Microon - A Tool for Identification and Validation of Oncomirs

THESIS

Submitted in partial fulfilment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

RAM K

Under the Supervision of

Dr. Sumit Biswas

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

2015

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

CERTIFICATE

This is to certify that the thesis entitled **Microon - A Tool for Identification and Validation of Oncomirs** and submitted by **Ram K**, ID No **2011PHXF013G** for award of Ph.D. degree of the Institute embodies original work done by him under my supervision.

Sumit Biswas

Signture of the Supervisor

Name in Capital Block Letters: **Dr. SUMIT BISWAS**

Designation: **Assistant Professor**

(Department of Biological Sciences)

Date: 13 APR 2015

Abstract

Since the first discovery in early 1990's, the predicted and validated population of microRNAs (miRNA) has grown significantly. These small (∼22 nucleotide long) regulators of gene expression have been implicated and associated with several genes in the cancer pathway as well. Globally, the identification and verification of miRNAs as biomarkers for cancer cell types has been the area of thrust for most miRNA biologists. However, there has been a noticeable vacuum when it comes to identifying a common signature or trademark that could be used to demarcate a miRNA to be associated with development or suppression of cancer. *In vivo* identification and analysis of miRNA expression profiles associated with various cancer cell lines are still laborious processes. On the other hand, *in silico* procedures (particularly machine learning approaches) are gaining in importance in miRNA-based studies by making the process faster and economically favourable. However, most predictive algorithms suffer from class imbalance problems and the techniques utilised to overcome the problems need more optimisations. Utilising randomly generated dataset to overcome the class imbalance problem may discard instances with strong discrimination and increase the noise during the training process.

To answer these queries, we report an *in silico* study involving the identification of global signatures in experimentally validated miRNAs which have been associated with cancer. This study has thrown light on the presence of significant common signatures, *viz.,* sequential and hybridisation, which may distinguish a miRNA to be associated with cancer. Based on our analysis, we suggest the utility of such signatures in design and development of a Machine Learning (ML) algorithm based model (MicRooN) for the prediction of miRNAs involved in the cancer. Subsequently, a web-based user interface was developed to query the predictions obtained from the ML-based model.

In brief, the major highlights of the thesis are:

- Search for common signatures in miRNAs involved in the cancer pathway
- Training machine learning based models with features extracted from miRNAs associated with cancer versus those that are not.
- Overcoming class imbalance problem with cost-sensitive approaches.
- Construction of an ensemble-based classifier from three learning algorithms *viz.,* kernel-based Support Vector Machine (SVM), decision tree-based Random Forest (RF) and C4.5.
- Development of web-based user interface (MicRooN) to query prediction obtained from ensemble classifier stored in MicRooNdb.

Acknowledgements

First and above all, I praise the God, the almighty for providing me this opportunity and granting me the capability to proceed successfully. This thesis appears in its current form due to the assistance and guidance of several people, whom I met and interacted with during the course of my Ph.D thesis. I would, therefore like to offer my sincere thanks to all of them.

Foremost, I would like to express my sincere gratitude to my advisor Dr. Sumit Biswas for the continuous support of my Ph.D. study and research, for his patience, motivation, enthusiasm and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I would also like to thank Prof. Pinak Chakrabarti, Bose Institute, Kolkata and Prof. Sanghamitra Bandyopadhyay, Machine Intelligence Unit, Indian Statistical Institute, Kolkata for their critical advice.

My gratitude also extends to Prof. Ranjit Prasad Bahadur and his lab members from Computational Structural Biology Laboratory, Indian Institute of Technology, Kharagpur for scientific interaction and fruitful discussions.

I am extremely grateful to Prof. B. N. Jain (Vice Chancellor, BITS, Pilani), Prof. Sasikumar Punnekkat (Director, BITS, Pilani K K Birla Goa Campus), Prof. Ashwin Srinivasan (Deputy Director, BITS, Pilani K K Birla Goa Campus), Prof. K. E. Raman (Acting Director, BITS, Pilani K K Birla Goa Campus), Prof. S. K. Verma (Dean, Academic Research, Ph.D. Programme, BITS, Pilani), Prof. Sunil Bhand (Dean, Sponsored Research and Consultancy Division, BITS, Pilani), and Dr. Prasanta Kumar Das (Associate Dean, Academic Research Division, BITS, Pilani K K Birla Goa Campus) for

providing me with the facilities to conduct my research work at BITS, Pilani K K Birla Goa Campus.

I express my gratitude to the members of the Doctoral Advisory Committee, Dr. Biju.K.Raveendran, Department of Computer Science and Information Systems and Prof. Dibakar Chakrabarty, Department of Biological Sciences for their guidance and co-operation. I also express my sincere thanks to Dr. Meenal Kowshik, Head, Department of Biological Sciences, Dr. Judith Braganca, Convener and members of the Departmental Research Committee for their support.

I wish to extend my thanks to my lab mates – Rahul Bhola, Malvika Sudhakar, Chirayu Garg, Atul.M.Sivan (IISER-Kolkatta), Divya Bandekar, Om Prakash Chouhan, Subhasish Sahoo; department colleagues and friends – Geetesh Mishra, Souvik Pal, Shamanth N and Gautam Kumar. I am also very thankful to the Department of Computer Science and Information Systems for helping me with High Performance Computing support.

I must say without financial support, it would have been impossible for me to carry out my research work so successfully till today. Therefore, I take this opportunity to thank Council for Scientific and Industrial Research for providing Senior Research Fellowship from April 2013-March 2015 and BITS, Pilani K K Birla Goa Campus for Institute Fellowship from August 2011-March 2013.

I warmly thank and appreciate my parents Shri. Kothandan and Smt. Shanthi; my mother-in-law Dr. Punitha Janagaraj and father-in-law Dr. V.O.S.Kalaimohan for their constant material and moral support in all aspects of my life.

I want to express my gratitude and deepest appreciation to my lovely daughter Nakshathira and my wife Smt. Janani Kalaimohan for their patience and understanding and may the almighty give all the best in return.

Ram K

Dedicated to my stars

Janani & Nakshathira

Contents

List of Figures

List of Tables

Abbreviations

ACC	Accuracy
AGO	Argonaute
AUC	Area Under the Curve
CDS	Coding region
CLD	Cancer Linker Degree
COSMIC	Catalog of Somatic Mutations in Cancer
cv	Cross validation
DGCR8	DiGeorge Syndrome Critical Region Gene 8
DT	Decision Tree
FN	False Negative
FP	False Positive
FPR	False Positive Rate
HMM	Hidden Markov Model
ID3	Interactive Dichotomiser 3
LOO	Leave One Out method
MCC	Matthew's Correlation Coefficient
miR*	miRNA star strand
miR	Mature miRNA
ML	Machine Learning
NCBI	National Centre for Biotechnology Information
ncRNA	Non-coding RNA
NGS	Next Generation Sequencing
nt	Nucleotides
OOB	Out Of Bag error
PACCMIT	Prediction of ACcessible and/or Conserved MIcroRNA Target
PCA	Principal Compound Analysis
qRT	Quantitative-Real Time PCR
RBF	Radial Basis Function
RFE	Recursive Feature Elimination
RISC	RNA Induced Silencing Complex
RLC	RISC Loading Complex
ROC	Receiver Operating curve
SMOTE	Synthetic Minority Oversampling Technique

TAG	Tumour Associated Gene
TF	Transcriptional Factor
TN	True Negative
TP	True Positive
TPR	True Positive Rate
WC	Watson-Crick Pairs

Symbols

b	Bias
c	Cost
γ	Gamma
χ^2	Chi-square
ξ	Slack variable
ϕ	Kernel function

Chapter 1

Introduction

Chapter 1

Introduction

1.1 MicroRNA - Discovery

MicroRNAs (miRNA) are a new class of non-coding RNAs which have been a hotbed of research activities for the last two decades. Lee, Feinbaum and Ambros first discovered miRNAs in 1993, during a study involving the various developmental stages of *Caenorhabditis elegans*. Earlier work done by Ambros with *C. elegans* had revealed the role of *lin-4*, in the larval development of the nematode (Figure 1.1). Worms with mutated *lin-4* could not repress the high levels of another protein LIN-14 (which in turn regulates the transition of L1 to L2 stage in the larval cycle) leading to developmental anomalies.

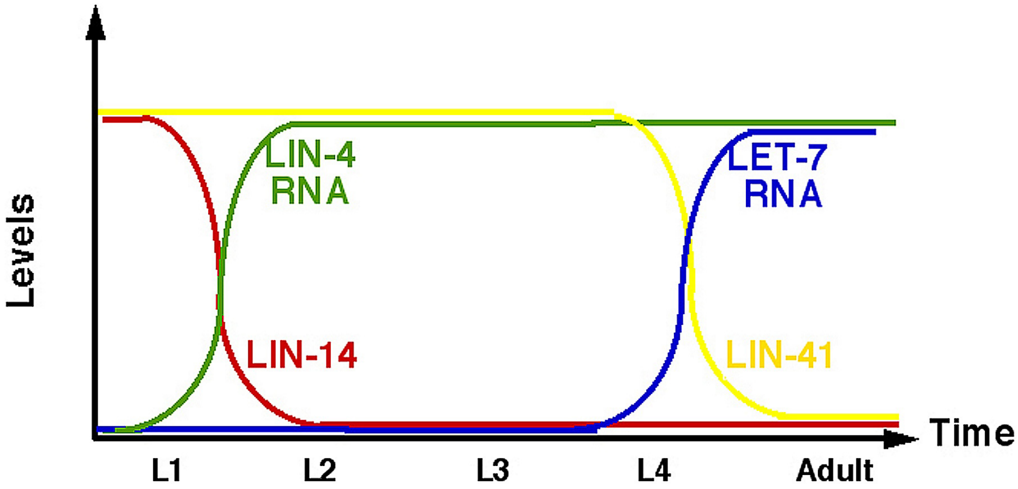


Figure 1.1: Regulation of miRNA in various stage of *C.elegans* larval development [1].

To their dismay and delight, Ambros and his colleagues (Ruvkun, in particular) discovered that the RNA coded by the *lin-4* gene did not code for any protein. Rather, this RNA was found to bind to seven complementary sites in the *lin-14* 3′UTR. Once bound to these specific sites, *lin-4* regulates the production of the nuclear protein (a putative transcription factor) encoded by *lin-14*. In the post developmental stages, it was also found that *lin-4* regulated *lin-28* in a similar fashion as that of *lin-14* [2]. Similar mechanism exists between *let-7* and *lin-14* in the L4 stages of the larval development. However, it was still considered that these small regulatory elements were worm-specific and not global.

Later over a year, almost hundred such regulatory small RNAs were reported in plants, human and other unicellular eukaryotes [2, 3]. Unlike in *let-4* and *let-7*, the gene expression of several newly found miRNA was specific to cell types and possessed differential gene expression. miRNAs mediate post-translational regulation by hybridising to target messenger RNA (mRNA). These small RNAs either bind to 3′ Untranslated Regions (UTR) or coding regions (CDS) or 5′ UTR of the mRNA [4–6] . By binding to these regions with near or perfect complementarity, they may induce either translational repression or complete cleavage of mRNA.

Systematic identification of miRNA by cloning and sequencing experiments was adopted for over a decade. Recently several experimental methods have been developed based on abundantly expressed miRNAs [7, 8]. Computational methods on the other hand, search for the location or homologs of the sequence in the entire genome considering the fact that most miRNA genes are present as tandem repeats separated by small distances. Methods utilising conserved region search (not necessarily inside the protein-coding region) were used to find several potential miRNA candidates [9–11]. With high sensitivity and specificity of these computational methods, the number of miRNA discovered in the recent years have contributed for 1% of the entire genome, a fraction similar to other regulatory gene families.

1.2 MicroRNA Biogenesis

MicroRNAs are non-coding RNAs averaging 22 nucleotides which regulate various steps in development, differentiation and physiological activities in a cell. In animals, miRNA biogenesis starts with the pri-miRNA being transcribed by RNA polymerase II, followed by processing by a microprocessor complex – Drosha and DiGeorge Syndrome Critical Region Gene 8 (DGCR8) (Pasha in invertebrates) to pre-miRNA (50-70nt long) with a 3′ overhang. Exportin 5 identifies 3′ overhang and transports the processed pre-miRNA to the cytoplasm (Figure 1.2).

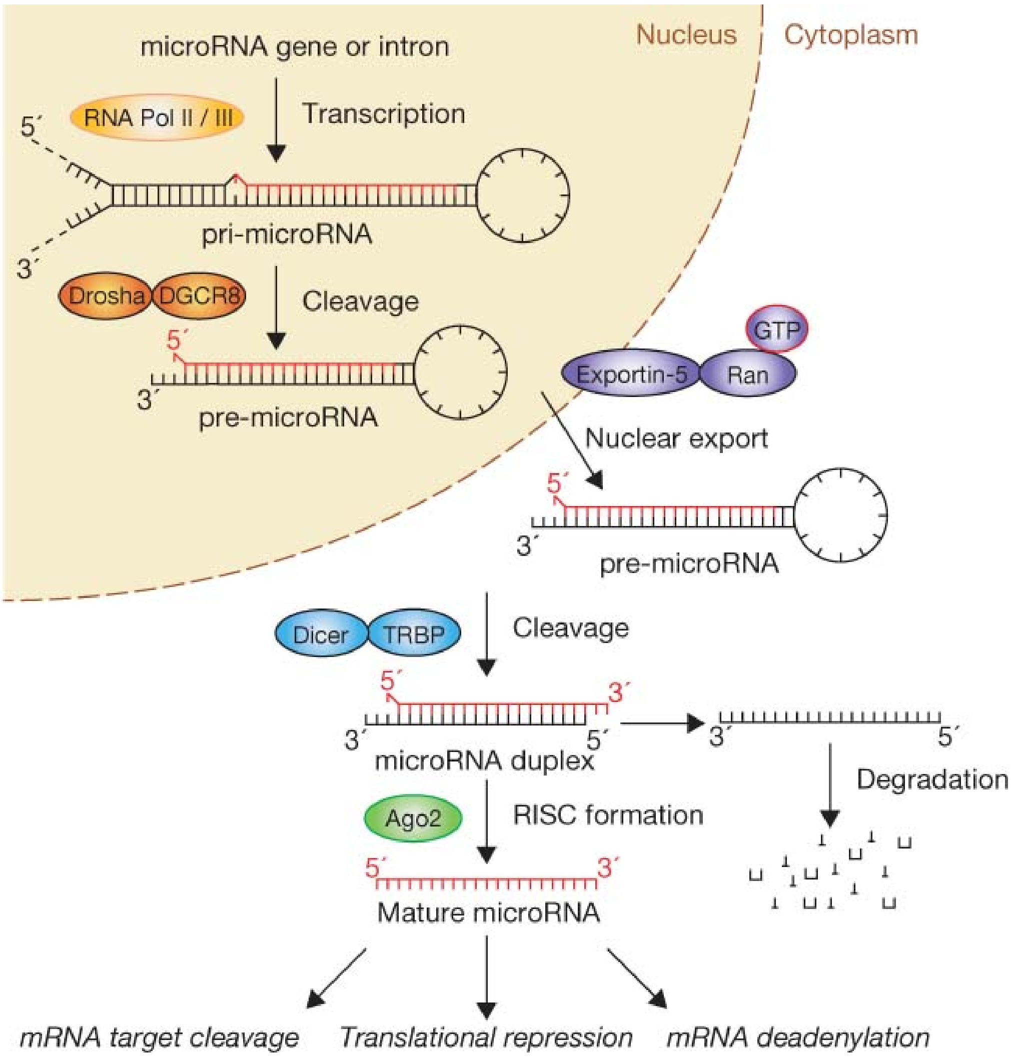


Figure 1.2: Biogenesis of human miRNA [12].

In the cytoplasm, the RISC (RNA Induced Silencing Complex) loading complex (RLC) mediates the miRNA processing and directs it to its target. The RLC contains an RNase named Dicer, the double-stranded RNA-specific Tar RNA Binding Protein (TRBP), Protein Activator of PKR (PACT) and Argonaute-2 (Ago2). TRBP and PACT do not actively participate in miRNA cleavage; rather they stabilise Dicer. Lower concentration of TRBP and PACT result in poor post-transcriptional gene silencing. Dicer belongs to the RNase III family that process miRNA with prominent double strand features. It recognises the $3'$ overhang on the double stranded miRNA and cuts $\sim$ 20 nt away to generate a short RNA duplex [13]. The end loop of the pre-miRNA is truncated and results in a duplex structure, containing a biologically active miR strand and carrier strand or the miRNA star *(miR*)*.

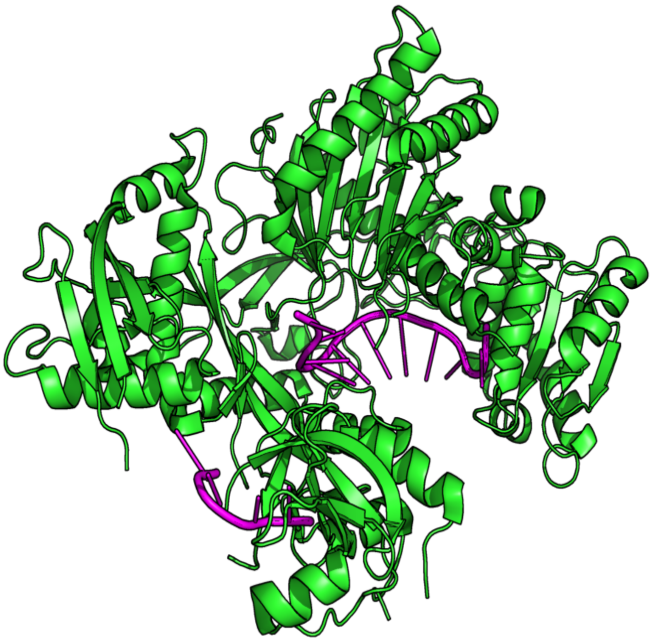


Figure 1.3: Human Argonaute-2 - *miR*-20a complex [4F3T.pdb]. hsa-*miR*-20a (magenta) bound with AGO protein (green).

Strand selection by RISC is not a stringent process [14], however it is entirely based on the inherent features of the duplex – usually the stability of base pairing in the $5'$ end determines the strand selection [15–17]. Additionally, the thermodynamics of duplex formation also plays an important role. Shortly after being processed by Dicer, the strands separate and the miR* is usually degraded. The mature miR strand binds to miRNA-Argonaute (AGO) protein complex, to form the final RISC. Mature miR bound to RISC mediates post-transcriptional regulation by binding to $3'$ or $5'$UTR or coding sequence. Once bound with AGO protein, this complex is more stable (half-life greater than 14 hours) [18] (Figure 1.3). Due to non-random strand selection process, it is also reported that some *miR** mature with higher gene expression levels [19]. miRNA stability mainly relies on the AU-richness in the sequence, a phenomenon similar to that of mRNA. Higher AU-richness signifies shorter miRNA half-life [20].

1.3 MicroRNA Regulation

In plants and animals, the regulatory activities of miRNA are highly dynamic and they depend on specific factors in them. The regulatory intensity in plants is dependent mostly on the sequence complementarity and target abundance. In plants, miRNA binds to the site with high degree of complementarity and initiates mRNA degradation. Once the mRNA degradation is complete, miRNA is released from RISC and it may target another site with complementary regions. However, in animals, the mechanism of regulation is entirely different. The miRNA-bound RISC binds to multiple targets separated by certain distance and induce translational regulation cooperatively.

Another important aspect of miRNA binding to mRNA in animals is that they bind to the targets with imperfections forming bulges, loops and mismatches [21]. Further, the translational intensity in animals is also based on the distance between the target sites [22]. Perfect pairing of mRNA and miRNA is not always sufficient for regulation. In certain cases a mismatch or loop/bulge formation may also be highly preferred [23]. miRNAs are found in clusters and may be sequence related – some miRNA clusters don't share sequence homology but still control the same functional process, suggesting the clustering of miRNA to be important for coordinated regulation. Clustering of genes is more common in case of animals than in plants, where $\sim$40% of the miRNA are found to be clustered. These cluster arrangements reflect distinct evolutionary style between plants and animals [24].

Additionally, structural analysis reveals that animal miRNA comprises highly variable size of stem and loop regions – these variations are mainly due to location of the independent transcriptional sites and the architecture of biogenesis (Figure 1.4). A reason for this may be the absence of Dicer processing from pri-miRNA to pre-miRNA in plants. However, several reports have suggested both plant and animals contain some degree of similarity in the primary structure of miRNA independent of their length. Site-specific nucleotide positions and the presence or absence of flanking regions confirms that common primary structures exist in both plant and animal miRNAs.

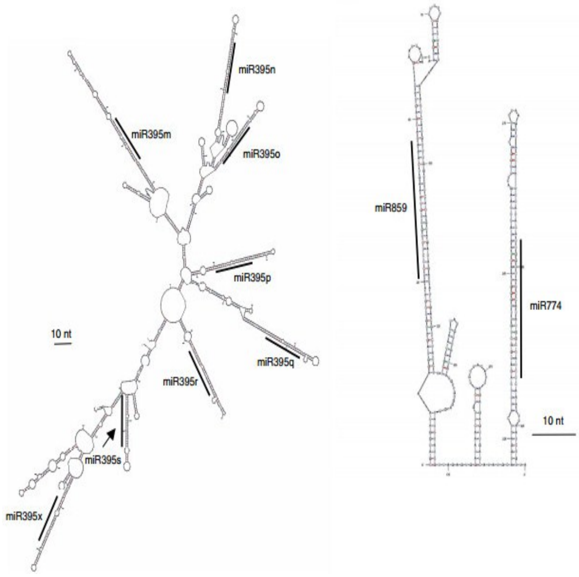


Figure 1.4: (a) Pre-miRNA cluster of *Oryza sativa* (Osa-MIR395) (b) *Arabidopsis thaliana* (Ath-MIR859-774). Mature miRNA are marked in bold lines.

1.4 MicroRNA and Cancer

MicroRNAs are known to be fine tuners of gene regulation; scientists have undertaken huge efforts in revealing the mechanism of these small non-coding RNAs (ncRNA) over the years. It is well known that these small ncRNAs are involved in various cellular pathways by interacting with their target mRNAs. In most cases, miRNAs combine with certain Transcriptional Factors (TFs) to alter the intensity of regulation. TFs are encoded by protein-coding genes and contain domains that participate in DNA binding, protein-protein interaction and transcriptional activation and repression.

Both TFs and miRNAs act in a highly coordinated fashion, the percentage of TFs coded by a protein-coding gene and the number of miRNA synthesised is entirely dependent on the organismal complexity. In case of binding, both TFs and miRNA bind to *cis*-regulatory elements and regulate gene expression. The jury is still out on the coordinated effects of TFs and miRNA. Only 5% of the protein-coding genes code for TF in flies and nematodes, whereas it is 10% in case of mouse and humans [25, 26]. Computational and experimental interaction data can be combined into functional network models to elucidate the system-level mechanisms of these gene regulators. However, miRNAs are known to function downstream of TFs, since their binding is possible only after mRNA is transcribed. Binding of TFs and miRNA to their targets plays a vital role in controlling the gene expression, particularly in development, apoptosis and several diseased states suggesting miRNA to be highly related with various stages of tumours [27] (Figure 1.5).

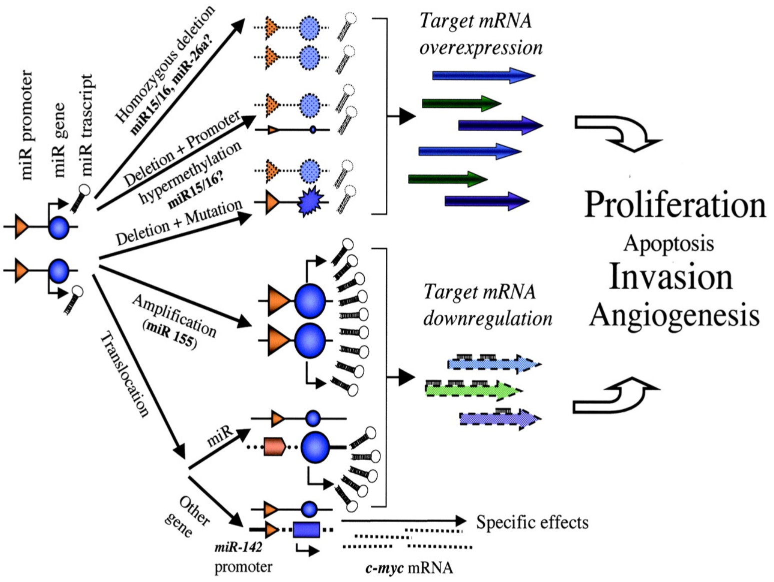


Figure 1.5: miRNA in cancer. Only experimentally validated interactions are shown. *miR-15/16a* involved in homozygous deletion in B-cell chronic lymphocytic leukemia and *miR143/miR-145* cluster down-regulation in colon cancer [28].

1.4.1 MicroRNA as Oncogenes and Tumour Suppressors

Cancer is a complex network of gene alteration, representing uncontrolled overgrowth of a particular cell type, initiated by either mutation or environmental factors. This initiation is followed by accumulation of defects in several other genes as a function of time progressively. Features related to cancer are loss of normal signals to stop proliferation, loss of signals for differentiation, sustained cell division and avoiding apoptosis. Inherently defective genes and environmental factors causing permanent damage to DNA are some of the main causes of cancer. These defective genes either sustain or increasingly mutate over year and finally cause cell specific cancer.

The significant role of miRNAs in gene expression has been verified in several tissue specific developments and in classification of tumours. However, there are no simple generic approaches in identifying or classifying new class of cancer with expression profile based studies. Gene expression is entirely based on the type of miRNA-mRNA interactions and they are variable. In human, miRNA do not bind with complete complementarity, rather they bind with several imperfections leading to different types of regulation. This kind of imperfect complementarity does not degrade the mRNA completely, but results in translational regulation with reduced protein levels. miRNA which cause increased gene expression can be loosely thought of as *oncogenes* and usually negatively inhibit tumour suppressors that are involved actively in cell differentiation and apoptosis. On the other hand, miRNA that down regulates tumour activity by negatively inhibiting oncogenes can broadly be termed *tumour suppressors* [21]. Infact, there are numerous instances which illustrate a single miRNA acting as both oncogene and tumour suppressor by involving in multiple molecular mechanisms and signaling pathways [29].

Most miRNA act as both oncogenes and tumour suppressors depending on the specific cell environment. In miRNA-mRNA interaction involving *miR-135b* and *miR-147*, there is down regulation in colorectal adenoma whereas, in carcinoma they are up regulated. Regulation of miRNA as oncogenes and tumour suppressors is mostly cooperative because of multiple binding in target sites. For example, *miR-155* is up regulated in Hodgkins lymphoma and targets genes such as ZIC3, AGTR1, ZNF537, FGF7 and IKBKE [30]. Other than the number of target sites in the binding region, features like local nucleotide context and regions around the target sites also contribute to the regulation significantly [23].

1.4.2 MicroRNA as Biomarkers

Being involved in several cellular functions like differentiation, development, metabolism and cell death, miRNAs are considered as powerful biomarkers. Tissue-specificity and distinctive signature profiles of miRNA in various cancers aid in sub-classifying cancer types and can be used as potential biomarker. However, considerable evidence suggest that within the cancer types, the expression profile varies during different stage of cancer and depends on the molecular mechanism involved [31]. Generally, miRNA have distinctive expression levels in normal and cancerous cells depending on the cell types they are associated with. In 2008, Rosenfeld *et al.* utilised distinctive expression profiles of 48 miRNAs in 22 different cancer types to associate miRNA with specific tissue types and provided some rudimentary tumour taxonomy [32]. However, the study failed to analyze cancers with unknown origin. Till date, this classification can account for only 2 to 5% of cancers of unknown origin making the diagnosis and treatment for these cancer types unclear. Several other groups utilised DNA microarray and gene signature to identify the origin of unknown cancer types [33] . Based on their studies, an unknown cancer type was classified with 80% accuracy based on site of origin and 73% accuracy for tissue origin [34].

The choice of profiling methods is also important in identifying potential biomarkers in cancer; mostly because they depend on the abundance of miRNA isolated. Quantification methods include *in situ* hybridisation, northern blotting, qRT-PCR, microarray, high-throughput sequencing and bead-based arrays [35]. However, these methods are laborious and are not appropriate for the fast prediction of involvement in cancers of unknown origin. For example, in the qPCR and sequencing method, large volumes of sample are required and runtime is higher. Similarly, in case of Next Generation Sequencing (NGS), the methods require larger volume of miRNA sample and are inappropriate for miRNA with low abundance. Thus, the choice of miRNA profiling methods depends on the specificity and sensitivity of expression profile required. Computational methods along with various gene regulatory network information can be utilised to identify miRNA, which could then be used as an efficient candidate for cancer biomarkers.

1.5 Tools for MicroRNA Analysis

Mirnomics, the study of miRNA structure and functional characterization of miRNA is gaining importance due to the involvement of miRNAs in several diseases. With the upsurge in the number of discovered miRNAs, the need for cataloguing and annotation of the miRNA is the need of the hour. Specialised databases are being constructed for categorizing sequential, structural and functional aspects of miRNA. Further, to process different user requests with the database, several tools have been built alongside these specialised databases, thus laying a fundamental support for identification, target prediction and functional analysis of novel miRNA.

1.5.1 MicroRNA Databases - miRBase

miRBase is a sequential database with searchable miRNA entries; they include both novel and computationally predicted miRNAs. miRBase follows a unique identifier and nomenclature for naming miRNA identified from different organism. miRBase includes both mature and stem loop sequence entries in it. Usually each new novel miRNA or predicted entry is given a sequential number as per submission (*e.g. miR-181* is followed by *miR-182*). Usually the identifiers are denoted by *hsa-miR-181*, where the first three letters signify the organism and the next three letters signify mature or stem loop sequence. When the 'miR' is denoted as 'mir', as in hsa-mir-181, it indicates the stem loop sequence.

Identical miRNA sequences from different precursors will be assigned the subscripts 1, 2 and so on (*hsa-miR-181-1* and *hsa-miR-181-2*). Alphabets at the end of the unique name indicate closely related mature sequences (*hsa-miR-181a, hsa-miR-181b).* In some cases, both the miRNA duplex sequence entries are cataloged, which will be denoted as *miR-181* and *miR-181**

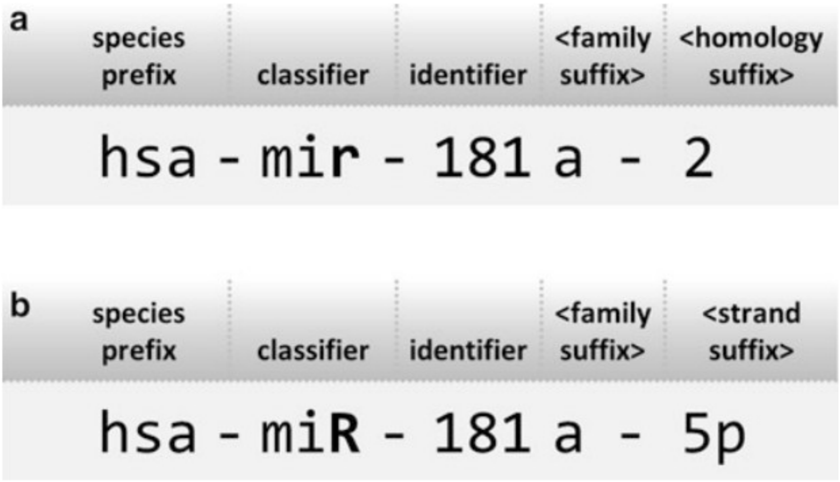


Figure 1.6: Naming nomenclature for **(a)** step loop and **(b)** mature miRNA sequence [36].

indicating predominant product and product obtained from opposite arm. Naming convention vary with the organism, for example plant miR are denoted with capitalised letters (MIR-181). A similar pattern is also followed in case of viral miRs. miRBase also have exceptional naming convention in case of *let-7* and *lin-4*, because these names are retained for historical reasons. Homologous sequences of these miRs will also attain the same naming convention [37].

With the advent of new computational prediction methods, the number of miRNA entries has grown exponentially year-by-year. Currently, miRBase *21.0* holds a total high confidence* entry of 28645 miRNA from human, mouse, fly, nematodes and *Arabidopsis*. MiRBase also provide links to experimental evidences and references to published results, which further enhance the confidence of predicted miRNA. Certain other repositories also contain some miRNA sequences, but are not exclusively dedicated to miRNA. Ensemble and National Center for Biotechnology Information (NCBI) Genbank contain annotation including phylogeny, gene and transcript information and splice variants. They provide annotation tools and hyperlinks to several other miRNA resources and several output formats including FASTA format for further analysis (Figure 1.7).

*High confidence sequences must either have atleast 10 reads mapping to each arm or have at least 5 reads mapping to each arm and at least 100 reads mapping in total using multiple deep sequencing data

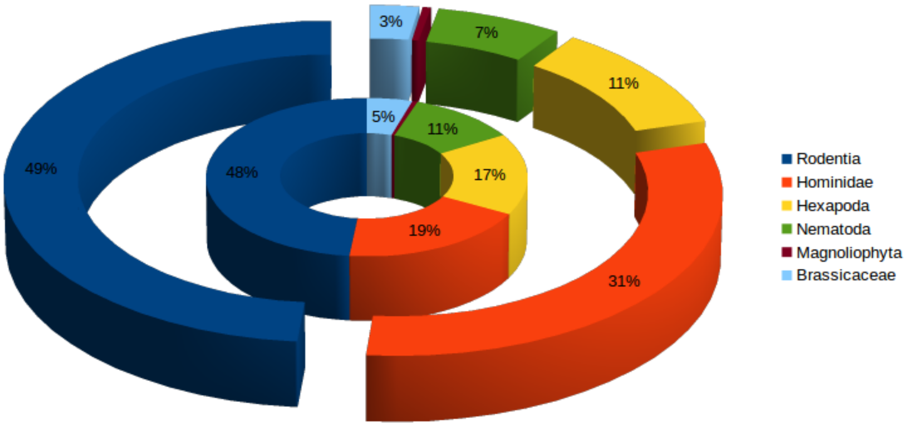


Figure 1.7: Categorical representation of miRBase entries [37]. The inner cirlce represents the pre-miRNAs and outer circle represents mature miRNAs.

Tools like *miRDeep*[7] predict novel miRNA from NGS data – they do so by searching for homologous sequences and identifying them. The identified homologous sequences obtained from precursor sequences are then folded into hairpin structures using RNA folding algorithm from Vienna package [38] and predicted sequences are further confirmed with PCR quantitatively.

Lately, some computational algorithms have been constructed and utilised effectively to identify miRNA coding gene in sequenced organisms. Regardless of the complexity within the organism, they consider hairpin secondary structure as their main target for miRNA gene identification. RNAfold [38] – a secondary structure-folding algorithm in Vienna package, folds pre-miRNA sequences into hairpin structure based on sequence complementarity and thermodynamics of folding. Structures are ranked based on a scoring system and/or the stability and thermodynamics of folding. Finally, top ranking structures are confirmed with *in vivo* experiments.

For the last decade, Machine Learning (ML) approaches are being used effectively for identifying miRNA gene even with low sequence homology. ML techniques utilise information from both experiments and knowledge obtained from previous identification procedures. The process starts with learning features being extracted from experimentally identified miRNA gene and followed by effective classification of real and pseudo miRNA structure. The classification accuracy of the entire process is dependent on the knowledge obtained from experimental procedure. With the vast amount of human genome data and the complexity involved, the miRNA gene identification method still needs improvements to make it accurate.

1.5.2 Target Prediction Tools and Databases

Predicting miRNA target is the primary step in functional analysis, involving the binding of mRNA with miRNA conjugated in RISC. The number of miRNA identified has grown exponentially with every new release of miRBase, considered as the primary database for miRNA-related studies. However, targets for several newly predicted miRNAs have not yet been identified and require more reliable and faster methods. Identification of targets in plants and animals vary significantly – in case of plants, sequence complementarity is the primary identifier for target binding since most plant-based miRNA-mRNA interactions show exact complementarity. On the other hand, animal miRNAs do not bind with perfect complementarity. Numerous algorithms have been proposed considering all the crucial parameters, but still they require improvement since the binding nature varies significantly with cell types and also based on the environment.

miRNA target multiple binding sites on mRNA, but not all binding would be regulatory. In human, most sites are conserved regulatory targets and additional regulatory function occurs through binding to non-conserved sites [39]. Another important factor to be considered is the UTR context, *i.e.,* at least not less than 6*nt* binding in the $3'$ UTR and also AU-richness in the particular region [23]. Cooperative effect due to multiple binding is also dependent on the distance between the binding sites [22]. Targets with rich A+U regions are observed to be more regulatory in nature [23]. TargetScan, a popular algorithm for target binding, considers all these parameters. The outcome of the prediction is also supported by experimental results (if available). The algorithm ranks all the miRNA-mRNA interactions based on total context scores, which is generally calculated from on local AU contribution, position contribution, target site abundance, seed pairing stability, site-type contribution and $3'$ pairing contribution. The tool considers $3'$ UTR as the most effective target-binding site than others.

Recent evidence also suggests that binding occurs at the $5'$ UTR and in the CDS as well [4, 5]. In case of miRNA binding to CDS, studies suggest that conserved sites occur in these regions and have been confirmed by comparing 700 human genes in 17 species [40]. Binding regions occur in CDS for certain miRs (*hsa-let-7a-5p, hsa-miR-9-5p, hsa-miR-125a-5p, and hsa-miR-153*). Experimental results also confirm that *let-7a-5p* down regulates Dicer, whose transcripts contain multiple target sites in CDS region for *let-7* binding. However, studies concerning binding stability in CDS region are scarce and more work needs to be done in this aspect.

Target prediction tools can be broadly categorised into sequence-based, structure/thermodynamics-based and homology-based predictors. Generally, most tools do not fit exactly into any single category because they consider at least two parameters to find an efficient target for the given miRNA. Tools like RNAhybrid (Vienna Package) consider both seed complementarity and thermodynamics of binding. Considering only sequence complementarity for target prediction may result in more non-functional miRNA target sites. Certain tools like mfold which utilise libraries from Vienna Package follow more stringent threshold values (*i.e.,* threshold values can be chosen based on the complexity of genome involved) in order to pick an accurate target. Most target prediction tools consider both sequence complementarity and thermodynamics of binding as the important parameters in predicting a miRNA target.

1.5.3 Filtering False Target Prediction

The approaches used for target prediction differ in the way they measure conservation, thermodynamics *etc.* Results from target prediction tools suffer from high false predictions; however, some filter algorithms have been successful in identifying true miRNA targets. Earlier reports of target filter algorithms employed machine learning approaches with experimentally validated miRNA target to obtain a high degree of accuracy. MiRTif (miRNA-mRNA interaction filter) utilised support vector machines to train models with real and pseudo-miRNA targets [41]. Only experimentally validated dataset was used for training and hence the tool could achieve a sensitivity and specificity of 83.59% and 73.68%, respectively. However, the tool did not consider any specific parameters from miRNA-mRNA interaction but learned from existing target interaction, thus resulting in poor performance with newly predicted interactions.

Recent developments in filter algorithms have been optimised for either seed conservation or thermodynamics of binding, but not for both. **P**rediction of **AC**cessible and/or **C**onserved **MI**croRNA **T**argets (PACCMIT), a target filter algorithm combines all three parameters (conservation, thermodynamics, or both) to obtain more precise miRNA targets [42]. RFMirTarget, a RF classifier based algorithm employs classification of real and pseudo-miRNA, based on the features extracted from structure, thermodynamics, conserved region and seed position and has shown consistently better results over other target prediction algorithms [43, 44].

1.5.4 Databases for miRNA-mRNA Interaction

With the growing number of target prediction tools, databases for storing the predicted and validated results are also important. Many target interaction databases have emerged using manual literature search for targets and based on high throughput screening techniques. miRTarBase 4.5, a database of experimentally validated miRNA-mRNA interactions provides 51,460 interactions for 1,232 miRNA with 17,520 target genes in 18 species. Over 2,636 published articles are linked with the validated result [45]. Similarly miRecords provides systematic and structured documentation of experimentally validated results, along with literature curation [46]. Additionally, miRecords provides information on the experiments used for validation and contains mainly animal miRNA entries. Certain databases like Argonaute [47] (upgraded to miRWalk [48]), use data mining techniques intensively for literature survey on mammalian miRNA and document them for easy access and referencing. Argonaute also collects validated information like miRNA-origin and families, tissue specific expression profiles and proposed function from other databases. Currently the database accumulates prediction from eight different target prediction tools, *viz.*, Diana-microT [49], miRanda [50], miRDB [51], PICTAR [52], PITA [53], RNA22 [52], RNAhybrid [38] and TargetScan [54].

1.5.5 Methods for Functional Analysis

miRNAs are post-transcriptional gene regulators and identifying expression profiles of miRNA reveal the molecular mechanism involved. Most functional analysis involved in miRNA introduce interference at the level of the miRNA or the target mRNA to restrict their interaction. This may result in the consequent loss of function if there is functional regulation associated. Once the interference is revoked, function is resumed and the relation of miRNA with target is confirmed.

Experimental approaches to validate miRNA-mRNA interaction include (i) Interference with miRNA levels by depleting pre-mirna followed by loss of function analysis with mature miRNA (ii) Blocking target site with anti-sense oligonucleotides and (iii) disruption of miRNA-mRNA base pairing through point mutation [55]. The major drawback with these methods is that they require a large amount of purified RNA and needs increased processing and handling. Recently, computational tools like MMpred has been used for predicting functional analysis [56]. The method requires miRNA-mRNA dataset and related microarray data and is successful in predicting true miRNA-mRNA pairs and approximate expression profile from microarray data.

1.6 Machine Learning in Biology

The term Machine Learning (ML) defines learning from existing data rather than following an explicit human instruction or a program. Machine learning has been employed for problems with complex relationships and provides a new insight into how input variables are mapped to output leading to pattern recognition or an effective classification process. ML starts with learning or a training process, followed by a construction of model, evaluation and finally optimisation of the constructed model for better performance (Figure 1.8). During the evaluation process, the constructed models are tested with novel inputs that are not a part of the training process.

Advancement in high throughput sequencing and other identification methods in biology have resulted deluge of data being generated. Introduction of ML in biology has offered a number of efficient solutions for (i) Annotating new genomic sequences, (ii) Functional analysis of macromolecules, (iii) Domain analysis, (iv) Target site identification in non-coding RNAs, (v) Biomarker identification and (vi) Genetic interaction networks. The initial use of machine learning (known as perceptron) was utilised in studying neuron behavior, and later Artificial Neural Network (ANN), one of the commonly used ML method was constructed based on neuron behavior. ANN was found to be efficiently used in identifying transcriptional start sites of microRNAs in *Escherichia coli (E.coli)* [57].

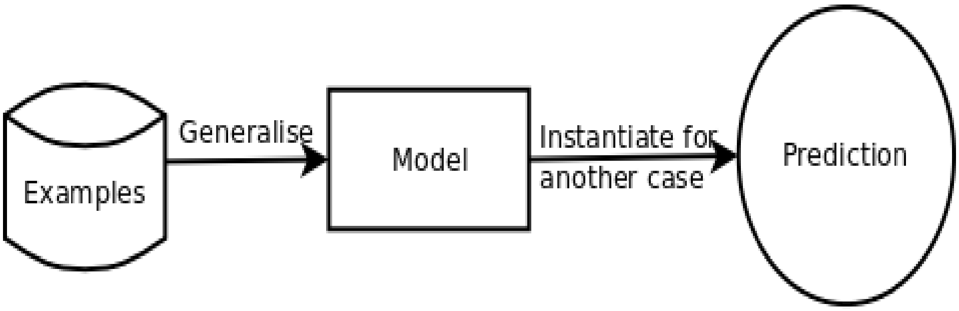


Figure 1.8: General pipeline of machine learning approaches.

The core objective of ML is to obtain generalization from the training process and to perform accurately on an unseen dataset. The construction of an efficient model depends on the training process, which in turn depends on the volume of the dataset. Training sets are stored in a feature-based format, i.e. converted to observable quantities best suitable for training purpose. These extracted features aid in mapping to the output in a much efficient way than learning from an uncategorised entire dataset. Only closely related feature set is considered for the training purpose because unrelated features may affect the entire training process and suppress performance. Selecting appropriate feature set is done iteratively and is considered an important process prior to training. The performance of the model is always more accurate with the most optimum feature set and with optimised algorithmic parameters. Fine tuning algorithm parameters is another critical task in obtaining a higher accuracy.

Additionally, based on the requirement of the training algorithm, data preprocessing is done for identifying empty variables; this is because, certain ML based algorithms do not identify empty variables and they may assign values randomly thus affecting performance. As a rule of ML (*no free lunch* theorem), there is no specific training algorithm available for a single problem. Mostly the selection of algorithm depends on the expected outcome or in certain cases depends on the complexity of the training dataset. Hence, in most cases, the choice of model is obtained by comparing several ML algorithms.

1.6.1 Algorithms in Machine Learning

Machine learning algorithms are broadly classified into (i) Supervised Learning and (ii) Unsupervised Learning algorithms.

1.6.1.1 Supervised vs Unsupervised Learning Algorithms

The choice of algorithm selection is entirely dependent on the type of input variable *i.e.,* either labeled or unlabeled dataset. In unsupervised learning, given a set of unlabeled input variables, the learning algorithm does not classify into individual classes; rather they cluster into groups. For example, in human eye, more than 106 photoreceptors are present. These photoreceptors learn on the basis of constantly changing environment and identify various parameters like light condition, object recognition, etc. In this example, photoreceptors can be assumed to learn through unsupervised methods. Neither there are prior labels nor is the feature set mapped to the output. Photoreceptors identify the object, cluster into groups and utilise them for future identification. In biology, unsupervised learning is used in problems where there is unavailability of feature sets to map input and output variables, *e.g.,* in gene regulatory networks and analysis of unknown gene expression. Additionally, in unsupervised learning, human interference is completely absent – hence bias on the output is completely eradicated and the output entirely is dependent on the algorithm used for clustering.

