Firoz Ahmed (2010). *Development of Bioinformatics tools for understanding RNA interference pathway.* Ph.D. Thesis. CSIR-IMTECH, Chandigarh/ Jawaharlal Nehru University, New Delhi: India.
Supervisor: Dr. G.P.S.Raghava Acc. No. : Th-240

SUMMARY OF THE THESIS

Recently RNA interference (RNAi) has gained tremendous attention in life science research because of their crucial role in gene regulation and as a major tool to knockdown gene of interest in order to understand its function. Proof-of-concepts are emerging stating that RNAi could also be used as a therapy. Despite several biochemical and computational study of RNAi, the mechanistic understanding remains obscure, which hampers the full exploitation of this mechanism as a therapeutic intervention. The rise of RNAi potential spurred the development of bioinformatics tools mainly for miRNA and their target prediction. In addition, siRNA designing algorithms also help to make effective molecules. Biogenesis of miRNAs involves various steps in which hairpin like transcripts expressed in the cell that is recognized and processed by RNase III enzyme to generate ~22 nt miRNA. The miRNA bind to complementary mRNA resulting in gene silencing either through cleavage of mRNA or repression of protein synthesis.

However, to harness RNAi as an effective tool for safer therapeutic purpose, there is a great need to uncover the different steps involved in RNAi pathway. The overall objective of this thesis is to understand each step of miRNA biogenesis pathway, from transcription to the sequential processing of pri-miRNA, pre-miRNA and generation of effective miRNA by using computational approach. In addition, several computational resources have been generated to make more effective siRNA and to explore potential of siRNAs against HIV.

Current literatures suggested that most miRNA genes are transcribed from intergenic region by RNA polymerase II as independent transcripts and undergo methylation at 5′-end and polyadenylation at 3′-end. However, the differences between promoter of miRNA and mRNA gene was not explored. Thus in present study features of nucleotide were used to develop a SVM-based model for predicting promoters belong to miRNA and achieved an accuracy of ~70%. In addition, a more accurate and robust method has been developed to predict various types polyadenylation signal in human genome to annotate the boundary of a gene and achieved an accuracy of 85.7%. The **PolyApred** tool is available at www.imtech.res.in/raghava/polyapred. Furthermore, the tool was implemented to elucidate the primary transcript of miRNA gene by

characterizing their 3′-end boundary. It was found that miRNA gene contain poly(A) signals at average distance of 3000 nt from miRNA hairpin located at last position.

Further, we extended our study towards selectivity and processing of pri-miRNAs and pre-miRNAs by Drosha and Dicer enzymes respectively. Earlier studies support the role of sequence and structural determinants for Drosha cleavage site selection. Thus, in this study a dataset of miRNA hairpin sequences with their flanking regions were taken. The sequence and structural characteristics of Drosha cleavage site was implemented to develop a SVM model that achieved highest accuracy of ~70% for predicting Drosha cleavage site. Various studies indicated that Dicer shows no specificity for cleavage site. To gain insight of Dicer specificity, large data of Dicer substrates (pre-miRNA) has been taken and analyzed their different components such as: 3′-end that interacts with PAZ domain of Dicer as well as 5′-end and found clear preferences of some dinucleotides at end position. Moreover, analysis of Dicer processing site showed the poor sequence conservation at both 5′ and 3′ arms. But surprisingly, when secondary structure and position specific nucleotides features associated with cleavage site were implemented in machine learning techniques, our model successfully differentiate between Dicer cleavage and non-cleavage site with an accuracy of 86%. This study indicates the relevance of position specific nucleotides and structural characteristics of cleavage site, which were not explored by earlier studies. Based on this study a method, **PHDcleav** (http://www.imtech.res.in/raghava/phdcleav/), has been developed for predicting Dicer cleavage site in human pre-miRNA.

A key step in the pathway is loading of miRNA into RISC complex, which bind to target in a sequence specific manner for gene silencing. However, only one strand of miRNA:miRNA* duplex (generates processed pre-miRNA by Dicer) bind to RISC while other strand degrades. Previous studies showed that thermodynamic asymmetry is mainly responsible for strand selection. However, this observation was based on studies conducted on siRNAs, and some sets of miRNAs which lacking complete information about their partner miRNA*. It has also been accepted that all functional strands, guide, do not follows the thermodynamic criteria. Therefore, first time statistical analysis has been carried out on experimental validated miRNA:miRNA* sequences and reported that miRNA and miRNA* are significantly different in their sequence features. In addition,

sequence as well as structural features of miRNA:miRNA* duplex was utilized to developed a robust SVM model, **RISCbinder**, for predicting miRNA guide strand. The tool achieved an accuracy of 80% and is available at http://crdd.osdd.net:8081/RISCbinder/. Our study also showed the structural asymmetry between miRNA:miRNA* and acts as a potential tool to predict the guide strand in siRNA duplex.

Short interfering RNA (siRNA) has become a major tool for sequence specific gene knockdown. Earlier it was believed that siRNA is highly sequence specific but later studies using microarray analysis showed that siRNA also suppress unintended target genes. Furthermore, studies indicate that siRNA could act like miRNA when not fully complement with the targets. In natural system it is not always possible to get an effective siRNA against a target. In this study weakness of siRNA (poor specificity) has been exploited to design highly effective siRNAs by making minimum mutations in it. The, ingenious approach has been divided into two parts. In the first part a SVM model has been developed for predicting siRNA efficacy and achieved a correlation coefficient of 0.67 between actual and predicted efficacy, which is comparable in performance with other well-known methods. In the second part bases were substituted at all possible sites of siRNA and these mutants were predicted with our SVM model. However, it is well known from the literature that making mismatches between siRNA and target affects the silencing efficacy. Therefore, the rules derived from base mismatches experimental data were incorporated to find out over all efficacy of mutated siRNAs. Based on above study a webserver **desiRm** has been developed, which is available at www.imtech.res.in/raghava/desirm/. It was found that three mutations could significantly change the efficacy of siRNA from non-effective to highly effective and vice versa.

Several experimental studies demonstrated that siRNA/shRNAs selectively inhibits proteins of Human Immunodeficiency Virus (HIV) and promising a potential therapeutic against HIV infection. There are numbers of databases of siRNAs with their target genes. However, these databases lack information of siRNAs targeting HIV genome. In order to accelerate potentiality of siRNAs against HIV infection, data of siRNAs/shRNAs targeting HIV genome has been collected and compiled from literatures to develop a database named **"HIVsir".** The database has been implemented on web-portal at http://crdd.osdd.net/raghava/hivsir. The current version of database contains 609

siRNA/shRNA sequences targeted against different genomic location in various strains of HIV. The HIVsir covers information on experimentally determined siRNA/shRNA. Each entry provides detailed description of siRNA including HIV strain and its NCBI accession, target location, its sequence, siRNA efficacy, type of cell and method of test.