/shRNA design and target prediction

The siRNAs originate from transgenes, viruses, and transposons and form perfect complementary double strand RNA precursors (dsRNAs) [6, 21]. Due to the perfect complementarity, they can silence the same
gene from which they originate. The siRNA precursor generates multiple siRNAs that can be transmitted among cells [23, 24]. Processing of siRNA requires their dicing and sorting by Argonaut proteins. The siRNAs get incorporated into RISC and silence genes that are complementary to the RISC incorporated strand. Because of the strand bias showed by RISC, the strand which is more unstable, or whose end is easier to open, is taken up by the RISC [12, 14]. Also, efficient silencing also depends on the base composition along the siRNA active strand. However, the siRNAs can also show off-target effects and silence unintended targets. Thus, a rational design of siRNA is critical.

A unique siRNA selection program, siRNA Scanner, automatically selects siRNA from the given RNA data pool. This program is coded in PERL (5.8.8.6 Build 820), uses a fuzzy logic-based system to calculate siRNA qualities and is accessible in a command line interface [26]. Another program, siRNA scanner ensures minimum user interaction and faster algorithm, thus, is fit for selecting siRNA for gene expression studies. The program also evaluates extensively each possible siRNA candidate and checks for all possible candidates from the given template sequence.

siDirect (http://design.RNAi.jp/) is an online software for computing highly effective siRNA sequences with target-specificity for RNAi in mammalian systems [27]. siDirect software prevents off-target effects that the most BLAST programs may not achieve. siRNAs for some common genes are easily available and are routinely used in experiments to silence genes. To avoid the off-target effects of siRNA constructs, they must be checked rigorously for being unique on non-redundant mRNA database (excluding splice variants). The other prerequisites for the siRNA design are: Less than 30–60 percent GC content or the sequences 4 As, 4 Ts, 4 Gs or 4 Cs are eliminated and with TT overhangs on the 3' ends of each strand [4, 6, 12, 14]. There are several websites that follow these criteria and allow the design of siRNAs.

Most commercial siRNA libraries incorporate chemical modifications to the siRNA seed region to help reduce off-target effects (OTEs) [27] but data analysis and experimental follow-up of screen results are essential to identify false-positive hits due to OTEs. For example, software tools that can be used to identify seed-dependent OTEs in RNAi data sets include Genome-wide Enrichment of Seed Sequence matches (GESS) and Haystack [27], which identify putative transcripts that are off-targets, and both produce good results on using larger (genome-wide) data sets.

Short hairpin RNAs or shRNAs are artificial constructs that can be inserted into a genome and expressed like endogenous genes [6, 7, 12, 14]. The primary transcripts fold back to form a dsRNA, mimicking native pre-RNA structure. This is directly accepted and used by the RISC to find target mRNA. The context of a known miRNA and siRNA-based designs can be used to construct these shRNAs. Several shRNAs libraries are now commercially available [21, 26, 28]. The libraries allow large-scale studies with siRNAs that are otherwise prohibitively expensive. To start with, all the shRNAs that can putatively silence a gene of choice are picked from the library and are infected into the cell. Following the cell growth for a few generations, the shRNAs that have depleted are investigated. Finally, the mRNA targets of the depleted shRNAs are sought, and further studies are planned.

To start with, all the shRNAs that can putatively silence a gene of choice are picked from the vector library, and host cells are transfected. Following the cell growth for a few generations, the shRNAs that have depleted are investigated. Finally, the mRNA targets of the depleted shRNAs are aimed for, and further studies are planned.
There are several bioinformatics limitations associated with the use of shRNA. For a more accurate analysis, understanding outcomes of library screening are essential, which can be achieved by simulation of shRNA libraries when used in populations of cells [29]. A central repository of shRNA constructs like the RNAi Codex [29, 30], which can help track results, identify the performance patterns of shRNA facilitate in finding constructs from a variety of sources.

Also, microarray probes need to be designed for the more efficient study of shRNA expression in cell populations. Finally, techniques that can analyze results of silencing of genes either singly or in groups need to be developed.

**What the future holds**

There has been a constant need to develop new algorithms and tools for designing efficient and specific siRNAs [21, 26, 31]. A few of the important on-going developments includes formulating algorithms to find off-targets of siRNAs design for genome-wide studies. It is also essential to maintain a unified data source that presents optimal conditions as well as the design failures of siRNAs used for validation or experimentation. Easy access to such data library would definitely help in refining algorithms and would save much of effort that could otherwise go waste by designing non-functional siRNAs.