1.6.1.2 Performance Evaluation in Learning Algorithms

Class labeling is absent in unsupervised algorithms, therefore, they can be used for large datasets where feature mapping to variable is laborious. Direct evaluation of performance is not possible since there is no prior class labeling, rather evaluation is done based on the quality of the cluster and the cluster density– they are calculated based on the variance in the distance between the centroid and the actual data (silhouettes). Silhouettes [58] are graphical representations of how data is distributed within the clustering and are represented as the Silhouette scores. The score describes the location of each data point, whether they are located well within the cluster or in between two/several clusters.

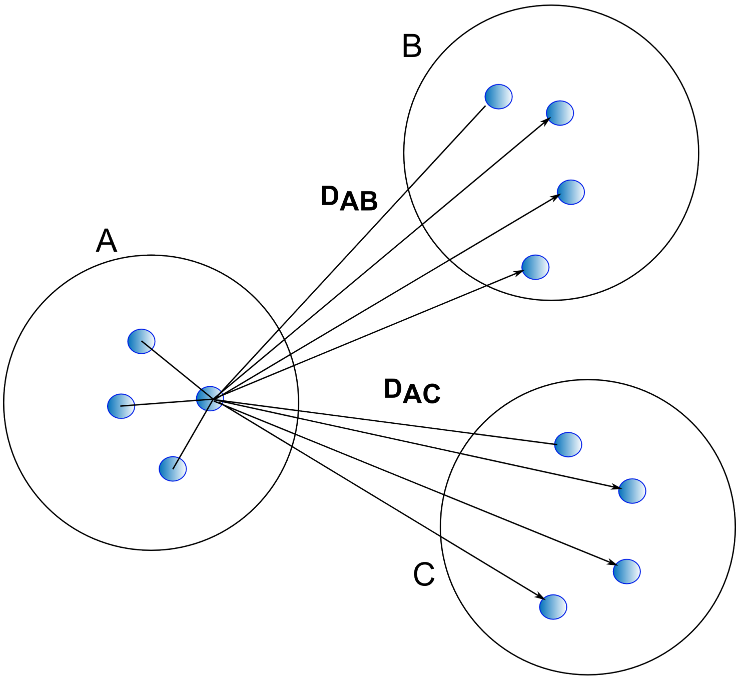


Figure 1.9: Calculation of silhouette score between three cluster A, B and C, where D_{AB} and D_{AC} is the distance between the point of interest from one cluster to the other.

Silhouette score is expressed in ratio scale as that of Euclidean distance. Let us consider three clusters (A, B and C) obtained after unsupervised learning. In figure 1.9, the distance of a point of interest from cluster A to B is denoted by D_{AB}. Silhouette scores start with calculation of average distance from all the points from cluster A to B and are denoted as AvgD_{AB}. Similarly, scores are calculated from all points in the cluster to the other clusters. If the ratio obtained between Cluster A and B is minimum, then B is referred as the neighbour of Cluster A, which indicates that data point may appear in Cluster A or B. The ratio obtained between the average distances calculated between the clusters is referred to as dissimilarity. If a cluster contains only a single point then the ratio is set to zero.

For example in figure 1.10, an unsupervised learning involving identification/clustering (*k*-mean algorithm) of species based on structure related dataset is emphasised. The initial clustering starts with defining the number of clusters (user-specific) and also the centroid of each cluster. The number of clusters is equal to the number of species classification required and the choice of centroid is usually a random process. Assigning closest centroid is dependent on whether the species appear closely related to each other. Once the number of clusters and closest centroid are assigned, the process starts. The clusters obtained in the initial step are recomputed with different centroid values iteratively until they obtain identical results.

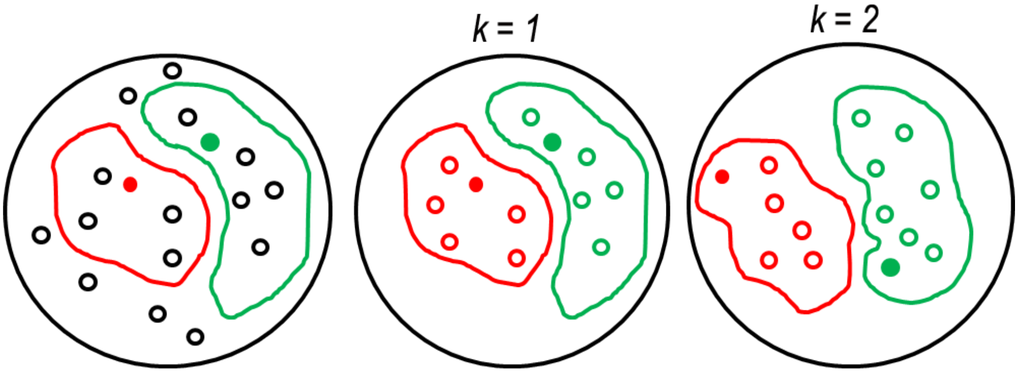


Figure 1.10: Steps in k-mean algorithm; empty dots indicate training set, dots filled with green and red are centroids for respective clusters. If we consider the number of clustering step as *(k= 2)*; then in *(k =1)*, centroid is arbitrarily assigned for clusters in red and green. Clustering is done based on the closeness to the centroid. In consecutive steps *(k=2)*, centroid is moved to a different point and clustering is done until stable clusters are obtained.

Hierarchical clustering is applied in cases of closely related clusters. The output of hierarchical clustering is usually a dendrogram and the process of clustering is terminated at any point of time based on the user request. Hierarchical clustering is usually employed to cluster closely related species of unknown origin. Commonly used unsupervised algorithms are *k*-mean algorithm, Neural Network and Hidden Markov Model (HMM). The main disadvantage of unsupervised algorithms is they are sensitive to the number of clusters assigned.

On the other hand, in a supervised algorithm, each training instance is mapped with known output values. Training process results in either a regression function (if the output is continuous) or a discrete model (a class classifier) with a well-defined feature set. Performance evaluation is carried out only with novel inputs and not with the dataset used in the training. Identification of tumour subtypes from gene expression profile is a classic example of supervised learning, where the input variables are mapped with the known set of tumour types (output variable).

In certain cases, the origin of the tumour is completely unknown, hence classification is done with semi-supervised learning. These algorithms make use of certain available labels (supervised learning) and use mostly unlabeled dataset (unsupervised learning). The main disadvantage of using semi-supervised learning algorithm, particularly in biology, is that they consider a hypothesis, which is not suitable for biological problems with large unlabeled dataset [59].

Classifiers are generally binary, but the need for multi-class classifier has also grown due to the complexity of the output. For example, the identification of pre-mirna as real or pseudo is a typical problem solved by a binary classifier. However in the case of tumour classification, where multiple subtypes are involved, a binary classifier is not a suitable choice. ML algorithms like decision trees can handle multi-class classification more efficiently than the kernel-based methods. Recent advance in Support Vector Machines have resulted in multi-class classification, where a multi-class problem is divided into several binary classifiers and results are accumulated either by averaging or by majority vote.

1.6.1.3 Support Vector Machines (SVM)

Support vector machines [60], as formulated by Vapnik in 1992 is a kernel-based binary classifier and used widely in biology for high dimensional data analysis, *viz.,* gene expression analysis and complex species classification. SVM is applied to problems with unknown distribution whose class boundary is quite unpredictable. To overcome the problem of class boundaries, SVM uses kernel functions that compute a dot product of the data points and map them to higher dimensional space. Construction of SVM is based on the type of dataset used in the training process, *i.e.,* either linear or non-linear dataset. For example, in tumour classification, the problem is usually not linearly separable, because most tumours share some common properties. Hence, a non-linear kernel function is generally preferred to map the feature representation into higher dimensional space where they are linearly separable. The choice of kernel function plays a critical role in the classification process, which in turn depends on the optimised kernel parameter, gamma *(γ)* and the soft margin parameter, cost *(c)*. Commonly used kernel functions are linear, Radial Basis Function (RBF) and polynomial.

Most real world datasets are not separated by a single linear hyperplane even after converting to higher dimensional space. Some data points may fall just on the hyperplane or in close proximity within the hyperplane identified. Generally, SVMs rely mostly on maximizing margin[†] and minimizing classification errors to select the best hyperplane. Maximizing the margin may accommodate data points that are close to the support vector. Therefore, SVM is also called maximum margin classifier. Additionally, there are situations when data points are misclassified by a linear hyperplane. The concept of soft margin is introduced in such cases specifying a trade-off between hyperplane violations (slack variable, ξ) and the size of the margin *(b)*. For a non-linear problem (Figure 1.11), with soft margin identified, the dot product of weight vector *(w)* with the training set is replaced by kernel function (ϕ), which

[†] A margin is defined as the smallest distance from the decision boundary, on which the data points (called as support vectors) are located.

allows mapping of each non-separable data point from two-dimensional space to higher dimensional space, making them linearly separable.

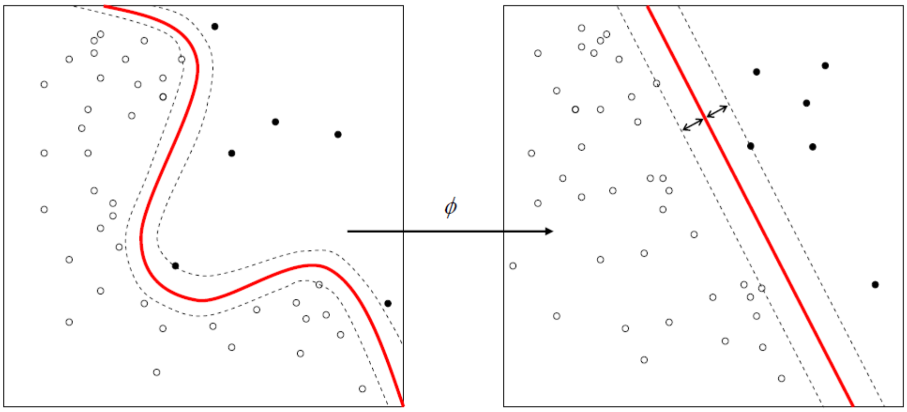


Figure 1.11: Applying kernel function for non-linear classification – mapping from low dimensional to higher dimensional space.

In an example of a linear case, where a simple hyperplane classifies ovarian cancer and normal tissue, the discriminant function *f(x)* can be formulated as

$$f(x) = w.x + b = 0 \tag{1.1}$$

where *w* is the weight vector and *b* is the bias. However, in case of large margin, the discriminant function is formulated as

$$Class1, \quad f(x) = w.x(i) + b \geq 1 \tag{1.2}$$

$$Class2, \quad f(x) = w.x(i) + b \leq 1 \tag{1.3}$$

For non-linear cases, soft margins are introduced with a bias in the classification, thereby allowing misclassification of a point that does not fall within the class boundary. Slack variable (ξ) (also called as the margin error) allows to overcome the misclassification and is defined as, when (ξ>0), data point is on the margin and if (ξ<0) then it is misclassified. After determining the soft margin, dot product of the support vector with the kernel function (ϕ) aids in constructing a maximum margin with minimum error in a higher dimensional space (Figure 1.12). Choice of suitable kernel and optimising kernel parameters is also an important process in obtaining lower misclassification error. Kernel

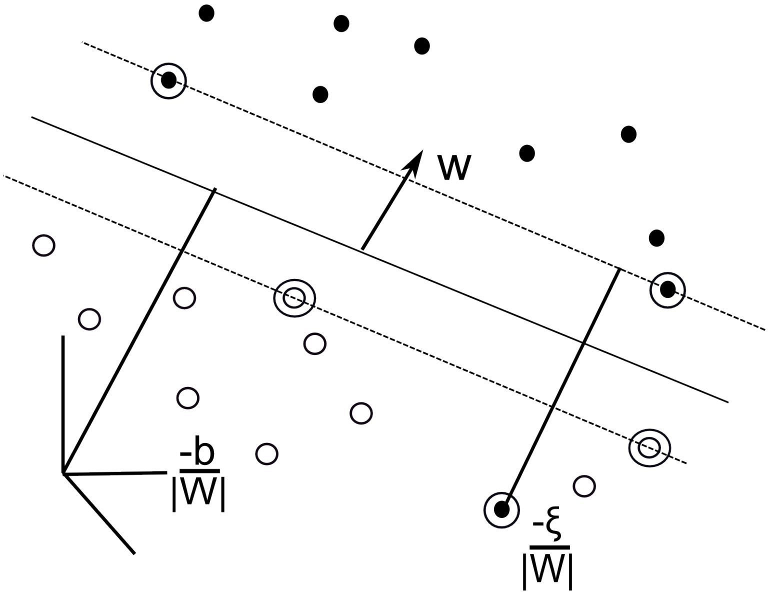


Figure 1.12: Hyperplane construction with slack variable (ξ) in higher dimensional space. ξ<0 indicates misclassified data points, ξ>0 indicates data points on the decision boundaries and *w, b* indicates weight vector and the distance between the margins of hyperplane respectively [61].

parameters along with the soft margin parameters determine the flexibility of the SVM boundary in fitting the dataset.

1.6.1.4 Decision Trees – Random Forest and C4.5

Decision trees (DT) are ensemble based methods involving the construction of multiple decision trees during the training process and output is either obtained by bagging [43] or majority voting [62]. Information entropy of individual features extracted from training set plays a crucial role in constructing an efficient tree. A decision tree has a structure consisting of internal nodes (where a decision function is executed), external nodes, connected by branches and finally end nodes or leaves populated with class labels. Trees are usually constructed in a bottom-up approach, where a feature is selected at each node and bifurcated into branches (starting at root and splitting until leaf node). The output of a decision tree represents rules, used in any knowledge system to predict new inputs. Decision trees can handle both numerical and categorical datasets and provide a clear indication of variable importance in prediction or classification. Generalization error during model construction is completely reduced as the number of trees increases. Generalization error in decision tree-based algorithms is however dependent on the strength of the individual tree and correlation between the trees. The main advantage of decision tree is the ability to handle multi-class classification in an efficient way than the kernel-based method.

Most popular decision tree algorithms are Random Forest (RF) [43] and C4.5 [63]. Both algorithms work in a similar fashion except in the final decision process. Random Forest algorithm involves two major steps: (i) random selection of features and (ii) bagging. Given a training data of size N and with feature subset F_n, RF starts with the random selection of features (M) based on information entropy ($M \ll F_n$). The number of randomly selected features for each tree construction is kept constant. During the actual training process, a bootstrap selection of sample (with replacement) is done followed by the propagation of the tree. Each tree is constructed based on the information entropy of the feature selected. The depth of each tree and number of features considered for tree construction (log M/log 2, where M is the number of features selected) are optimised for best performance or kept as user-specific. Generally, no pruning is done in Random Forest method – thus, when a test set is supplied the decisions are obtained by majority voting from different trees. Error estimates in RF are carried out by Out-Of-Bag error (OOB) [43][‡]

C4.5 is an improvement of the previously used Interactive Dichotomiser 3 (ID3) developed by Quinlan owing to the sensitivity of ID3 to features with a large population of values [63]. In C4.5, this limitation is accounted for by pruning the outliers and achieving a hypothesis with higher accuracy. Tree construction is similar to that of RF, depending entirely on the information gain of the individual features. C4.5 works in two phases: (i) Growth phase, in which the dataset is split into several small clusters segregated on the basis of several attributes. Tree construction starts with the feature having highest normalised information gain (to be labeled the root), and all the other features are sublisted as aids in the splitting process. Each node holds the criterion for the splitting and the leaves are populated with the labels (ii) Pruning phase, which

[‡] Out-of-bag error: Each tree is trained on about 2/3 of the total training data. As the forest is built, each tree can thus be tested (similar to leave one out cross validation) on the samples not used in building that tree. This is the out of bag error estimate - an internal error estimate of a RF as it is being constructed.

involves generalization of fully-grown trees and removing the outlier data. C4.5 can handle continuous attributes by portioning values into discrete set of intervals (called as discretization) but for most Decision Tree (DT) algorithms, categorical attributes are the prerequisite for tree construction.

Over fitting of data (propagating much deeper into tree in a sense to obtain perfect classification) is the main disadvantage with most DT algorithms. Over fitting arises due to uncommon characteristics among the selected features resulting in empty and insignificant branches. This is common in RF algorithm and results in high variance trees with low prediction accuracy. RF entirely depends on the dataset and the selected features during the propagation and even one removal/addition of feature may change the entire prediction of the tree. C4.5 avoids over fitting by pruning – initially the trees are allowed to propagate until they reach a maximal point of perfect classification and then a post-pruning step removes all the outliers and noise to obtain the best tree [64]. Both decision trees can handle missing attributes – in RF, it is either done with nearest neighbour imputation or mean substitution; in C4.5, probability values are used rather than assigning existing most common value for that attribute. Both DT algorithms can handle large datasets efficiently by allowing parallelization for faster computation.

1.6.1.5 Ensemble Methods

Ensemble methods aggregate predictions of multiple classifiers with the goal of improving accuracy. Predictions from multiple classifiers are pooled together either by weighted or unweighted voting. Initial construction of ensembles involved Bayesian averaging to combine the predictions, though bagging [65] and boosting [66] techniques are being used recently. The performance of an ensemble method is usually higher than that of a single model trained from the entire dataset. In fact, prediction from a single classifier often contains prediction errors (if the training dataset contains inequal distribution of instances), which can be totally removed by ensemble construction.

Ensemble methods provide improved flexibility and accuracy in prediction. Construction of ensembles involves (i) building individual models from different learning algorithms and performance optimisation

(ii) combination of constructed models to form an ensemble and the result is obtained by either bagging or boosting (Figure 1.13). Boosting improves model with high dimensional predictors, whereas bagging is prominent in improving tree-based algorithms.

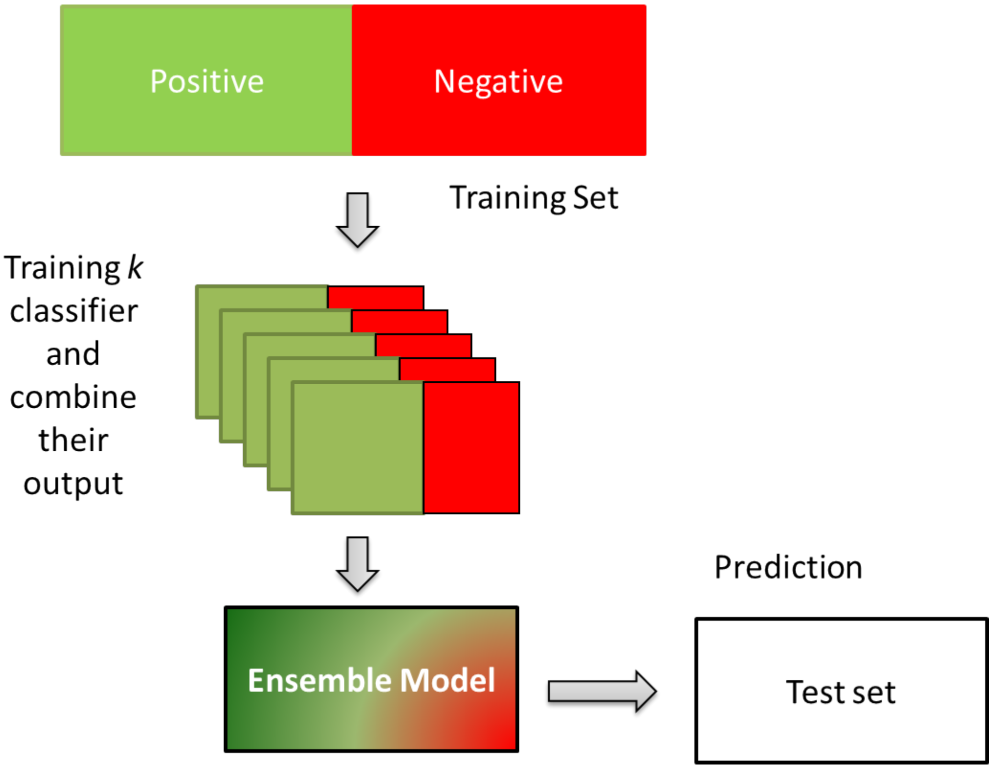


Figure 1.13: Framework of ensemble classifier. Ensemble involves training dataset simultaneously with k-classifiers. Results obtained from classifier are aggregated either by averaging or by voting.

Bagging (**B**oostrap **Agg**regation) and boosting are useful techniques to improve the predictive performance of models (Figure 1.14 (A) & (B)).

Creation of a bagging ensemble involves:

- Construction of bootstrap set of size N with replacement from training set M (N $\ll$ M).
- The training observations that are not chosen in a specific bootstrap set are referred to as Out-Of-Bag (OOB) observations, N^T. Each base learner can report errors on the OOB observations.

- Creation of base model T from each of the bootstrap training sets. Each base model is allowed to make prediction.

- Majority vote is taken as the final prediction from the base model's prediction.

In boosting (Adapt[at]ive Resampling and Combining), a set of weak learners are combined to perform as a strong learner. In the present context, a weak learner does not imply a learning algorithm with poor performance, rather they denote algorithms with prediction performance slightly more than random guessing. Different types of boosting algorithms are available *viz.,* adaboost, logitboost, rankboost, coboost etc. These algorithms differ only on the way they weigh each instance during training and hypotheses generation. Creating a boosting algorithm involves the following steps:

- Drawing a bootstrap training set D_m from training set D according to the weight w_i

- Generation of a classifier C_m using training set D_m

- Measurement of error of C_m on D.

    · For next iteration: Increasing weights for misclassified training points and decreasing weights of correctly classified points

- Iteration is continued until (C$_{boost}$ = ΣC_m) has low error.

Overall classification is given by

$$C_{boost}(X) = \pm(\Sigma_m \alpha_m C_m(X)) \tag{1.4}$$

where α is the measure of quality of classifier C$_m$. Boosting algorithms do not overfit and are highly sensitive to outlier/noise.

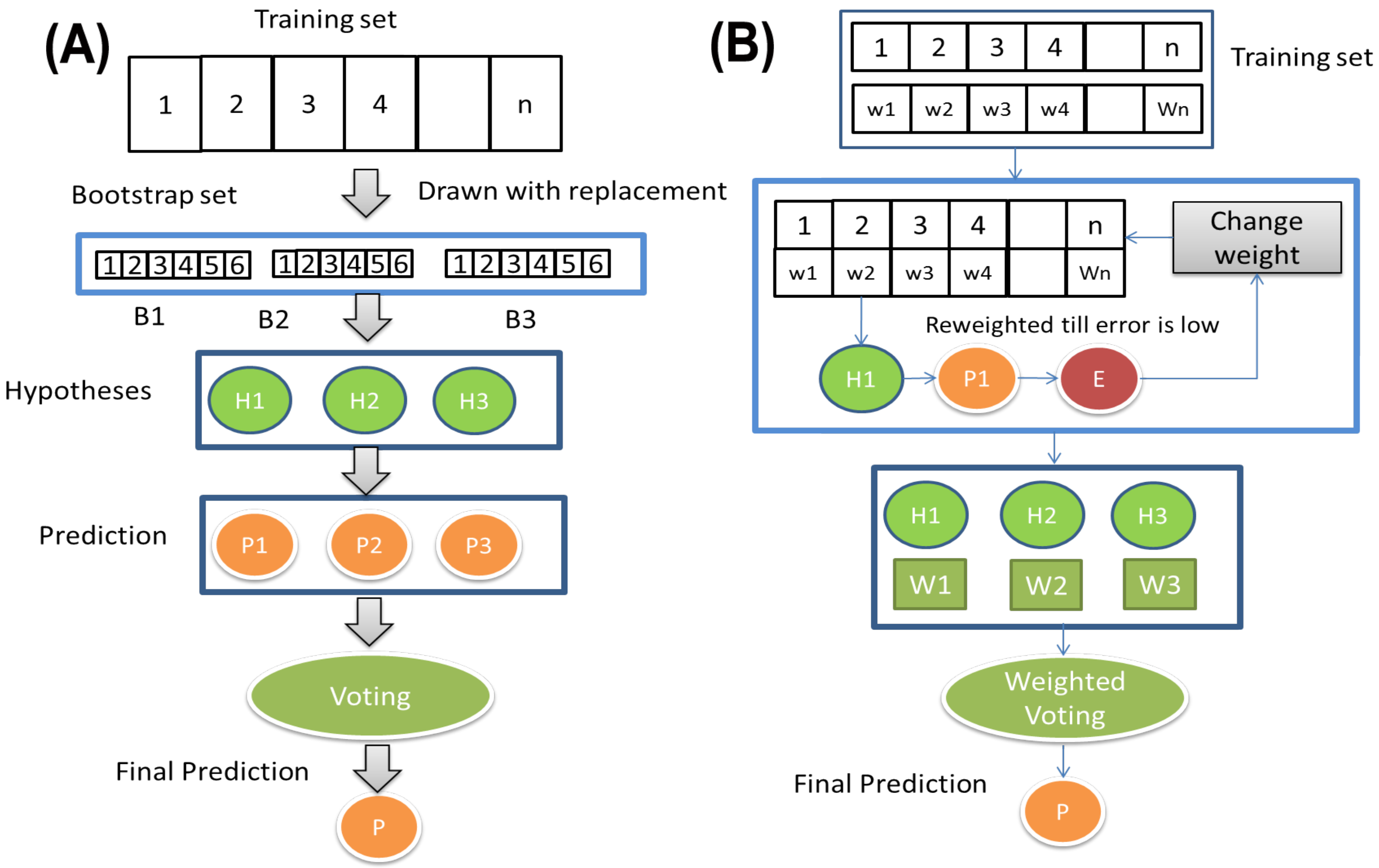


Figure 1.14: Steps involved in bagging and boosting algorithms. (A) Bagging involves drawing bootstrap set from training set with replacement (B1, B2 and B3) and then constructing hypotheses (H1, H2 and H3). Predictions (P1, P2 and P3) are combined by voting and final prediction is obtained. (B) Boosting involves drawing bootstrap samples with weights. Individual models with predetermined weights are used for final prediction.

1.6.2 Imbalance in Dataset

A dataset is imbalanced if the classification categories are not equally represented [67]. Class imbalance or skewed dataset mainly arises when most of the instances are labeled as one class (majority class), while very few are labeled as the other class (minority class). Traditional classifiers utilising the entire training set for prediction are not suitable to deal with imbalanced datasets because they show bias towards the majority class due to over-prevalence.

Machine learning algorithms are evaluated mostly based on the prediction accuracy but in imbalanced datasets, accuracy cannot be an absolute measure of error. For example, prediction of cancer from mammography image dataset [68] which contains 98% of the normal pixel and 2% of the abnormal pixels. Measuring prediction accuracy of the trained classifier using a dataset like this will result in 98% accuracy since the dataset is already skewed towards the majority class. In order to achieve good performance for a classifier with an imbalanced dataset, a high degree of error correction may be incorporated in the minority class and fewer corrections in the majority class.

The main problem in training a classifier with an imbalanced dataset is that the minority class is often considered a noisy dataset and hence suppressed or overlooked by the majority class. Class imbalance in the dataset deteriorates the performance of a classifier. To overcome the problem of imbalance in dataset, machine learning algorithms utilise two major methods, *viz.,* (i) assigning cost to the training set and (ii) re-sampling the training set, *i.e.,* either undersampling the majority class and/or oversampling the minority class (Figure 1.15). These resampling methods work at the data level, hence the choice of method is entirely data driven. Class imbalance is ignored at the algorithm level by certain methods *viz.,* adjusting the cost of the classes to counter imbalance, adjusting the probabilistic estimates (in case of decision trees) and adjusting decision threshold (in case of one class). In certain situations, both resampling and cost based methods are used in combination, *i.e.,* individual models are adjusted with these methods and combined as an ensemble to provide better performance.

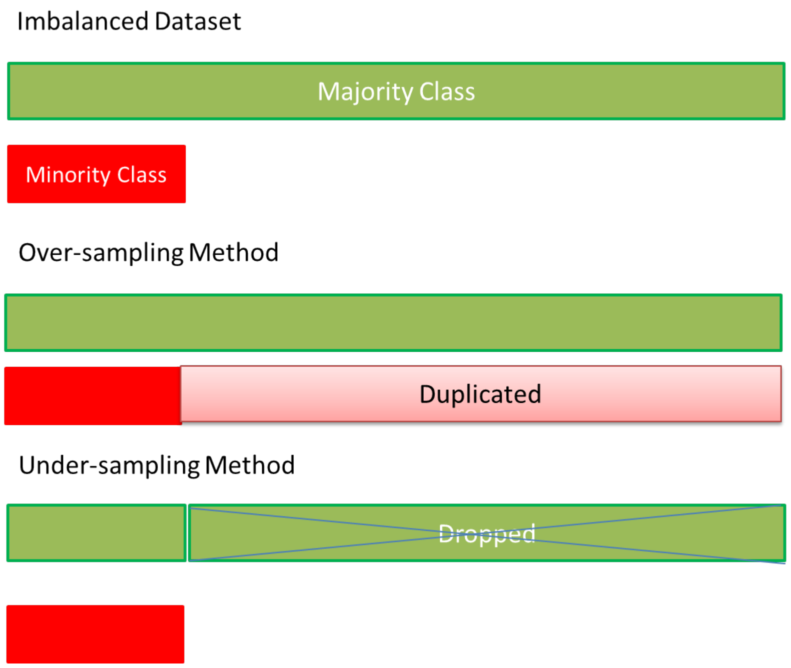


Figure 1.15: Over-sampling and undersampling to overcome imbalance in dataset. In over-sampling methods, minority class is over-sampled to generate duplicates and in under-sampling the instances are reduced in majority class or only a subset of the training dataset is used.

1.6.2.1 Oversampling and Undersampling Methods

Over-sampling methods balance a training dataset by increasing the number of minority class data points, while under-sampling methods balance a training class by decreasing the number of majority class data points. The most common method in over-sampling is **S**ynthetic **M**inority **O**versampling **TE**chnique (SMOTE), in which the minority class is over-sampled with synthetic samples as proposed by Chawla [69] and Japkowicz [70]. SMOTE centers more on a specific region in the feature space as the decision region for the minority class, than increasing the overall number of instances. New instances are obtained by nearest neighbour method – the number of neighbours (X_i) are chosen depending on the number of data points required.

Let us consider the mammography dataset in the previous example, where the outcome of the study is a decision tree-based model. Majority class samples are shown by *blue squares* and the minority class samples are shown by *red circles* in the Figure 1.16. Decision region chosen for resampling minority class

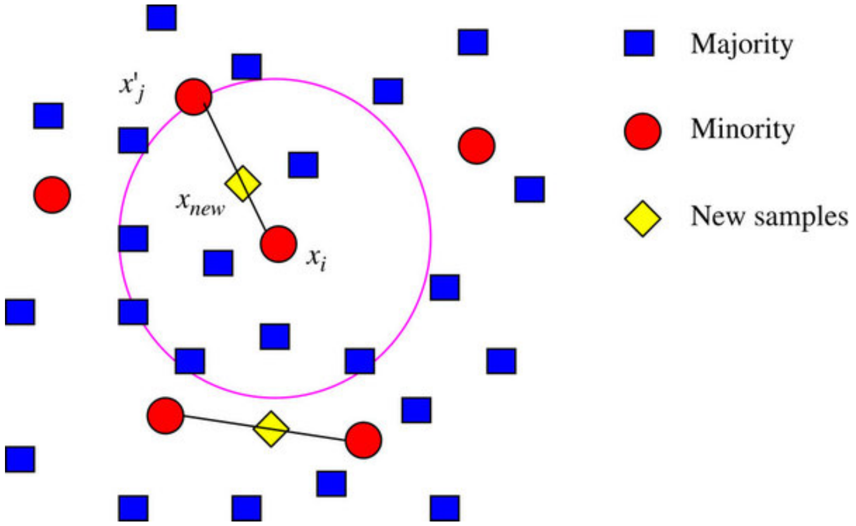


Figure 1.16: Generating synthetic samples using SMOTE [71].

is marked with a pink circle. Replicating the minority class in this area by choosing the perfect nearest neighbour will result in more terminal nodes in the final decision tree. The main disadvantage of this resampling method is that it is sensitive to over-fitting because random samples are generated. On the other hand, under-sampling methods utilise a subset of majority class to train the classifier. Since only a part of the training set is utilised, the dataset is highly balanced and the computation is faster than over-sampling methods. Undersampling methods though ignore a large part of the training set thereby making such methods vulnerable to miss many discriminative features present in the ignored parts. To overcome this deficiency in under-sampling methods, methods like easy ensemble and balance cascade as proposed by Liu *et al.,* [72] are widely used.

Easy ensemble is an unsupervised learning strategy and uses random sampling with replacement, whereas supervised learning is applied in balance cascade method. Both these methods use adaboost algorithm to train several weak classifiers and combine them into a single ensemble.

In the Figure 1.17, an imbalanced training dataset (T) containing positive (P) and negative (N) sample is considered, where (P $\gg$ N). Random sampling of N_i from P is done such that (N_i = N). For each subset, classifier H_i is trained until (i = T). Finally instead of collecting votes from the weak

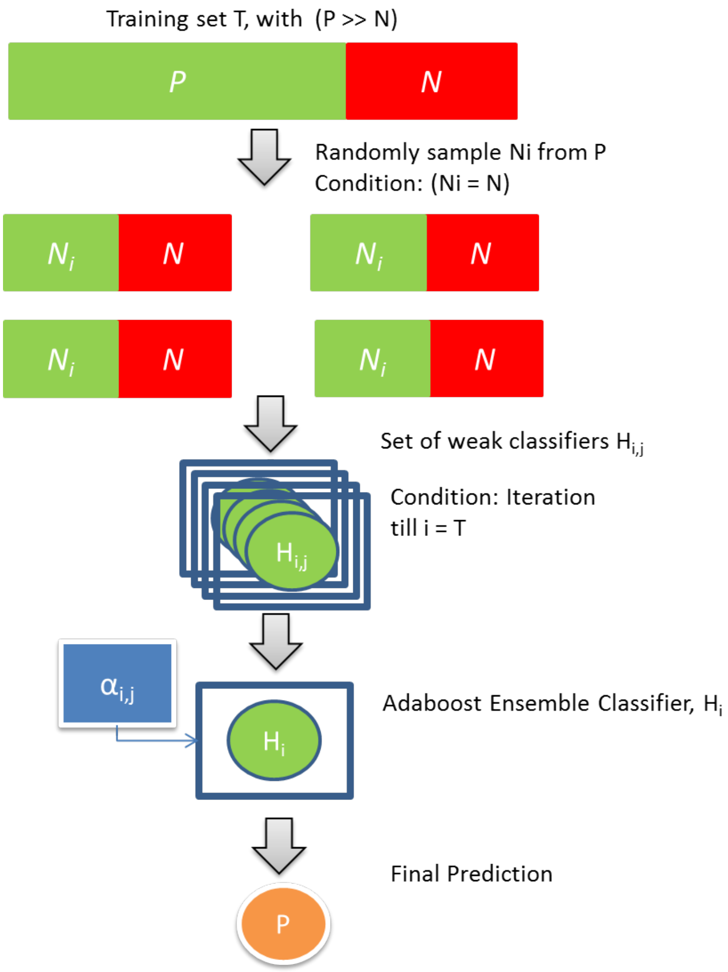

Training set T, with (P >> N)
P
N
Randomly sample Ni from P
Condition: (Ni = N)
N_i
N
N_i
N
N_i
N
N_i
N
Set of weak classifiers H_{i,j}
Condition: Iteration
till i = T
H_{i,j}
α_{i,j}
H_i
Adaboost Ensemble Classifier, H_i
Final Prediction
P

Figure 1.17: Easy-ensemble method for imbalanced dataset.

classifiers (H$_{i,j}$), features are collected and an adaboost ensemble classifier (S$_i$) is constructed. Hence, the final prediction using easy ensemble is given by

$$H(x) = \pm(\sum_{i=1}^{T}\sum_{i=1}^{S_i} \alpha_{i,j}\, h_{i,j}(x) - \sum_{i=1}^{T}\theta_i) \qquad (1.5)$$

where θ is the threshold of the ensemble generated.

Undersampling in balance cascade is carried out in a similar fashion, with the exception being the removal of correctly classified N$_i$ in each iteration. This is followed till a state is reached where the majority class is classified to be the minority class (achieved by calculating false positive rate at each iteration).

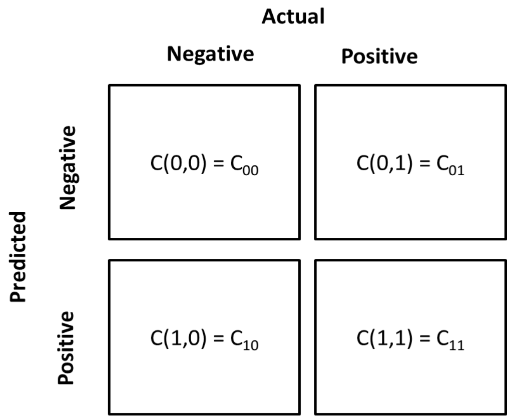
False positive rate is given as

$$f = (T - 1)\sqrt{\frac{|P|}{|N|}} \tag{1.6}$$

The only disadvantage with the balance cascade method is the computation time which is a little on the higher side. Additionally, the method is likely to suffer from over-fitting since the number of minority samples are limited in each iteration.

1.6.2.2 Cost-sensitive Methods

In most learning algorithms, there is an attempt to minimise error rate in classification by ignoring the difference between types of misclassification errors. However in real world problems, this assumption does not hold true. For example, let us consider a cancer diagnosis containing an imbalanced dataset of 99.5% positive instances (cancerous) and 0.5% negative instance (healthy or non-cancerous). When the classification is carried out without considering the imbalance, it was found that more healthy patients were predicted to be positive for cancer. In cost-sensitive learning, models are constructed considering misclassification costs and other costs (*viz.,* instance and attribute cost, active learning cost, computation cost). Among these, the misclassification cost is most important in cost-sensitive learning.



		Actual	
		Negative	Positive
Predicted	Negative	$C(0,0) = C_{00}$	$C(0,1) = C_{01}$
	Positive	$C(1,0) = C_{10}$	$C(1,1) = C_{11}$

Figure 1.18: Cost-matrix for imbalanced dataset; where C_{00}, C_{11}, C_{10} and C_{01} are the cost associated with the prediction of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) respectively.

Misclassification cost can be applied to both binary and multi-class classification problems. Considering a binary classification, in a cost-sensitive learning, the costs of false positive (actual negative but predicted as positive), false negative (actual positive predicted as negative), true positive (positive predicted as positive) and true negative (negative predicted as negative) can be given as a cost matrix (Figure 1.18).

C$_{i,j}$ denotes the misclassification cost for classifying an instance from its actual class *j* into the predicted class *i* (usually positives are denoted as 1 and negatives as 0) (Figure 1.18). Cost-sensitive methods are extension of non-cost-sensitive methods with an addition of bias to error-based classification. Common methods to introduce bias into the error-based classification are:

- Changing the class distribution by resampling, instance weights and meta-cost (incorporates cost into the preprocessing steps).
- Modifying learning algorithms – requires cost calibration for each dataset.
- Introducing boosting approaches *viz.,* adaboost, costboost *etc.,*
- Applying direct cost-sensitive learning approaches (*viz.,* Laplace correction, smoothing) by introducing cost as a function of probability estimates.

1.6.3 Performance Evaluation in Machine Learning

For a given machine-learning problem, *n* number of multiple models/hypotheses can be trained and the performance of a specific model depends on various parameters like complexity of the dataset (in terms of relationship between the input and output variable), the size of the dataset used as training data and finally the computational complexity involved (*i.e.,* time and memory). To evaluate the performance of the model, the entire dataset is divided into three subsets *viz.,* a training set (used to construct a generalized model or hypothesis), a validation set (used to measure the complexity of the generated model) and finally a test set (used to evaluate the performance of the model). The model with the least misclassification error is considered the best. In addition, performance evaluation aids in comparing different learning algorithms used in construction and to optimise the constructed model. Thus, performance evaluation is defined as the trade-off of correctly classifying all the data points of the same class and making sure that each class contains points of only one class.

1.6.3.1 Choice of Performance Metrics

The aim of constructing a machine based model is to obtain a generalised prediction. In order to evaluate the generalization ability of the constructed model, several performance metrics are widely used. The choice of performance metrics is based on the type of the constructed model – whether it is a discrete classifier (predicts either positive or negative) or a regression (predicts continuous values). Most commonly used performance metrics for evaluation are accuracy, precision, recall, Receiver Operating Curve (ROC) and Area Under the Curve (AUC).

In machine learning, the output of the classification is given as a confusion matrix, which contains four possible situations that can occur while classifying data points (*i.e.,* actual and predicted points)

Based on the confusion matrix (Figure 1.19), several other parameters are also calculated, *viz.,* accuracy, precision, recall *etc.* Accuracy (ACC) or overall classification rate is the most commonly known performance metrics and is defined as the ratio of instances that are correctly classified over the total number of instances. Let us consider the following example involving cancer diagnosis from mammography images. If the constructed discrete classifier results in 100% accuracy in predicting healthy and diseased candidates, then the model is considered best. If the model results in accuracy lower than the threshold value then the model is not suitable for diagnosis.

Accuracy is given by the formula,

$$ACC = \frac{(TP+TN)}{N} \tag{1.7}$$

		Predicted outcome		
		p	n	Total
Actual Outcome	p ’	True Positive	False Negative	P’
	n ’	False Positive	True Negative	N’
	Total	P	N	

Figure 1.19: Prediction outcome in form of confusion matrix.

where TP is True positive,TN is true negative and N is the total number of samples The error rate is given as,

$$Error\, rate = (1 - ACC) \qquad (1.8)$$

Both accuracy and error rate are based on the overall generalization performance of the dataset and they are not suitable for measuring the performance of individual class distribution. Further, in case of a skewed dataset, accuracy and error rate measurement are biased towards the dominant class. In terms of misclassification costs, since the measurement is very generalised, they have to limit all the misclassification errors equally and thus fail to distinguish different cost during the classification.

On the other hand, precision and recall are used for measuring the performance of the regression, where the output is continuous. Let us see what happens for a model that performs query-based search from a cancer related database that results in both relevant and irrelevant records. To measure the effectiveness of the search, precision and recall are used, which are given by the formula

$$Precision = \frac{|Relevant\,records \cap Retrieved\,records|}{|Retrieved\,records|} \tag{1.9}$$

$$Recall = \frac{|Relevant\,records \cap Retrieved\,records|}{|Relevant\,records|} \tag{1.10}$$

Both precision and recall are inversely proportional. For the above example, F-score can also be used as a single measure of performance and is given as

$$F\text{-}score = 2\left[\frac{(Precision \times Recall)}{(Precision + Recall)}\right] \tag{1.11}$$

In case of binary classification, precision is given as sensitivity or True Positive Rate (TPR) and recall as specificity or False Positive Rate (FPR). Both the metrics are calculated from confusion matrix and are given by the formulae:

$$Sensitivity = \frac{(TP)}{(TP + FN)} \tag{1.12}$$

$$Specificity = \frac{(TN)}{(TN + FP)} \tag{1.13}$$

Considering the limitations of accuracy and error rate in a skewed dataset, both TPR and FPR provide a complete description of individual class performance. However, they are dependent on the arbitrary choice of threshold values. Hence, the Receiver Operating Curve (ROC) is used for imbalanced datasets where the choice of threshold is critical.

In a binary classification, ROC aids in visualizing the performance of the learning algorithm graphically over varying thresholds (or decision criteria), usually drawn between TPR and FPR (Figure 1.20). ROC is used in identifying the optimal threshold level, optimal behaviour region, model selection and comparative evaluation of different learning algorithms. A ROC curve analysis depicts the relationship between FPR and TPR. For a classification, ROC is obtained by plotting FPR on the x-axis and TPR on the y-axis.

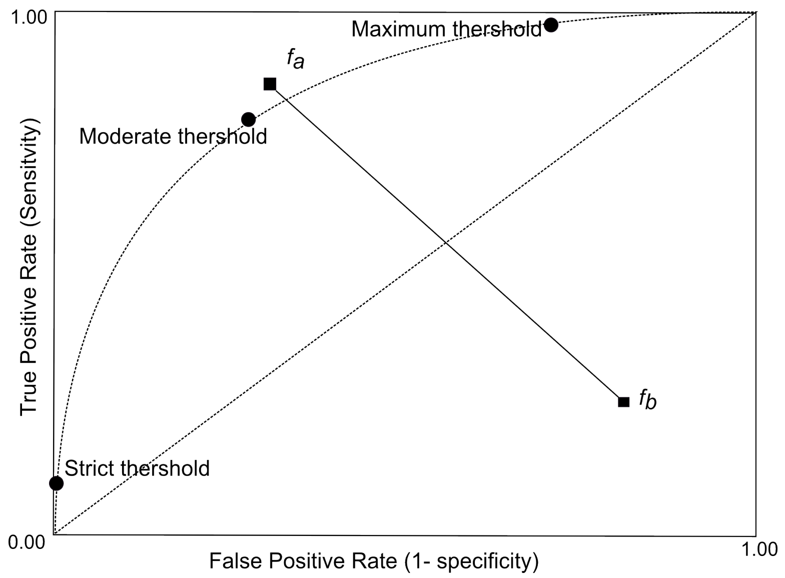


Figure 1.20: Receiver Operating Curve (ROC) of a binary classification. f_a and f_b denote operational points of classifier a and b. f_a indicates the best classifier and f_b indicates the worst classifier.

For a binary classification identifying patients diagnosed with cancer, TPR and FPR are calculated for all the instances predicted and plotted as ROC. The point (0, 0) denotes that all instances are classified as negative instances (TPR = FPR = 0) and the point (1,1) denotes that all the instances are classified as positive instances (TPR = FRP = 1). The diagonal line connecting the two points denotes random classification and hence (TPR = FPR). Thus in a ROC space, the output of the classifier results as a single point (f_a or f_b). The classifier whose prediction falls above the diagonal line are considered best classifiers, whereas predictions below the diagonal line are considered the poor classifications.

In order to obtain a comparison between various learning algorithms on the same dataset, Area Under the Curve (AUC) is widely used (Figure 1.21). AUC defines the probability that a randomly chosen positive instance has a higher decision function value than a random negative instance [73]. AUC is a single scalar value deduced from the ROC curve and it measures discrimination *i.e.,* the ability to classify the positives and negatives in a testset.

The total area of the grid represented by the AUC is always 1 since TPR and FPR ranges between zero to one. Thus, an AUC=1, indicates the best classifier and AUC=0 indicates the prediction as purely random .

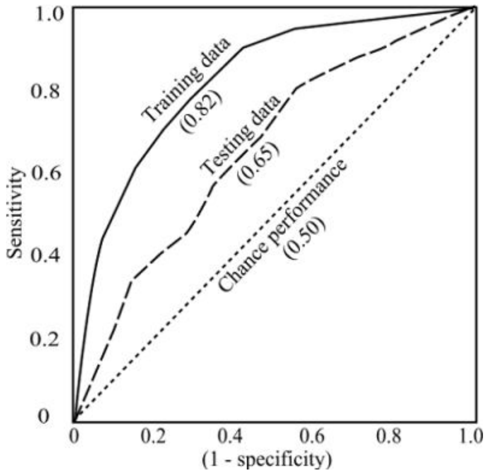


Figure 1.21: AUC of a classification process.

1.6.3.2 Error Estimates in Classification

The main drawback with disease-related classification processes is the inadequate amount of data sources. In cancer-related studies particularly, the sizes of experimentally validated positive and negative dataset are highly imbalanced and inadequate. To overcome this, oversampling or undersampling techniques may be used which generates a good amount of representative dataset. Obviously, these dataset are representative and indicate random instances generated by either averaging or from nearest neighbours. Thus making error estimates in randomly generated dataset is complex. More recently, *k-fold* cross-validation and holdout methods have been used for error estimates. These methods are simple and use most of the data for error estimates.

In holdout sampling method, one part of the entire dataset is used for testing the performance of the classifier (called as testset). The classifier is trained and a discrete classifier obtained; the performance and the error estimates of the classifier is calculated based on the separate set reserved for testing. The advantage of using the holdout method is that it gives a concrete generalization on the classifier. However, the main disadvantage of using holdout method is that it requires a large dataset and does not consider the whole dataset for generalization (*i.e.,* considering that testset may contain some amount of discriminative features), which may result in the loss of overall performance or accuracy.

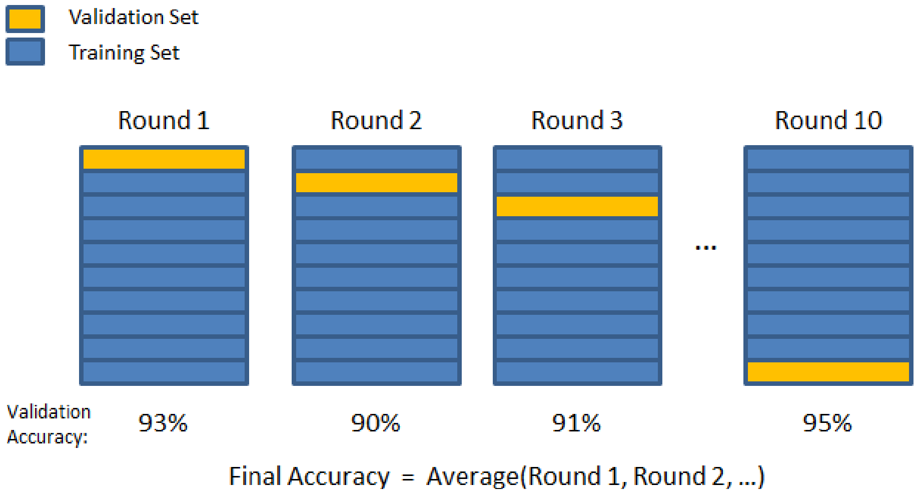


Figure 1.22: *k-fold* cross-validation method. In each training process, one subset is considered as testset and others as training set. For each training process *cv*-rate is calculated and finally averaged over all the *k-fold* to give the overall cross-validation rate. (*cv*-rates are just shown for illustration).

To overcome this, resampling methods (*viz.,* *k-fold* cross-validation, Leave-One-Out) are used commonly in classifications involving smaller datasets, where a separate dataset cannot be reserved as testset. In *k-fold* cross-validation method (*k-fold cv*), the entire dataset is divided into k subsets of equal size from m samples. Each subset is called a fold. The training is carried out on *(k-1)* subsets together and then tested on the k^{th} subset. The entire training process is repeated k times on different k^{th} subsets, thus considering all the *k-fold* subsets for testset as once. For each training iteration, the error estimate is calculated and finally the averaged error estimates are obtained for the entire *k-fold* training process (Figure 1.22). The default k-value is set to k =10, considering all the computational complexity related with the calculation involved in the training process.

The Leave-One-Out method (LOO) is similar to *k-fold cv* method except the fact that *k-fold* is equal to m samples (*i.e.,* the training process is carried out $k = m$ times). The advantage of LOO method is that it utilises the entire dataset for training and test set and hence an unbiased classifier is obtained. In case of the cross-validation method, the final classifier may be biased since in each iteration there is a high chance that the samples in the training set overlap.

1.6.4 Feature Selection and Training process

The aim of any feature selection process is to improve the prediction accuracy of the classifier by choosing subset of features that are relevant to discriminate the class labels. The process chooses the best subset of features, thus decreasing the structure and complexity of the dataset without loss of overall performance. In practice, there are four major steps involved in the feature selection process as illustrated [74] in figure 1.23.

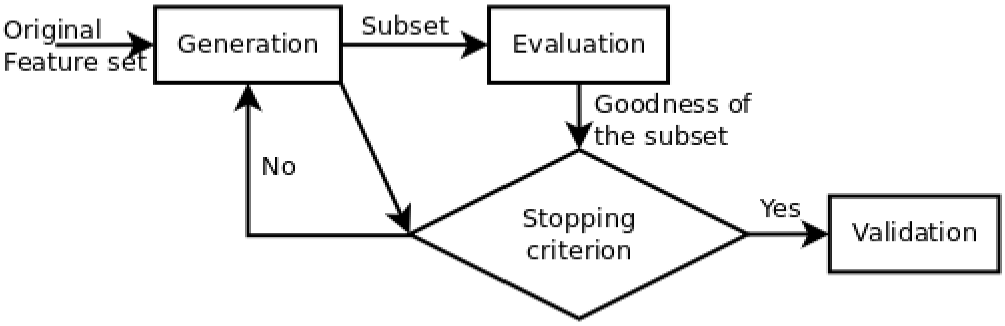


Figure 1.23: Feature selection process [75].

In generation step, optimal subset of features is selected for evaluation. In general, the number of subsets that can be generated from a sample size of N is given as 2^N. Such an exhaustive number may increase the computation cost significantly even with a medium sized dataset. Hence, generation step is done using exhaustive or heuristic search or by using random selection of subsets. In exhaustive search, complete search for optimal subset is done using evaluation function. The method is considered as a complete search for features since it allows backtracking techniques to guarantee the selected subset as optimal.

In heuristic approach (iterative method), the search is incremental based on the performance of the selected features in the subset. In each step, a feature is added and evaluated subsequently. If there is a drop in performance, then the feature is rejected and the iteration moves with the next set of features. In heuristic search, a threshold value is set during the evaluation; features above threshold value are considered for optimal feature set. The method works well with both continuous and nominal datasets but results in a non-optimal subset if the features are redundant. A random search for optimal subsets is possible only if there are certain values assigned to the features. The number of features chosen is always less in case of random search but utilise maximum number of iteration for finding optimal subset.

Feature selection methods can be grouped into two categories (i) Filter and (ii) Wrapper methods.

1.6.4.1 Filter Approach

In the filter-based approach, the learning algorithm is ignored completely during the feature selection process and this may result in some amount of performance degradation during the training process even with the optimal subset. Most common algorithms using filter methods are FOCUS, Relief and decision tree-based algorithms. In FOCUS algorithm, a minimal subset of features is chosen such that it is sufficient to discriminate the labels in the training set. This selection of minimalistic feature set may result in “MIN-FEATURES” bias. For example, let us consider again the cancer infected patient dataset containing Social Security Number (SSN) of patient as one of the labels. When FOCUS method is applied for feature selection, it chooses SSN as one of the important features reflecting the label in the training set. Thus, when classification is done with subset containing SSN, it results in large misclassification error.

In Relief method, a weight representing the relevance of feature with respect to the target label is assigned. The weight applied is based on the significance between the nearest neighbours in the dataset. Relief method shows good performance with weight-based learning algorithms and is not suitable for large datasets with higher number of features representing the target labels. Again, it results in poor performance when the feature set is redundant. In most cases, it chooses the strongest features that discriminate the target labels as their optimal subset. Decision tree-based approach is also employed for feature selection process and is mainly used along with nearest-neighbour algorithm. The major drawback with this approach is that it results in data fragmentation and results in only few splitting nodes (*i.e.,* utilisation of only few features) for constructing a decision tree.

1.6.4.2 Wrapper Approach

Wrapper approach utilises the learning algorithm itself as a part of the evaluation function in selecting the optimal subset. The method considers the entire feature space to select the optimal subset and evaluates the performance of the learning algorithm with the selected optimal subset [76]. For an efficient feature selection process, wrapper approach requires a state space — an initial state, a termination condition and a search engine (Figure 1.24). For example, let us consider the feature selection process, with n features in a state and each state consisting of m bits, where a bit represents the presence or absence in the optimal feature subset. Operators determine the connectivity between the state and for the given problem the search space is given by $O(2^n)$. It is impractical to do an exhaustive search for features unless the number of features is less. Hence, search engines like hill-climbing and best-first search are utilised.

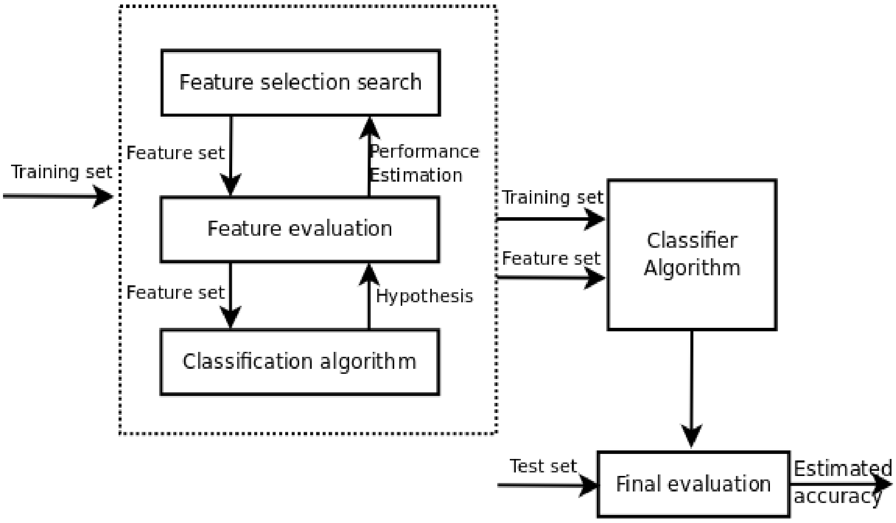


Figure 1.24: Wrapper approach for feature subset selection [76].

The hill-climbing (or greedy search or steepest ascent) starts with an empty feature set and starts adding features sequentially. Adding new features to the already selected feature set increases performance of the learning algorithm. The iteration continues to add new features until there is no improvement over the performance even after addition/deletion of features. In the best-first search, the process works similar to a decision tree algorithm. For each iteration, the best features are added to empty feature space and evaluated. Iteration stops when there are no further nodes generated or there is no further improvement in the performance. Usually there is no difference in the accuracies obtained between these methods, except the fact that the best-first method produces a larger feature subset.

Utilising learning algorithms for searching optimal subset may reduce the computational efficiency and slow down the classification process. To overcome this, algorithm specific feature selection techniques are widely used. Among them, Recursive Feature Elimination (RFE) [77] used along with SVM classifier reduces the computational complexity by only ranking the individual features based on the influence in the class assignment. RFE follows a simple iterative process as given below (Figure 1.25).

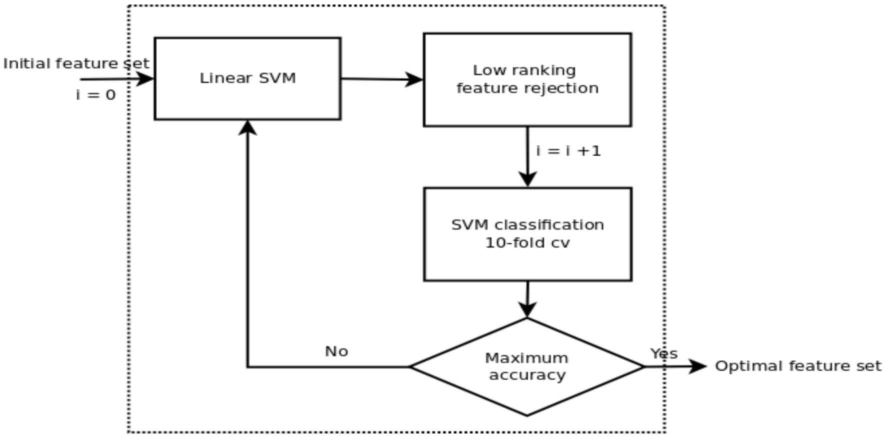


Figure 1.25: Recursive Feature Elimination using Support Vector Machines.

The iteration is continued until no further performance improvement is obtained with the given set of optimal features. RFE was used initially in handwritten digit recognition [78] and it is now used widely with several other learning algorithms.

1.6.5 Data Preprocessing

Several factors determine the generalization of a learning algorithm; the foremost among them is the quality of the instances. The result of machine learning is dependent on the structural complexity of the data associated with it. If the data is too noisy or irrelevant for the learning algorithm, then the training process is more difficult. Hence, data preprocessing is done meticulously before the training process. The steps involved in data preprocessing are:

- Data cleaning: involves handling missing values, smoothening noisy data, identification or complete removal of outliers and resolving redundancy in dataset,
- Data transformation: involves normalization and aggregation,
- Data reduction: reducing the volume but producing same or similar analytical results,
- Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

Outliers represent instances that deviate extensively containing too many null feature values. Removal of outliers is a complex process associated with initial data cleaning [79].

Data representation is yet another major issue associated with initial cleaning, since in a dataset there may be *n* number of features that are unrelated to the actual class labeling. Redundant features representing similar information may also be neglected during the data cleaning process. Feature selection process aids in distinguishing both the redundant and the non-redundant features and reduces the dimensionality of the data. Reducing the complexity associated with the data structure may reduce the training time significantly.

Incomplete datasets are unavoidable in most cases and they appear because of forgotten values during data entry or represent values that are irrelevant to the given instance. Some common methods employed in handling missing values are:

- Use of mean or median values to fill missing values. Usually for balanced dataset, mean is employed and for skewed dataset, median is employed.
- Use of global constant for all missing values.
- Use of probable values as determined by regression, inference-based tools, Bayesian formalism or using decision-tree induction.

Noise or the outliers indicate random error or wide variance in the instances. Noise in the dataset may result in inefficient classification and can be handled by the method of *binning* (*i.e.,* smoothening a sorted data values based on nearest neighbours). The process begins with sorting the entire data range in small bins or buckets. Each noisy value is then replaced by the mean or the median identified by the bin boundaries. This process is also called local smoothing since it considers only the neighbouring values. Data smoothing can be done using regression guided by a function. Both linear and multi linear regression are used widely for data smoothing based on the dimensionality of the dataset.

In certain cases, there is a need for collecting data from multiple datasets. In such cases, data integration should be done very carefully to avoid redundancy and inconsistencies in the resulting dataset. The major concern during the data integration process is the heterogeneity and the structure of data, which result in an entity identification problem. Let us consider data integration from two different miRNA databases where the primary key varies between the database considered *viz.,* miRID and miRnumber. During data integration, difference in the field names will pose great risk of inconsistency and redundancy of data. Hence, to avoid the problem, metadata attributes like species name, primary miRname and other information are also considered during integration to reduce errors. Another important concern with the data integration is the difference in representation in databases and this is mainly because of the dissimilarity in the scaling of data or in encoding.

The structure of the database is also a major concern while integrating attributes from two different databases. Redundancy during data integration is analysed using χ^2 (chi-square) test for nominal datasets and correlation coefficient and covariance are used for numeric datasets [80]. Both the methods aid in testing the attribute variation between the two databases considered for integration. In many real world problems, the dataset is either complex or huge, making data analysis quite impractical. Hence, data reduction is applied to reduce the representation in the dataset without losing the integrity of the original dataset. Data reduction involves dimensionality reduction, numerosity reduction and data compression.

Dimensionality reduction methods include wavelet transforms [81] and Principal Component Analysis (PCA) [82], which project the original dataset into a smaller space. Attribute reduction is also done during the dimensionality reduction, where irrelevant or weakly relevant attributes are removed. Numerosity reduction involves representing the original dataset into smaller representations. It includes both parametric (*e.g.,* regression and log-linear method) and non-parametric methods (*e.g.,* histogram, clustering, sampling and data cube aggregation [83]). In data compression, a transformation involving compression or reduction of complete dataset is done. The compression may be lossless or inferior based on the information loss during the compression. Both dimension reduction and numerosity reduction is considered as a form of data compression. Thus, the success of the machine learning algorithm relies on the quality of data which in turn depends on adequate data and relevant features. Thus, data preprocessing is an important step in machine learning process.

1.6.6 Machine Learning in miRNA studies

The discovery of miRNA have changed our perspective view on eukaryotic gene regulation completely. However there is still a lot of work to do between the discovery and functional annotation of the miRNA. In the past two decades, the number of miRNA identified from several sequenced genomes have produced a massive amount of data which in turn rely on several algorithms, particularly machine based learning algorithms for functional annotation. On the other hand, newly developed experimental protocols has influenced machine learning with a large amount of validated dataset backed by experimental support. Initially, these machine learning algorithms utilised only the most dominating features obtained from experimental results for functional annotation but the results obtained from such methods are rather susceptible and do not give a clear idea about the functional aspects of the miR being investigated. Hence, there is a need for a robust machine learning model which will consider both experimental results and other features (*viz.,* structural, thermodynamics of binding).

1.7 Gaps in the Existing Research

With the deluge of miRNA numbers in miRBASE, the need for functional analysis of miRNA is also growing rapidly. miRNA regulation is defined by its binding with the target mRNA. Several target prediction tools are available to predict potent functional and non-functional miRNA targets. In turn, these tools rely on some specified characteristics obtained from previous annotation (*viz.,* base pairing in seed region, thermodynamics of binding) to find the potential miRNA target. However, there are no clear guidelines or a concrete algorithm which might predict a miRNA to be involved with the cancer pathway. To overcome this deficit the following objectives were proposed aiming for

- Identifying global signatures in miRNA associated in cancer pathway.
- Constructing and optimizing a machine-learning algorithm to predict miRNA associated with cancer.
- Constructing a database of oncogenically involved miRNAs – Microondb and design of web-based user interface – MicRooN.

Chapter 2

Search for Signatures in miRNAs Associated with Cancer

Chapter 2

Search for Signatures in miRNAs Associated with Cancer

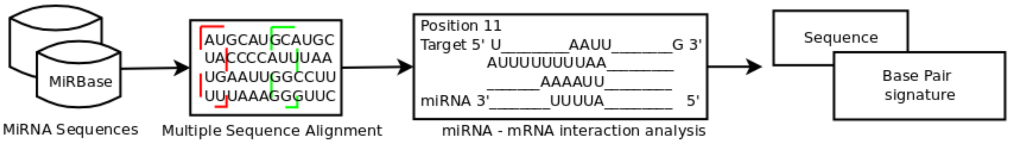


Figure 2.1: Process pipeline for extracting signatures in miRNA associated with cancer.

Globally, the identification and verification of miRNAs as biomarkers for cancer cell types has been the area of thrust for several miRNA biologists. However, there has been a noticeable vacuum when it comes to identifying a common signature or trademark that could be used to demarcate a miRNA to be associated with cancer or not. Additionally, studies aimed at identifying cancer specific miRNA signatures are rather sketchy and specific to a group of related cancerous cells. To answer these queries, we undertook an *in silico* study involving the identification of global signatures in experimentally validated miRNAs which have been associated with cancer. This chapter deals with the first step in the process *viz.*, the construction of a database and the search for features that would distinguish miRNA associated with cancer from those which are not.

2.1 Materials and Method

2.1.1 Dataset Preparation

For the purpose of generating a classifier, the first step needed to be undertaken is the construction of a miRNA dataset which has been experimentally validated to be associated with cancer. To begin with, a list of genes involved in cancer was downloaded from the Catalogue of Somatic Mutations (COSMIC)[84]. A total of 488 genes were thus listed, which could be further segregated into oncogenes and tumour suppressors by cross-referring with the Tumour Associated Gene database (TAG)[85]. Experimentally validated miRNA interactions with target mRNA can be obtained from miRecords[46] and miRTarBase[45]. Therefore, the list of genes obtained from COSMIC was curated with miRecords and miRTarBase to obtain a list of experimentally validated targets. This process finally yielded a set of targets for miRNA which have been experimentally validated to be associated with cancer. A total of 2578 miRNAs were extracted from miRBase 20.0 (May 2013)[86], and these were compared with the experimentally validated miRNA-mRNA interactions obtained as above, yielding a final set of 239 microRNAs which have been conclusively implicated in the cancer pathway (Appendix A). These 239 miRNAs were manually checked with their available literature and revalidated. $3'$UTR mRNA sequences involved in the interaction of these 239 miRNAs with their targets were obtained from BIOMART- Ensemble[87] (Figure 2.2).

Mature miRNa sequences were extracted from miRBase 20.0, a total of 2578 miRNAs were extracted and were compared with the experimentally validated miRNA-mRNA interactions obtained as above, yielding a final set of 239 non-redundant miRNAs which have been evidentially associated in the cancer pathway (Table 2.1). These 239 miRNAs were manually checked with their available literature and revalidated.

2.1.2 Positive and Negative Dataset

The 239 experimentally validated miRNA sequences were considered as a positive dataset and 100 randomly generated miRNA sequences with an average

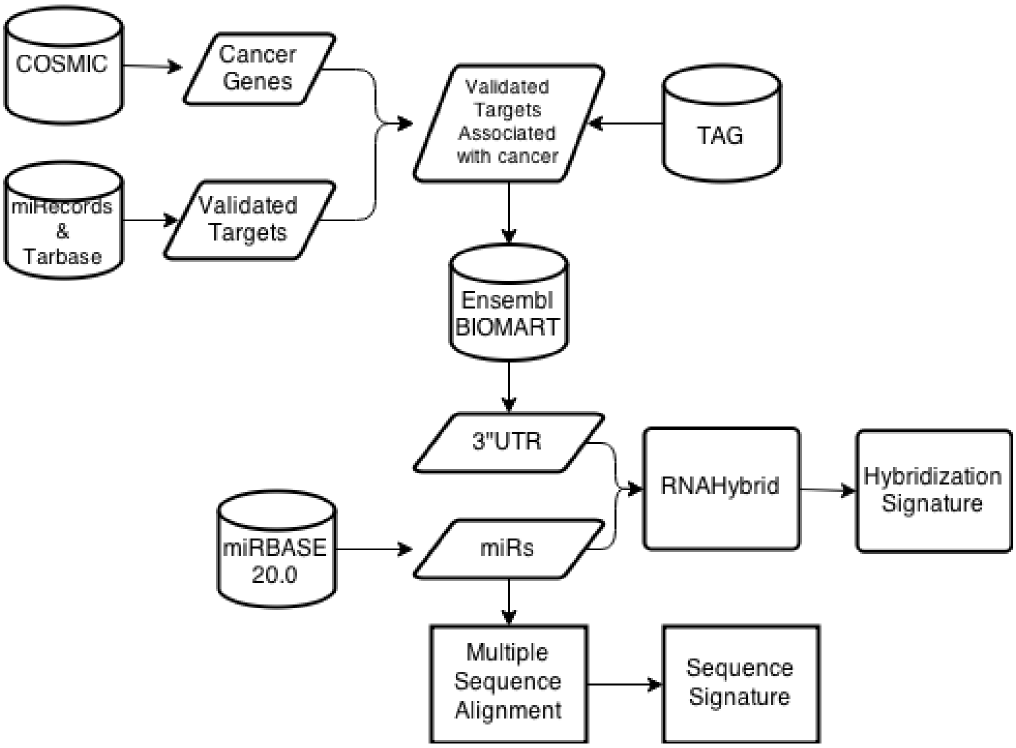


Figure 2.2: Data preparation workflow for identifying sequence and hybridisation-based signatures.

length of 22*nt* as negative set. Only non-redundant miRNA sequences were considered for sequential analysis both the datasets (Table 2.1 & 2.2).

2.1.3 Search for Signatures

2.1.3.1 Multiple Sequence Alignment

Searching for signatures, we started by looking into the mature sequence of miRNAs in the datasets. Multiple sequence alignment was done in search of specific miRNA sequence signature in both positive and negative dataset. miRNA sequences were aligned in MATLAB with *Multialign* function. Existing Gap Alignment method was used for alignment to maintain the existing sequence conservation. Regional position conservation scores were calculated manually to recheck the signature obtained from the sequence alignment.

Table 2.1: Positive instances - Experimentally validated miRNAs (239 instances)

hsa-*miR*-203a	hsa-*miR*-1	hsa-*miR*-127-5p	hsa-*miR*-20a-5p
hsa-*miR*-125b-5p	hsa-*miR*-133b	hsa-*miR*-146a-5p	hsa-*miR*-26a-5p
hsa-*miR*-149-3p	hsa-*miR*-204-5p	hsa-*miR*-24-1-5p	hsa-*miR*-103a-3p
hsa-*miR*-185-5p	hsa-*miR*-335-5p	hsa-let-7e-5p	hsa-*miR*-107
hsa-*miR*-199a-3p	hsa-*miR*-630	hsa-*miR*-195-5p	hsa-*miR*-15b-5p
hsa-*miR*-451a	hsa-*miR*-181a-5p	hsa-*miR*-19b-1-5p	hsa-*miR*-16-5p
hsa-*miR*-184	hsa-*miR*-21-5p	hsa-*miR*-34b-3p	hsa-*miR*-26b-5p
hsa-*miR*-708-3p	hsa-*miR*-34c-5p	hsa-*miR*-520b	hsa-*miR*-424-5p
hsa-*miR*-122-5p	hsa-*miR*-365a-3p	hsa-let-7a-5p	hsa-*miR*-503-5p
hsa-let-7b-5p	hsa-*miR*-449a	hsa-*miR*-15a-5p	hsa-*miR*-124-5p
hsa-*miR*-7-5p	hsa-*miR*-18b-5p	hsa-*miR*-100-5p	hsa-let-7g-5p
hsa-*miR*-125a-5p	hsa-*miR*-193b-5p	hsa-*miR*-99a-5p	hsa-*miR*-98-5p
hsa-*miR*-331-3p	hsa-*miR*-206	hsa-*miR*-138-5p	hsa-*miR*-143-5p
hsa-*miR*-548d-5p	hsa-*miR*-20b-5p	hsa-*miR*-101-5p	hsa-*miR*-663a
hsa-*miR*-559	hsa-*miR*-221-5p	hsa-*miR*-197-5p	hsa-*miR*-30a-5p
hsa-*miR*-205-5p	hsa-*miR*-222-5p	hsa-*miR*-200b-5p	hsa-*miR*-146b-5p
hsa-*miR*-22-5p	hsa-*miR*-29b-1-5p	hsa-*miR*-324-5p	hsa-*miR*-10b-5p
hsa-*miR*-19a-5p	hsa-*miR*-302c-5p	hsa-*miR*-326	hsa-*miR*-135b-5p
hsa-*miR*-302d-5p	hsa-*miR*-199a-5p	hsa-*miR*-17-3p	hsa-*miR*-25-5p
hsa-*miR*-130a-5p	hsa-*miR*-199b-5p	hsa-let-7d-5p	hsa-*miR*-181c-5p
hsa-let-7f-5p	hsa-*miR*-378a-5p	hsa-*miR*-27a-3p	hsa-*miR*-183-5p
hsa-*miR*-151a-5p	hsa-*miR*-224-5p	hsa-*miR*-106a-5p	hsa-*miR*-186-5p
hsa-*miR*-28-5p	hsa-*miR*-497-5p	hsa-*miR*-106b-5p	hsa-*miR*-21-5p
hsa-*miR*-708-5p	hsa-*miR*-31-5p	hsa-*miR*-147a	hsa-*miR*-1
hsa-*miR*-373-5p	hsa-*miR*-183-5p	hsa-*miR*-330-5p	hsa-*miR*-124-5p
hsa-*miR*-30e-5p	hsa-*miR*-569	hsa-*miR*-361-5p	hsa-*miR*-204-5p
hsa-*miR*-150-5p	hsa-*miR*-181b-5p	hsa-*miR*-520h	hsa-*miR*-101-5p
hsa-*miR*-29a-5p	hsa-*miR*-192-5p	hsa-*miR*-93-5p	hsa-*miR*-34a-5p
hsa-*miR*-17-5p	hsa-*miR*-144-5p	hsa-*miR*-519a-5p	hsa-*miR*-122-5p
hsa-*miR*-371a-5p	hsa-*miR*-633	hsa-*miR*-29a-3p	hsa-*miR*-141-5p
hsa-*miR*-34b-5p	hsa-*miR*-145-5p	hsa-*miR*-345-5p	hsa-*miR*-214-5p
hsa-*miR*-34c-5p	hsa-*miR*-146b-5p	hsa-*miR*-363-5p	hsa-let-7b-5p

Table 2.1 Continued: Positive Instances

hsa-*miR*-25-5p	hsa-*miR*-17-5p	hsa-let-7g-5p	hsa-*miR*-129-5p
hsa-*miR*-9-5p	hsa-*miR*-182-5p	hsa-*miR*-125b-5p	hsa-let-7i-5p
hsa-*miR*-92a-1-5p	hsa-*miR*-20a-5p	hsa-*miR*-492	hsa-*miR*-107
hsa-*miR*-106a-5p	hsa-*miR*-20b-5p	hsa-*miR*-424-5p	hsa-*miR*-223-5p
hsa-*miR*-106b-5p	hsa-*miR*-28-5p	hsa-*miR*-503-5p	hsa-*miR*-27a-5p
hsa-*miR*-10b-5p	hsa-*miR*-298	hsa-*miR*-137	hsa-*miR*-205-5p
hsa-*miR*-125a-5p	hsa-*miR*-299-5p	hsa-*miR*-181b-5p	hsa-let-7f-5p
hsa-*miR*-132-5p	hsa-*miR*-302a-5p	hsa-*miR*-197-5p	hsa-*miR*-224-5p
hsa-*miR*-429	hsa-*miR*-675-5p	hsa-*miR*-19a-3p	hsa-*miR*-222-3p
hsa-*miR*-373-3p	hsa-*miR*-155-3p	hsa-*miR*-19b-3p	hsa-*miR*-25-3p
hsa-*miR*-106b-3p	hsa-*miR*-130b-3p	hsa-*miR*-204-3p	hsa-*miR*-30d-3p
hsa-*miR*-17-3p	hsa-*miR*-125b-1-3p	hsa-*miR*-93-3p	hsa-*miR*-559
hsa-*miR*-192-3p	hsa-*miR*-324-3p	hsa-*miR*-125a-3p	hsa-*miR*-661
hsa-*miR*-20a-3p	hsa-*miR*-326	hsa-*miR*-1285-3p	hsa-*miR*-92a-3p
hsa-*miR*-23b-3p	hsa-*miR*-338-3p	hsa-*miR*-15a-3p	hsa-*miR*-30a-3p
hsa-*miR*-26a-5p	hsa-*miR*-21-3p	hsa-*miR*-16-1-3p	hsa-*miR*-212-3p
hsa-*miR*-145-5p	hsa-*miR*-217	hsa-*miR*-200c-5p	hsa-*miR*-200a-3p
hsa-*miR*-302a-5p	hsa-*miR*-96-5p	hsa-*miR*-192-5p	hsa-*miR*-132-3p
hsa-*miR*-34a-5p	hsa-*miR*-223-5p	hsa-*miR*-146a-5p	hsa-*miR*-532-5p
hsa-*miR*-34b-5p	hsa-*miR*-218-5p	hsa-*miR*-15a-5p	hsa-*miR*-302b-3p
hsa-*miR*-126-5p	hsa-*miR*-214-5p	hsa-*miR*-16-5p	hsa-*miR*-612
hsa-*miR*-155-5p	hsa-*miR*-23b-5p	hsa-*miR*-212-5p	hsa-*miR*-335-5p
hsa-*miR*-140-5p	hsa-*miR*-340-5p	hsa-*miR*-24-1-5p	hsa-*miR*-18a-3p
hsa-*miR*-18a-5p	hsa-*miR*-449b-5p	hsa-*miR*-103a-2-5p	hsa-*miR*-221-3p
hsa-*miR*-200a-5p	hsa-*miR*-562	hsa-*miR*-200a-5p	hsa-*miR*-200b-5p
hsa-*miR*-18a-3p	hsa-*miR*-23b-5p	hsa-*miR*-24-1-5p	hsa-*miR*-335-5p
hsa-*miR*-145-5p	hsa-*miR*-340-5p	hsa-*miR*-103a-2-5p	hsa-*miR*-18a-3p
hsa-*miR*-302a-5p	hsa-*miR*-449b-5p	hsa-*miR*-138-5p	hsa-*miR*-221-3p
hsa-*miR*-34a-5p	hsa-*miR*-562	hsa-*miR*-155-5p	hsa-let-7a-5p
hsa-*miR*-34b-5p	hsa-*miR*-200a-5p	hsa-*miR*-29a-5p	hsa-*miR*-181a-5p
hsa-*miR*-126-5p	hsa-*miR*-200b-5p	hsa-let-7a-5p	hsa-*miR*-124-3p
hsa-*miR*-155-5p	hsa-*miR*-200c-5p	hsa-*miR*-181a-5p	hsa-*miR*-194-5p

Table 2.1 Continued: Positive Instances

hsa-*miR*-140-5p	hsa-*miR*-192-5p	hsa-*miR*-124-3p	hsa-*miR*-200b-3p
hsa-*miR*-18a-5p	hsa-*miR*-146a-5p	hsa-*miR*-194-5p	hsa-*miR*-200c-3p
hsa-*miR*-200a-5p	hsa-*miR*-15a-5p	hsa-*miR*-200b-3p	hsa-*miR*-138-5p
hsa-*miR*-542-5p	hsa-*miR*-16-5p	hsa-*miR*-200c-3p	hsa-*miR*-155-5p
hsa-*miR*-18a-3p	hsa-*miR*-212-5p	hsa-*miR*-200a-3p	hsa-*miR*-29a-5p
hsa-*miR*-217	hsa-*miR*-96-5p	hsa-*miR*-132-3p	hsa-*miR*-542-5p
hsa-*miR*-218-5p	hsa-*miR*-223-5p	hsa-*miR*-532-5p	hsa-*miR*-214-5p
hsa-*miR*-302b-3p	hsa-*miR*-612		

Table 2.2: Negative Dataset - Randomly generated miRNA dataset (100 instances)

>Sequence1	UCCGGCUGCGGAACUAUAAUUU	>Sequence51	GGUACGUAGCGUGGUCGCACAA
>Sequence2	GCCGUUGCAAUCCUUUAAUGGA	>Sequence52	GCACGGUGGAUCCUCCCCGCGC
>Sequence3	CCCGCGAAAUAGAUUUGCGCUG	>Sequence53	ACCCCACCUAUCGAGUCGGUCC
>Sequence4	CUGUCCGCGUGAGGAGUCCGGU	>Sequence54	UAUGGCAGCACGGUCACACGCG
>Sequence5	GUAGCGAAGGAUGAGGGCGACC	>Sequence55	GGCGGGCAGUGGCCGGCAGCCG
>Sequence6	CUAGGUGGCAACCGCCGGCUCC	>Sequence56	CACGCCUGCCGCGGCGCUCAAC
>Sequence7	GGCGGCGAGGCAUCACUCAGGG	>Sequence57	GGCCGGGGCUGGAGAGGCGGGG
>Sequence8	AGCAGGCGCGGAAAGGCACGGU	>Sequence58	GCCAUGGCGUGUGACCCGUCAC
>Sequence9	CCAGCGGACCGUCUAUCGGCUG	>Sequence59	GCUCGAGUUCGGUCAGGGCGUC
>Sequence10	GGCCAAAUGGGGCGCUCCGGUA	>Sequence60	ACCGCGAGUGGUCGACUGCUUU
>Sequence11	UCAGCGUGUCCAGCCUUAGGAC	>Sequence61	CCCAAUCUCCGAGCGAUUUAGC
>Sequence12	UCGGCCCAGCGCGCUGGCCUGG	>Sequence62	GUGGCGGCCCCGGGGGACCCAC
>Sequence13	GUCGAGGUGAAAUCACCGGCGC	>Sequence63	GAAAUGCGGUCGCAGCCCACCC
>Sequence14	CCAAGACCAGGCGGGCCCGCCG	>Sequence64	GACCGUACACGGAAGGGAGGGU
>Sequence15	CGUUGGCCAACCCCGGUACACC	>Sequence65	CCCCGUACGCCGACGCGCCUGC
>Sequence16	CUGUAAUCGGCGUUCAGGGGGA	>Sequence66	UCGCACGUCGUAUGCAUAAACG
>Sequence17	AGCCCGUGCCAGGGGGACGAGC	>Sequence67	GGCCGCACGAACCGGAGAGCGC
>Sequence18	CACCACGUGCCAGCGGCGGCAA	>Sequence68	AGGGAGGACCCCUAGCUCCUUU
>Sequence19	CGAUCGGUCGGACUAUUCAUCG	>Sequence69	ACAAAGCGCAGGCUCGCCCGCC
>Sequence20	CGGUGGUGGCGCUCGGAUCGCG	>Sequence70	GCCGGGACGCCUUACCUAGACG
>Sequence21	CGGGAAAGGUGCCUGUGUCCCG	>Sequence71	CGAUGACGGGCGCACUCCUCUG
>Sequence22	GCAGGCUAGGGCACGGCGCCGG	>Sequence72	GCCUCAACGGUUCCUGCUCCCG

Table 2.2 Continued: Negative Instances

>Sequence23	GGCGCUGCCCCAACCGUCCGGC	>Sequence73	CUGGGAUCCAAGGUUGGCGGCC
>Sequence24	GUGGGGUUCGCUACGACUUCCG	>Sequence74	GAGGCCGCCUCUCCGAAGUGAG
>Sequence25	AGUGCCGCGUGUGCGAGACCAC	>Sequence75	UCCUUCGUCCGUGGCUAACCGU
>Sequence26	GUUAUGUGCGCACAAGGCCGGC	>Sequence76	GCCAGAUCGCCUCGCAGACUCC
>Sequence27	AAUAGGACGUGGCCUUCGGGCU	>Sequence77	CGACCCGGUUUAACCCGCCAGG
>Sequence28	CUAUAGCCGCACAGGCCCGAAU	>Sequence78	GAAAGGGCUUGAGGCACGCCAA
>Sequence29	CCUGAGCCGUGUCGCGCGACCG	>Sequence79	UUCGCACCGCCGGGGUCGCCUG
>Sequence30	GCCCCUGCUCAACUUCUGUGCC	>Sequence80	GGUGUUUUGCGCCACCGUCGGG
>Sequence31	CGGGGGUUCUGGUCCGCCCGGG	>Sequence81	UGCGCUGGCAUGCGCCCUUCCU
>Sequence32	CGGCGCAGCCGAUUGGGGCCAU	>Sequence82	CCCAGGGGCAUGCGGCUGCGUG
>Sequence33	CUAGUGCACUUGCUGCAAGACU	>Sequence83	GUCAAGGGUGCGGCAUUCGUAU
>Sequence34	CCUUUCGGACACCCUCUCCCUG	>Sequence84	UUGCCCCCCGUGCUUGCUCUCA
>Sequence35	CCCAGUGGCGGAUGGUGGCGGC	>Sequence85	CGAGCCCGACCUGGAGAUCGAG
>Sequence36	UGUUGCCAGCCGGCGUGGAAGG	>Sequence86	GAGAUGCUUCCCGUGGAACCGG
>Sequence37	UAGCGGCACCGGCGCGAGCCUA	>Sequence87	GCGGCGCGCCAACGCAACGGAU
>Sequence38	GCGCGCCGUCUCCACCAACACA	>Sequence88	CUGCGCUACAGCGCGCAUAGCG
>Sequence39	GGGCCGUCCGGUCGCAUAGUGG	>Sequence89	AGAGCGGAGUUGCCGACGACGA
>Sequence40	GGCCCCGCGACGGGGUUGGCAA	>Sequence90	AGGCGACGCUGGGAUCCGUCCG
>Sequence41	GUGGCAGCCCAAACGAUGCCGG	>Sequence91	CCGCCACCCGCGGAAAGCAUCC
>Sequence42	GGGCUCGCGACGCACACGCUCU	>Sequence92	GCUCACGAGGCGGGCACCGAUU
>Sequence43	GGUCCAAUACACGCGUGACCCG	>Sequence93	GACACGGUCUUGCAGAGGGUCA
>Sequence44	GCGGUUAUCCUGCACCGGAACG	>Sequence94	GGGGGGUAGGUCAAAUUGGGUG

Table 2.2 Continued: Negative Instances

>Sequence45	CCGAUCGUGCAUCGGGCCAGCG	>Sequence95	GCUUGAAAACGCCGUGUCCGGG
>Sequence46	UGAUCGUGUCAUCUGGGAGGCG	>Sequence96	GUUAGGGUGCAGUAGACCGCGG
>Sequence47	GCCGUAGGGUGGAUAGUUCAAC	>Sequence97	GAGGAUGUCGUCCUGCCAGUGU
>Sequence48	GCGCCUGGGCGUCACCCGCCAU	>Sequence98	CCCUGUGCGGGCGGGCCGGCGA
>Sequence49	GACGCUGCCCCUGAUCUCUCCG	>Sequence99	CCGCUGAUAGCGCACACGGGGC
>Sequence50	AGUACCAGCACAAGCCAGUCUC	>Sequence100	GCGGGGGCGCUCGUCAGCACAC

2.1.3.2 miRNA-mRNA Interaction

In order to analyse the base pairing distribution between the two datasets, we utilised miRNA-mRNA interaction data obtained from RNAhybrid program (Vienna Package) [88]. For positive dataset, we utilised only experimentally validated miRNA-mRNA interactions (Table 2.1). Negative dataset containing genes not involved in cancer pathway was constructed by calculating Cancer Linker Degree (CLD) [89]. A jack-knife selection of 100 genes from 1025 genes obtained by CLD served as our negative dataset. For our analysis, we considered miRNA binding in 3′UTR only; mRNA sequences involved in interaction for both positive and negative were obtained from ENSEMBL-BIOMART [87]. In both dataset, miRNA sequence along with their specific 3′UTR sequences were hybridised with RNAhybrid to obtain hybridised structure with lower p-value. Generally, a single miRNA can bind to multiple mRNA targets or to different positions in the same target. Thus, p-values are assigned to individual hits or multiple hits of the same miRNA to one target or to multiple targets. They provide a guide to confident target prediction. Small p-value indicates good binding[90]. In a hybridised miRNA-mRNA structure, regions of complementarity having atleast continuous four base pairing was considered as ‘seed’ region and regions outside the seed were considered ‘outseed’ region. GU wobble base pairing was allowed in hybrid structure since GU wobble pairing is significant to miRNA function and are essential in preserving target specificity[91–95].

2.1.3.3 Analysing miRNA-mRNA Interaction Data

For parsing a seed vs outseed region an indigenous perl script *PairFinder*, which identifies seed, outseed regions, mismatches and bulges was scripted (Appendix B). Using Pairfinder, we delimited the hybrids into regions of matches and mismatches (Figure 2.3). The match regions were further demarcated as seed pairs based on their constitution. Regions of matches where the numbers of base pairings were less than four were not considered as seed pairs.

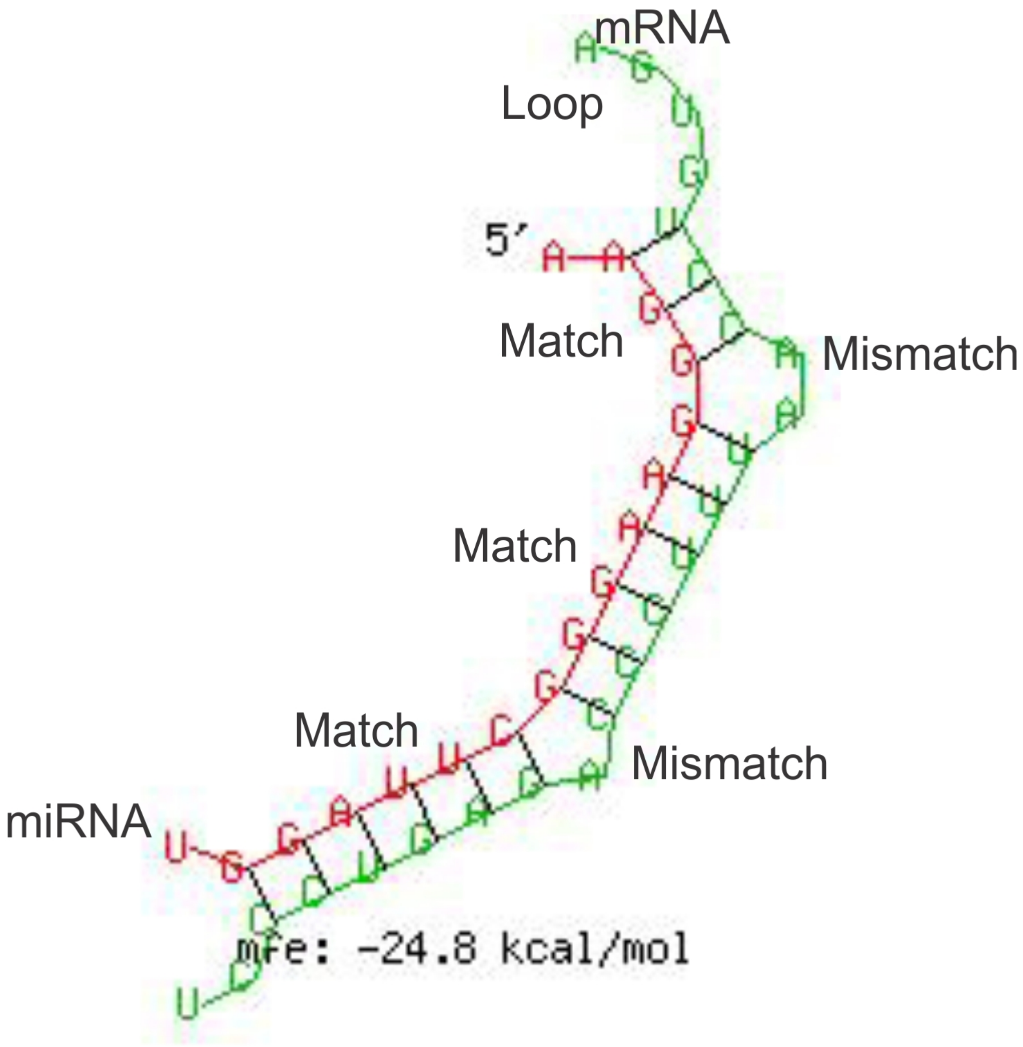


Figure 2.3: Formation of hybrids between hsa-*miR*-125a and Ataxia Telangiectasia Mutated gene (ATM) indicating matches and mismatches obtained from RNAhybrid.