The information gained by studies on siRNA and miRNA designs can also be used to improvise on the other studies. Using siRNAs in experiments to mimic effects of drugs, a field that exhibits immense potential because it could reduce the cost of large-scale screening.

Many programs have been developed to predict the target of miRNA and design siRNAs. These algorithms assess the sequence similarity between of miRNA and target RNA and the free energy change that occurs during the interaction. For target predictions, the seed sequences are looked upon the untranslated region (UTR) of mRNA. In humans, most of the seed sequence are found to be located at the 3' end of the mRNA where it can bind and lead to either degradation of translation inhibition of the mRNA. In the list of web applications reviewed by Akhtar et al. [32], one can find a well-categorised list of tools and databases to study or different aspects of miRNA. These are for validation of miRNA (DIANA-TarBase, miRTarBase, miRecords) [33], correlating miRNA expression (StarBase, MiRonTop) metabolic pathways (EiMMO) regulator network (MiGator), transcription factor interactions (TransmiR) and prediction of cellular target of host and viral miRNA are few examples from the list.

Computational analysis can correct some experimental biases that may exist while screening the experimental data. For this purpose, a web-based platform named Comprehensive Analysis of RNAi Data (CARD) has been developed that integrates analysis and visualization of RNAi screen data [34]. Other tools like R and Python code9 seed analysis predicts off-target siRNAs by testing whether siRNAs with shared seed sequence have significantly higher/lower scores compared with the rest [34]. The genes that are potentially off-targeted can be identified by MATLAB code and web-service (http://www.flyrnai.org/gess/) [34]. The screen data can be integrated with PPI networks, and optimal threshold for hit selection can be identified by RNAiCut [35]. Expression of genes is used as a parameter and also as a screening filter for flagging and filtering of genes.

A completely automated search program for siRNAs, inSiDE, contains a comprehensive set of rules for selecting of siRNA candidates (mentioned earlier in the “siRNAs” section) can be applied and allows for significantly high specificity [36]. After selecting potential siRNA
candidates with the optimal functional properties, putative unspecific matches, which can cause cross-hybridization, are checked in the databases containing a unique entry for each gene. These truly non-redundant databases are constructed from the genome annotations (Ensembl) which is followed by examining other factors such as the presence of single nucleotide polymorphisms, intron/exon boundaries, transcript specificity, before the design of siRNAs [35, 36].

SPICE is another web-based tool that supports RNAi screening through comprehensive inquiry and automates several challenging tasks during evaluation of the short RNAs obtained by RNAi screening. This tool identifies the shRNA sequence in the input sequence, the target genes as well as the biological information in the database, and finally, prepares search result files that can be accessed locally. The system facilitates RNAi screening, which requires sequence analysis after screening [37].

In addition to carrying out statistical analyses and experimental follow-up studies of RNAi screen data, the application of various bioinformatics approaches can greatly aid in distinguishing between high-confidence and low-confidence screen hits. One common approach is to analyze the data in aggregate using a gene-set enrichment or related algorithms, such as using the DAVID database (the database for annotation, visualization, and integrated discovery) from the US National Institutes of Health, COMPLEAT and other software to detect gene ontology terms, protein complexes or signaling pathways that are enriched in the screen hits compared to controls [38]. These results can be used to confirm low confidence hits, such as hits with borderline statistical scores or hits for which not all RNAi reagents targeting the gene were positive and, conversely, to rule out hits that are the sole representatives of a category, which might suggest that they are false positives.

**Discussion**

While the RNAi effector molecules function in the regulation of a variety of biological processes and can have an immense impact on various applications, much information about their mechanisms with respect to several physiological and regulatory functions still remain to be clarified. High-throughput methods have significantly advanced our knowledge of these small RNAs, and bioinformatics tools play a significant role in parallel to address the different aspects of RNAi research. Today, a vast range of tools are available for prediction of miRNA gene and target, network regulation and for siRNA design. While the tools can deliver to the best of their capacity, the predictions need to be linked and validated with experimental evidence.

**References**


5. Genschik, P., et al., *Selective protein degradation: a rheostat to modulate...*


Editor: Dr. Prashant Pant, University of Delhi, Delhi, INDIA
Reviewed by: Muniba Faiza, PhD., South China University of Technology, Guangzhou, CHINA