All the hybridised results were parsed and analysed using *PairFinder* to give a complete list – total number of base pairing (includes both Watson-Crick and non-Watson Crick base pairing), number of seed and outseed region, number of bulges, number of mismatches in seed and outseed region, minimum free energy of binding (kcal/mol) and p-value of miRNA-mRNA binding. The hybrids were classified based on the total number of matches – some appear as continuous hybrid structures, whereas some possess upto six matches intercepted by mismatch regions. To understand the effect on non-Watson-Crick base pairing in the hybridised structure, a scoring system was followed for both seed and outseed region.

Scores (H_s) was obtained by the formula

$$H_s = (AU)_n + (GC)_n - (GU)_n \tag{2.1}$$

2.1.3.4 Thermodynamics of Binding

Most of the previous work concerning miRNA thermodynamics concentrated on miRNA folding but such results were mostly predictive and did not account for target acceptability[96]. Hence, we felt the need for a thermodynamic signature that would distinguish an miRNA associated with cancer and those that are not involved in cancer based on the minimum free energies (MFEs) of hybridisation between the miRNA and the 3′UTR of the mRNA sequence, using the RNAhybrid program[88]. The energy score for miRNA-mRNA interaction was obtained according to the formula,

$$\Delta\Delta G = \Delta G_{duplex} - \Delta G_{open} \tag{2.2}$$

where ΔG_{duplex} is the Minimum Free Energy (MFE) value obtained from RNAhybrid and ΔG_{open} is the energy required to make the target region accessible for the miRNA. ΔG_{open} is calculated by the RNAfold program of the Vienna package[38].

2.2 Results and Discussion

For the present study, only experimentally validated miRNA-mRNA interactions were considered. Predicted or non-validated miRNA-mRNA interactions were removed during the dataset construction.

2.2.1 Sequence Analysis

Initial analysis was focused on the distribution of nucleotide in and around the seed region in miRNA sequences. In positive dataset, it was observed uracil was the most preferred base in the seed region whereas cytosine was least preferred. Additionally, in positive dataset, AU-richness around the seed region increases and drops quickly proportional to the distance from the seed (Figure 2.4). However, in negative dataset, it was observed that GC-richness was prominent throughout the miRNA sequence.

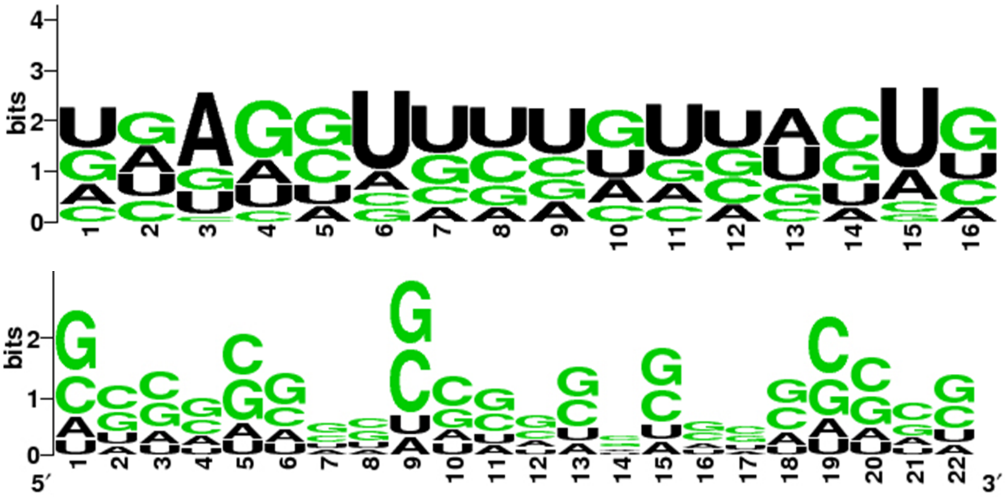


Figure 2.4: Sequence conservation in positive and negative dataset. Each stack of bases represents the relative frequecy of the bases at that position. The letter at the top of the stack is also the tallest and implies the relative abundance at that position.

Multiple sequence analysis with the *MultiAlign* function and *ExistingGapAdjust* option showed that miRNA associated with cancer (positive dataset) has a sequence signature that can be generalised as **AG-UU-U-U–CU**. This result was verified manually with the regional percentage conservation score data and found to be true (Figure 2.5). Additionally the region of consensus falls exactly in the seed region within the position 2-13*nt*. Thus confirming that consensus as a signature for positive dataset. However, the sequence contents outside the seed region also plays a vital role in the regulation, hence analysis of outseed region was also carried out for matches and mismatches.

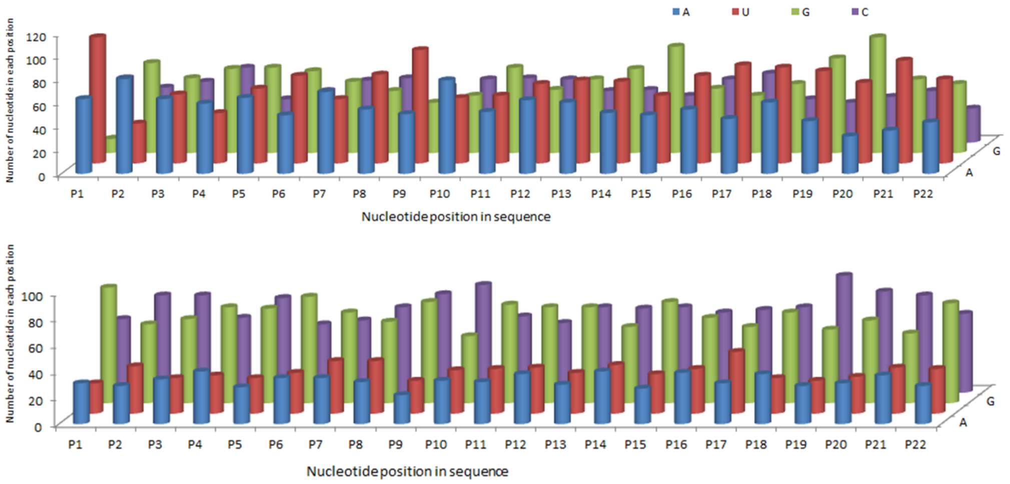


Figure 2.5: Regional percentage conservation score of miRNA sequences in positive and negative dataset.

2.2.2 miRNA-mRNA Interaction

2.2.2.1 Thermodynamics of miRNA-mRNA Binding

The average $\Delta\Delta G$ values for miRNAs associated with cancer and miRNAs not associated with cancer do not show much deviation. Comparing the MFE of both datasets revealed a common value around -25 kcal/mol. Therefore, the values of individual seed pairs were compared for both the datasets. The results of this comparison show that the $\Delta\Delta G$ values appear to be uniform (-24.9 kcal/mol) in positive datasets, whereas in the case of negative dataset it vary significantly (ranging from -27 kcal/mol to -23 kcal/mol) .

2.2.2.2 Analysing Base Pairing Interaction

Pairfinder was used to identify and categorise the seed, outseed, mismatches and bulges in the miRNA interacting with the mRNA. Patches of complementarity (PC) are demarcated as the seed regions, as well as the regions outside seeds where base pairings can occur (but in less than four pairs). All bases outside the PCs are unpaired bases. Quantitatively, the number of unpaired bases in miRNAs not involved in the cancer pathway was quite higher than those in the cancer pathway dataset (Figure 2.6). For a miRNA-mRNA interaction which has a single patch of complementarity to those which have multiple PCs, it was always observed that the number of unpaired bases is more in the interactions

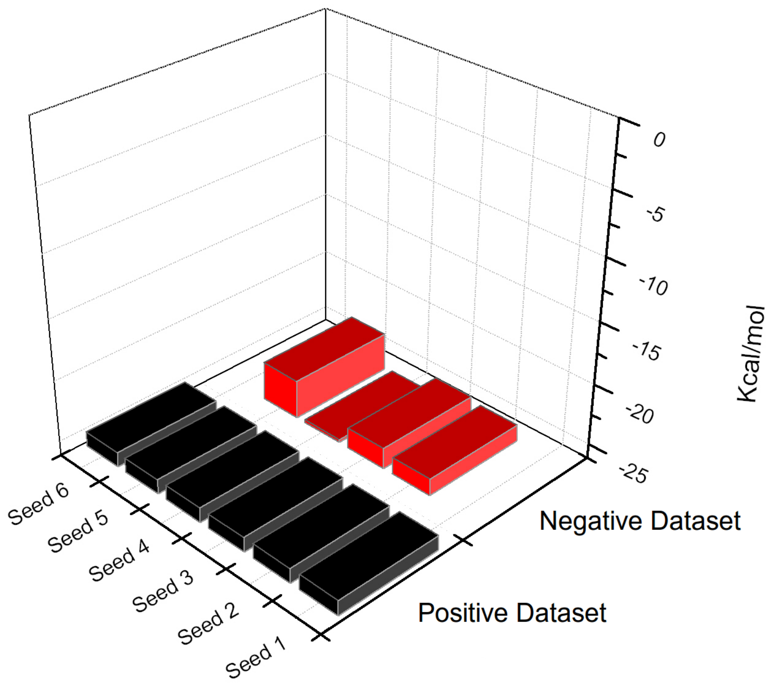


Figure 2.6: Comparison of average minimum free energies of individual seed regions in positive and negative multi-seed hybrids.

involving miRNAs not associated with the cancer pathway (Figure 2.7). This was a pointer to the better complementarity of the miRNA while binding to the respective mRNA of genes associated with cancer.

Additionally, analysing the base pairing distribution in the datasets emphasise some of the common signatures obtained by site depletion analysis [23]. Prior to base pairing distribution analysis, average number of seed in both the dataset was calculated. In positive dataset, the average number of seed region formed is six whereas in case of negative dataset it does not extend beyond four. This is a clear indication of poor complementarity binding and lack of site efficacy. Base pairing in the outseed region also contribute to the site efficacy, hence we calculated base pairing distribution in both the dataset[23].

Base pairing distributions were also analysed in context of Watson-Crick (WC) and non-WC base pairs to study the effect of non-WC base pairing in the hybridised structure (Figure 2.8). In positive dataset, the percentage of AU-base

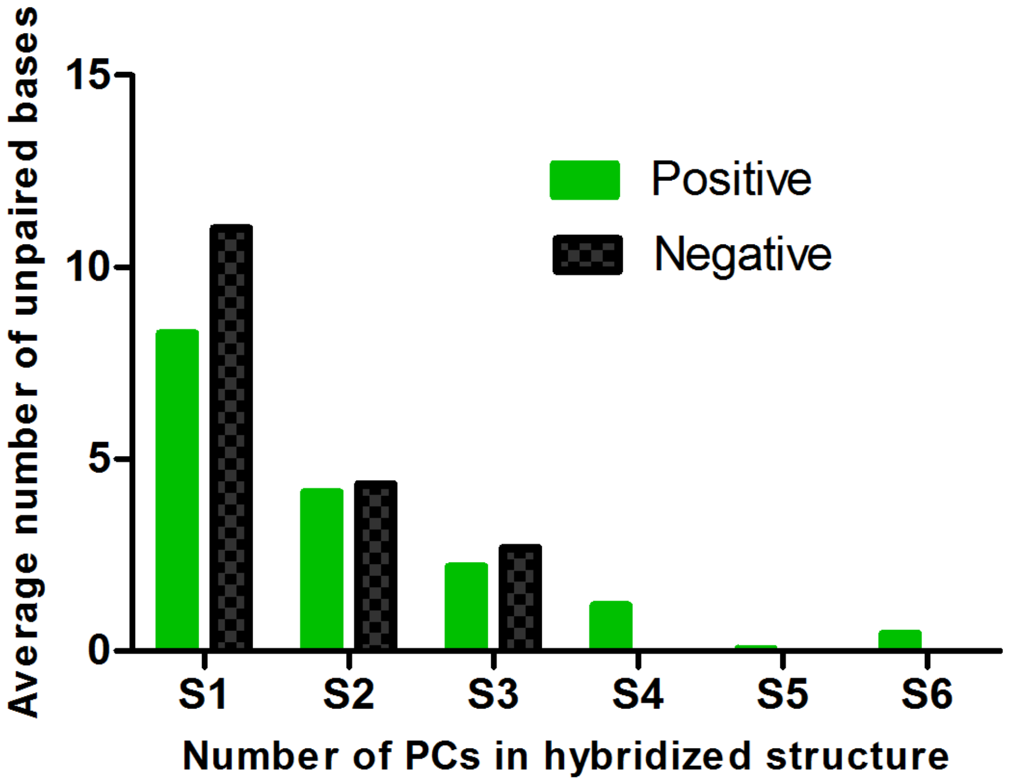


Figure 2.7: Variation in the number of unpaired bases in positive and negative dataset. The first pair of bar stands for the variation in the hybrids having a single patch of complementarity (PC), the second hybrids having two patches and so on.

pairing with in the seed region was much higher, signifying functional binding sites, which in turn indicate signs of complete miRNA destabilization and protein expression level. On the other hand, in outseed region AU counts were consistent throughout the sequence in case of negative dataset but in positive dataset it was found that AU counts are higher around the seed region and drops significantly.

The number of GC in the negative dataset was much higher when compared to positive dataset. Seed sites with higher local GC content will have better pairing with higher folding-free energy due to its stronger bonding however, higher GC content do not signify higher site efficacy[23].

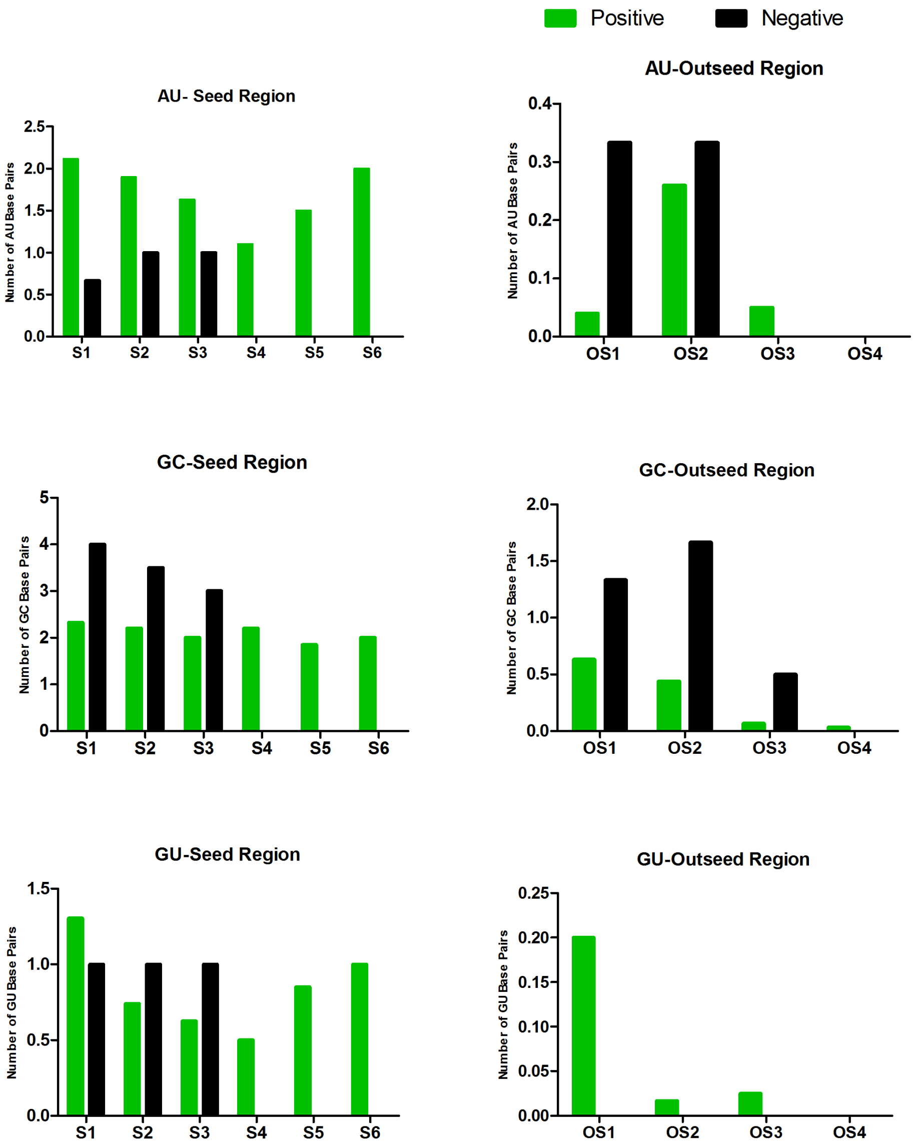


Figure 2.8: Distribution of the Watson-Crick (WC) and non-WC base pairings between positive and negative dataset. The panels on the left are for the pairings in the seed region, while the panels on the right are pairings in the regions outside the seed (OS).

Analysis on non-Watson-Crick pairs revealed that the number of GU pairs is quite higher in positive than the negative. However, presence of wobble base pairs in the hybridised structure does not have any impact on the site efficacy. Consequently, seed score analysis between the two dataset shows higher seed score in case of positive (H_s = 4.108 ± 1.67), which is a clear indication of better stability of hybridisation compared to the negative dataset (H_s = 2.151 ± 1.16). It was also observed that the final seed pair of each individual hybrid shows a noticeable difference in the seed score between positive and negative dataset. In other words, the seed score for the last seed of a multi-seed hybrid is always higher in the case of miRNAs associated with cancer as compared to the miRNAs not associated with cancer. The rise in seed score is accounted by an increase in the number of AU/GC pairings rather than a significant decrease in the number of GU pairs (Table 2.3). This may justify a progressive increase in stability of binding between the mRNA and miRNA in the case of oncomiRs and can be looked upon as a novel binding signature.

Table 2.3: Frequency distribution of base pairs according to seed region in positive and negative hybrids. Frequency distribution of negative dataset have been indicated in parentheses. Bp freq stands for base pairing frequencies

Bp Freq	**Seed 1**			**Seed 2**			**Seed 3**			**Seed 4**			
# Seeds	**AU**	**GU**	**GC**	**AU**	**GU**	**GC**	**AU**	**GU**	**GC**	**AU**	**GU**	**GC**	**SC**
Single	3.1 (2.9)	1.4 (1.1)	4.1 (3.8)	-	-	-	-	-	-	-	-	-	5.6 (5.6)
Two	2.3 (2.2)	1.2 (0.9)	2.8 (3.1)	3.2 (2.6)	1.2 (1)	3.4 (3.3)	-	-	-	-	-	-	9.4 (9.3)
Three	1.9 (1.9)	1.1 (0.9)	2.4 (2.6)	2.1 (2)	0.8 (0.8)	2.3 (2.3)	2.8 (2)	1.1 (1)	3.2 (2)	-	-	-	11.7 (15)
Four	1.7 (1.6)	0.9 (0.4)	2.4 (2.6)	1.7 (1.6)	0.8 (1.4)	2.2 (1.6)	1.7 (1.4)	0.7 (0.7)	2.2 (2.1)	2.6 (2.4)	0.9 (0.6)	3.2 (2.6)	14.2 (9.8)

2.3 Chapter Summary

- It was observed that in miRNA associated with cancer (positive dataset), uracil was the most preferred base in the seed region whereas cytosine was least preferred.
- In terms of hybridisation, the average number of seed region formed in miRNA associated with cancer is six whereas in case of miRNAs not associated with cancer, it does not extend beyond four. This is a clear indication of poor complementarity binding and lack of site efficacy.
- In terms of thermodynamics of binding with respect to $\Delta\Delta$G during seed formation, uniformity of the MFE in case of positive datasets. Randomly varying free energies of binding generally tend to associate with the appearance of sudden mismatches where non-Watson-Crick or wobble base pairings dominate, a trend prevalent in miRNAs not associated with cancer.
- Additionally, miRNA-mRNA interaction data reveals that AU bases are more predominant around the seed region in miRNA associated with cancer.

In this chapter, we obtained somewhat distinguishable signatures that discriminate a miRNA associated with cancer from others. However, verification of miRNA-mRNA test sets with these signatures did not always result in a classification which matched the experimental data. The cause attributed to the mismatch would probably be the consideration of a negative dataset which was randomly generated. Secondly, the number of features chosen for the classification seemed to be insufficient for discrimination. Hence, the onus of the work was shifted to the generation of an unbiased negative set and extraction of more features for the classification process. This prompted the direction of the work to be integrated with a machine learning approach, as documented in the following chapters.

Chapter 3

Identifying miRNAs Involved in Cancer Pathway using Support Vector Machines

Chapter 3

Identifying miRNAs Involved in Cancer Pathway using Support Vector Machines

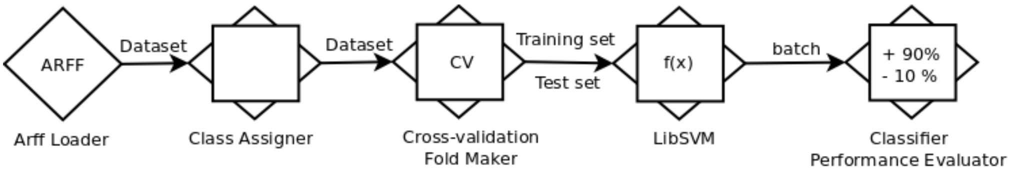


Figure 3.1: Process pipeline for constructing SVM model.

Owing to the shortcomings of the classification process experimented in the previous chapter, the focus of the methodology was shifted to the application of learning algorithms for our problem. The prerequisite would be the construction or catenation of a negative dataset, and extraction of features for the classification. Feature extraction and selection has been the primary focus for most classification processes. Both these attributes are capable of improving learning performance, lowering computational complexity in case of large dataset and building a better classification model. However, unavailability of experimentally validated miRNA dataset is often a limiting factor in finding the most effective features and constructing an optimal classifier. Training on a limited range of dataset introduces bias in the performance of the classifier relative to that trained with a large dataset. Additionally, class imbalance in the dataset results in skewed performance, in particular towards the samples belonging to minority classes. Hence, this chapter deals with a two-step support vector machine-based learning model utilising various features extracted from the dataset of miRNAs associated with cancer versus those which are not involved in cancer.

3.1 Methods

3.1.1 Dataset Preparation

For the purpose of generating a classifier, the first step needed to be undertaken is the construction of a positive and negative miRNA dataset which has been experimentally validated to be associated with cancer (Figure 3.2).

3.1.1.1 Positive and Negative Dataset

Positive datasets for training and testing the classifier were carried over from the previous chapter (the same 239 obtained from Chapter 2-Table 2.1 would serve as the positive dataset). For negative dataset, we utilised the dataset employed in TargetMiner [97]. We extracted 59 negative instances from TargetMiner containing entries from *Mus musculus, Drosophila melanogester* and *Homo sapiens*. Only human mature miRNAs were culled from the dataset and finally, we obtained 32 non-redundant human miRNAs as negative instances (Table 3.1). The negative dataset was constructed on the basis of specific experimental evidence presented in literature for miRNAs whose binding to a target mRNA does not involve gene regulation [35, 98–106]. Selection of random samples for negative dataset was strictly avoided since they may increase the false positives thereby decreasing the performance of the classifier.

To identify miRNAs associated with cancer, we constructed a two-step classifier (i) miRSEQ – Identifies miRNAs associated with cancer based on sequence-based features and (ii) miRINT – validates the prediction obtained from miRSEQ based on features extracted from miRNA-mRNA hybrids. Positive and negative dataset were constructed individually for miRSEQ and miRINT. For miRSEQ, experimentally validated miRNAs (239 instances) and 32 non-redundant human miRNAs (obtained from TargetMiner) served as positive and negative dataset respectively. For miRINT, experimentally validated miRNA-mRNA interactions were further segregated as oncogene interactions (129 instances) and tumour suppressor interactions (110 instances) and considered as the positive and negative dataset (Table 3.2 & 3.3). Class imbalance in the datasets was overcome by applying Synthetic Minority Oversampling Technique (SMOTE) [107] (Chapter 1 section 1.6.2.1).

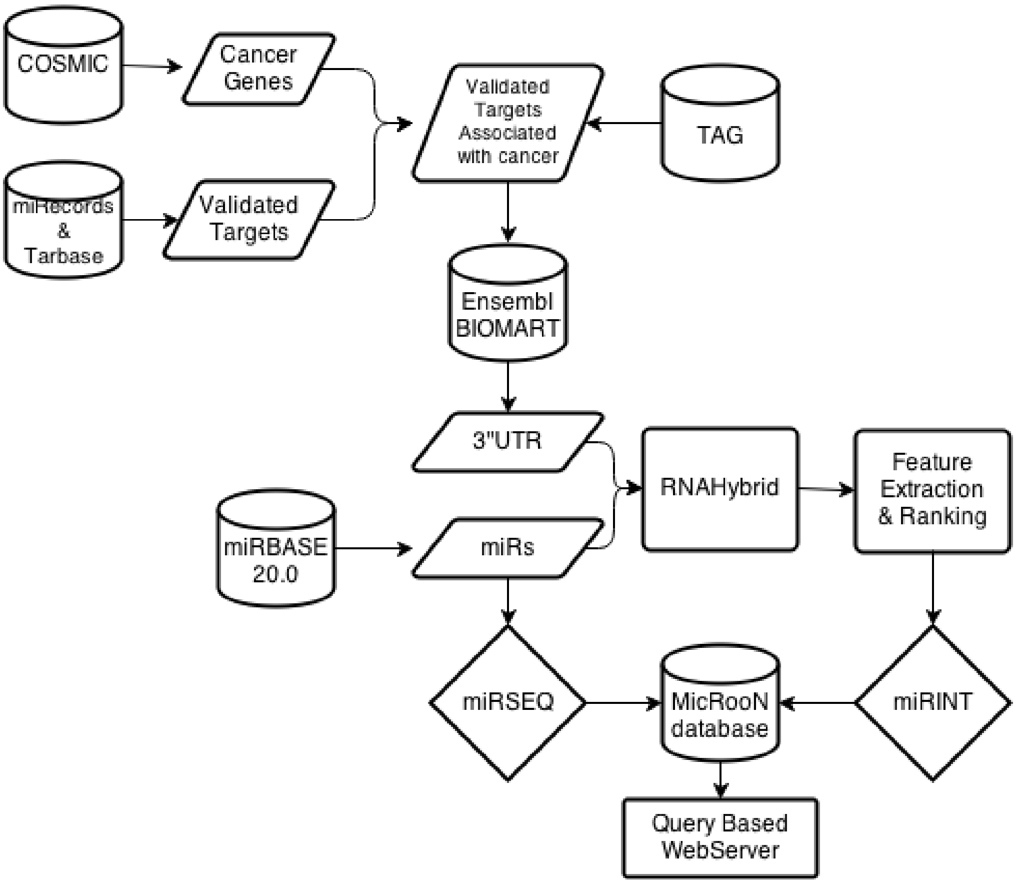


Figure 3.2: Flowchart for data preparation – miRSEQ and miRINT.

To obtain unbiased result in performance measurement, an independent test dataset not utilised in training purpose is required for miRSEQ and miRINT. For miRSEQ, non-validated miRNA dataset obtained from miRBase [37] is considered as test dataset. For miRINT, non-validated miRNAs were allowed to hybridise in RNAhybrid [88] with the list of cancer genes obtained from COSMIC and only the most energetically favoured structures were considered. False predictions were removed using a post-processing filter, MiRTif algorithm [41] and then let into the classification process. MirTif serves as a target interaction filter by providing SVM scores for each prediction obtained from RNAhybrid that distinguishes true targets from false ones (chapter 1 section 1.5.3).

Table 3.1: Negative instances: experimentally validated miRNAs for miRSEQ.

hsa-*miR*-429	hsa-*miR*-145-5p	hsa-*miR*-126-5p	hsa-*miR*-19b-3p	hsa-*miR*-145-3p
hsa-*miR*-141-3p	hsa-*miR*-124-3p	hsa-*miR*-124-5p	hsa-*miR*-155-5p	hsa-*miR*-16-5p
hsa-*miR*-128-3p	hsa-let-7e-3p	hsa-*miR*-24-2-5p	hsa-*miR*-103a-3p	hsa-let-7b-5p
hsa-*miR*-29a-3p	hsa-*miR*-138-5p	hsa-*miR*-29a-5p	hsa-*miR*-29c-3p	hsa-*miR*-29c-5p
hsa-*miR*-375	hsa-*miR*-155-3p	hsa-*miR*-16-1-3p	hsa-*miR*-200a-3p	hsa-*miR*-302a-5p
hsa-let-7b-3p	hsa-*miR*-1	hsa-*miR*-16-2-3p	hsa-*miR*-138-1-3p	hsa-*miR*-15a-3p
hsa-*miR*-126-3p	hsa-*miR*-103a-2-5p	hsa-*miR*-200a-5p	hsa-*miR*-19b-2-5p	hsa-let-7e-5p
hsa-*miR*-19a-3p	hsa-*miR*-128-2-5p	hsa-*miR*-19a-5p	hsa-*miR*-302a-3p	hsa-*miR*-15a-5p
hsa-*miR*-24-1-5p	hsa-*miR*-24-3p	hsa-*miR*-19b-1-5p	hsa-*miR*-141-5p	hsa-*miR*-128-1-5p
hsa-*miR*-138-2-3p				

Table 3.2: Oncogene associated miRNAs: Positive dataset for miRINT.

hsa-*miR*-203a	hsa-*miR*-19b-1-5p	hsa-*miR*-125b-5p	hsa-*miR*-34b-3p	hsa-*miR*-149-3p
hsa-*miR*-451a	hsa-*miR*-20a-5p	hsa-*miR*-184	hsa-*miR*-26a-5p	hsa-*miR*-708-3p
hsa-*miR*-103a-3p	hsa-*miR*-122-5p	hsa-*miR*-107	hsa-let-7b-5p	hsa-*miR*-15b-5p
hsa-*miR*-1	hsa-*miR*-16-5p	hsa-*miR*-133b	hsa-*miR*-26b-5p	hsa-*miR*-204-5p
hsa-*miR*-181a-5p	hsa-*miR*-145-5p	hsa-*miR*-21-5p	hsa-*miR*-302a-5p	hsa-*miR*-34c-5p
hsa-*miR*-127-5p	hsa-*miR*-155-5p	hsa-*miR*-146a-5p	hsa-*miR*-140-5p	hsa-*miR*-24-1-5p
hsa-*miR*-7-5p	hsa-*miR*-200b-5p	hsa-*miR*-125a-5p	hsa-*miR*-324-5p	hsa-*miR*-331-3p
hsa-*miR*-205-5p	hsa-let-7g-5p	hsa-*miR*-22-5p	hsa-*miR*-98-5p	hsa-*miR*-19a-5p
hsa-*miR*-18b-5p	hsa-*miR*-146b-5p	hsa-*miR*-193b-5p	hsa-*miR*-10b-5p	hsa-*miR*-206
hsa-*miR*-222-5p	hsa-*miR*-18a-3p	hsa-*miR*-29b-1-5p	hsa-*miR*-217	hsa-*miR*-302c-5p
hsa-*miR*-100-5p	hsa-*miR*-214-5p	hsa-*miR*-99a-5p	hsa-*miR*-23b-5p	hsa-*miR*-138-5p
hsa-let-7f-5p	hsa-*miR*-224-5p	hsa-*miR*-151a-5p	hsa-*miR*-497-5p	hsa-*miR*-28-5p
hsa-*miR*-30e-5p	hsa-*miR*-181b-5p	hsa-*miR*-150-5p	hsa-*miR*-192-5p	hsa-*miR*-29a-5p
hsa-*miR*-378a-5p	hsa-*miR*-106a-5p	hsa-*miR*-106b-5p	hsa-*miR*-520	hsa-*miR*-147a
hsa-*miR*-93-5p	hsa-*miR*-330-5p	hsa-*miR*-519a-5p	hsa-*miR*-361-5p	

Table 3.3: TSG associated miRNAs: Negative dataset for miRINT.

hsa-*miR*-29a-3p	hsa-*miR*-92a-1-5p	hsa-*miR*-183-5p	hsa-*miR*-106a-5p	hsa-*miR*-186-5p
hsa-*miR*-21-5p	hsa-*miR*-10b-5p	hsa-*miR*-1	hsa-*miR*-125a-5p	hsa-*miR*-124-5p
hsa-*miR*-204-5p	hsa-*miR*-145-5p	hsa-*miR*-101-5p	hsa-*miR*-146b-5p	hsa-*miR*-34a-5p
hsa-*miR*-122-5p	hsa-*miR*-182-5p	hsa-*miR*-141-5p	hsa-*miR*-20a-5p	hsa-*miR*-200a-5p
hsa-*miR*-200b-5p	hsa-*miR*-28-5p	hsa-*miR*-200c-5p	hsa-*miR*-298	hsa-*miR*-192-5p
hsa-*miR*-146a-5p	hsa-*miR*-302a-5p	hsa-*miR*-15a-5p	hsa-*miR*-345-5p	hsa-*miR*-16-5p
hsa-*miR*-212-5p	hsa-let-7g-5p	hsa-*miR*-24-1-5p	hsa-*miR*-125b-5p	hsa-*miR*-103a-2-5p
hsa-*miR*-34b-5p	hsa-*miR*-424-5p	hsa-*miR*-34c-5p	hsa-*miR*-503-5p	hsa-*miR*-25-5p
hsa-*miR*-9-5p	hsa-*miR*-181b-5p	hsa-*miR*-197-5p	hsa-*miR*-17-3p	hsa-*miR*-214-5p
hsa-let-7b-5p	hsa-*miR*-20a-3p	hsa-*miR*-129-5p	hsa-*miR*-23b-3p	hsa-let-7i-5p
hsa-*miR*-107	hsa-*miR*-335-5p	hsa-*miR*-223-5p	hsa-*miR*-675-5p	hsa-*miR*-27a-5p
hsa-*miR*-205-5p	hsa-*miR*-130b-3p	hsa-let-7f-5p	hsa-*miR*-532-5p	hsa-*miR*-224-5p
hsa-*miR*-138-5p	hsa-*miR*-324-3p	hsa-*miR*-155-5p	hsa-*miR*-326	hsa-*miR*-29a-5p
hsa-let-7a-5p	hsa-*miR*-21-3p	hsa-*miR*-181a-5p	hsa-*miR*-18a-3p	hsa-*miR*-124-3p
hsa-*miR*-194-5p	hsa-*miR*-19b-3p	hsa-*miR*-200b-3p	hsa-*miR*-204-3p	hsa-*miR*-200c-3p
hsa-*miR*-200a-3p	hsa-*miR*-93-3p	hsa-*miR*-429	hsa-*miR*-125a-3p	hsa-*miR*-373-3p
hsa-*miR*-106b-3p	hsa-*miR*-15a-3p	hsa-*miR*-132-3p	hsa-*miR*-16-1-3p	hsa-*miR*-221-3p
hsa-*miR*-222-3p	hsa-*miR*-92a-3p	hsa-*miR*-25-3p	hsa-*miR*-30a-3p	hsa-*miR*-30d-3p

Table 3.3 Continued: TSG associated miRNAs.

hsa-*miR*-612	hsa-*miR*-559	hsa-*miR*-106b-5p	hsa-*miR*-132-5p	hsa-*miR*-20b-5p
hsa-*miR*-299-5p	hsa-*miR*-363-5p	hsa-*miR*-492	hsa-*miR*-137	hsa-*miR*-192-3p
hsa-*miR*-26a-5p	hsa-*miR*-155-3p	hsa-*miR*-125b-1-3p	hsa-*miR*-338-3p	
hsa-*miR*-302b-3p	hsa-*miR*-1285-3p	hsa-*miR*-661	hsa-*miR*-212-3p	
hsa-*miR*-19a-3p				

3.1.2 Feature Extraction

Construction of an efficient classifier depends on meticulous feature extraction, since the quality of the feature reflects upon the effective performance of the classifier. For our classifier, features were identified and extracted based on a survey of previous studies and our own indigenous parameters [44, 108–110]. For miRSEQ, 26 features were considered, which included nucleotide positions and repeat information (Figure 3.3). The maximum length of the miRNA used in training was restricted to 22*nt*.

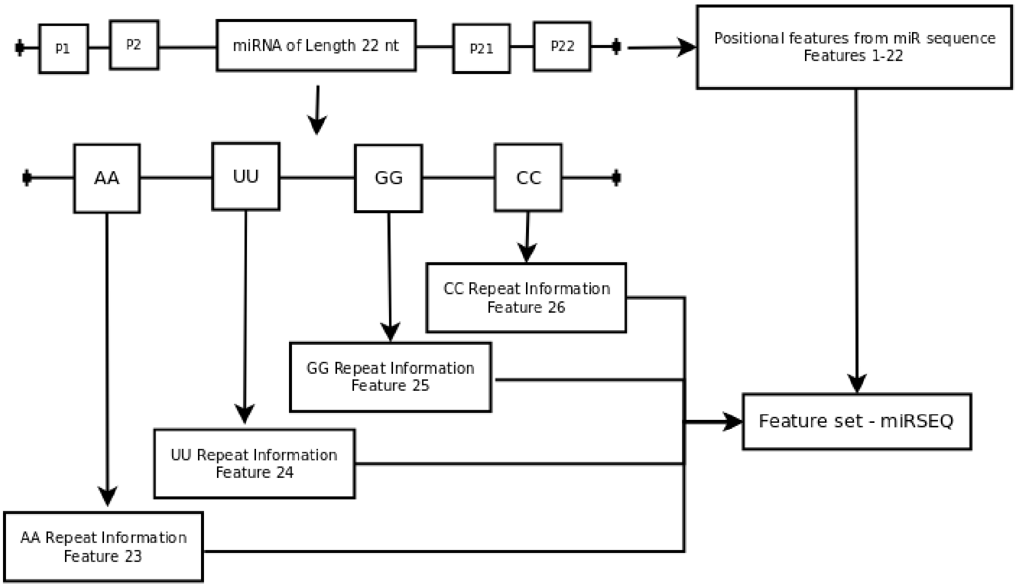


Figure 3.3: Flowchart for feature set preparation– miRSEQ. Only miRNA of length 22 nt is considered for feature selection, where P1 to P22 are the respective nucleotide position in a miRNA sequence and AA, UU, GG, CC are the repeat information of $2w$ size.

For miRINT, a total of 34 features based on the hybridisation profile (miRNA-mRNA interactions) were utilised. miRNA-mRNA hybrids having the best fit in terms of free energy, were obtained using RNAhybrid– ViennaRNA package [88]. A total of 2926 hybrid structures were generated and considered for feature extraction. miRNA-mRNA hybridisation using RNAhybrid may contain false target site predictions. Hence, a post-processing filter was applied to miRNA-mRNA interactions in order to remove the false predictions. An indigenous Perl script, *PairFinder* was used to parse and analyse the hybrids for seeds, regions outside seeds, mismatches and bulges (Appendix B). Seed regions have been defined according to the convention followed in Lekprasert *et al.,* [96] and our previous work [109]. A detailed list of all the 60 features has been presented in Table 3.4 & 3.5.

Table 3.4: miRSEQ - feature set.

Features	Description
Position 1 -22	miRNA nucleotide position
AA, UU,GG,CC	Nucleotide repeat information in miRNA

Table 3.5: miRINT - feature set.

Feature 1 Minimum Free energy of the hybridised structure in kcal/mol.
Feature 2 TB, Total number of base pairing in hybrid structure.
Feature 3 G+C%, Percentage of GC base pairing in the hybridised structure.
Feature 4-9 AU% UA% UG% GU% GC% CG% Percentage base pair composition in all combination in hybridised structure.
Feature 10-12 $\frac{\lvert A-U\rvert}{L}, \frac{\lvert G-U\rvert}{L}, \frac{\lvert G-C\rvert}{L}$ Base Composition per length of the miRNA with which mRNA is hybridised.
Feature 13 Average base pair per stem region.
Feature 14-19 $\frac{\%AU}{S}, \frac{\%UA}{S}, \frac{\%GU}{S}, \frac{\%UG}{S}, \frac{\%GC}{S}, \frac{\%CG}{S}$ Percentage of base pair per seed.
Feature 20 Number of bulges in the hybridised structure.
Feature 21 Unpaired bases in the hybridised structure.

Feature 22

Minimum Free Energy Index **1**, MFEI1 = $\frac{dG}{(G+C\%)}$

Feature 23

Minimum Free Energy Index **2** , MFEI2 = $\frac{dG}{Number\ of\ seeds}$

Feature 24

Minimum Free Energy Index **3**, MFEI3 = $\frac{dG}{Number\ of\ bulges}$

Feature 25

Minimum Free Energy Index **4**, MFEI4 = $\frac{MFE}{Total\ base\ pairing}$

Feature 26

$\frac{Minimum\ Free\ Energy}{(G+C)\%}$

Feature 27

Normalised Minimum free energy of folding, dG = $\frac{Minimum\ Free\ Energy}{Length\ of\ the\ miRNA}$

Feature 28

Normalised base pairing propensity, dP = $\frac{Total\ bases}{L}$

Feature 29

Normalised Base Pairing Probability, dQ = $\sum \frac{(p.\log(p))}{L}$

Feature 30

Normalised Base Pairing Distance, dD = $\sum \frac{(p-p^2)}{L}$

Feature 31

Z-score = $\frac{(MFE-Mean_{MFE})}{SD_{MFE}}$

Feature 32 Number of Mismatches in the hybridised structure.

Feature 33 Number of Watson-crick base pairing in the hybridised structure.

Feature 34 Number of Wobble base pairing in the hybridised structure.

3.1.3 Feature Selection

Features extracted were ranked based on F-score and eventually prioritised for training miRSEQ and miRINT (Figure 3.4). F-score (Fisher score) is the measure of discrimination between the feature and the label (Equation (3.1)). For a given instance x_i = *{1....n}*, the F-score of the j^{th} feature is calculated as

$$F(j) = \frac{(\bar{x}_j^{(+)} - \bar{x}_j)^2 + (\bar{x}_j^{(-)} - \bar{x}_j))^2}{\frac{1}{n_+-1}\sum_{i=1}^{n_+}(x_{i,j}^{(+)} - \bar{x}_j^{(+)})^2 + \frac{1}{n_--1}\sum_{i=1}^{n_-}(x_{i,j}^{(-)} - \bar{x}_j^{(-)})^2} \tag{3.1}$$

where, n_+ and n_- are the number of positive and negative instances, $\bar{x}_j$, $\bar{x}_j^{(+)}$, $\bar{x}_j^{(-)}$ are the average of j^{th} feature, positive-labelled and negative-labelelled instances. The numerator denotes the interclass variance and the denominator is the sum of the variance within each class. A larger F-score indicates that the feature is more discriminative [111].

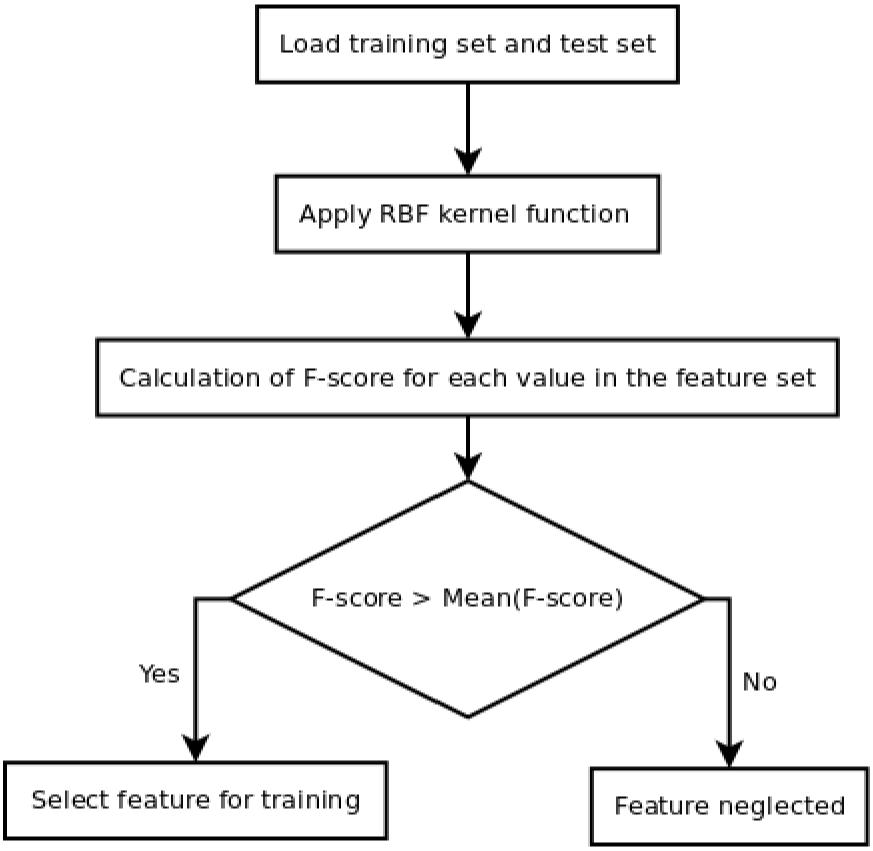

Load training set and test set
Apply RBF kernel function
Calculation of F-score for each value in the feature set
F-score > Mean(F-score)
Yes
No
Select feature for training
Feature neglected

Figure 3.4: Feature selection with F-score ranking. Radial Basis Function (RBF) is used to convert data from low dimensional to high dimensional space with SVM training. Only features with F-score > mean(F-score) of all the features are selected for training process.

Two sets of features for miRSEQ and miRINT were finalised as described before. Additionally, for miRINT, models were built based on the number of seeds (*viz.,* Seed1, Seed2 and Seed3 models), since a miRNA may form a completely complementary hybrid with target miRNA or form Patches of Complementarity (PC) with mismatches or bulges in its hybrid structure (Section 2.2.2.2). If the pattern of hybridisation is not accounted for and all binding considered together, a realistic picture would not encourage. We considered a maximum of three-seed hybrids for the training. Feature ranking was done individually for each of the models and individually trained. To find the optimum subset for the classifier, we followed Recursive Feature Elimination (RFE) for both miRSEQ and miRINT during the training process. Low ranking features were removed one by one iteratively and the performance of the classifier measured until saturation. Removing all the low ranking features at a glance may degrade the performance of the classifier completely; hence the process of optimum feature subset selection was carried out iteratively [78].

3.1.4 Training – miRSEQ and miRINT

In this study, we used LibSVM package for constructing classifier models [112]. Radial Basis Function (RBF) was chosen as the kernel function for the classification process. Parameters for RBF (cost (c) and gamma (γ)) were found using a grid search, which involved the construction of a mesh grid allowing a search for best c and γ ($=\frac{1}{N}$), where N is the number of features. The main disadvantage of training a disease related dataset is the inadequate number of training instances that are experimentally validated and it is important that the same training set should never be used as a test set in any of the experiment because they may lead to over fitting in the model generated. In order to circumvent these problems, a *10-fold* cross validation (*cv*) method was used to evaluate the performance of the classifier during training process.

3.1.5 Performance Evaluation

Due to the difference in numbers between the positive and the negative sets, class imbalance existed in the dataset; so, accuracy could not be chosen as a direct measure of performance for such sets [113]. Hence, performance measures were chosen in compliance with the cross validation rate (*cv-rate*) and Matthew's Correlation Coefficient (MCC). MCC ranges from -1 to 1; a MCC value of 1 indicates the best prediction and a negative value indicates imperfect classification. Matthew's correlation co-efficient (Equation (3.2)) can be calcualted from confusion matrix

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{3.2}$$

3.2 Results and Discussion

Dataset preparation was carried out individually for the classifiers miRSEQ and miRINT (Figure 3.2). Consequently, a total of 271 miRNAs were used in the miRSEQ training. Class imbalance problem in the dataset was overcome by the SMOTE (*k*-Nearest Neighbour (*kNN*) algorithm with no replacement)) method which generated sufficient number of negative instances for the training set. Like most SVM classification problems related to miRNAs, our dataset was also not linearly separable as it was too complicated in nature. RBF was applied to convert all non-linear data from lower dimensional space to linearly separable higher dimensional space.

For miRSEQ, nucleotide position conservation was used initially as the main feature set. However, poor performance of the classifier (*cv-rate* of 45%) prompted us to use nucleotide repeat information with appropriate window size to boost the performance. Selection of appropriate window size (*w*) for nucleotide repeat information was done by measuring the performance of the classifier keeping a sliding window size ranging from 2 to 5. Performance was measured from the plot between *cv-rate* and window size, depicting a clear drop in *cv-rate* when the window size exceeded (*w* = 2) (Figure 3.5). Hence, a *2w* sized repeat was considered for training miRSEQ.

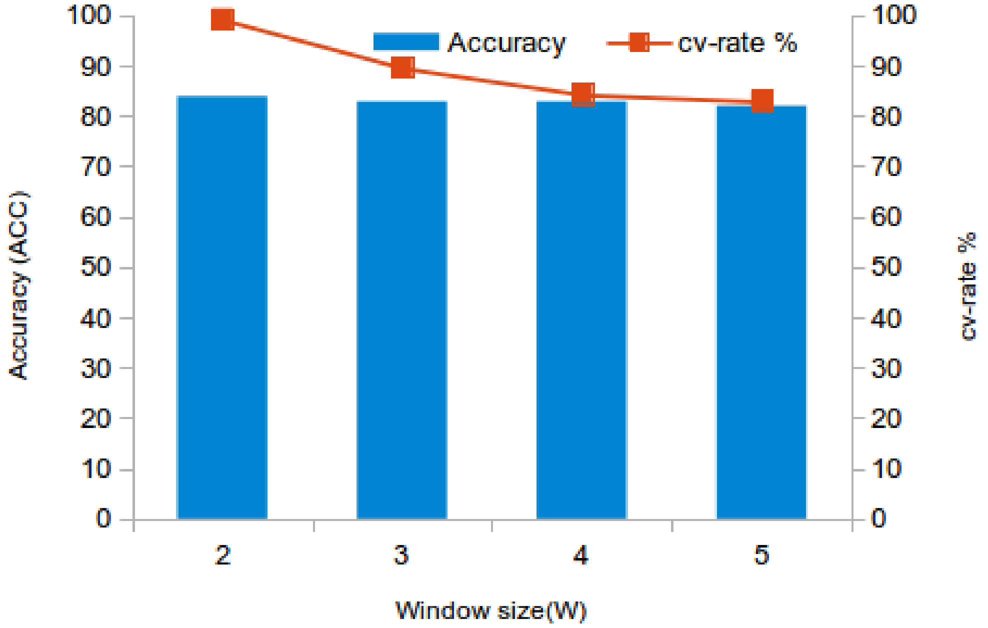


Figure 3.5: Decrease in performance and *cv-rate* with increase in window size(w) for miRSEQ. Accuracy (ACC) has been depicted as bars while the cv-rate is the curve.

The 26 features chosen were ranked by F-score method and Recursive Feature Elimination was performed to find the best subset of features for the dataset as well as retain all the features with very low classification error, respectively (Table 3.6). Optimum subset of features which were finally selected has been depicted in Figure 3.6. Judging by the thickness of the bands in the Circos diagram, the following features yielded the best subset for the classification – Position 1, GG repeat, CC repeat, Position 6, Position 19 and Position 10, in sequence of their relative importance. These features were prioritised to construct the optimal feature subset for miRSEQ and performance measures were carried out which yielded a *cv-rate* of 91.15% and MCC of 0.803 (Table 3.7). Model generation and performance estimation were carried out with the training set (only experimentally validated miRNA sequences) with a *10-fold* cross validation method. The model generated was used on an unseen test set for a primary prediction of the association of those miRNAs with cancer.

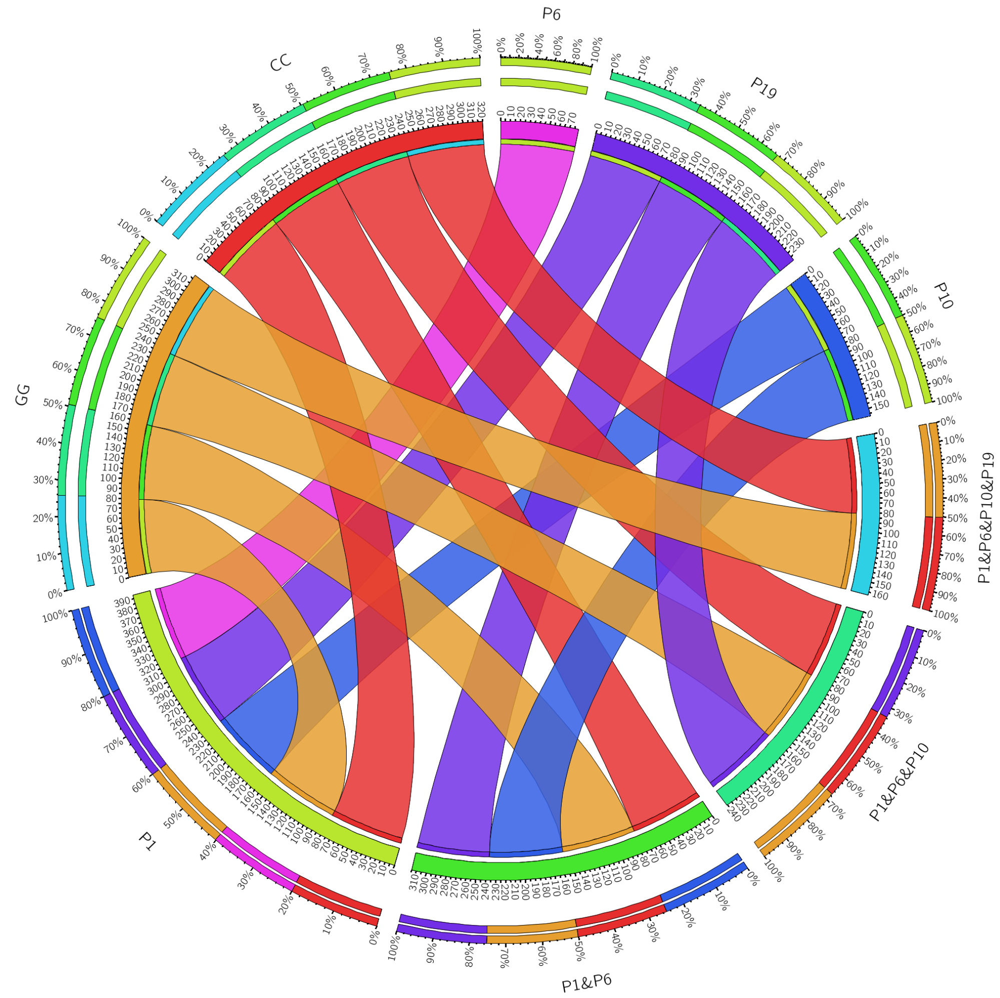


Figure 3.6: Overlap between features subsets (ranked by F-score) selected for miRSEQ. The outer ticks denotes the maximum accuracy of the classifier (in the scale of 100%). The inner ticks denote the accuracy of individual features in the subset in combination with other features. The width of the ribbon denotes the individual accuracy in those combinations.

Table 3.6: F-score ranking with miRSEQ feature set

Features	F-score Ranking	Features	F-score Ranking
GG	0.011026	Position 17	0.001587
Position 10	0.0107	AA	0.001517
UU	0.010417	Position 3	0.001297
Position 22	0.006124	Position 2	0.000924
Position 6	0.005507	Position 19	0.000838
Position 1	0.004927	Position 4	0.000596
Position 13	0.003754	Position 14	0.000496
Position 5	0.002739	Position 20	0.000408
Position 12	0.002508	Position 9	0.000303
Position 11	0.002007	Position 18	0.000289
Position 21	0.001982	Position 16	0.000138
Position 15	0.001788	Position 8	0.000124
Position 26	0.001685	Position 7	0.000089

Table 3.7: miRSEQ - Performance measurement with *10-fold cv-rate*

	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5	AVERAGE
TP	160	160	160	160	160	
TN	77	77	77	77	77	
FP	15	11	12	11	14	
FN	11	11	11	11	12	
MCC	0.7809	0.8106	0.803	0.81	0.7805	**0.803**
ACC	90.11	91.505	91.153	91.505	90.11	**91.153**

For the second classifier miRINT, the choice of features were initially centered around the results of the hybridisation – number of unpaired bases, Watson-Crick and non-Watson-Crick base pairing in and around the seed region, to name a few. The generated model had a very poor performance with low *cv-rate* (<40%). Addition of normalised base pairing and normalised free energy features raised the total number of features to 34 for miRINT and showed a marked improvement in the performance of the classifier, but not to expected levels.

Table 3.8: F-score ranking with miRINT feature set

Seed 1		**Seed 2**		**Seed3**	
Features	**F-score**	**Features**	**F-score**	**Features**	**F-score**
MFE2	10.04584	dQ	590.7773	$\frac{\|G-C\|}{L}$	5.81492
$\frac{\|G-C\|}{L}$	0.731177	Z-score	12.85279	$\frac{\%CG}{S}$	0.524597
$\frac{\|A-U\|}{L}$	0.571601	$\frac{\%UG}{S}$	0.697078	CG%	0.524597
MFE3	0.281471	MFE3	0.25297	AU%	0.504312
$\frac{\|G-U\|}{L}$	0.103704	CG%	0.013531	$\frac{AU\%}{S}$	0.504312
GU%	0.102834	$\frac{\%CG}{S}$	0.013531	MFE3	0.261375
$\frac{\%GU}{S}$	0.102834	GC%	0.008815	Z-Score	0.05433
nWC	0.049696	$\frac{\%GC}{S}$	0.008815	dQ	0.024891
dQ	0.044225	UG%	0.007426	dD	0.02484
Z-score	0.044181	WC	0.006423	dP	0.009802
dD	0.035635	AU%	0.005979	GU%	0.008921

It was therefore, decided to have different models for hybridisation structures with different numbers of seed formation. For each of the different classes, the method of ranking by F-score and prioritisation (as with miRSEQ) was carried out to achieve three different optimal feature subsets (Table 3.8). Precaution was taken to utilise only the non-redundant informative features for model construction. This improved the performance of all the three models of the classifier with good *cv-rate* of 92.19% for single seed (MCC 0.821), 89.54% for two seed (MCC 0.765) and 87.61% for three seed (MCC 0.722) hybrids. The effect of number of features versus the accuracy measurement is given in the graph for all three models (Figure 3.7). Feature selection not only improved the classification but also optimised the total time taken for training the model. The resulting classifier model not only predicts the association of a miRNA with cancer, but also gives an output about that association with either a tumour suppressor gene or an oncogene. Performance measurement carried out on the independent test dataset for miRINT is shown in the Table 3.9.

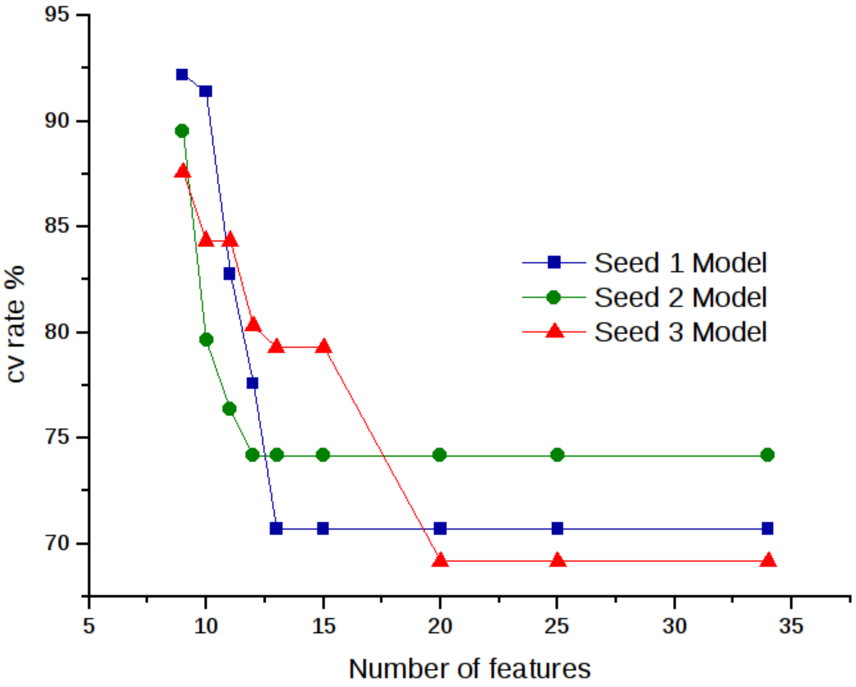


Figure 3.7: Feature selection and effect on *cv-rate* for miRINT for various seed types.

Table 3.9: miRINT - Performance measurement with *10-fold cv-rate*

		RUN 1	RUN 2	RUN 3	RUN 4	RUN 5	AVERAGE
Seed 1	MCC	0.747	0.7977	0.8258	0.8258	0.8211	0.8211
	ACC	88.05	90.76	92.187	92.187	92.187	92.187
Seed 2	MCC	0.765	0.777	0.8244	0.752	0.752	0.765
	ACC	89.54	90.1315	92.255	88.961	88.961	89.54
Seed 3	MCC	0.722	0.712	0.7301	0.782	0.722	0.722
	ACC	87.61	87.224	88	90.41	87.61	87.61

3.3 Chapter Summary

- A new negative dataset was constructed and a Support Vector Machine based binary classifier was employed to predict a miRNA associated with cancer.

- The imbalance in the positive and negative datasets was offset by using Synthetic Minority Over-sampling Technique.

- A two-step classifier was developed on the basis of features extracted from mature miRNA sequences and miRNA-mRNA interactions, named miRSEQ and miRINT.

- Initial classification process was quite complex mainly due to the unique behaviour of miRNA-mRNA interactions. So in order to suppress the complexity, we considered features both from within and outside the seed regions. Features extracted outside the seed region along with several site specific features provided quite a good classification performance.

- A non-validated miRNA would go through the two-step classifier, first through miRSEQ and then through miRINT, before a final prediction.

With the available training datasets, the LIBSVM model constructed performed satisfactorily. However, during the feature selection process it was observed that several low ranking features have been eliminated completely – which may possess a better discrimination when combined with other features. Expectedly, the accuracy of predictions derived from the kernel-based LIBSVM left more room for improvements. Hence, in Chapter 4, work involving a search for better learning algorithm that utilises all the informative features and also provide a better prediction accuracy is undertaken.

Chapter 4

MicRooN - An Ensemble Classifier for Identifying miRNAs Associated with Cancer

Chapter 4

MicRooN - An Ensemble Classifier for Identifying miRNAs Associated with Cancer

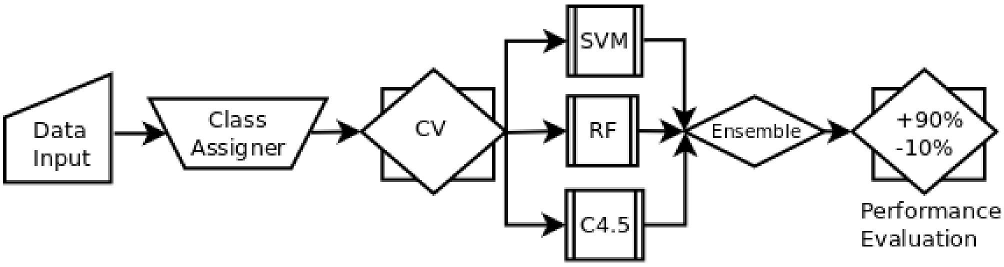


Figure 4.1: Process pipeline for comparing three different algorithms and construction of ensemble.

Performance of a constructed model depends on the nature of the dataset used for training and the choice of the learning algorithm. Apart from selecting a suitable algorithm for training, feature subset selection is also important since the features selected not only represent discrimination in the dataset but also provide a path for scaling the performance of the classifier. Class imbalance in the miRNA dataset and the methods utilised to overcome it affected the prediction performance of the previously constructed SVM classifier. Average performance with the independent test dataset, high computational complexity (kernel transformation) associated with SVM and resampling at the data level to overcome class imbalance in the miRNA dataset collectively left room for improvement and search for a better learning algorithm. In this chapter, we provide an in-depth analysis in terms of comparing three different algorithms *viz.,* Support Vector Machines (SVM), Random Forest (RF) and C4.5 to subsequently construct an ensemble-classifier for the prediction of miRNAs associated with cancer.

4.1 Methods

4.1.1 Dataset Preparation

For the present study, we utilised both the training and test datasets from Chapter 3.

4.1.2 Feature Extraction

For miRSEQ, we utilised positional information of experimentally validated miRNAs involved in cancer. Additionally, a 2 window ($2W$) repeat information was also utilised to train miRSEQ. For miRINT, we utilised 34 features extracted from miRNA-mRNA hybrid structures. Complete feature extraction process was discussed elaborately in Chapter 3.

4.1.3 Optimal Feature Selection

In the previous Chapter, features were ranked solely based on the information entropy and several low-ranked features were neglected during the actual training process. Abruptly removing several low ranking features may degrade the performance of the classifier, hence we employed Recursive Feature Elimination process (RFE) [78] to obtain optimal subset of features for the training process (Figure 4.2).

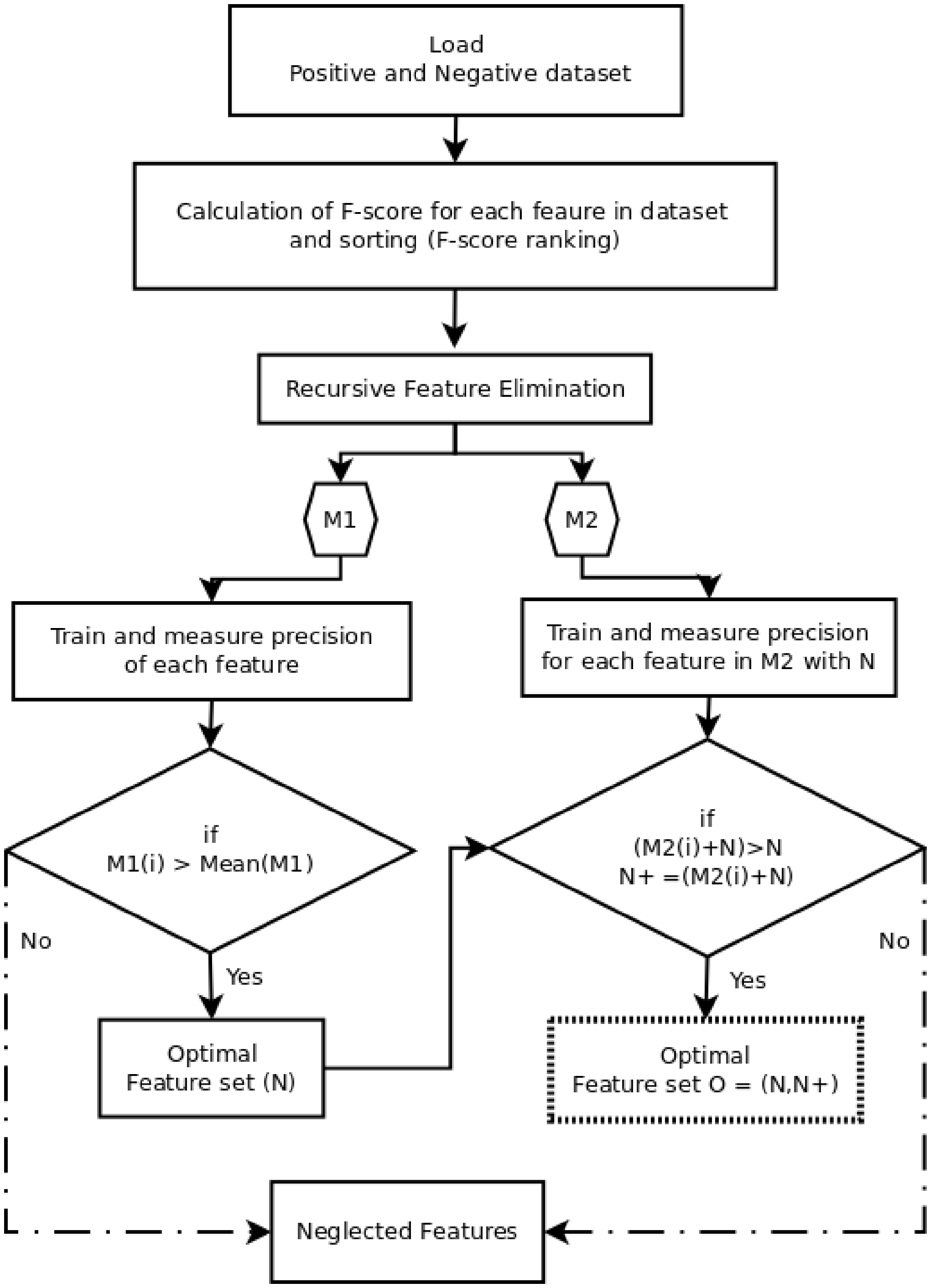

Load
Positive and Negative dataset
Calculation of F-score for each feaure in dataset and sorting (F-score ranking)
Recursive Feature Elimination
M1
M2
Train and measure precision of each feature
Train and measure precision for each feature in M2 with N
if
M1(i) > Mean(M1)
if
(M2(i)+N)>N
N+ =(M2(i)+N)
No
Yes
Yes
No
Optimal
Feature set (N)
Optimal
Feature set O = (N,N+)
Neglected Features

Figure 4.2: Optimal feature selection with Recursive Feature Elimination (RFE).

Finding the optimal subset includes generation of feature subset, evaluation of the feature subset, stopping criterion, and result validation [114]. Initially, features extracted were ranked based on F-score, later RFE was employed in two stages for identifying the optimal subset of features. The steps involved are :

Let $\{M\}$ be the feature subset extracted from training dataset. Initially features were ranked {*Rank $M*} and divided into two subsets *viz.,* $\{M_1\}$ and $\{M_2\}$. Let $\{M_1\}$ contain the most discriminative features and the rest in $\{M_2\}$. Information gain in F-score for individual feature is calculated based on the entropy formula:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \tag{4.1}$$

In stage I, an optimal subset $\{N\}$ was constructed by removing lowest ranked discriminative feature in each iteration from $\{M_1\}$ until $\{N\}$ reaches ACC$_{max}$. In stage-II, optimal subset $\{N^+\}$ was constructed from $\{M_2\}$. These are features that boost the performance of optimal subset $\{N\}$. Finally, an optimal subset $\{O\}$ containing features from $\{N\}$ and $\{N^+\}$ is obtained and utilised for training. For both miRSEQ and miRINT, we employed the above described feature selection process to obtain optimal subset of features.

4.1.4 Class Imbalance

In chapter 3, we employed SMOTE to oversample the training set, which obtained quite a good accuracy but with a increase in the False Negative Rates (FNR). In this chapter, we employed cost-sensitive methods for reweighing instances rather than randomly resampling the entire dataset. A cost-matrix was constructed prior to each training process based on the total cost assigned to each class. We constructed a 2×2 cost matrix for both miRSEQ and miRINT. Cost matrix employed during training process to overcome class imbalance in miRSEQ and miRINT is given in Table 4.1.

Table 4.1: Example of a cost-sensitive matrix

	Predicted class		
		+ve	-ve
Actual class	+ve	$C(0{,}0)$	$C(0{,}1)$
	-ve	$C(1{,}0)$	$C(1{,}1)$

where *C*(0,0) is the cost associated to the prediction of TP, *C*(1,1) to the prediction of TN, *C*(1,0) to the prediction of FP and *C*(0,1) for prediction of FN. Cost is calculated by assuming the condition *C*(1,1) = *C*(0,0) = 0 (*i.e.,* the costs associated to predict TP and TN is zero).

(A).

$$Cost\,matrix\,employed\,in\,miRSEQ = \begin{pmatrix} 0 & 32 \\ 239 & 0 \end{pmatrix}$$

(B).

$$Cost\,matrix\,employed\,in\,miRINT - Seed\,1 = \begin{pmatrix} 0 & 17 \\ 41 & 0 \end{pmatrix}$$

(C).

$$Cost\,matrix\,employed\,in\,miRINT - Seed\,2 = \begin{pmatrix} 0 & 86 \\ 189 & 0 \end{pmatrix}$$

(D).

$$Cost\,matrix\,employed\,in\,miRINT - Seed\,3 = \begin{pmatrix} 0 & 61 \\ 131 & 0 \end{pmatrix}$$

$$Total\,Cost = FNrate \times C(0,1) + FPrate \times C(1,0) \tag{4.2}$$

where *C*(1,0) and *C*(0,1) are the costs associated to the prediction of FP and FN respectively.

4.1.5 Algorithm Selection and Training

Selection of the learning algorithm is dataset specific. In our study, we employed three learning algorithms *viz.,* the kernel-based SVM and the tree-based Random Forest and C4.5. Learning algorithms were compared and implemented in WEKA (Waikato Environment for Knowledge Analysis) environment [115]. C4.5 was implemented as J48 in WEKA.

4.1.5.1 Hyperparameter Selection

Hyperparameters for the three learning algorithms were optimised individually (Table 4.2). In order to find the best optimised hyperparameters as well as to lower the computational cost, we employed random search method rather than dimension-based grid search. In case of LIBSVM, we chose Radial Basis Function (RBF) as kernel function; hyperparameters cost (c) and gamma (γ) were obtained from random search method [116]. A similar random search method was employed for both RF and C4.5 to identify the optimum number of trees constructed per training process and the number of attributes utilised during individual tree construction. In C4.5, pruning was enabled (by default) to avoid overfitting during training.

Table 4.2: Parameters used in training algorithms– SVM, RF and C4.5

Algorithm	**Parameters**
Support Vector Machines (SVM)	Kernel Type = Radial Basis Function (RBF)
	C = 32; Gamma (γ) = 0.01; Degree (for kernal) = 3
	coeff0 = 0.0; Shrink = True
	Replace Missing Values = True; Loss = 0.1
	Normalize = False; Probability Estimates = False
Random Forest (**RF**)	Maxdepth = 0
	Numfeatures = 0 (Utilises all features for bootstrapping)
	Numtrees = 100
	Printtrees = False (Optional)
C4.5	Binary Split = False
	Confidence Factor = 0.25; minobj = 2 ; numfold = 3
	Sub-tree Raising = true ; unpruned = False ; use Laplace = False
	use MDL Correction = True

4.1.6 Comparing Algorithms

Initially we compared all three learning algorithms to assess the performance of an individual algorithm with selected features and training set. Later, an ensemble model was constructed by aggregating results with majority voting technique.

4.1.6.1 Kernel-based Classifier SVM

Kernel methods are known to learn from instances — these methods learn from the training process, rather than from fixed parameters [117]. Construction of Support Vector Machines (SVM) was pioneered by Vapnik *et al.,* [60], generalising on the previously derived hyperplane method. Linear hyperplane method separates training instances based on their weights and stores them as subsets; when a test dataset arrives, it is either classified above or below the hyperplane. Construction of the SVM is based on the type of dataset used in the training process *i.e.,* either linear or non-linear. Most cancer related datasets are not linearly separable, hence optimal hyperplane is not always obtained. Therefore, introducing a non-linear kernel function to map the feature representation into higher dimensional space and separation with a maximum margin hyperplane (soft margins) were undertaken. Generally, these soft margins classify the testset based on the nearest mapped feature representation obtained from the training set [118]. The choice of kernel function plays a critical role in the classification process, which in turn depends on the optimised kernel parameter gamma (γ) and the soft margin parameter cost (C). Optimisation of γ and C is usually done using a random search, each combination of parameters validated using *10-fold* cross validation (*cv*) methods and then selected.

4.1.6.2 Decision Trees

Decision trees, a very commonly used method to classify a dataset, use a set of binary rules applied to calculate a target value based on several attributes obtained. The classification is carried out based on the weights of the attributes. A decision tree has a structure consisting of internal nodes (where a decision function is executed) and external nodes, connected by branches. Two decision tree methods, *viz.,* Random Forest [43] and C4.5 [63] were considered for comparison with the kernel methods. Although both these algorithms are decision tree-based, the process of constructing the trees are completely different (Section 1.6.1.4).

4.1.6.3 Adaboost-meta Classifier

In order to obtain a good performance measure with the existing algorithms, we boosted the classifier with Adaboost algorithm [66]. Adaboost is employed when the trained base classifier performs poorly with the optimal feature set and optimising the hyperparameters for the respective algorithm leads to saturation. Adaboost trains a given weak or a base classifier repeatedly in series of iterations say *i = {1... n}* with user-defined weights. Initially, the weight for all instances are kept constant and in the consequent steps, weight of the misclassified instance is increased so that the weak classifier is trained more on misclassified instances only. Thus the overall classification of an adaboost algorithm is given by

$$C_{boost}(X) = \pm(\Sigma_m \alpha_m C_m(X)) \tag{4.3}$$

where α is the measure of quality of classifier C_m. Boosting algorithms do not overfit and are highly sensitive to outlier/noise (Section 1.6.1.5).

4.1.6.4 Evaluating Classifier Performance

The performance of the classifier was evaluated using various measures, given that our main concern was to identify the best classifier with low misclassification error. Training and testing process was carried out with a *10-fold* cross validation (*cv*) for both miRSEQ and miRINT in WEKA. Cross validation method was also employed during feature selection and ranking. Since our dataset used for training process was highly imbalanced, utilising Accuracy (ACC) would not project the actual classification performance. Hence, in our study we utilised precision, recall, Area Under the Receiver Operating curve (AUC) and F-Measure [69].

In general, F-measure is defined as the harmonic mean of precision and recall and it is derived as

$$Precision = \frac{(TP)}{(TP + FP)} \tag{4.4}$$

$$Recall = \frac{(TP)}{(TP + FN)} \tag{4.5}$$

$$\textit{F-measure} = \frac{2\ \times Recall\ \times Precision}{Recall + Precision} \tag{4.6}$$

AUC is calculated by

$$AUC = \int_0^1 (1 - f(y))dy = 1 - \int_0^1 f(y)dy \tag{4.7}$$

4.1.7 Ensemble Classifier

One major concern with miRNA dataset is the class imbalance problem. Training individual classifier with cost-sensitive based approach and optimising hyperparameters may boost the performance of the classifier to a certain extent. However, bias due to class imbalance plays a critical role in deteriorating the performance. Additionally, the misclassification error generated by the three individually optimised algorithms are not common [119]. To address the problem, we employed ensemble-based classifier approach to obtain a generalised prediction and low misclassification error. An ensemble generates a hypothesis that is not necessarily within the hypothesis generated by individual learning algorithm. Majority voting technique was used as a final predictor to aggregate the results obtained from the three classifiers [120].

4.2 Results and Discussion

4.2.1 Optimal Feature Selection

Recursive Feature Elimination (RFE) process was employed in two-stages to select the optimal subset of features for both miRSEQ and miRINT. Revalidation was done concurrently with *10-fold cv* to observe the effect of feature removal on the performance of the classifier (Figure 4.3)

For miRSEQ, we ranked 26 sequence-based features initially with F-score. Ranked features were divided into two subsets $\{M_1\}$ and $\{M_2\}$. Search for optimal subset initiated with $\{M_1\}$ containing most discriminative features, whereas $\{M_2\}$ contained all the low ranked features. Features from $\{M_2\}$ were considered in the later stage in the feature selection process for boosting the performance of the classifier obtained as a result of optimal feature

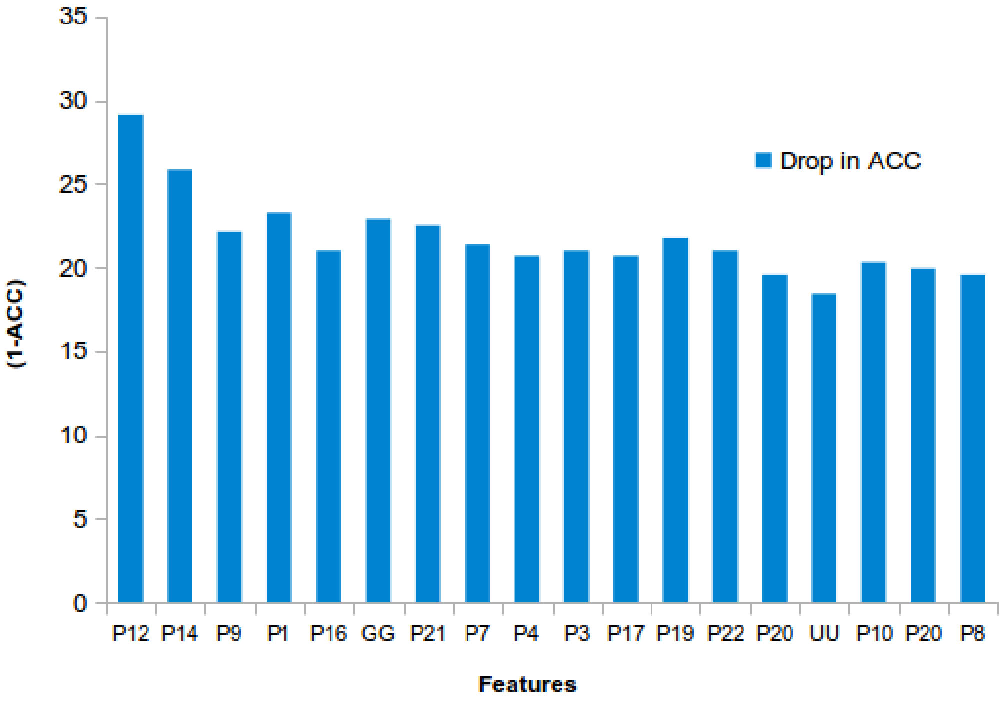


Figure 4.3: Effect on performance with feature removal during optimal feature selection with *10-fold cv.* Only features with maximum accuracy variation are shown.

from $\{M_1\}$. Features were removed iteratively one at a time until ACC$_{max}$ was reached. We employed meta-cost-sensitive Adaboost-LibSVM algorithm for training and evaluating the performance during the entire feature selection process (Figure 4.3). We obtained an optimal subset of features containing position {P2, P3, P5, P6, P7, P9, P11, P13, P15, P16, P18, P20, P21} and base repeat of {AA, GG, CC}.

For miRINT, a similar two-step RFE method was employed. However, due to difference in the number of seed, presence or absence of imperfection in the structure (*viz.,* bulges, mismatches *etc.,*) different optimal feature subsets were obtained for different hybrid models. Although, 60 features were identified and extracted from the miRNA dataset, it was observed that the features were highly correlated. Generally, correlated features do not add any discrimination to the classification process but enhance already existing information in the feature set. Reinforcing the same information may boost or degrade the performance of the classifier, depending on the context of the information. Hence, features with negligible class discrimination were manually removed.

The optimal feature set chosen (after manual culling) for training the three different classifier models are given in (Table 4.3). *Meta*-Adaboost algorithm was employed to improve the prediction obtained from several weak classifiers. In order to overcome class imbalance in the miRNA dataset, we employed cost-sensitive approach by constructing a 2×2 cost matrix [119] as per the Equation (4.2).

Table 4.3: Optimal feature subset for different seed-based hybrid.

Seed 1	Seed 2	Seed 3
(G+C)%	MFE	MFE
GC%	(G+C)%	(G+C)%
CG %	UA%	UA%
$\frac{\lvert A-U \rvert}{L}$	UG%	GC%
$\frac{\lvert G-C \rvert}{L}$	GU%	CG%
$\frac{\%AU}{S}$	GC%	$\frac{\lvert A-U \rvert}{L}$
$\frac{\%GC}{S}$	CG%	$\frac{\lvert G-U \rvert}{L}$
$\frac{\%CG}{S}$	$\frac{\lvert G-C \rvert}{L}$	$\frac{\lvert G-C \rvert}{L}$
MFE1	$\frac{\%AU}{S}$	$\frac{UA}{S}$
MFE2	$\frac{\%UA}{S}$	$\frac{GC}{S}$
MFE4	$\frac{\%UG}{S}$	$\frac{CG}{S}$
$\frac{MFE}{GC\%}$	$\frac{\%GC}{S}$	UP
dQ	$\frac{\%CG}{S}$	MFE2
Z-score	MFE3	MFE3
	Z-score	$\frac{MFE}{GC\%}$
		dQ
		Z-score

4.2.2 Comparison of Learning Algorithms

For the kernel-based SVM, we used Radial Basis Function(RBF) as a kernel function. Both kernel parameters gamma (γ) and the soft margin parameter (C) were found using random search and revalidated with a *10-fold cv* method. Pruning was set to be true in case of C4.5 by default to avoid over-fitting of data. whereas, in the case of RF, trees were naturally grown without pruning. Total cost was calculated using the Equation (4.2) and utilized for training the imbalanced dataset. Since there are no specific rules suggested for applying the cost ratios on imbalanced dataset [121], we applied cost only for misclassified instances (rare cases). Rare instances were identified and segregated manually. Applying cost for correctly classified instances did not have any effect on the performance of the classifier. In our present study, misclassified instances generally belonged to the negative class and imbalance in the dataset occurred due to the scarcity of experimentally identified miRNAs not associated with cancer. Total cost assigned for miRSEQ and miRINT is given in Table 1. Adaboost was needed to boost the performance of the prediction models, to further improve performance with the ranked features and the cost matrix utilized [119].

4.2.2.1 Performance metrics for miRSEQ

A comparative analysis of the performance of the three classifiers (Table 4.4) revealed that the model generated for miRSEQ with RF during the training process performed better than the other two learning algorithms (with less False Negative prediction). While precision for the RF method (0.8) was better than those obtained with SVM (0.7) and C4.5 (0.7), the AUC curve also returned the best measures for RF as well. However, the difference between the two decision trees was not so pronounced when compared with the SVM method. The reason for the comparatively better prediction efficacy of RF than the other two classifiers may be due to its inherent ensemble method. The AUC curve (Figure 4.4) in this instance was constructed between the False Positive Rate (FPR) and True Positive Rate (TPR) in order to achieve an unbiased and non-parametric measurement. AUC curve is typically a function, that explains how much evidence is necessary for the model to predict a response, and what is the outcome of these responses.

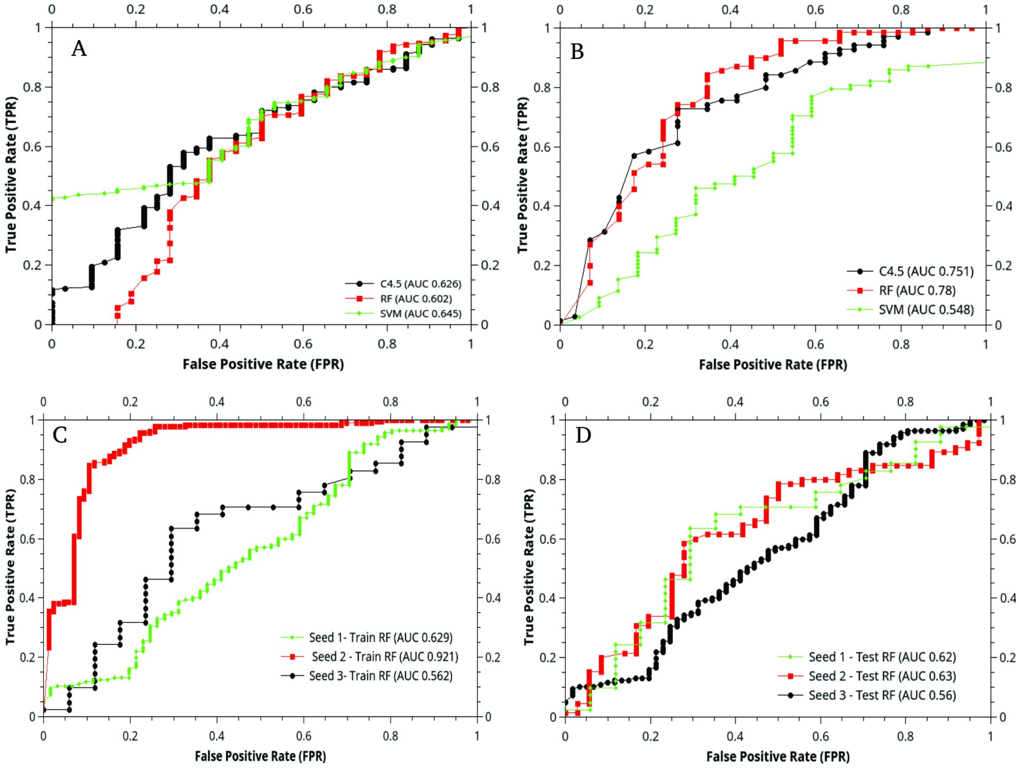


Figure 4.4: Area under the Receiver Operating Curve for miRSEQ and miRINT. For miRSEQ, AUC of all three classifiers with (A) training set (B) test dataset is plotted. For miRINT, AUC of RF classifier with (C) training set and (D) test dataset is plotted; the other two classifiers showed a poor performance with an average AUC of 0.5 for all seed models.

Table 4.4: Performance evaluation with *10-fold cv* for miRSEQ.

	Precision	**Recall**	**F-measure**	**AUC**
		Training set - miRSEQ		
SVM	0.804	0.713	0.750	0.645
RF	0.805	0.797	0.801	0.602
C4.5	0.801	0.762	0.780	0.626
		Test set - miRSEQ		
SVM	0.71	0.680	0.693	0.548
RF	0.802	0.778	0.780	0.780
C4.5	0.748	0.747	0.692	0.751

The efficacy of the RF method is primarily due to its inherent majority voting process. RF involves random selection of features to split each node, growing an ensemble of trees and voting for the most popular class and there is a considerable reduction of error rates. In case of C4.5, the classification efficiency is slightly lower in terms of precision and AUC, which may be due to the abrupt labeling method followed during model generation by this algorithm. The feature with the highest information gain is chosen as root and other branch splits are based on the search for the next higher information gain attribute [43]. For a non-linear dataset (like ours), decision tree building is halted when a single instance can fit into more than one of the attributes. The pause in the decision tree building process results in the final branch being given the classification label neglecting rest of the features, therefore, leading to marginally lower performance when compared to RF.

For the kernel-based SVM, non-linearity is handled quite easily by plotting non-linear data to a higher dimensional space with Radial Basis Function (RBF) or any other kernel function that suits the dataset. Even when the features are not very discriminative, mapping them to a higher dimensional space increases performance of the classifier. However, SVM is often beleaguered with the problem of over-fitting of data, which might have been the reason behind the precision (0.71) and very low AUC of 0.548.

4.2.2.2 Performance metrics for miRINT

In case of miRINT, performance measure was evaluated individually for seed based models. This was done to avoid generalization of the classification process which results from variation in hybridization patterns in different seed models [122]. Hence, hybrids which formed a single seed were considered differently from those that form two or three seed regions. Even after accounting for the correlated parameters, the performance of all three learning algorithms was not discernible and moderate overall.

On closer inspection, the values of precision for the three algorithms show a comparative edge for RF over the other two classifiers. AUC values though vary significantly; for seed 1 and seed 2 models, RF reported an AUC of 0.627 and 0.62 respectively (Figure 4.4), whereas the AUC values did not have any distinction in case of seed 3 models. Low precision and AUC are attributed to the number of correlated features. Among the three learning algorithms, C4.5 performed very poorly in all the seed models (Precision 0.365, 0.414 and 0.173), while SVM performed marginally better (Table 4.5). Though the feature selection was extensive, and the method of oversampling rational to the best possible limits, none of the classifiers in isolation seemed to return gold standard performance. This is a pointer to the fact that a single learning algorithm might not be efficient in predicting the association of miRNA with cancer.

Table 4.5: Performance evaluation with *10-fold cv* for miRINT.

	Precision	Recall	F-measure	AUC
Seed 1 - Training Set				
SVM	0.500	0.707	0.580	0.540
RF	0.657	0.638	0.646	0.629
C4.5	0.526	0.534	0.530	0.428
Seed 1 - Test Set				
SVM	0.562	0.426	0.485	0.520
RF	0.637	0.554	0.593	0.627
C4.5	0.558	0.557	0.557	0.484
Seed 2 - Training Set				
SVM	0.794	0.705	0.749	0.559
RF	0.889	0.884	0.878	0.921
C4.5	0.558	0.557	0.557	0.484
Seed 2 - Test Set				
SVM	0.414	0.644	0.504	0.500
RF	0.539	0.624	0.0.578	0.620
C4.5	0.414	0.644	0.504	0.500
Seed 3 - Training Set				
SVM	0.569	0.374	0.331	0.497
RF	0.663	0.662	0.641	0.562
C4.5	0.521	0.369	0.500	0.444
Seed 3 - Test Set				
SVM	0.341	0.584	0.431	0.500
RF	0.541	0.574	0.556	0.500
C4.5	0.173	0.416	0.244	0.500

4.2.3 MicRooN - an Ensemble Model

An ensemble classifier was constructed by aggregating three individually optimized learning algorithms with majority voting technique as final predictor. The performance of the classifier thus constructed was evaluated with independent test dataset. It was observed that in case of miRSEQ, the ensemble classfier performed marginally equal to that of RF; whereas in case of miRINT, ensemble classifier constructed outperformed all the three learning models (Table 4.6).

Comparison of three learning algorithms along with ensemble model for individual seed-based models are illustrated (Figure 4.5).

Table 4.6: Performance measures of miRSEQ and miRINT ensemble models with independent test dataset

		Precision	**Recall**	**F**-measure	AUC
miRSEQ-Ensemble		0.802	0.778	0.790	0.780
	Seed 1	0.703	0.672	0.687	0.648
miRINT -Ensemble	Seed 2	0.927	0.900	0.913	0.930
	Seed 3	0.584	0.618	0.601	0.531

The list of novel miRNA to be associated with cancer and their predicted interaction obtained from ensemble classifier is given in Appendix C.

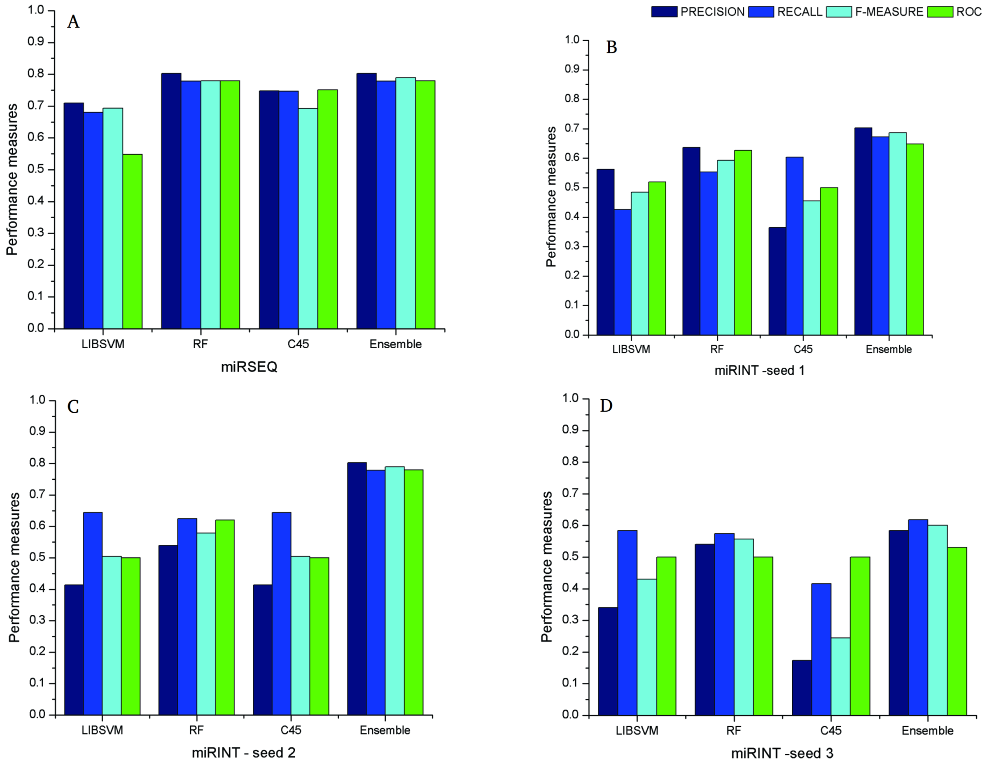


Figure 4.5: Comparison of performance measures – individual learning algorithm and ensemble model. (A) miRSEQ model (B) miRINT - Seed1 model (C) miRINT - Seed2 model and (D) miRINT - Seed3 model

4.3 Chapter Summary

- Comparison of kernel-based and Decision Tree-based learning algorithms revealed that RF algorithm performed better with miRNA dataset, irrespective of the number of seeds formed with target mRNA. The major reason for higher performance in RF is due to its inherent ensemble-based prediction.
- C4.5 algorithm performed the worst due to its abrupt labelling followed during the training process. SVM performed marginally lower than the RF and the computational cost to transform non-linear data from lower to higher dimensional space was quite higher, when compared to tree-based algorithms.
- An ensemble classifier was constructed from three different learning algorithms *viz.,* Support Vector Machines, Random Forest and C4.5. Majority voting technique was employed to aggregate the results obtained from individual classifiers.
- A two-stage RFE was employed to obtain optimal subset of features and the selected features showed more discrimination in terms of increased performance when compared with models generated solely with F-score ranking.
- Cost-sensitive methods utilised to overcome class imbalance increased the performance of the constructed ensemble classifier significantly. Random sample generated to overcome class imbalance with SMOTE performed poorly when challenged with independent test datasets.

Although the data resampling method employed in the previous chapter reported a higher performance measure, the number of FN predicted with independent test dataset was much higher when compared with cost-sensitive method utilised. Thus, we conclude that for miRNA dataset with high class imbalance (like ours), resampling at the algorithm level (cost-sensitive methods) performs better than data-level resampling (SMOTE) and an individual learning algorithm is inefficient in predicting miRNA associated with cancer, hence ensemble classifier is preferred.

Chapter 5

MicRooN - A Webserver for Identifying miRNAs Associated with Cancer

Chapter 5

MicRooN - A Webserver for Identifying miRNAs Associated with Cancer

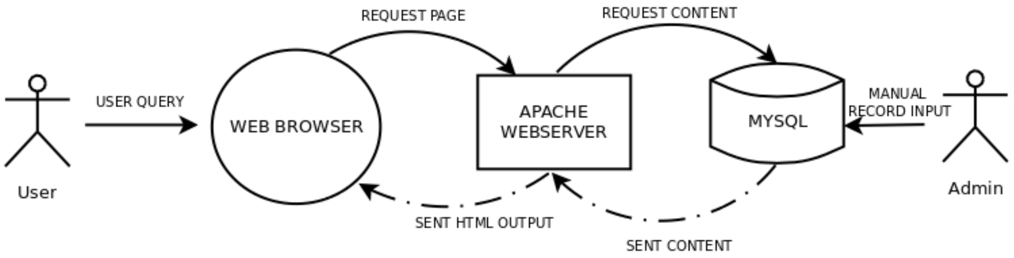


Figure 5.1: Process pipeline for identifying miRNAs associated with cancer using MicRooN.

Predicting potential miRNA function from the vast amount of biological data is an important problem for the miRNA biologist. This is compounded by the unique behaviour in miRNA-mRNA binding and the economically unfavourable experimental studies associated with miRNAs. On the other hand, existing algorithms to unwind the function of miRNA are guided by certain parameters/functionalities and additionally involve several manual processes which are time consuming. Hence in this chapter, we present a machine learning based model MicRooN, which predicts miRNAs associated with cancer. The predictions obtained from MicRooN are sorted and documented in a database, MicRooNdb. The tool provides a minimalistic web-based user interface and offers two search modes – simple miRNA-based search and an advanced miRNA-mRNA based search. The database is customized based on the number of seeds in the hybrid structure. Currently the database contains miRNA entries for *Homo sapiens* binding to 3′UTR only.

5.1 Methods and Results

5.1.1 MicRooN

The MicRooNdb is constructed based on the predictions obtained from MicRooN – a machine based ensemble learning model. In brief, MicRooN predicts miRNA associated with cancer based on several features extracted from mature miRNA sequences and miRNA-mRNA interactions. Predictions obtained from MicRooN are sorted based on the number of seeds formed in the hybrid structure and are recorded in MicRooNdb.

Predictions are obtained from trained ensemble model constructed with 60 features including sequence, thermodynamics and miRNA-mRNA based interactions. Sequence features mainly include nucleotide position and repeat based information (Section 3.1.2) and miRNA-mRNA interaction features include features extracted from base pairing frequencies in the entire hybrid structure, base-pairing in the seed region and outseed region, thermodynamics of binding, minimum free energy based features and normalized free energy parameters (Figure 5.2).

Currently, the tool is constructed based on the binding of miRNA to 3$'$UTR region only and considers a maximum of three seed regions. Hybridized miRNA-mRNA structures are obtained from RNAhybrid – a Vienna Package. Only hybrids with best fit in terms of binding energy are considered for feature selection and for further analysis. MicRooN allows GU wobble base pairing (non-Watson-Crick) in seed and outseed regions since they are essential in preserving target specificity in miRNA-mRNA interactions [91–95].

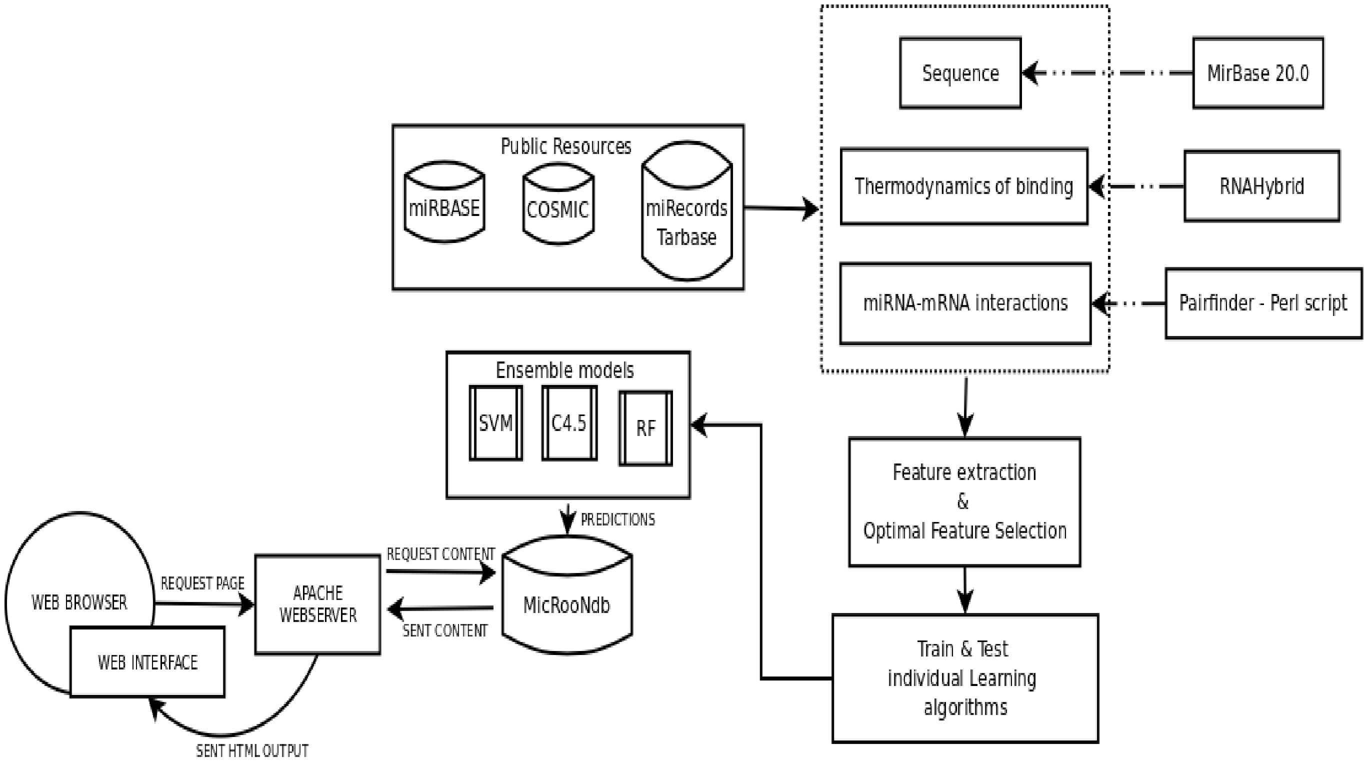


Figure 5.2: Flowchart for prediction of miRNAs associated with cancer using MicRooN

5.1.2 MicRooNdb

5.1.2.1 Data Sources

Mature miRNA sequences and miRNA-mRNA interaction data were retrieved from public database only. Mature miRNAs were extracted from miRBase 20.0 [37]. List of genes involved in cancer was extracted from COSMIC [84]. miRNA-mRNA interaction data were retrieved from miRecords [46] and miRTarbase [45]. TargetMiner [97] provided us with a list of miRNA not associated with cancer. List of 60 features utilized for training classifiers were obtained based on a survey of previous studies and our own indigenous parameters [44, 108–110].

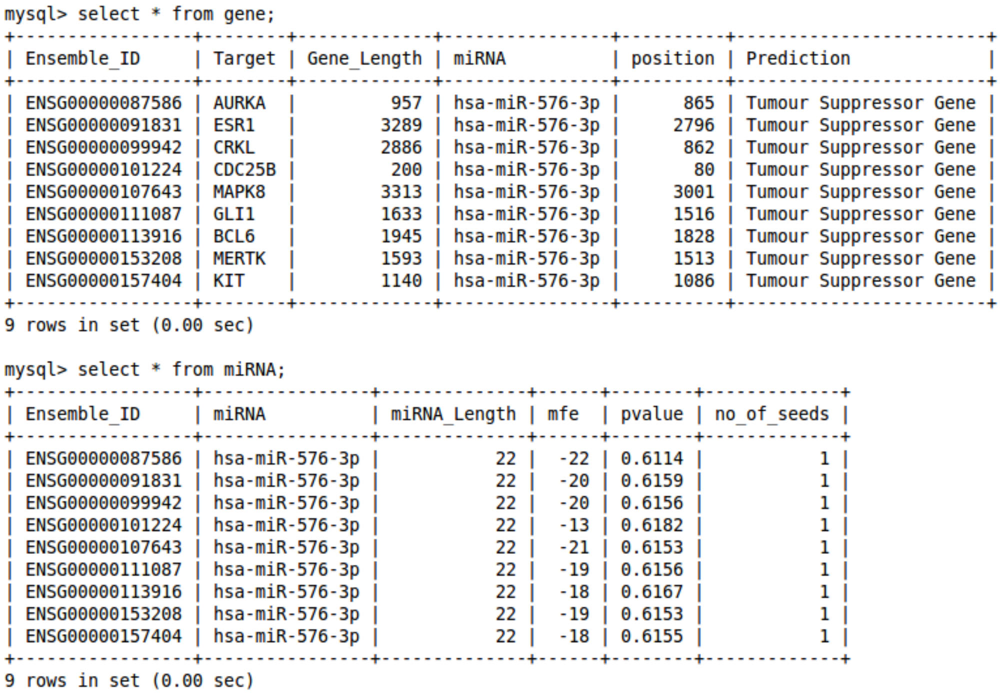
Currently, the tool is built with dataset obtained from the following database version only (Table 5.1). Regular updates will be done as per the release of datasets.

Table 5.1: Dataset used in construction of MicRooN.

Data Extracted	Database Version
Mature miRNA	miRBase 20.0
Cancer Genes	COSMIC *v70*
Validated miRNA-mRNA targets	miRecords (April 27, 2013) miRTarBase (Release 4.5: Nov 1, 2013)

5.1.2.2 MicRooNdb – Database Design

MicRooN database was built using MySQL – an open-source database management system. MySQL provides a fast, flexible, secure and stable medium for retrieving, updating and entering information into the database by an authorized user. The database houses information about the various miRNA-mRNA interactions and their associations with cancer in the form two tables as shown in the flat files (Figures 5.2 & 5.3). The table “GENE”, stores details of Ensemble ID, target name, gene length, miRNA binding, position at which interaction occurs and finally the prediction as either oncogenes or tumor suppressor genes. Since, a single miRNA can bind to multiple positions on a mRNA, we considered ensemble ID along with miRNA to be the primary key. Considering either miRNA or ensemble ID as primary key will result in data redundancy and hence we assigned both fields as primary key. The table “MIRNA”stores ensemble ID, miRNA, miRNA Length, Minimum Free Energy (MFE), p-value (confidence of binding) and Number of seeds formed during a miRNA-mRNA interaction. Similar to table “GENE”, in this table also we assigned primary key as ensemble ID and miRNA.



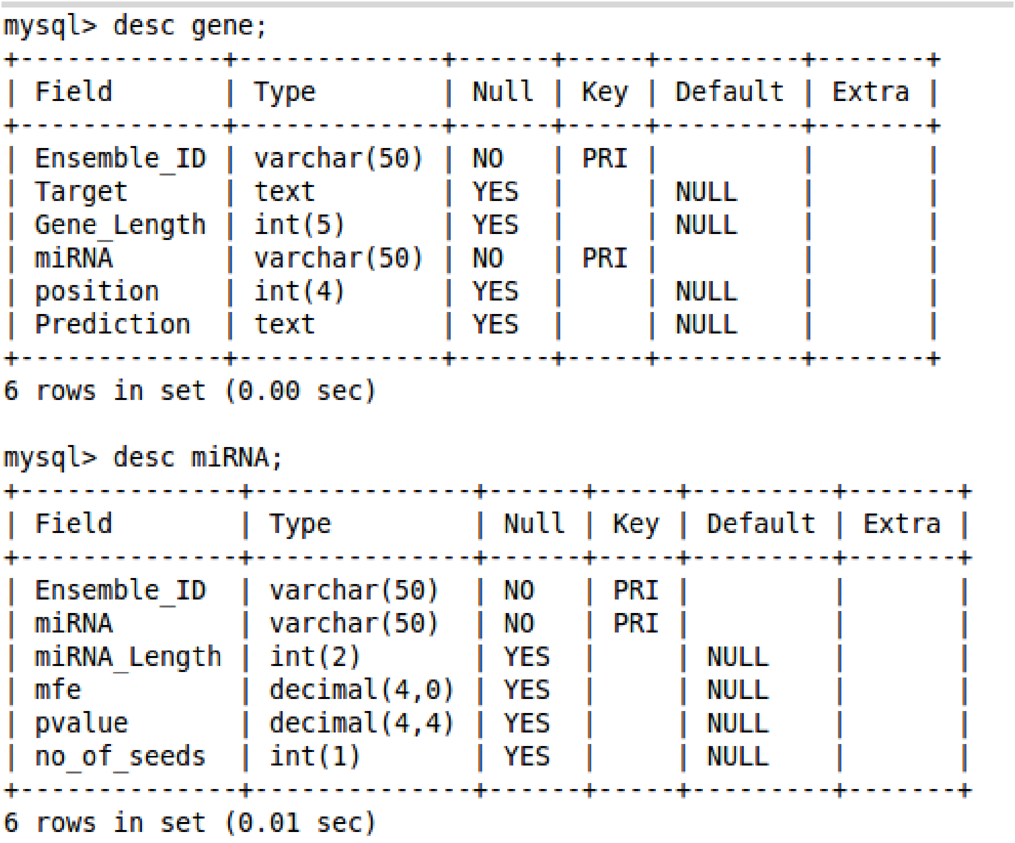
```
mysql> select * from gene;
+-----------------+--------+-------------+-----------------+----------+-------------------------+
| Ensemble_ID     | Target | Gene_Length | miRNA           | position | Prediction              |
+-----------------+--------+-------------+-----------------+----------+-------------------------+
| ENSG00000087586 | AURKA  |         957 | hsa-miR-576-3p  |      865 | Tumour Suppressor Gene  |
| ENSG00000091831 | ESR1   |        3289 | hsa-miR-576-3p  |     2796 | Tumour Suppressor Gene  |
| ENSG00000099942 | CRKL   |        2886 | hsa-miR-576-3p  |      862 | Tumour Suppressor Gene  |
| ENSG00000101224 | CDC25B |         200 | hsa-miR-576-3p  |       80 | Tumour Suppressor Gene  |
| ENSG00000107643 | MAPK8  |        3313 | hsa-miR-576-3p  |     3001 | Tumour Suppressor Gene  |
| ENSG00000111087 | GLI1   |        1633 | hsa-miR-576-3p  |     1516 | Tumour Suppressor Gene  |
| ENSG00000113916 | BCL6   |        1945 | hsa-miR-576-3p  |     1828 | Tumour Suppressor Gene  |
| ENSG00000153208 | MERTK  |        1593 | hsa-miR-576-3p  |     1513 | Tumour Suppressor Gene  |
| ENSG00000157404 | KIT    |        1140 | hsa-miR-576-3p  |     1086 | Tumour Suppressor Gene  |
+-----------------+--------+-------------+-----------------+----------+-------------------------+
9 rows in set (0.00 sec)

mysql> select * from miRNA;
+-----------------+-----------------+--------------+------+--------+-------------+
| Ensemble_ID     | miRNA           | miRNA_Length | mfe  | pvalue | no_of_seeds |
+-----------------+-----------------+--------------+------+--------+-------------+
| ENSG00000087586 | hsa-miR-576-3p  |           22 |  -22 | 0.6114 |           1 |
| ENSG00000091831 | hsa-miR-576-3p  |           22 |  -20 | 0.6159 |           1 |
| ENSG00000099942 | hsa-miR-576-3p  |           22 |  -20 | 0.6156 |           1 |
| ENSG00000101224 | hsa-miR-576-3p  |           22 |  -13 | 0.6182 |           1 |
| ENSG00000107643 | hsa-miR-576-3p  |           22 |  -21 | 0.6153 |           1 |
| ENSG00000111087 | hsa-miR-576-3p  |           22 |  -19 | 0.6156 |           1 |
| ENSG00000113916 | hsa-miR-576-3p  |           22 |  -18 | 0.6167 |           1 |
| ENSG00000153208 | hsa-miR-576-3p  |           22 |  -19 | 0.6153 |           1 |
| ENSG00000157404 | hsa-miR-576-3p  |           22 |  -18 | 0.6155 |           1 |
+-----------------+-----------------+--------------+------+--------+-------------+
9 rows in set (0.00 sec)
```

Figure 5.3: MicRooNdb - Flat file for table GENE and MIRNA

5.1.2.3 Database Normalization

The basic goal of database normalization is to ensure that the key data elements are maintained without redundancy from table to table within the database [123]. Generally, database normalization is done to obtain an internally consistent and accurate records. In our study, 1NF normalization was achieved by segregating gene and miRNA based information to construct two distinct tables of uniform size *i.e.,* related information were grouped together (Figure 5.4).

Multiple value column indicating target ID were parsed as ensemble ID, database ID and target name. Ensemble ID along with specific miRNA



```
mysql> desc gene;
+-------------+-------------+------+-----+---------+-------+
| Field       | Type        | Null | Key | Default | Extra |
+-------------+-------------+------+-----+---------+-------+
| Ensemble_ID | varchar(50) | NO   | PRI |         |       |
| Target      | text        | YES  |     | NULL    |       |
| Gene_Length | int(5)      | YES  |     | NULL    |       |
| miRNA       | varchar(50) | NO   | PRI |         |       |
| position    | int(4)      | YES  |     | NULL    |       |
| Prediction  | text        | YES  |     | NULL    |       |
+-------------+-------------+------+-----+---------+-------+
6 rows in set (0.00 sec)

mysql> desc miRNA;
+--------------+--------------+------+-----+---------+-------+
| Field        | Type         | Null | Key | Default | Extra |
+--------------+--------------+------+-----+---------+-------+
| Ensemble_ID  | varchar(50)  | NO   | PRI |         |       |
| miRNA        | varchar(50)  | NO   | PRI |         |       |
| miRNA_Length | int(2)       | YES  |     | NULL    |       |
| mfe          | decimal(4,0) | YES  |     | NULL    |       |
| pvalue       | decimal(4,4) | YES  |     | NULL    |       |
| no_of_seeds  | int(1)       | YES  |     | NULL    |       |
+--------------+--------------+------+-----+---------+-------+
6 rows in set (0.01 sec)
```

Figure 5.4: MicRooNdb - tables GENE and MIRNA

was considered as the primary key. In order to achieve 2NF normalization, we removed fields that are not dependent on the primary key. We removed database ID which was irrelevant and also introduced a large data redundancy. To achieve 3NF, the database should meet the requirements of both 1NF and 2NF. Usually, columns that are not fully dependent upon the primary key were removed. However, in our study there were no such irrelevant fields. The Entity relationship diagram for the two tables is shown in Figure 5.5

5.1.2.4 Database Access and Web Interface

The MicRooN is designed and developed on an Apache webserver with PHP-HTML. User query is processed by MySQL and is passed as HTML output to user interface. The user interface connects with the MicRooNdb through *mysqli* function, a prepared statement used in PHP– a must for web application security as they protect it from MySQL injection vulnerability.

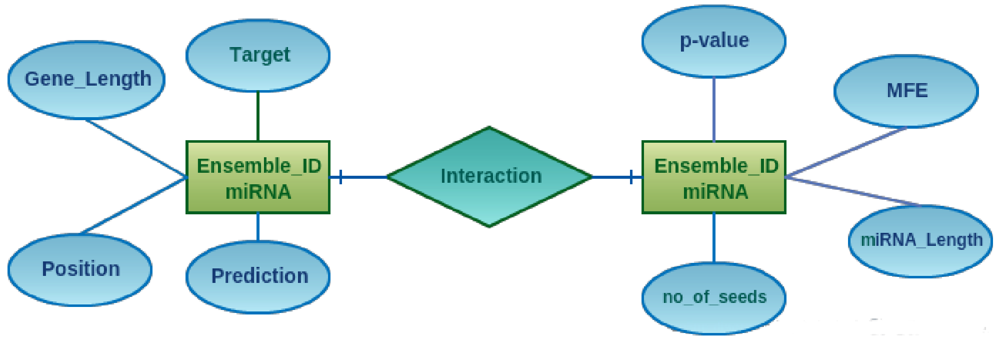
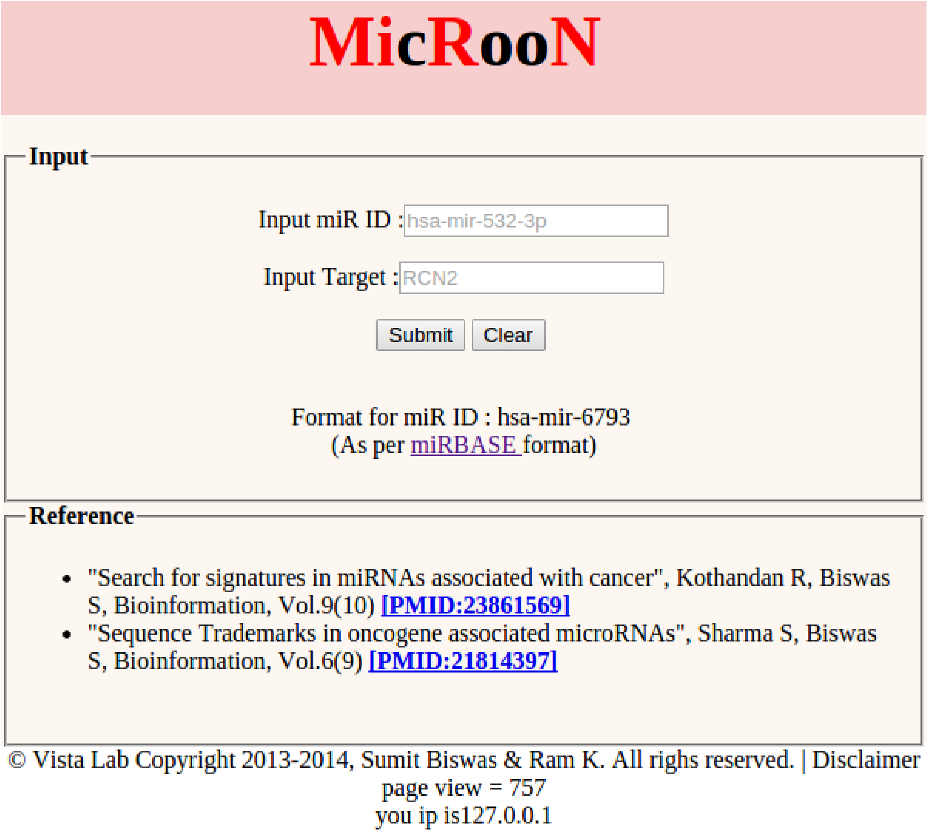


Figure 5.5: MicRooNdb - Entity Relationship (ER) diagram for table GENE and MIRNA.

The web interface connects to the MicRooNdb database *via mysqli* function as shown below

```
$mysqli = new mysqli("localhost", "root", "password", "mysql");
```

The web interface is extremely user friendly with two search option (Figure 5.5). The user can either enter the miRNA ID to query targets to which the miRNA binds. Usually, a miRNA can bind to multiple positions in mRNA, hence the result is always more than one hit and they are ordered based on the best p-value. p-value indicates the confidence of binding in a miRNA-mRNA interaction and is obtained from RNAhybrid hybridization.



MicRooN

Input

Input miR ID : hsa-mir-532-3p

Input Target : RCN2

Format for miR ID : hsa-mir-6793
(As per miRBASE format)

Reference

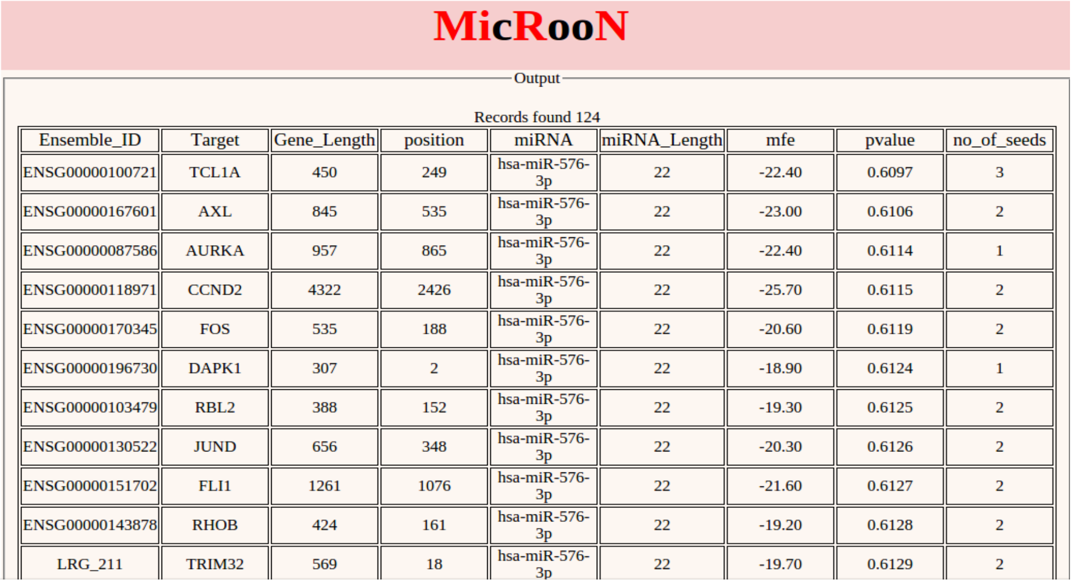
- "Search for signatures in miRNAs associated with cancer", Kothandan R, Biswas S, Bioinformation, Vol.9(10) **[PMID:23861569]**
- "Sequence Trademarks in oncogene associated microRNAs", Sharma S, Biswas S, Bioinformation, Vol.6(9) **[PMID:21814397]**

© Vista Lab Copyright 2013-2014, Sumit Biswas & Ram K. All righs reserved. | Disclaimer
page view = 757
you ip is127.0.0.1

Figure 5.6: MicRooNdb - User interface with two search options. User can search based on miRNA ID or based on target name.

5.1.2.5 Querying MicRooNdb

For a given miRNA based search, the web interface displays details about ensemble ID of the target it binds, target name, position at which the interaction occurs, miRNA ID, miRNA Length, miRNA-mRNA prediction to be either oncogenes or TSG and the number of seeds it forms during the interaction (Figure 5.7). The total number of hits obtained is also provided for a user query. In case, if the user is more specific about the interaction *i.e.,* if user requires a particular miRNA binding to a specific mRNA target, then the target name can be provided in the search option. A more precise miRNA-mRNA interaction is displayed with all the additional features describing the interaction (Figure 5.8).



MicRooN

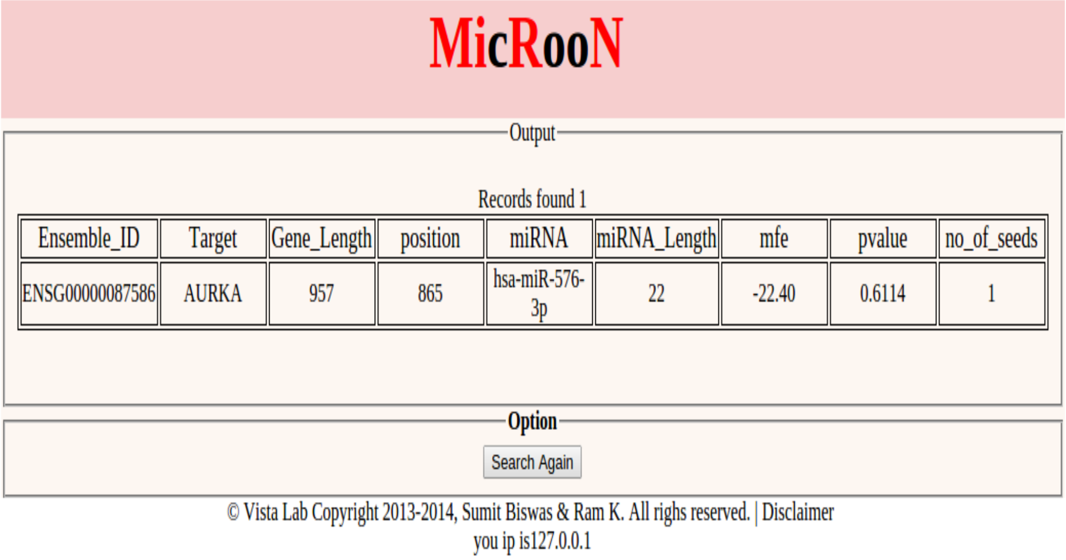
Output

Records found 124

Ensemble_ID	Target	Gene_Length	position	miRNA	miRNA_Length	mfe	pvalue	no_of_seeds
ENSG00000100721	TCL1A	450	249	hsa-miR-576-3p	22	-22.40	0.6097	3
ENSG00000167601	AXL	845	535	hsa-miR-576-3p	22	-23.00	0.6106	2
ENSG00000087586	AURKA	957	865	hsa-miR-576-3p	22	-22.40	0.6114	1
ENSG00000118971	CCND2	4322	2426	hsa-miR-576-3p	22	-25.70	0.6115	2
ENSG00000170345	FOS	535	188	hsa-miR-576-3p	22	-20.60	0.6119	2
ENSG00000196730	DAPK1	307	2	hsa-miR-576-3p	22	-18.90	0.6124	1
ENSG00000103479	RBL2	388	152	hsa-miR-576-3p	22	-19.30	0.6125	2
ENSG00000130522	JUND	656	348	hsa-miR-576-3p	22	-20.30	0.6126	2
ENSG00000151702	FLI1	1261	1076	hsa-miR-576-3p	22	-21.60	0.6127	2
ENSG00000143878	RHOB	424	161	hsa-miR-576-3p	22	-19.20	0.6128	2
LRG_211	TRIM32	569	18	hsa-miR-576-3p	22	-19.70	0.6129	2

Figure 5.7: MicRooNdb - miRNA ID based search. An user query with hsa-mir-576-3p results in 124 hits.

In certain cases,if the user provides a invalid query *i.e.,* if the miRNA ID or target is not found in the MicRooNdb, then the result will be displayed as *miRNA record not found* (Figure 5.9)

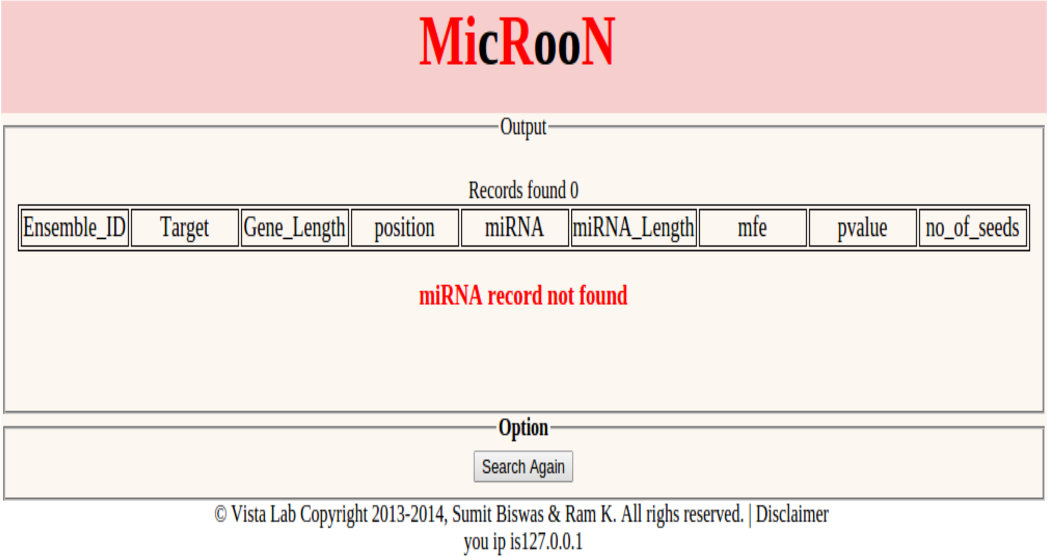


Records found 1

Ensemble_ID	Target	Gene_Length	position	miRNA	miRNA_Length	mfe	pvalue	no_of_seeds
ENSG00000087586	AURKA	957	865	hsa-miR-576-3p	22	-22.40	0.6114	1

© Vista Lab Copyright 2013-2014, Sumit Biswas & Ram K. All righs reserved. | Disclaimer
you ip is127.0.0.1

Figure 5.8: MicRooNdb - target based search. User query with hsa-miR-576-3p with target AURKA.



Records found 0

Ensemble_ID	Target	Gene_Length	position	miRNA	miRNA_Length	mfe	pvalue	no_of_seeds

miRNA record not found

© Vista Lab Copyright 2013-2014, Sumit Biswas & Ram K. All righs reserved. | Disclaimer
you ip is127.0.0.1

Figure 5.9: Query results obtained when miRNA ID or target name do not match any record.

The query constructed to retrieve a particular record from MicRooNdb is shown below.

```
<?php
$a = $_REQUEST["miRNA"];
$b = $_REQUEST["target"];
$mysqli = new mysqli("localhost", "root", "password", "mysql");
$a = $mysqli -> escape_string($a);
$b = $mysqli -> escape_string($b);
$a = $mysqli->query("SELECT * FROM GENE INNER JOIN MIRNA
on GENE.Ensemble_ID = MIRNA.Ensemble_ID
WHERE  MIRNA.miRNA = '$a' anTarget LIKE '%$b'
ORDER BY MIRNA.pvalue
for($a=0;$a<sizeof($result);$a++){
print '<tr>
<td>'.htmlentities($result[$a]["Ensemble_ID"]).'</td>
<td>'.htmlentities($result[$a]["Target"]).'</td>
<td>'.htmlentities($result[$a]["Gene_Length"]).'</td>
<td>'.htmlentities($result[$a]["position"]).'</td>
<td>'.htmlentities($result[$a]["miRNA"]).'</td>
<td>'.htmlentities($result[$a]["miRNA_Length"]).'</td>
<td>'.htmlentities($result[$a]["mfe"]).'</td>
<td>'.htmlentities($result[$a]["pvalue"]).'</td>
<td>'.htmlentities($result[$a]["no_of_seeds"]).'</td>
</tr>';
?>
```

The complete code for constructing the MicRooN– web interface with MySQL query is given in Appendix D.

5.2 Chapter Summary

- MicRooNdb is currently constructed with miRNA from miRBase 20.0, experimentally validated miRNA-mRNA interaction from miRecords (April 27, 2013) and miRTarbase (Release 4.5, Nov 1, 2013) and cancer genes from COSMIC (v70).

- Ensemble model generated in chapter 4 was employed for prediction of miRNAs associated with cancer. Predictions are recorded with their miRNA-mRNA interaction data in MicRooNdb.

- MicRooN - a web-based user interface was constructed with minimalistic design and utilised potentially for querying MicRooNdb.

- MicRooN allows user to query both on miRID and miRNA targets and the results are sorted based on the p-value (a measure of confident binding in miRNA-mRNA hybrids).

Chapter 6

Conclusions

Chapter 6

Conclusions

Identifying the involvement of miRNA in cancer is a major obstacle for researchers striving to understand the basis of the disease and to generate new therapies against particular cancer types. miRNAs regulate the molecular pathways in cancer by either upregulating or downregulating various oncogenes and tumour suppressors, and sometimes acting as oncogenes themselves. The functional annotation of miRNAs in cancer is still a painstaking process, though cancer therapies using miRNA has been picking up lately. So in an attempt to aid cancer biologists, we employed a machine learning based binary classifier to predict miRNAs associated with cancer. During this thesis work, several observations were documented and it is being summarized as chapter-wise.

In **chapter 1 (Introduction)**, a detailed description about miRNA biogenesis, regulation, miRNA and cancer, miRNA as biomarkers and experimental strategies employed for identifying miRNA involved in cancer pathway has been described. The chapter also focuses on the existing problems faced by cancer biologists due to massive growth of miRNA data in the recent decades. In terms of various approaches to achieve the objectives, we have discussed about the machine learning algorithms employed in biology, their pitfall, class imbalance problem, data preprocessing and finally the present situation in miRNA studies with machine learning approaches was also described.

In **chapter 2 (Search for signatures in miRNAs associated with cancer)**, the study was aimed at search for signatures in miRNA associated with cancer. We utilised experimentally validated miRNAs as positive dataset and randomly generated as negative dataset. Within the two datasets, a search for sequence and hybridization-based signature was carried out. It was observed that in miRNAs associated with cancer, uracil is the most preferred base in the seed region whereas cytosine was least preferred, a result which is in complete agreement with the experimental result obtained from site depletion analysis [23]. In terms of hybridization, the average number of seed regions formed in miRNAs associated with cancer is six whereas in case of miRNAs not associated with cancer, it does not extend beyond four. This is a clear indication of poor complementarity of binding and lack of site efficacy. Additionally, we observed AU base pairing was more predominant around the seed region in miRNAs associated with cancer [109, 110]. Thus, in chapter 2, signatures discriminating miRNA associated with cancer and those that are not associated with cancer was obtained. However, utilising randomly generated miRNAs as negative dataset may mimic several miRNAs in the positive dataset. When trained with distinguishable signatures/features obtained from this dataset, it was observed that the trained model had a good performance, but when challenged with an independent test dataset, the model performed very poorly due to overfitting problem. Hence, utilising randomly generated dataset for identifying miRNAs associated with cancer was strictly avoidable and extensive search for experimentally validated miRNAs not involved in cancer was undertaken.

In **chapter 3 (Identifying miRNAs involved in cancer pathway using Support Vector Machines)**, construction of a two-step SVM based binary classifier, utilising 60 features extracted from miRNA involved in cancer and those not involved in cancer was carried out. Radial Basis Function (RBF) was used as a kernel-function to map instances from low dimensional to higher dimensional space (since non-linearity exists). Features were extracted from mature miRNA sequences, free-energy of miRNA-mRNA binding and their interaction profiles. Out of the 60 features extracted, 26 features contain nucleotide position information and a 2-window size ($2W$) sequence repeat information for training miRSEQ. The remaining 34 extracted features were utilised for training miRINT which contains free-energy of miRNA-mRNA binding and their interaction profiles.

Features were initially ranked based on information gain and utilised for training. During the training process, it was observed that the performance was skewed towards the positive instances due to class imbalance in the dataset. To overcome imbalance in the dataset, we employed Synthetic Minority Over-Sampling Technique (SMOTE). The model constructed in this chapter is based on the experimentally annotated interactions of a miRNA when bound to a particular mRNA only. Since either the oncogene or TSG may switch invariantly between each other depending on the cell stimuli, the tool considers a training set with experimentally validated data only. The two step classifier model – miRSEQ and miRINT had reasonably good performance measures with fairly high values of Mathew's Correlation Coefficient (MCC), ranging from 0.72 to 0.82.

The major pitfall with the constructed model is that only features with higher order (*i.e.,* higher discrimination) was utilised during the training process. Several low ranking features were completely eliminated – which may have boosted the performance of the model constructed. Additionally, to overcome class imbalance in the dataset, we utilised SMOTE – an oversampling technique. Utilising oversampling for disease related prediction may result in speculative predictions. Expectedly, the constructed classifier's performance with an independent dataset, left room for further improvement.

In **chapter 4 (MicRooN – an ensemble classifier for identifying miRNAs associated with cancer)** precautionary steps were taken to involve all the informative features identified. We employed Recursive Feature Elimination (RFE) in two-stages to select optimal subset of features for training process. It was observed in miRSEQ training that features corresponding to nucleotide position {P2, P3, P5, P6, P7, P9, P11, P13, P15, P16, P18, P20, P21} and base repeat {AA, GG, CC} are the features with higher discrimination. In miRINT, the features varied based on the number of seeds the miRNA formed with the mRNA. To overcome class imbalance in the dataset and to avoid oversampling of instances, we employed cost-sensitive approaches for both miRSEQ and miRINT. The performance of each learning algorithm was evaluated with precision, recall, AUC and F-measure to adapt the cost-sensitive learning approaches employed. We compared three learning algorithms *viz.,* Support

Vector Machines, Random Forest (RF) and C4.5 to subsequently construct an ensemble-based system for predicting miRNA associated with cancer. It was observed in terms of prediction efficacy that the RF algorithm outperformed kernel-based SVM and decision tree-based C4.5. We observed that SVM classified about 25.6% of the total instances as False Negative, whereas C4.5 misclassified more than one third of the entire instances. Thus, we concluded that RF is the best of the three due to its inherent ensemble prediction.

Additionally, we concluded that, for miRNA datasets with high class imbalance, cost-sensitive based approaches perform better than the oversampling methods. While considering the performance of individual learning algorithms, although RF performed with higher precision (miRSEQ 0.802 and miRINT$_{(average\, of\, seeds)}$ 0.738) and lower FN, the AUC of all three learning algorithms were found to marginally equal except for seed-2 based models. This, emphasizes the fact that a single learning algorithm is inefficient in generalising a model for predicting miRNAs associated with cancer. Hence, an ensemble for miRSEQ with precision 0.802, AUC 0.780 and for miRINT (Precision 0.703, AUC 0.648), (Precision 0.927, AUC 0.930) and (Precision 0.584, AUC 0.531) for seeds 1, 2 and 3 hybrid model respectively.

In **chapter 5 (MicRooN – a web server for identifying miRNAs associated with cancer)**, construction of database (MicRooNdb) for documenting predictions from ensemble models and a web based user interface (MicRooN) was carried out. MicRooN allows user to query based on miRID and mRNA target. Along with miRNAs associated with cancer, it also provides a detailed description about the mRNA target, minimum free energy of miRNA-mRNA binding, position of binding in $3'$UTR and the type of association they posses with the mRNA. MicRooN will be updated with the number of novel miRNAs identified and documented in miRBase. MicRooN is particularly helpful for cancer biologists for screening miRNAs associated with cancer rather than employing time-consuming and economically unfavorable experimental procedures.

Future scope of the work

- Search for tissue specific oncogenic signatures in human.
- Protocols for identifying experimentally validated negative instances *i.e.,* miRNAs not involved in cancer for improvising predictive performance.
- Search for oncogenic signatures in virus infecting humans.

References

[1] Eisenmann, D. M. Wnt signaling. *WormBook* **25**, 1–17 (2005).

[2] Reinhart, B. J. *et al.* The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403**, 901–906 (2000).

[3] Olsen, P. H. & Ambros, V. The *lin-4* regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Developmental Biology* **216**, 671–680 (1999).

[4] Hausser, J., Syed, A. P., Bilen, B. & Zavolan, M. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Research* **23**, 604–615 (2013).

[5] Zhou, X., Duan, X., Qian, J. & Li, F. Abundant conserved microRNA target sites in the 5′untranslated region and coding sequence. *Genetica* **137**, 159–164 (2009).

[6] Didiano, D. & Hobert, O. Molecular architecture of a miRNA-regulated 3′UTR. *RNA* **14**, 1297–1317 (2008).

[7] An, J., Lai, J., Lehman, M. L. & Nelson, C. C. miRDeep*: An integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Research* **41**, 727–737 (2013).

[8] Chiang, H. R. *et al.* Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes & Development* **24**, 992–1009 (2010).

[9] Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B. & Bartel, D. P. Vertebrate microRNA genes. *Science* **299**, 1540–1540 (2003).

[10] Lim, L. P. *et al.* The microRNAs of *Caenorhabditis elegans*. *Genes & Development* **17**, 991–1008 (2003).

[11] Rajewsky, N. microRNA target predictions in animals. *Nature Genetics* **38**, S8–S13 (2006).

[12] Winter, J., Jung, S., Keller, S., Gregory, R. I. & Diederichs, S. Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nature Cell Biology* **11**, 228–234 (2009).

[13] Zeng, Y. & Cullen, B. R. Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. *Journal of Biological Chemistry* **280**, 27595–27603 (2005).

[14] Kim, V. N., Han, J. & Siomi, M. C. Biogenesis of small RNAs in animals. *Nature Reviews Molecular Cell Biology* **10**, 126–139 (2009).

[15] Schwarz, D. S. *et al.* Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115**, 199–208 (2003).

[16] Okamura, K. *et al.* The regulatory activity of microRNA* species has substantial influence on microRNA and 3′UTR evolution. *Nature Structural & Molecular Biology* **15**, 354–363 (2008).

[17] Khvorova, A., Reynolds, A. & Jayasena, S. D. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**, 209–216 (2003).

[18] Hwang, H.-W., Wentzel, E. A. & Mendell, J. T. A hexanucleotide element directs microRNA nuclear import. *Science* **315**, 97–100 (2007).

[19] Ghildiyal, M., Xu, J., Seitz, H., Weng, Z. & Zamore, P. D. Sorting of *Drosophila* small silencing RNAs partitions microRNA* strands into the RNA interference pathway. *RNA* **16**, 43–56 (2010).

[20] Cuperus, J. T., Fahlgren, N. & Carrington, J. C. Evolution and functional diversification of miRNA genes. *The Plant Cell Online* **23**, 431–442 (2011).

[21] Koscianska, E. *et al.* Prediction and preliminary validation of oncogene regulation by miRNAs. *BMC Molecular Biology* **8**, 79 (2007).

[22] Sætrom, P. *et al.* Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Research* **35**, 2333–2342 (2007).

[23] Grimson, A. *et al.* MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell* **27**, 91–105 (2007).

[24] Axtell, M. J., Westholm, J. O. & Lai, E. C. *Vive la différence*: biogenesis and evolution of microRNAs in plants and animals. *Genome Biology* **12**, 221 (2011).

[25] Martinez, N. J. & Walhout, A. J. The interplay between transcription factors and microRNAs in genome-scale regulatory networks. *Bioessays* **31**, 435–445 (2009).

[26] Hobert, O. Gene regulation by transcription factors and microRNAs. *Science* **319**, 1785–1786 (2008).

[27] Volinia, S. *et al.* A microRNA expression signature of human solid tumors defines cancer gene targets. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 2257–2261 (2006).

[28] Calin, G. A. *et al.* Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proceedings of the National Academy of Sciences* **101**, 2999–3004 (2004).

[29] Wuchty, S., Arjona, D., Bozdag, S. & Bauer, P. O. Involvement of microRNA families in cancer. *Nucleic Acids Research* **40**, 8219–8226 (2012).

[30] Gibcus, J. H. *et al.* Hodgkin lymphoma cell lines are characterized by a specific miRNA expression profile. *Neoplasia* **11**, 167–169 (2009).

[31] Krutovskikh, V. A. & Herceg, Z. Oncogenic microRNAs (OncomiRs) as a new class of cancer biomarkers. *Bioessays* **32**, 894–904 (2010).

[32] Rosenfeld, N. *et al.* Micrornas accurately identify cancer tissue origin. *Nature Biotechnology* **26**, 462–469 (2008).

[33] Gormley, M., Dampier, W., Ertel, A., Karacali, B. & Tozeren, A. Prediction potential of candidate biomarker sets identified and validated on gene expression data from multiple datasets. *BMC Bioinformatics* **8**, 415 (2007).

[34] Oien, K. A. & Evans, T. J. Raising the profile of cancer of unknown primary. *Journal of Clinical Oncology* **26**, 4373–4375 (2008).

[35] Zhao, Y., Samal, E. & Srivastava, D. Serum response factor regulates a muscle-specific microRNA that targets during cardiogenesis. *Nature* **436**, 214–220 (2005).

[36] Schmitz, U. & Vearasilp, K. MiRBase. In Dubitzky, W., Wolkenhauer, O., Cho, K.-H. & Yokota, H. (eds.) *Encyclopedia of Systems Biology*, 1363–1366 (Springer New York, 2013).

[37] Griffiths-Jones, S., Grocock, R. J., Van Dongen, S., Bateman, A. & Enright, A. J. MiRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research* **34**, D140–D144 (2006).

[38] Hofacker, I. L. Vienna RNA secondary structure server. *Nucleic Acids Research* **31**, 3429–3431 (2003).

[39] Garcia, D. M. *et al.* Weak seed-pairing stability and high target-site abundance decrease the proficiency of *lsy-6* and other microRNAs. *Nature Structural & Molecular Biology* **18**, 1139–1146 (2011).

[40] Forman, J. J., Legesse-Miller, A. & Coller, H. A. A search for conserved sequences in coding regions reveals that the *let-7* microRNA targets Dicer within its coding sequence. *Proceedings of the National Academy of Sciences* **105**, 14879–14884 (2008).

[41] Yang, Y., Wang, Y.-P. & Li, K.-B. MiRTif: a support vector machine-based microRNA target interaction filter. *BMC Bioinformatics* **9**, S4 (2008).

[42] Marín, R. M. & Vaníček, J. Optimal use of conservation and accessibility filters in microRNA target prediction. *PloS one* **7**, e32208 (2012).

[43] Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).

[44] Mendoza, M. R. *et al.* RFMirTarget: predicting human microRNA target genes with a Random Forest classifier. *PloS one* **8**, e70153 (2013).

[45] Hsu, S.-D. *et al.* MiRTarBase: a database curates experimentally validated microRNA–target interactions. *Nucleic Acids Research* gkq1107 (2010).

[46] Xiao, F. *et al.* MiRecords: an integrated resource for microRNA–target interactions. *Nucleic Acids Research* **37**, D105–D110 (2009).

[47] Shahi, P. *et al.* Argonaute – a database for gene regulation by mammalian microRNAs. *Nucleic Acids Research* **34**, D115–D118 (2006).

[48] Dweep, H., Sticht, C., Pandey, P. & Gretz, N. miRWalk–database: prediction of possible mirna binding sites by "walking"the genes of three genomes. *Journal of Biomedical Informatics* **44**, 839–847 (2011).

[49] Maragkakis, M. *et al.* Diana-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Research* gkp292 (2009).

[50] Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. miRBase: tools for microRNA genomics. *Nucleic Acids Research* **36**, D154–D158 (2008).

[51] Wang, X. miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA* **14**, 1012–1017 (2008).

[52] Creighton, C. J., Nagaraja, A. K., Hanash, S. M., Matzuk, M. M. & Gunaratne, P. H. A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions. *RNA* **14**, 2290–2296 (2008).

[53] Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. The role of site accessibility in microRNA target recognition. *Nature Genetics* **39**, 1278–1284 (2007).

[54] Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).

[55] Vasudevan, S. Functional validation of microRNA-target RNA interactions. *Methods* **58**, 126–134 (2012).

[56] Stempor, P. A., Cauchi, M. & Wilson, P. MMpred: functional miRNA–mRNA interaction analyses by miRNA expression prediction. *BMC Genomics* **13**, 620 (2012).

[57] Chien, C.-H. *et al.* Identifying transcriptional start sites of human micrornas based on high-throughput sequencing data. *Nucleic Acids Research* **39**, 9345–9356 (2011).

[58] Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987).

[59] Xu, R., Anagnostopoulos, G. C. & Wunsch, D. Multi-class cancer classification by semi-supervised ellipsoid ARTMAP with gene expression data. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, vol. 1, 188–191 (IEEE, 2004).

[60] Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* **20**, 273–297 (1995).

[61] Hepworth, P. J., Nefedov, A. V., Muchnik, I. B. & Morgan, K. L. Broiler chickens can benefit from machine learning: support vector machine analysis of observational epidemiological data. *Journal of The Royal Society Interface* **9**, 1934–1942 (2012).

[62] James, G. *Majority vote classifiers: theory and applications*. Ph.D. thesis, Stanford University (1998).

[63] Quinlan, J. R. *C4.5: programs for machine learning*, vol. 1 (Morgan kaufmann, 1993).

[64] Mazid, M. M., Ali, S. & Tickle, K. S. Improved C4.5 algorithm for rule based classification. In *Proceedings of the 9th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases*, 296–301 (World Scientific and Engineering Academy and Society (WSEAS), 2010).

[65] Breiman, L. Bagging predictors. *Machine learning* **24**, 123–140 (1996).

[66] Freund, Y., Schapire, R. & Abe, N. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence* **14**, 1612 (1999).

[67] Chawla, N. V. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, 853–867 (Springer, 2005).

[68] Woods, K. S. *et al.* Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence* **7**, 1417–1436 (1993).

[69] Chawla, N. V., Cieslak, D. A., Hall, L. O. & Joshi, A. Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery* **17**, 225–252 (2008).

[70] Japkowicz, N. & Shah, M. *Evaluating Learning Algorithms* (Cambridge University Press, 2011).

[71] Sun, J., Shang, Z. & Li, H. Imbalance-oriented SVM methods for financial distress prediction. *Journal of the Operational Research Society* **65**, 1905–1919 (2014).

[72] Liu, X.-Y., Wu, J. & Zhou, Z.-H. Exploratory undersampling for class-imbalance learning. *Trans. Sys. Man Cyber. Part B* **39**, 539–550 (2009).

[73] Brefeld, U. & Scheffer, T. AUC maximizing support vector learning. In *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning* (Citeseer, 2005).

[74] Keivanfard, F., Teshnehlab, M., Aliyari Shoorehdeli, M., Nie, K. & Su, M.-Y. Feature selection and classification of breast cancer on dynamic Magnetic Resonance Imaging by using Artificial Neural Networks. In *Biomedical Engineering (ICBME), 2010 17th Iranian Conference of*, 1–4 (IEEE, 2010).

[75] Dash, M. & Liu, H. Feature selection for classification. *Intelligent Data Analysis* **1**, 131–156 (1997).

[76] Kohavi, R. *et al.* A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, vol. 14, 1137–1145 (1995).

[77] Hermes, L. & Buhmann, J. M. Feature selection for support vector machines. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 2, 712–715 (IEEE, 2000).

[78] Zeng, X., Chen, Y.-W., Tao, C. & van Alphen, D. Feature selection using recursive feature elimination for handwritten digit recognition. In *Intelligent Information Hiding and Multimedia Signal Processing, 2009. IIH-MSP'09. Fifth International Conference on*, 1205–1208 (IEEE, 2009).

[79] Kotsiantis, S., Kanellopoulos, D. & Pintelas, P. Data preprocessing for supervised learning. *International Journal of Computer Science* **1**, 111–117 (2006).

[80] Zhang, X., Pan, F., Wang, W. & Nobel, A. Mining non-redundant high order correlations in binary data. *Proceedings of the VLDB Endowment* **1**, 1178–1188 (2008).

[81] Qu, Y. *et al.* Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. *Biometrics* **59**, 143–151 (2003).

[82] Yamamoto, H. *et al.* Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables. *Chemometrics and Intelligent Laboratory Systems* **98**, 136–142 (2009).

[83] Rawsthorne, J., Roshier, D. A. & Murphy, S. R. A simple parametric method for reducing sample sizes in gut passage time trials. *Ecology* **90**, 2328–2331 (2009).

[84] Higgins, M. E., Claremont, M., Major, J. E., Sander, C. & Lash, A. E. CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Research* **35**, D721–D726 (2007).

[85] Chen, J.-S., Hung, W.-S., Chan, H.-H., Tsai, S.-J. & Sun, H. S. *In-silico* identification of oncogenic potential of *fyn*-related kinase in hepatocellular carcinoma. *Bioinformatics (Oxford, England)* **29**, 420–427 (2013).

[86] Kozomara, A. & Griffiths-jones, S. MiRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research* **39**, D152–D157 (2011).

[87] Kinsella, R. J. *et al.* Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database : the journal of biological databases and curation* (2011).

[88] Krüger, J. & Rehmsmeier, M. RNAHybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Research* **34**, W451–W454 (2006).

[89] Aragues, R., Sander, C. & Oliva, B. Predicting cancer involvement of genes from heterogeneous data. *BMC Bioinformatics* **9**, 172 (2008).

[90] Rehmsmeier, M., Steffen, P., Höchsmann, M. & Giegerich, R. Fast and effective prediction of microRNA/target duplexes. *RNA* **10**, 1507–1517 (2004).

[91] Doench, J. G. & Sharp, P. A. Specificity of microRNA target selection in translational repression. *Genes & Development* **18**, 504–511 (2004).

[92] Enright, A. J. *et al.* MicroRNA targets in *Drosophila*. *Genome Biology* **5**, R1–R1 (2004).

[93] Brennecke, J., Stark, A., Russell, R. B. & Cohen, S. M. Principles of microRNA–target recognition. *PLoS Biology* **3**, e85 (2005).

[94] MacFarlane, L.-A. & Murphy, P. R. MicroRNA: biogenesis, function and role in cancer. *Current Genomics* **11**, 537 (2010).

[95] Wuchty, S., Fontana, W., Hofacker, I. L., Schuster, P. *et al.* Complete suboptimal folding of rna and the stability of secondary structures. *Biopolymers* **49**, 145–165 (1999).

[96] Lekprasert, P., Mayhew, M. & Ohler, U. Assessing the utility of thermodynamic features for microRNA target prediction under relaxed seed and no conservation requirements. *PloS one* **6**, e20622 (2011).

[97] Bandyopadhyay, S. & Mitra, R. TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics* **25**, 2625–2631 (2009).

[98] Hebert, C., Norris, K., Scheper, M. a., Nikitakis, N. & Sauk, J. J. High mobility group A2 is a target for miRNA-98 in head and neck squamous cell carcinoma. *Molecular Cancer* **6**, 5 (2007).

[99] Kiriakidou, M. *et al.* A combined computational-experimental approach predicts human microRNA targets. *Genes & Development* **18**, 1165–1178 (2004).

[100] Lewis, B. P., Shih, I.-h., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115**, 787–798 (2003).

[101] Musiyenko, A., Bitko, V. & Barik, S. Ectopic expression of *miR-126**, an intronic product of the vascular endothelial EGF-like 7 gene, regulates prostein translation and invasiveness of prostate cancer LNCaP cells. *Journal of Molecular Medicine (Berlin, Germany)* **86**, 313–322 (2008).

[102] Robins, H., Li, Y. & Padgett, R. W. Incorporating structure to predict microRNA targets. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 4006–4009 (2005).

[103] Schultz, J., Lorenz, P., Gross, G., Ibrahim, S. & Kunz, M. MicroRNA *let-7b* targets important cell cycle molecules in malignant melanoma cells and interferes with anchorage-independent growth. *Cell Research* **18**, 549–557 (2008).

[104] Sethupathy, P. *et al.* Human microRNA-155 on chromosome 21 differentially interacts with its polymorphic target in the AGTR1 3′ untranslated region: a mechanism for functional single-nucleotide polymorphisms related to phenotypes. *American Journal of Human Genetics* **81**, 405–413 (2007).

[105] Skalsky, R. L. *et al.* Kaposi′s sarcoma-associated herpesvirus encodes an ortholog of *miR*-155. *Journal of Virology* **81**, 12836–12845 (2007).

[106] Visvanathan, J., Lee, S., Lee, B., Lee, J. W. & Lee, S. K. The microRNA *miR-124* antagonizes the anti-neural REST/SCP1 pathway during embryonic CNS development. *Genes and Development* **21**, 744–749 (2007).

[107] Bowyer, K. W., Chawla, N. V., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *CoRR* **abs/1106.1813** (2011).

[108] Batuwita, R. & Palade, V. MicroPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics* **25**, 989–995 (2009).

[109] Kothandan, R. & Biswas, S. Search for signatures in miRNAs associated with cancer. *Bioinformation* **9**, 524–527 (2013).

[110] Sharma, S. & Biswas, S. Sequence trademarks in oncogene associated microRNAs. *Bioinformation* **6**, 364–365 (2011).

[111] Chen, Y.-W. & Lin, C.-J. Combining SVMs with various feature selection strategies. In *Feature extraction*, 315–324 (Springer, 2006).

[112] Chang, C.-C. & Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2**, 27 (2011).

[113] Batuwita, R. & Palade, V. Efficient resampling methods for training Support Vector Machines with imbalanced datasets. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, 1–8 (IEEE, 2010).

[114] Novakovic, J. Using information gain attribute evaluation to classify sonar targets. In *17th Telecommunications forum TELFOR* (2009).

[115] Holmes, G., Donkin, A. & Witten, I. H. WEKA: A machine learning workbench. In *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, 357–361 (IEEE, 1994).

[116] Bergstra, J. & Bengio, Y. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* **13**, 281–305 (2012).

[117] Aizerman, M. A., Braverman, E. A. & Rozonoer, L. Theoretical foundations of the potential function method in pattern recognition learning. In *Automation and Remote Control,*, no. 25 in Automation and Remote Control,, 821–837 (1964).

[118] Boser, B. E., Guyon, I. M. & Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, 144–152 (1992).

[119] Sun, Y., Kamel, M. S., Wong, A. K. & Wang, Y. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition* **40**, 3358–3378 (2007).

[120] ElGokhy, S. M., ElHefnawi, M. & Shoukry, A. Ensemble-based classification approach for micro-RNA mining applied on diverse metagenomic sequences. *BMC Research Notes* **7**, 286 (2014).

[121] Veropoulos, K., Campbell, C., Cristianini, N. *et al.* Controlling the sensitivity of support vector machines. In *Proceedings of the International Joint Conference on AI*, 55–60 (1999).

[122] Kothandan, R. & Biswas, S. Identifying microRNAs involved in cancer pathway using Support Vector Machines. *Computational Biology and Chemistry* **55**, 31–36 (2015).

[123] Date, C. J., Date, C. J. & Date, C. J. *An introduction to database systems*, vol. 7 (Addison-wesley Reading, Mass., 1986).

Appendix A

miRNA-mRNA Interaction Data

Source: https://sites.google.com/site/sumitslab/tools/miR-mRNAInteractionData.zip

miRNA-mRNA interaction dataset are segregated as oncogene interaction data and tumour suppressor gene interaction data

Appendix B

PairFinder - code for calculating basepairing in seed and outseed regions in hybrid structures

Coded by Ram Kothandan & Malvika Sudahar

```
print "Enter file name";

#### Takes filename to be read and parsed
chomp ($file=<STDIN>);
open (FD, $file);
@contents=<FD>;
close FD;

#####change value for line no to 1 if you want it to be
#####printed to the file
$lno1=1;
$lno2=1;
$lno3=1;
$lno4=1;

##### Create output file
$filename=$file;
$filename=~s/txt$/xls/;
open (HD,">$filename");
print HD "\t\t\t\t\t\t\t
```

```
Seed 1\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t
Seed 2\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t
Seed 3\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t
Seed 4\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t
Seed 5\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t
Seed 6\n";
print HD "Target\tGene length\tmiRNA\tmiRNA
length\tmfe\tp-value\tposition\tAA\tAU\tAG\tAC
\tUA\tUU\tUG\tUC\tGA\tGU\tGG\tGC\tCA\tCU\tCG\tCC\t
Gap in mRNA\tGap in miRNA\tAA\tAU\tAG\tAC\tUA\tUU
\tUG\tUC\tGA\tGU\tGG\tGC\tCA\tCU\tCG\tCC\t
Gap in mRNA\tGap in miRNA\tAA\tAU\tAG\tAC\tUA\tUU
\tUG\tUC\tGA\tGU\tGG\tGC\tCA\tCU\tCG\tCC\t
Gap in mRNA\t
Gap in miRNA\tAA\tAU\tAG\tAC\tUA\tUU\tUG\tUC\tGA
\tGU\tGG\tGC\tCA\tCU\tCG\tCC\t
Gap in mRNA\tGap in miRNA\tAA\tAU\tAG\tAC\tUA\tUU
\tUG\tUC\tGA\tGU\tGG\tGC\tCA\tCU\tCG\tCC\t
Gap in mRNA\tGap in miRNA\tAA\tAU\tAG\tAC\tUA\tUU
\tUG\tUC\tGA\tGU\tGG\tGC\tCA\tCU\tCG\tCC\t
Gap in mRNA\tGap in miRNA\n";

#####initialise count
$AA=0;
$AU=0;
$AG=0;
$AC=0;
$UA=0;
$UU=0;
$UG=0;
$UC=0;
$GA=0;
$GU=0;
$GG=0;
$GC=0;
```

```
$CA=0;
$CU=0;
$CG=0;
$CC=0;
$GapmRNA=0;
$GapmiRNA=0;
@range=();

#### Reads each line of the file and parses it
foreach $line(@contents)
{
if ($line=~/\$target\:\s(.*)/)
     {
     $targetname=$1;
     $t1=1;
     chomp($targetname);
     }
elsif ($line=~/length\:\s(.*)/)
     {
     if ($t1==1)
          {
          $genelength=$1;
          chomp($genelength);
          $t1=0;
          }
     elsif ($m1==1)
          {
          $mirnalength=$1;
          chomp($mirnalength);
          $m1=0;
          }
     }
elsif ($line=~/miRNA\s\:\s(.*)/)
     {
     $m1=1;
     $miRNA=$1;
     chomp($miRNA);
     }
elsif ($line=~/mfe\:\s(.*)/)
     {
     $mfe=$1;
  $mfe = substr($mfe,0,6);
     chomp($mfe);
     }
elsif ($line=~/p\-value\:\s(.*)/)
     {
     $pvalue=$1;
     chomp($pvalue);
     }
elsif ($line=~/^position\:\s(.*)/)
{
$position=$1;
chomp($position)
}
elsif ($line=~/^target\s5’\s/)
{
#print "target";
$target=1;
$targetnonseed=$line;
$targetnonseed=~s/^target\s5’\s//;
$targetnonseed=~s/\s3’//;
#          if ($lno1==1)
#               {print HD $targetnonseed;}
}
elsif ($line=~/^miRNA\s\s3’\s/)
{
#print "mirna";
$mirnanonseed=$line;
$mirnanonseed=~s/^miRNA\s\s3’\s//;
$mirnanonseed=~s/\s5’//;
```

```
#if ($lno4==1)
#{print HD $mirnanonseed;}

####### Creating arrays
@targets=split('',$targetseed);
@targetns=split('',$targetnonseed);
@mirnas=split('',$mirnaseed);
@mirnans=split('',$mirnanonseed);
###### calculating length
     $totallen=length($targetseed);
###### creating line to be printed
     $printline="$targetname\t$genelength\t$miRNA\t
     $mirnalength\t$mfe\t$pvalue\t$position\t";
#######Count the bases
$i=0;
while($targetseed=~/\s*([AUGC]{4,})\s*/g)
{
$len=length($1);
$start=(index($targetseed,$1));
$end=$start + $len -1;
$range[$i][0]=$start;
$range[$i][1]=$end;
for ($j=$start;$j<=$end;$j++)
{
if ($targets[$j] eq 'A' && $mirnas[$j] eq 'U')
{$AU++;}
if ($targets[$j] eq 'U' && $mirnas[$j] eq 'A')
{$UA++;}
if ($targets[$j] eq 'U' && $mirnas[$j] eq 'G')
{$UG++;}
if ($targets[$j] eq 'G' && $mirnas[$j] eq 'U')
{$GU++;}
if ($targets[$j] eq 'G' && $mirnas[$j] eq 'C')
{$GC++;}
if ($targets[$j] eq 'C' && $mirnas[$j] eq 'G')
{$CG++;}
}
$printline=$printline."$AA\t$AU\t$AG\t$AC\t$UA\t$UU
\t$UG\t$UC\t$GA\t$GU\t$GG\t
$GC\t$CA\t$CU\t$CG\t$CC\t$
GapmRNA\t$GapmiRNA\t";
$AA=0;
$AU=0;
$AG=0;
$AC=0;
$UA=0;
$UU=0;
$UG=0;
$UC=0;
$GA=0;
$GU=0;
$GG=0;
$GC=0;
$CA=0;
$CU=0;
$CG=0;
$CC=0;
$GapmRNA=0;
$GapmiRNA=0;
$i++;
}
print HD "$printline\n";
}
elsif ($line=~/\s{10,}/)
{
if ($target==1)
{
$targetseed=substr($line,10);
$targetseed=~s/\s{3}$//;
$target=0;
```

```
#if ($lno2==1)
#{print HD $targetseed;}
}
else
{
$mirnaseed=substr($line,10);
$mirnaseed=~s/\s{3}$//;
#if ($lno3==1)
#{print HD $mirnaseed;}
}
}
}
close HD;
```

Appendix C

Novel miRs predicted to be associated with cancer by MicRooN

Source: https://sites.google.com/site/sumitslab/tools/NovelmiRNAPredictions.zip

Appendix D

MicRooN - Codes for Web Interface Design

```
<!CODE - MICROON - WEB INTERFACE>
<! INPUT EITHER MIRNA_ID OR TARGET NAME>
<! HOSTED VIA AN APACHE SERVER>
<style type="te<td>xt/css">
.style29 tr tbody tr td p strong {
font-family: "Courier New", Courier, monospace;
}
</style>

<TABLE class=style29 cellSpacing=0 cellPadding=0 width=600
bgColor=#fdf7f2 border=0 align = "center">
<TR>
<form name="MicRooN">
<tr> <td align ="center" bgcolor = #f6cece>
<h1><big><big><strong>
<B style = "COLOR:#FF0000">Mi</B>c<B style = "COLOR:#FF0000">R</B>oo
<B style = "COLOR:#FF0000">N</B>&nbsp;</strong></big></big></h1>
</td>
<! this is for the about the tool >
<tr >
<td >
<!--
<fieldset><legend><B>About the Tool</B></legend>
MicRooN (an acronym of miRNA), a tool for identification
of miRs associated with cancer.
```

```
Its a ensemble based classifier[LIBSVM,C4.5 & Random Forest],
built on experiementally validated set of miRs associated with cancer.
The tool aims in identifying novel miR associated in cancer pathway.
&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;
&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;
&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;
&nbsp;&nbsp;&nbsp;
-->

<br></fieldset>
<! this is for the Input>
<tr>
<td><fieldset><legend><B>Input</B></legend>
<p align = "center">
<form method = "post" action="/">
<center>Input miR ID :<input name = "miRNA" id ="textfeild"
type = "text" placeholder="hsa-mir-532-3p" /></p>
Input Target :<input name = "target" id ="textfeild"
type = "text" placeholder="RCN2" /></p></center>
<p align="center"> <input id="Submit" value="Submit" type="submit" />
</form>
<button onclick="window.location.assign('http:\\')" />Clear</button>
<br><br>
<p style="text-align: center;">Format for miR ID :
hsa-mir-6793&nbsp;<br />
(As per <A href  = "http://www.mirbase.org">miRBASE </A>format)</p>
 </fieldset>  </td>
</tr>
<! this is for the output>




<?php
if (isset($_REQUEST["miRNA"])){
    $dummy = true;
}
else{
    goto foot;
}
?>



<tr>
<td align="center" ><fieldset><legend>Output</legend><br>
<style>
#output tr td{
    border:1px solid black;
}
.bold{
    font-size:18px;
    text-decoration:bold;
}
</style>
<table id="output" style="border:1px solid black;">
 <tr class="bold">
 <td>Ensemble_ID</td>
 <td>Target</td>
 <td>Gene_Length</td>
 <td>position</td>
 <td>miRNA</td>
 <td>miRNA_Length</td>
 <td>mfe</td>
 <td>pvalue</td>
 <td>no_of_seeds</td>
</tr>
<?php
$a = $_REQUEST["miRNA"];
$b = $_REQUEST["target"];
// $result = db::table("‘table‘") -> pluck("*")
-> where("miRNA",$a) -> select() -> get();
```

```
$mysqli = new mysqli("localhost", "root", "password", "mysql");
$a = $mysqli -> escape_string($a);
$b = $mysqli -> escape_string($b);
$a = $mysqli->query("SELECT * FROM GENE INNER JOIN MIRNA
on GENE.Ensemble_ID = MIRNA.Ensemble_ID
WHERE  MIRNA.miRNA = '$a' aTarget LIKE '%$b'
ORDER BY MIRNA.pvalue ASC;");
$i = 0;
while ($row = $a -> fetch_assoc()) {
$result[$i] = $row;
$i++;


        }
        $mysqli->close();
        for($a=0;$a<sizeof($result);$a++){
print '<tr>
<td>'.htmlentities($result[$a]["Ensemble_ID"]).'</td>
<td>'.htmlentities($result[$a]["Target"]).'</td>
<td>'.htmlentities($result[$a]["Gene_Length"]).'</td>
<td>'.htmlentities($result[$a]["position"]).'</td>
<td>'.htmlentities($result[$a]["miRNA"]).'</td>
<td>'.htmlentities($result[$a]["miRNA_Length"]).'</td>
<td>'.htmlentities($result[$a]["mfe"]).'</td>
<td>'.htmlentities($result[$a]["pvalue"]).'</td>
<td>'.htmlentities($result[$a]["no_of_seeds"]).'</td>
    </tr>';
        }
    ?>
</table>
<style type="text/css">
    table{
        text-align: center;
    }
    td{
        min-width: 100px;
    }
</style>
<?php
if(sizeof($result) == 0){
    print "<h3 style='color:red;'>miRNA record not found</h3>";
}
?>
<br><br></fieldset> </td>
</tr>

<?php

foot:
?>

<! this is for the reference>
<tr>
<td>
<fieldset><legend><B>Reference</B></legend>
<ul>
<li> "Search for signatures in miRNAs
associated with cancer", Kothandan R,
Biswas S, Bioinformation, Vol.9(10)
<A href = "http://www.ncbi.nlm.nih.gov/pubmed/?term=23861569">
<B> [PMID:23861569] </B></A> </li>
<li> "Sequence Trademarks in oncogene associated microRNAs",
Sharma S, Biswas S, Bioinformation,
Vol.6(9)
<A href = "http://www.ncbi.nlm.nih.gov/pubmed/?term=21814397">
<B> [PMID:21814397] </B></A> </li>
</ul>
<br></fieldset>
</td>
</tr>
</table>
```

```
<footer align = "center">
 &#169 Vista Lab Copyright 2013-2014, Sumit Biswas
 & Ram K. All righs reserved.
| Disclaimer </footer>
<?php
$x = file_get_contents("./view/counter");
file_put_contents("./view/counter", $x+1);
print "<center>page view = $x<br>";
$ip = $_SERVER["REMOTE_ADDR"];
print "you ip is".$_SERVER["REMOTE_ADDR"]."</center>";
file_get_contents("./view/ip");
$str = "Access from ".$ip."\n";
$log = file_get_contents("./view/ip");
$log = $log.$str;
file_put_contents("./view/ip", $log);
?>
```

Appendix E

List of Publications and Conference Presentations

Publications

- **Ram K** and Sumit Biswas, Identifying microRNAs involved in cancer pathway using Support Vector Machine, *Journal of Computational Biology and Chemistry* **55**, 31-36 (2015).

- **Ram K** and Sumit Biswas, Search for signatures in miRNAs associated with cancer, *Bioinformation* **9**(10), 524-527 (2013).

- **Ram K**, Handling class imbalance problem in miRNA dataset associated with cancer, *Bioinformation* **11**(1), (2015).

- **Ram K** and Sumit Biswas, Comparison of kernel and decision tree-based algorithms for the prediction of microRNAs associated with cancer. [Under final review with *Current Bioinformatics*; BSP-CBIO-2014-420].

- Sumit Biswas and **Ram K**, MicRooN and Microondb: A software suite and a database for prediction and validation of miRs involved in cancer. *International work-conference on Bioinformatics and Biomedical Engineering* (2015).

Poster Presentations

- **Ram K** and Sumit Biswas, Search for signatures in miRNA associated with cancer, *BIOFEST-2013* (*won the Best Poster award*)
- Sumit Biswas, **Ram K** and Sumit Sharma, MicRooN – A tool for identification and validation of oncomiRs. *Conference on Informatics & Integrative Biology* (CIIB-2011), **63** (2011)

Appendix F

Biography of the Candidate

Ram K is currently serving as a CSIR- Senior Research Scholar in department of Biological Sciences at BITS-PILANI K K BIRLA Goa Campus, Goa, India. He received B.Tech degree in Industrial Biotechnology from Anna University, Chennai in 2007 and M.Tech degree in Biotechnology from Anna University, Coimbatore in 2009. Soon after his M.Tech., he joined Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu as an Assistant Professor and worked till May, 2011. Later he joined as a full time research scholar at BITS-PILANI K K BIRLA Goa Campus in August 2011. He is recipient of CSIR-SRF Fellowship from April 2012 to April 2015. He has published three research paper in international journal and communicated one research paper in reputed international journals. He has presented one poster in international conference for which he won the best poster award (BIOFEST-2012) and co-authoured one poster presentation in national conference in CIIB-2011.

Appendix G

Biography of the Supervisor

Dr. Sumit Biswas completed his Ph.D in Bose Institute, Kolkata, under the supervision of Prof. Pinak Chakrabarti, as a CSIR fellow in 2008. His doctoral work elucidated the interfaces of protein-nucleic acid interactions, as well as the structure determination of two very important proteins. He went on to work as a DBT Research Associate in the DBT initiative, “Setting up of National Facility on Interactive Graphysics Computer System for Biomolecular Modelling, Molecular Dynamics & Structures” till 2009. Dr. Biswas joined BITS-Pilani, K K Birla Goa Campus as an Assistant Professor in 2009. He has since been involved as the Principal Investigator of four research projects funded by BRNS, DAE, DBT and DST, as well as the co-investigator of a UGC project. His work involves the molecular mechanism and biology of the *Vibrio* life cycle, bioinformatics of non-coding RNA and protein-nucleic acid interactions, and therapeutic biology of natural products. Dr. Sumit Biswas has 13 publications in reputed journals and several conference publications to his name. He is also working on a book on Biophysics sanctioned by Prentice Hall of India.

Dr. Biswas has acted as the convenor for symposia and workshops funded by DST, DSTE-Goa and BRNS. He is a life member of the Indian Crystallography Association, and a member of CholdInet (a WHO initiative for cholera research) and the Proteomics Society of India. He has received several awards and honours, the most recent being the prestigious EMBL Scholarship for presenting paper at EMBL conference on Cancer Genomics, held at Heidelberg. Besides, he has delivered invited talks at different international conferences as well as institute of repute like IIT, Kharagpur. He has been actively involved as a reviewer of international journal from OUP, Elsevier, *etc.,* as well as a question setter of DBT. Presently, he has three registered Ph.D students under his tutelage and numerous thesis dissertation and project students working with him. Recently, he has been certified as the approved Radiological Safety Officer for the Institute.

Appendix H

Reprints of Published Articles

Computational Biology and Chemistry 55 (2015) 31–36

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Identifying microRNAs involved in cancer pathway using support vector machines

Ram Kothandan, Sumit Biswas [*]

VISTA Lab, Department of Biological Sciences, BITS, Pilani - K K Birla Goa Campus, Zuarinagar, Goa 403726, India

ARTICLE INFO

Article history:
Received 25 July 2014
Received in revised form 26 January 2015
Accepted 28 January 2015
Available online 29 January 2015

Keywords:
Signatures
miRNA:mRNA interaction
Machine based learning
Feature selection

ABSTRACT

Since Ambros' discovery of small non-protein coding RNAs in the early 1990s, the past two decades have seen an upsurge in the number of reports of predicted microRNAs (miR), which have been implicated in various functions. The correlation of miRs with cancer has spurred the usage of this class of non-coding RNAs in various cancer therapies, although most of them are at trial stages. However, the experimental identification of a miR to be associated with cancer is still an elaborate, time-consuming process. To aid this process of miR association, we undertook an *in-silico* study involving the identification of global signatures in experimentally validated microRNAs associated with cancer. Subsequently, a support vector machine based two-step binary classifier system has been trained and modeled from the features extracted from the above study. A total of 60 distinguishing features were selected and ranked to form the feature set for classification – 26 of these extracted from the miR sequence itself, and the remainder from the thermodynamics of folding and the hybridized miRNA–mRNA structure. The two step classifier model – miRSEQ and miRINT had reasonably good performance measures with fairly high values of Matthew's correlation coefficient (MCC) values ranging from 0.72 to 0.82 (availability: https://sites.google.com/site/sumitslab/tools).

1. Introduction

miRNA (miR) are small non-coding, single stranded RNAs (about 22 nucleotides in length) involved in several regulatory pathways in the cell cycle. They bind to the untranslated regions (UTRs) of mRNA, (particularly the 3'UTR) and play an important role in the post-transcriptional regulation of gene expression (Bartel, 2004; Filipowicz et al., 2008). Recent studies suggest that these noncoding RNAs can bind to 5'UTRs (Ragan et al., 2009) and coding regions (Hausser et al., 2013) of mRNA as well, but little is known about the mechanism of binding and their regulation. Binding of a miR to a specific target in an UTR with complete complementarity either leads to degradation of the mRNA itself or induce translational repression (Esquela-Kerscher and Slack, 2006). In tissues associated with various tumors, it has been observed that the expression pattern of miRs is altered considerably (Cummins et al., 2006; Zhang et al., 2006). Additionally, gene mapping reveals that most of the human miRs are located in chromosomal positions which are susceptible to rearrangements (Calin and Croce, 2007). Hence, it can be asserted that miRs in humans play a major role in the cancer pathway.

Previous studies by several authors have investigated the involvement of different types of base pairing in miR–mRNA interactions and target prediction algorithms have been formulated based on these precincts. These algorithms predominantly considered Watson Crick base pairing between the miR and its respective mRNA – especially with the 2nd to the 8th nucleotide positions of miR – as the potential target sites. However, in later studies, it was found that animal miRs do not bind to mRNA with perfect complementarity (unlike in plants); rather their binding leaves several imperfections like loops, mismatches or bulges and often involves GU(non-Watson Crick) base pairing as well (Axtell et al., 2011; Didiano and Hobert, 2008). Other than these determinants, AU richness around the seed regions and folding of mRNA play a vital role in target binding (Grimson et al., 2007; Robins et al., 2005). All these factors need to be considered, not in isolation but together to hypothesize miR:mRNA interactions.

Some of the computational methods used in the functional annotation of miRs involved in cancer mainly rely on the expression profile of various cancer cell types and statistical analysis for further classification (Jayaswal et al., 2011). These methods utilize the expression profile but they fail to consider the fact that a single miR can bind to several mRNA target sites and regulate the cell differently. Our aim at feature selection was, therefore, to embrace all these redundancy checks. Other attempts to classify miRs into oncogenes and tumor suppressor genes (TSGs)

[*] Corresponding author. Tel.: +91 832 2580178.
E-mail address: sumit@goa.bits-pilani.ac.in (S. Biswas).

http://dx.doi.org/10.1016/j.compbiolchem.2015.01.007
1476-9271/© 2015 Elsevier Ltd. All rights reserved.

were based on functional and evolutionary features (Wang et al., 2010) like conservation, expression levels, chromosome distribution, etc.

The present study involved a search and analysis of features involved in the interaction of a miR:mRNA associated with cancer. These features encompassed sequential, hybridization and thermodynamics of validated miR:mRNA interactions only. Based on the curated and prioritized features, we developed a two-step machine based classifier model – miRSEQ and miRINT, which will identify a miR to be associated with cancer and also classify the type of its association, i.e., either with an oncogene or a tumor suppressor. Prioritization of the features and a diversification of the models according to the number of seed regions drastically improved the performance of the classifier, as compared to generalized features and holistic hybridization. The incorporation of seed based classification in the determination of features is a novel approach in our algorithm. The final classifier thus developed had good performance with experimentally validated datasets giving good prediction accuracy (cross validation (cv-rate) ranging from 92% to 87%).

2. Methods

2.1. Dataset preparation

For the purpose of generating a classifier, the first step needed to be undertaken is the construction of a microRNA dataset which has been experimentally validated to be associated with cancer. To begin with, a list of genes involved in cancer was downloaded from the catalog of somatic mutations (COSMIC) (Higgins et al., 2007). A total of 488 genes were thus listed, which could be further segregated into oncogenes and tumor suppressors by cross-referring with the tumor associated gene database (TAG) (Chen et al., 2013). Experimentally validated miRNA interactions with target mRNA can be obtained from miRECORDS (Xiao et al., 2009) and miRTARBASE (Hsu et al., 2011). Therefore, the list of genes obtained from COSMIC was curated with miRECORDS and miRTARBASE to obtain a list of experimentally validated targets. This process finally yielded a set of targets for miRNA which have been experimentally validated to be associated with cancer. A total of 2578 miRNAs were extracted from miRBASE 20.0 (May 2013, (Griffiths-jones et al., 2006)), and these were compared with the experimentally validated miR–mRNA interactions obtained as above, yielding a final set of 239 microRNAs which have been conclusively implicated in the cancer pathway (Supplementary data S1). These 239 miRNAs were manually checked with their available literature and revalidated. 3′UTR mRNA sequences involved in the interaction of these 239 miRNAs with their targets were obtained from BIOMART – Ensemble (Kinsella et al., 2011) (Fig. 1).

Positive and negative datasets for training and testing the classifier were built separately for miRSEQ and miRINT. For miRSEQ, experimentally validated mature miR sequences (the same 239 obtained as above) would serve as the positive dataset. The negative dataset was built in accordance with the method employed in (Bandyopadhyay and Mitra, 2009) (Supplementary data S3). The negative dataset was constructed on the basis of specific experimental evidence presented in literature for miRNAs whose binding to a target mRNA does not involve gene regulation (Hebert et al., 2007; Kiriakidou et al., 2004; Lewis et al., 2003; Musiyenko et al., 2008; Robins et al., 2005; Schultz et al., 2008; Sethupathy et al., 2007; Skalsky et al., 2007; Visvanathan et al., 2007; Zhao et al., 2005). Selection of random samples for negative dataset was strictly avoided since they may increase the false positives thereby decreasing the performance of the classifier. For miRINT, experimentally validated miRNA:mRNA interactions, further segregated as oncogene interactions (129 instances) and tumor suppressor interactions (110 instances) were considered as the positive and negative datasets, respectively (Supplementary data S7). Class imbalance in the datasets was overcome by applying the synthetic minority over sampling technique (SMOTE) (Chawla et al., 2002).

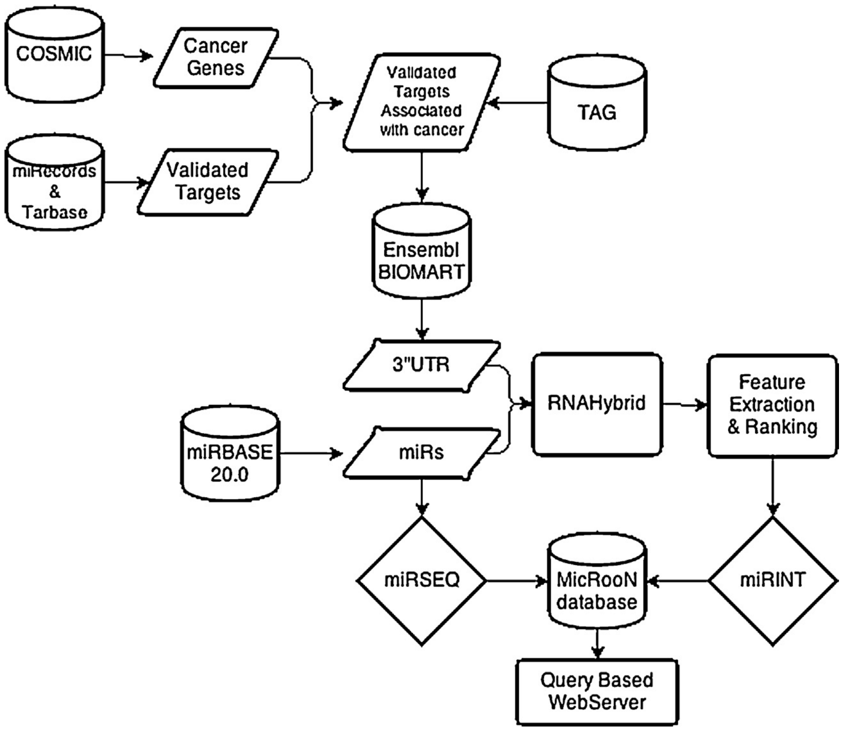


Fig. 1. Flowchart for the process representing the generation of dataset for miRSEQ and miRINT.

2.2. Feature selection

Construction of an efficient classifier depends on meticulous feature extraction, since the quality of the feature reflects upon the effective performance of the classifier. For our classifier, features were identified and extracted based on a survey of previous studies and our own indigenous parameters (Batuwita and Palade, 2009; Kothandan and Biswas, 2013; Mendoza et al., 2013; Sharma and Biswas, 2011). For miRSEQ, 26 features were considered, which included nucleotide positions and repeat information. The maximum length of the miR used in the training was restricted to 22 nucleotides (Supplementary data S4 and S5).

For miRINT, a total of 34 features based on the hybridization profile (miR:mRNA interactions) were utilized. miR:mRNA hybrids having the best fit in terms of free energy, were obtained using RNAHybrid – ViennaRNA package (Krüger and Rehmsmeier, 2006). A total of 2926 hybrid structures were generated and considered for feature extraction. miR:mRNA hybridization using RNAHybrid may contain false target site predictions. Hence, a post-processing filter was applied to miR–mRNA interactions in order to remove the false predictions. An indigenous Perl script, "PairFinder" (https://sites.google.com/site/sumitslab/tools) was used to parse and analyze the hybrids for seeds, regions outside seeds, mismatches and bulges (Kothandan and Biswas, 2013). Seed regions have been defined according to the convention followed in (Lekprasert et al., 2011) and our previous work (Kothandan and Biswas, 2013). A detailed list of all the 60 features has been summarized in the supplementary files (Supplementary data S5 and S6).

For the construction of test datasets, a non-validated miR is allowed to hybridize in RNAHybrid with the list of genes obtained from COSMIC and the most energetically favored structure was considered. False interactions were removed using the same post processing filter and then let into the classification process.

2.3. Training – miRSEQ and miRINT

In this study, we used LibSVM package for constructing classifier models (Chang and Lin, 2011). Radial basis function (RBF) was chosen as the kernel for the classification process. Parameters for RBF (cost and gamma) were found using a grid search, which involved the construction of a mesh grid allowing a search for best cost (c) and gamma ($g = 1$/number of features). The main disadvantage of training a disease related dataset is the inadequate number of training instances that are experimentally validated and it is important that the same training set should never be used as a test set in any of the experiment because they may lead to over fitting in the model generated. So in order to overcome these hassles we used 10-fold cross-validation step (by default) to evaluate the performance of the classification.

Features extracted were ranked based on F-score (Supplementary Tables T1 and T2) and eventually prioritized. The F-score method has been described in detail in Supplementary section S9. Two sets of features – for miRSEQ and miRINT – were finalized as has been described before. Additionally, for miRINT, models were built based on the number of seeds they form in the hybrid (Seed 1, Seed 2 and Seed 3 model). This is because the parameters which play crucial roles for miRNA binding to mRNA differ when the interaction involves the formation of a single seed compared to the interaction where more than one seeds are formed. We considered a maximum of three-seed hybrid for the training. Feature ranking was done individually for each of the models and individually trained. This was done to prevent the dilution or extrapolation of some features when all the differently-seeded hybrids were taken together. To find the optimum subset for the classifier, we followed recursive feature elimination (RFE) for both miRSEQ and miRINT during the training process. Low ranking features were removed one by one iteratively and the performance of the classifier measured until saturation. Removing all the low ranking features at a glance may degrade the performance of the classifier completely; hence the process of optimum feature subset selection was carried out iteratively (Zeng et al., 2009). As a result of the difference in binding parameters, the features that dominate in the optimal feature set for each seed model differ as well (Supplementary Table 2).

Due to the difference in numbers between the positive and the negative sets, class imbalance existed in the dataset; so, accuracy could not be chosen as a direct measure of performance (Batuwita and Palade, 2010) for such sets. Hence, performance measures were chosen in compliance with the cross-validation rate (cv-rate) and Matthew's correlation coefficient (MCC). MCC ranges from −1 to 1; a MCC value of 1 indicates the best prediction and a negative value indicates imperfect classification.

3. Results

Dataset preparation was carried out individually for the classifiers miRSEQ and miRINT (Fig. 1). Consequently, a total of 263 miRs were used in the miRSEQ training. Class imbalance problem in the dataset was overcome by the SMOTE (k-nearest algorithm with no replacement) method which generated sufficient number of negative instances for the training set. Like most SVM classification problems related to miRNAs, our dataset was also not linearly separable as it was too complex in nature. RBF was applied to convert all non-linear data from lower dimensional space to linearly separable higher dimensional space.

For miRSEQ, nucleotide position conservation was used initially as the main feature set. However, poor performance of the classifier (cv-rate of 45%) prompted us to use nucleotide repeat information with appropriate window size to boost the performance. Selection of appropriate window size (W) for nucleotide repeat information was done by measuring the performance of the classifier keeping a sliding window size ranging from 2 to 5. Performance was measured from the plot between cv-rate and window size, depicting a clear drop in cv-rate when the window size exceeded 2 (Fig. 2). Hence, a 2-window sized repeat was considered for training miRSEQ.

The 26 features chosen were ranked by F-score method and recursive feature elimination was performed to find the best subset of features for the dataset as well as retain all the features with very low classification error, respectively. Optimum subset of features which were finally selected has been depicted in Fig. 3.

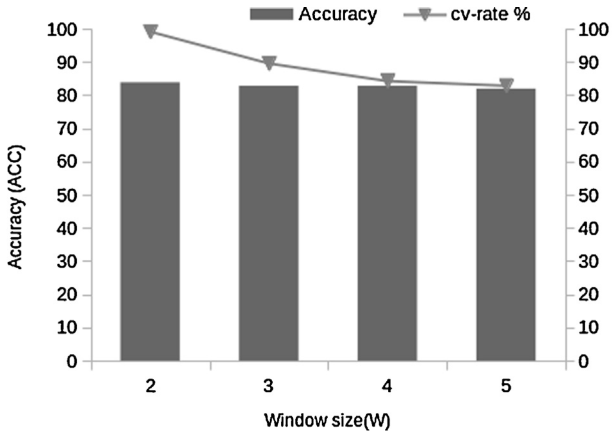


Fig. 2. Decrease in performance and cv-rate with increase in window size (W) for the classifier miRSEQ. Accuracy (ACC) has been depicted as bars, while the cv-rate is the curve. The value for the same has been included in the graph.

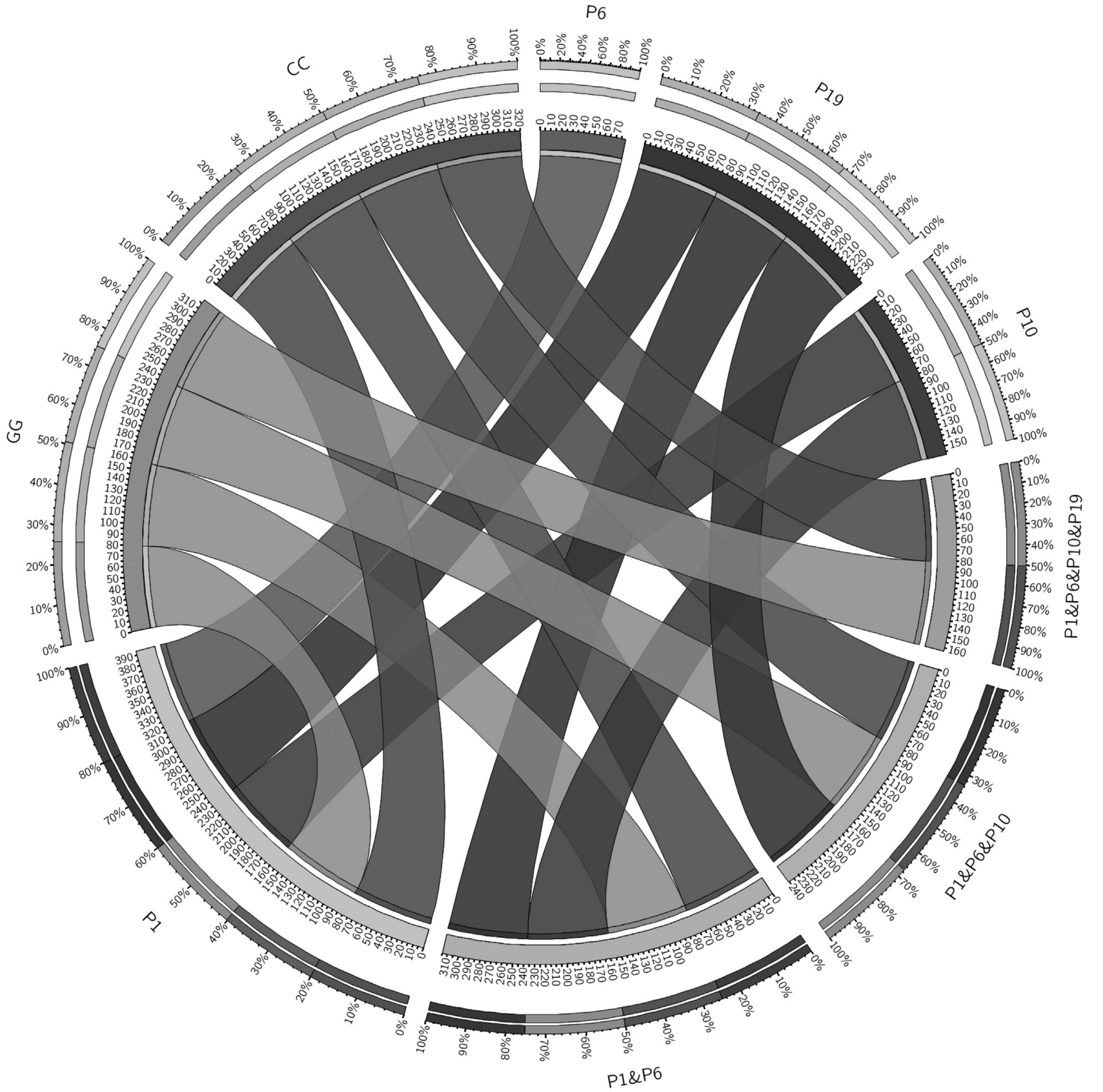


Fig. 3. Overlap between features subsets (ranked by F-score) selected for miRSEQ. Outer ticks denote the maximum accuracy of the classifier (in a scale of 100%). The inner ticks denote the accuracy of individual features in the subset in combination with other features. The width of the ribbon denotes the individual accuracy in those combinations. Example: For Position 1, the inner ticks total 390, which when divided by the number of overlapping features gives an accuracy of 78% (390/5).

Judging by the thickness of the bands in the Circos diagram, the following features yielded the best subset for the classification – Position 1, GG repeat, CC repeat, Position 6, Position 19 and Position 10, in sequence of their relative importance. These features were prioritized to construct the optimal feature subset for miRSEQ and performance measures were carried out which yielded a cv-rate of 91.15% and MCC of 0.803. Model generation and performance estimation were carried out with the training set (only validated miR sequences) with a 10-fold cross validation method (Table 1). The model generated was used on an unseen test set for a primary prediction of the association of those miRs with cancer.

Table 1
miRSEQ – performance measurement using 10-fold cross validation.

	Run 1	Run 2	Run 3	Run 4	Run 5	Average
TP	160	160	160	160	160	
TN	77	77	77	77	77	
FP	15	11	12	11	14	
FN	11	11	11	11	12	
MCC	0.7809	0.8106	0.803	0.81	0.7805	0.803
ACC	90.11	91.505	91.153	91.505	90.11	91.153

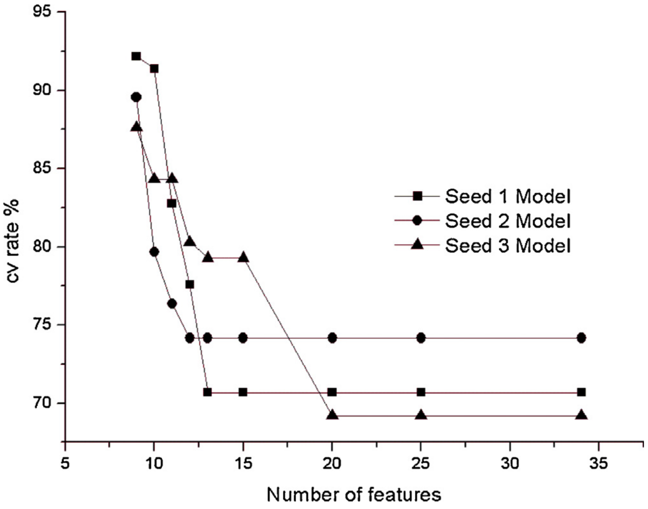


Fig. 4. Feature selection and effect on cross validation rate for miRINT for various seed types.

For the second classifier miRINT, the choice of features were initially centered around the results of the hybridization – number of unpaired bases, Watson–Crick and non-Watson–Crick base pairing in and around the seed region, to name a few. The generated model had a very poor performance with low cv-rate (<40%). Addition of normalized base pairing and normalized free energy features raised the total number of features to 34 for miRINT and showed a marked improvement in the performance of the classifier, but not to expected levels.

It was therefore, decided to have different models for hybridization structures with different numbers of seed formation. For each of the different classes, the method of ranking by F-score and prioritization (as with miRSEQ) was carried out to achieve three different optimal feature subsets. As with miRSEQ, the highest ranked feature was not exclusive, but considered in conjunction with other features as well during the construction of the optimal feature set. Precaution was taken to utilize only the non-redundant informative features for model construction. This improved the performance of all the three models of the classifier with good cv-rate of 92.19% for single seed (MCC 0.821), 89.54% for two seed (MCC 0.765) and 87.61% for three seed (MCC 0.722) hybrids. The effect of number of features versus the accuracy measurement is given in the graph for all three models (Fig. 4). Feature selection not only improved the classification but also optimized the total time taken for training the model. The resulting classifier model not only predicts the association of a miRNA with cancer, but also gives an output about that association with either a tumor suppressor gene or an oncogene. Performance measurement carried out on the independent test dataset for miRINT is shown in Table 2.

Table 2
miRINT – performance measurement using 10-fold cross validation.

		Run 1	Run 2	Run 3	Run 4	Run 5	Average
Seed 1	MCC	0.747	0.7977	0.8258	0.8258	0.8211	0.8211
	ACC	88.05	90.76	92.187	92.187	92.187	92.187
Seed 2	MCC	0.765	0.777	0.8244	0.752	0.752	0.765
	ACC	89.54	90.1315	92.255	88.961	88.961	89.54
Seed 3	MCC	0.722	0.712	0.7301	0.782	0.722	0.722
	ACC	87.61	87.224	88	90.41	87.61	87.61

4. Discussion

Identifying miR involvement in cancer is a major obstacle for researchers striving to understand the basis of the disease and to generate new therapies against particular cancer types. miRNAs regulate the molecular pathways in cancer by either upregulating or downregulating various oncogenes and tumor suppressors, and sometimes acting as oncogenes themselves. The functional annotation of miRNAs in cancer is still a painstaking process, though cancer therapies using miRNA has been picking up lately. So, in an attempt to aid the cancer biologist, we employed a support vector machine based binary classifier system to predict a miR associated with cancer.

The tool described in this study is based on the experimentally annotated interactions of a miRNA when bound to a particular mRNA only. Since either the oncogene or TSG may switch invariantly between each other depending on the cell stimuli, the tool considers a training set with experimentally validated data only. Cross verification performed on test datasets with our classification model proved to be consistent with experimentally validated data.

During the initial training process, although a number of features have been extracted and used, performance improved only after systematic ranking and prioritization were introduced. Of these features, some again could be used to discriminate binding against oncogenes and TSG while the rest, in combination with the above features boosted the discrimination. Initial classification process was quite complex mainly due to the unique behavior of miR:mRNA interactions. So in order to suppress the complexity, we considered features both from within and outside the seed regions. Features extracted outside the seed region along with several site specific features provided quite a good classification performance. With the available training datasets, the tool performed satisfactorily and prediction performance should improve as the number of experimentally validated data increases. Further work involving a multiple algorithm based model (apart from SVM), in order to utilize all the informative features extracted from the validated dataset is being undertaken to check for better performance efficiency.

Acknowledgements

The authors wish to thank Dr. Ranjit Prasad Bahadur, Indian Institute of Technology – Kharagpur, India for his initial assistance in machine learning approaches. Ram K. was supported by a scholarship from Council of Scientific Research and Industrial Research, Govt. of India.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compbiolchem.2015.01.007.

References

Axtell, M.J., Westholm, J.O., Lai, E.C., 2011. Vive la différence biogenesis and evolution of microRNAs in plants and animals. Genome Biol. 12, 221. doi:http://dx.doi.org/10.1186/gb-2011-12-4-221.

Bandyopadhyay, S., Mitra, R., 2009. TargetMiner:microRNA target prediction with systematic identification of tissue-specific negative examples. Bioinformatics 25, 2625–2631. doi:http://dx.doi.org/10.1093/bioinformatics/btp503.

Bartel, D.P., 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116, 281–297. doi:http://dx.doi.org/10.1016/S0092-8674(04)00045.

Batuwita, R., Palade, V., 2009. microPred: effective classification of pre-miRNAs for human miRNA gene prediction. Bioinformatics 25, 989–995. doi:http://dx.doi.org/10.1093/bioinformatics/btp107.

Batuwita, R., Palade, V., 2010. Efficient resampling methods for training support vector machines in imbalanced datasets. Imbalanced Learning: Foundations

Algorithms Applications, Proceedings of the International Joint Conference on Neural Networks, Barcelona, pp. 1–20. doi:http://dx.doi.org/10.1002/9781118646106.index.

Calin, G.A., Croce, C.M., 2007. Chromosomal rearrangements and microRNAs: a new cancer link with clinical implications. J. Clin. Invest. 117. doi:http://dx.doi.org/10.1172/JCI32577.(18).

Chang, C.C., Lin, C.-J., 2011. LIBSVM. ACM Trans. Intell. Syst. Technol. 1–27. doi:http://dx.doi.org/10.1145/1961189.1961199.

Chawla, N.V., Bowyer, K.W., Hall, L.O., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357. doi:http://dx.doi.org/10.1613/jair.953.

Chen, J.S., Hung, W.-S., Chan, H.-H., Tsai, S.-J., Sun, H.S., 2013. In silico identification of oncogenic potential of fyn-related kinase in hepatocellular carcinoma. Bioinformatics 29, 420–427. doi:http://dx.doi.org/10.1093/bioinformatics/bts715.

Cummins, J.M., He, Y., Leary, R.J., Pagliarini, R., Diaz, L.A., Sjoblom, T., Barad, O., Bentwich, Z., Szafranska, A.E., Labourier, E., Raymond, C.K., Roberts, B.S., Juhl, H., Kinzler, K.W., Vogelstein, B., Velculescu, V.E., 2006. The colorectal microRNAome. PNAS 103, 3687–3692. doi:http://dx.doi.org/10.1073/pnas.0511155103.

Didiano, D., Hobert, O., 2008. Molecular architecture of a miRNA-regulated 3′ UTR Molecular architecture of a miRNA-regulated 3' UTR 1297–1317. http://dx.doi.org/10.1261/rna.1082708.

Esquela-Kerscher, A., Slack, F.J., 2006. Oncomirs – microRNAs with a role in cancer. Nat. Rev. Cancer 6, 259–269. doi:http://dx.doi.org/10.1038/nrc1840.

Filipowicz, W., Bhattacharyya, S.N., Sonenberg, N., 2008. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat. Rev. Genet. 9, 102–114. doi:http://dx.doi.org/10.1038/nrg2290.

Griffiths-jones, S., Grocock, R.J., Dongen, S., Van, A., Bateman, Enright, A.J., van Dongen, S., 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 34, D140–D144. doi:http://dx.doi.org/10.1093/nar/gkj112.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., Bartel, D.P., 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol. Cell 27, 91–105. doi:http://dx.doi.org/10.1016/j.molcel.2007.06.017.

Hausser, J., Syed, A.P., Bilen, B., Zavolan, M., 2013. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. Genome Res. 23, 604–615. doi:http://dx.doi.org/10.1101/gr.139758.112.

Hebert, C., Norris, K., Scheper, M.A., Nikitakis, N., Sauk, J.J., 2007. High mobility group A2 is a target for miRNA-98 in head and neck squamous cell carcinoma. Mol. Cancer 6, 5. doi:http://dx.doi.org/10.1186/1476-4598-6-5.

Higgins, M.E., Claremont, M., Major, J.E., Sander, C., Lash, A.E., 2007. CancerGenes: a gene selection resource for cancer genome projects. Nucleic Acids Res. 35, D721–D726. doi:http://dx.doi.org/10.1093/nar/gkl811.

Hsu, S.-D., Lin, F.-M., Wu, W.-Y., Liang, C., Huang, W.-C., Chan, W.-L., Tsai, W.-T., Chen, G.-Z., Lee, C.-J., Chiu, C.-M., Chien, C.-H., Wu, M.-C., Huang, C.-Y., Tsou, A.-P., Huang, H.-D., 2011. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res. 39, D163–D169. doi:http://dx.doi.org/10.1093/nar/gkq1107.

Jayaswal, V., Lutherborrow, M., Ma, D.D.F., Yang, Y.H., 2011. Identification of microRNA–mRNA modules using microarray data. BMC Genomics 12, 138. doi:http://dx.doi.org/10.1186/1471-2164-12-138.

Kinsella, R.J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., Flicek, P., Ka, A., 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford). http://dx.doi.org/10.1093/database/bar030.

Kiriakidou, M., Nelson, P.T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., Hatzigeorgiou, A., 2004. A combined computational–experimental approach predicts human microRNA targets. Genes Dev. 18, 1165–1178. doi:http://dx.doi.org/10.1101/gad.1184704.

Kothandan, R., Biswas, S., 2013. Search for signatures in miRNAs associated with cancer. Bioinformation 9, 524–527. doi:http://dx.doi.org/10.6026/97320630009524.

Krüger, J., Rehmsmeier, M., 2006. RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res. 34, W451–W454. doi:http://dx.doi.org/10.1093/nar/gkl243.

Lekprasert, P., Mayhew, M., Ohler, U., 2011. Assessing the utility of thermodynamic features for microRNA target prediction under relaxed seed and no conservation requirements. PLoS One 6, e20622. doi:http://dx.doi.org/10.1371/journal.pone.0020622.

Lewis, B.P., Shih, I., Jones-Rhoades, M.W., Bartel, D.P., Burge, C.B., 2003. Prediction of mammalian microRNA targets. Cell 115, 787–798.

Mendoza, M.R., Fonseca, G.C., Loss-morais, G., Alves, R., Margis, R., Bazzan, A.L.C., 2013. RFMirTarget: predicting human microrna target genes with a random forest classifier. PLoS One 8. doi:http://dx.doi.org/10.1371/journal.pone.0070153.

Musiyenko, A., Bitko, V., Barik, S., 2008. Ectopic expression of miR-126*, an intronic product of the vascular endothelial EGF-like 7 gene, regulates prostein translation and invasiveness of prostate cancer LNCaP cells. J. Mol. Med. (Berl.) 86, 313–322. doi:http://dx.doi.org/10.1007/s00109-007-0296-9.

Ragan, C., Cloonan, N., Grimmond, S.M., Zuker, M., a Ragan, M., 2009. Transcriptome-wide prediction of miRNA targets in human and mouse using FASTH. PLoS One 745 doi:http://dx.doi.org/10.1371/journal.pone.0005745.

Robins, H., Li, Y., Padgett, R.W., 2005. Incorporating structure to predict microRNA targets. Proc. Natl. Acad. Sci. U. S. A. 102, 4006–4009. doi:http://dx.doi.org/10.1073/pnas.0500775102.

Schultz, J., Lorenz, P., Gross, G., Ibrahim, S., Kunz, M., 2008. MicroRNA let-7b targets important cell cycle molecules in malignant melanoma cells and interferes with anchorage-independent growth. Cell Res. 18, 549–557. doi:http://dx.doi.org/10.1038/cr.2008.45.

Sethupathy, P., Borel, C., Gagnebin, M., Grant, G.R., Deutsch, S., Elton, T.S., Hatzigeorgiou, A.G., Antonarakis, S.E., 2007. Human microRNA-155 on chromosome 21 differentially interacts with its polymorphic target in the AGTR1 3′ untranslated region: a mechanism for functional single-nucleotide polymorphisms related to phenotypes. Am. J. Hum. Genet. 81, 405–413. doi:http://dx.doi.org/10.1086/519979.

Sharma, S., Biswas, S., 2011. Sequence trademarks in oncogene associated microRNAs. Bioinformation 6, 364–365.

Skalsky, R.L., Samols, M.A., Plaisance, K.B., Boss, I.W., Riva, A., Lopez, M.C., Baker V, H., Renne, R., 2007. Kaposi's sarcoma-associated herpesvirus encodes an ortholog of miR-155. J. Virol. 81, 12836–12845. doi:http://dx.doi.org/10.1128/JVI.01804-07.

Visvanathan, J., Lee, S., Lee, B., Lee, J.W., Lee, S., 2007. The microRNA miR-124 antagonizes the anti-neural REST/SCP1 pathway during embryonic CNS development. Genes Dev 21, 744–749. doi:http://dx.doi.org/10.1101/gad.1519107.

Wang, D., Qiu, C., Zhang, H., Wang, J., Cui, Q., Yin, Y., 2010. Human microRNA oncogenes and tumor suppressors show significantly different biological patterns: from functions to targets. PLoS One 5 doi:http://dx.doi.org/10.1371/journal.pone.0013067.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., Li, T., 2009. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. 37, D105–D110. doi:http://dx.doi.org/10.1093/nar/gkn851.

Zeng, X., Chen, Y.-W., Tao, C., Alphen Van, D., 2009. Feature selection using recursive feature elimination for handwritten digit recognition. Fifth International Conference Intelligent Information Hiding Multimedia Signal Process 1205–1208. doi:http://dx.doi.org/10.1109/IIH-MSP.2009.145.

Zhang, Huang, J., Yang, N., Greshock, J., Megraw, M.S., Giannakakis, A., Liang, S., Naylor, T.L., Barchetti, A., Ward, M.R., Yao, G., Medina, A., Brien-jenkins, A.O., Katsaros, D., Hatzigeorgiou, A., Gimotty, P.A., Weber, B.L., Coukos, G., 2006. microRNAs exhibit high frequency genomic alterations in human cancer. PNAS 103, 9136–9141. doi:http://dx.doi.org/10.1073/pnas.0508889103.

Zhao, Y., Samal, E., Srivastava, D., 2005. Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. Nature 436, 214–220. doi:http://dx.doi.org/10.1038/nature03817.

BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 9(10)

Search for signatures in miRNAs associated with cancer

Ram Kothandan & Sumit Biswas*

Department of Biological Sciences, BITS, Pilani – K K Birla Goa Campus, Zuarinagar, Goa- 403726; Sumit Biswas – Email: sumit@goa.bits-pilani.ac.in; *Corresponding author

Received May 14, 2013; Accepted May 17, 2013; Published June 08, 2013

Abstract:
Since the first discovery in the early 1990's, the predicted and validated population of microRNAs (miRNAs or miRs) has grown significantly. These small (~22 nucleotides long) regulators of gene expression have been implicated and associated with several genes in the cancer pathway as well. Globally, the identification and verification of microRNAs as biomarkers for cancer cell types has been the area of thrust for most miRNA biologists. However, there has been a noticeable vacuum when it comes to identifying a common signature or trademark that could be used to demarcate a miR to be associated with the development or suppression of cancer. To answer these queries, we report an *in silico* study involving the identification of global signatures in experimentally validated microRNAs which have been associated with cancer. This study has thrown light on the presence of significant common signatures, *viz.*, - sequential and hybridization, which may distinguish a miR to be associated with cancer. Based on our analysis, we suggest the utility of such signatures in the design and development of algorithms for prediction of miRs involved in the cancer pathway.

Keywords: MicroRNA, Signatures, Matches, Seeds, Hybridization.

Background:

The discovery of a short RNA product regulating the expression of the *lin-14* gene in *C. elegans* **[1]** opened the door to a new family of biologically important RNAs that proved to be crucial in fine-tuning the expression patterns of genes. MicroRNAs have later been identified as short sequences (18-22 nucleotides) of RNA, which act as post-transcriptional regulators by binding to complementary sequences on target messenger RNA transcripts, in both the plant and animal kingdoms **[2–6]**. The mature miR binds to the 3'Untranslated Region (UTR) **[7]**, 5' UTR **[8]** and CDS **[9]** of target mRNA sequences, thereby downregulating or upregulating the translation of these genes. This downregulation is achieved either by translational inhibition, or increased mRNA de-adenylation and degradation, or mRNA sequestration **[10–12]** and upregulation by translational enhancement **[8].** Recent evidence however suggests that the target mRNA may also regulate the level and function of miRNAs **[3].**

The extent of complementarity of the so-called "seed" region – generally positions 2-7 **[13,14]** of the miR, was thought to be the basis for identification of potential mRNA targets by a miR **[15, 16].** However, Chi, Hanon and Darnell **[17]** present a new alternative mode for miRNA target recognition involving transitional nucleation, which allows for bulge formation and consequent seed propagation. Recent studies **[18]** also suggest that the regions outside the so-called "seed" may also be important to consider while ascertaining miR-mRNA binding.

Several reviews and articles have been published relating the complicity of certain miRNAs to some cell types **[19–21].** Most studies aimed at identifying cancer specific miR signatures are rather sketchy and specific to a group of related cancerous cells. However, there is no literature or work on common "signatures" to distinguish a miR to be associated with cancer. In an attempt to fill up this void, we have undertaken an extensive exercise, involving all the experimentally validated mRNA targets and their corresponding microRNA interactions, where the mRNA has an established role in cancer development. This dataset was analysed with an aim to discover sequential, structural or hybridization properties to identify microRNAs associated with the cancer pathway. We infer that there are distinct signatures or trademarks that can enable us to demarcate a miR to be involved in the cancer pathway – features that are present in the mature sequences and in the selective arrangement of the seed regions as well.

Methodology:

For the purpose of the present study, construction of an extensive dataset is a prerequisite. A list of genes involved in cancer was obtained from Cancer Gene Census Database (COSMIC) **[22].** From the listed 488 genes, it was observed that they contained both oncogenes and tumor suppressors. List of genes which were not involved in cancer were obtained by calculating their Cancer Linker Degree (CLD) **[23]**. A jack-knife selection of 100 from the total list of 1025 genes would serve as the negative dataset. Further, a list of gene targets which have documented miR interactions was obtained from miRTARBASE (release 2.5) **[24],** which is accepted as the curated database of experimentally validated miRs. A comparison of the list obtained from COSMIC with the interaction data from miRTARBASE yielded the final list of miRNAs involved in cancer. MicroRNA sequences thus filtered were retrieved from miRBASE version 17.0 **[25],** and checked for redundancy. The final size of this dataset came to 2926 microRNAs, which were experimentally validated and unique. Since the 3'UTR regions of genes is the major site for microRNA interaction, we obtained the 3'UTR regions for all the 488 genes in question from the ENSEMBL-BIOMART portal **[26].**

A multiple sequence alignment was done using "MultiAlign" function of MATLAB with “ExitingGapAlignment” method to search for sequence signatures, following our previously published method **[27].** To find the hybridized structure with the best fit in terms of free energy, the miR sequence along with their specific 3’UTR sequence were hybridized using the RNAHybrid program **[28].** Hybridization results obtained from RNAHybrid were parsed and analyzed using an indigenous Perl script, “PairFinder”, which identifies seed, regions outside seeds, mismatches and bulges [http://universe.bits-pilani.ac.in/goa/sumit/Research]. Regions of complementarity having atleast four bases at a stretch were considered to be “seed” regions **[14].** Since regular Watson-Crick base-pairings, especially AU are found to be abundant in functional sites of miR-mRNA interactions **[18],** we wanted to investigate the nature of the base pairing both in the seed regions as well in the regions outside seed. Finally, seed scores, which are indicators of the relative stability of the miR-mRNA interaction were obtained by the formula n(AU)+ n(GC) – n(GU), where AU and GC are assigned positive scores and GU was assigned a negative score.

Results & Discussion:

Construction of the miR dataset was strictly based on the premise that predicted miR will not be, and only experimentally validated miR sequence will be considered. Similarly, all miRs which do not have an experimentally validated target were also excluded from the dataset. Looking for sequence preference in the dataset of oncogenically involved miRs, it was evident that Uracils are the most preferred nucleotides, whereas Cytosines are the least preferred **(Figure 1A)** a result which is in complete agreement to our previous work with a pilot dataset **[27]**. Each stack of bases in the figure represents the relative frequency of the bases at that position **[29].** The letter at the top of the stack is also the tallest and implies its relative abundance at that position. However, the sequence preference for the negative dataset **(Figure 1B)** shows a relative abundance of mainly Guanines, Cytosines are fairly represented as well, while Uracils are least preferred.

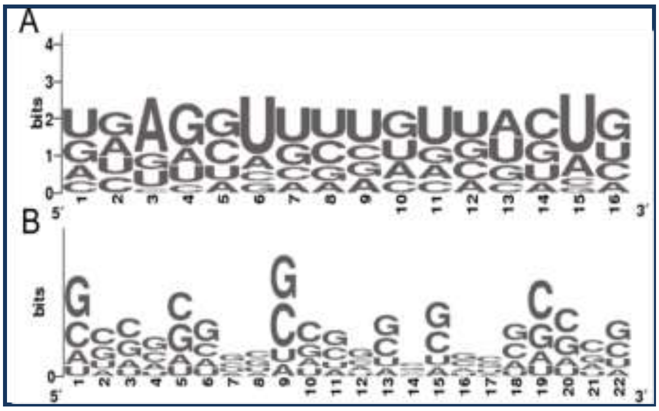


Figure 1: Sequence Conservation in miRs associated with cancer **(A)** and in the negative dataset **(B).**

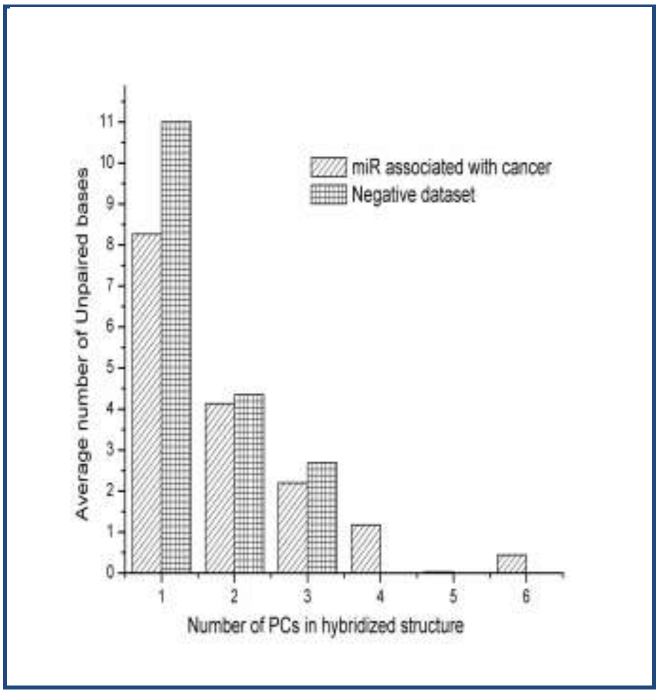


Figure 2: Variation in number of unpaired bases in miRs associated with cancer and the negative dataset. The first pair of bars stands for the variation in the hybrids having a single patch of complementarity (PC), the second for hybrids having two patches, and so on.

Multiple sequence analysis with the ‘MultiAlign’ function and ‘ExistingGapAdjust’ option showed that mature miRs associated with cancer have a sequence signature which can be generalized as 'AG-UU-U-U--CU'. This result was verified manually with the regional percentage conservation score data and found to be true. Additionally the region of consensus lies exactly in the seed region within the position 2-13 nt. This sequence pattern does not have any semblance to the sequences in the negative dataset.

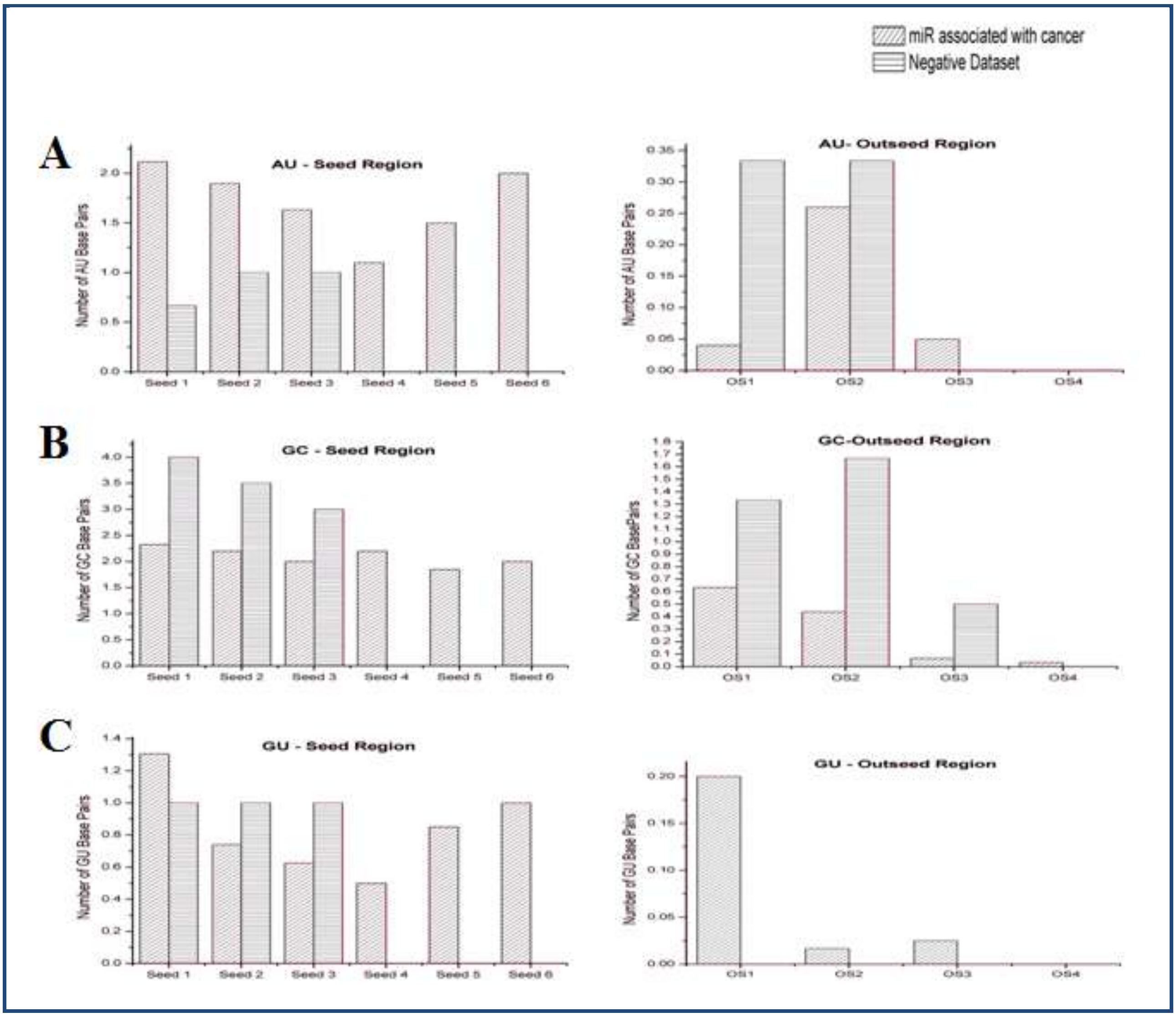


Figure 3: Distribution of the regular Watson – Crick (WC) and the non-WC base pairings between miR associated with cancer and the negative dataset. The panels on the left are for the pairings in the seed region, while the panels on the right are pairings in the regions outside the seed (OS).

Pairfinder was used to identify and categorise the seed, regions outside seeds, mismatches and bulges in the miRNA interacting with the mRNA. Patches of complementarity (PC) are demarcated as the seed regions, as well as the regions outside seeds where base pairings can occur (but in less than four pairs). All bases outside the PCs are unpaired bases. Quantitatively, the number of unpaired bases in miRs not involved in the cancer pathway was quite higher than those in the cancer pathway dataset **(Figure 2)**. For a miRNA-mRNA interaction which has a single patch of complementarity to those which have multiple PCs, it was always observed that the number of unpaired bases is more in the interactions involving miRNAs not associated with the cancer pathway. This was a pointer to the better complementarity of the miRNA while binding to the respective mRNA of genes associated with cancer. Looking for the distribution of the regular Watson – Crick (WC) and the non-WC base pairings, it was evident that AU pairs in the patches of complementarity were much higher in the miRs involved in the cancer pathway than in those which were not **(Figure 3A)**. Higher average of (A+U) % contents have already been cited as an indicator of higher stability **[18].** However, the scenario is reversed when we considered GC pairs. These are more abundant in the interactions of miRNA not associated with cancer **(Figure 3B)**, with the difference being more pronounced in the regions of complementarity outside seeds. The non-WC base pairing, again shows relative abundance in the seed regions of the negative dataset, but are negligible in the regions outside the seed when compared to the dataset of the miRs associated with cancer **(Figure 3C).** Consequently, the seed score of the cancer associated microRNAs is higher on an average (4.108 ± 1.67) than for those microRNAs which are not involved in the cancer pathway (2.151 ± 1.16). This provides a further confirmation to the stability of interactions of those miRNAs which have been experimentally validated to be involved with cancer.

Conclusion:

The work presented in this manuscript highlights the presence of trademarks or signatures that can be used to distinguish between a microRNA which is associated with cancer from one that is not. While sequence signatures show a clear bias towards Uracil usage and against Cytosine in cancer associated miRs, the trend is reversed in the case of non-oncogenically involved miRs. The regions of mRNA-miRNA interaction were categorized using the script "Pairfinder" and Patches of Complementarity were ascertained to distinguish between paired and unpaired regions. Unpaired bases, which contribute to weaker binding, were decidedly more abundant in the negative dataset. So, by the corollary, the miRs associated with the cancer pathway, were found to have stronger interactions with their binding mRNAs. To further augment this hypothesis, the nature of base pairings in the PCs was investigated and the number of AU pairs (which contribute to stability) in both the seed regions and the regions of complementarity outside the seeds was found to be higher in the cases of miRs involved in cancer.

The hypothesis is further strengthened by the seed score – again an indicator of stability of interactions – which is found to be significantly higher for miRNAs with oncogenic associations. Thus, we can safely conclude that miRNAs associated with cancer have more stable and stronger interactions with their mRNAs, as compared to those which are not associated with cancer. While this study was based on the interactions between the 3'UTR region of the gene and the microRNA, it is also true that some interactions in the 5'UTR and coding sequence of the genes need to be analysed as well, and work is being undertaken for the same. These findings, along with other ongoing searches for thermodyanamic signatures would be beneficial to the ultimate goal of constructing an algorithm for identification and validation of microRNAs which could be associated with cancer.

References:

[1] Lee RC *et al. Cell.* 1993 **75**: 843 [PMID: 8252621]
[2] Ambros V, *Nature.* 2004 **431**: 350 [PMID: 15372042]
[3] Reinhart BJ *et al. Nature.* 2000 **403**: 901 [PMID: 10706289]
[4] Nelson P *et al. Trends Biochem Sci*. 2003 **28**: 534 [PMID: 14559182]
[5] Jones Rhoades MW *et al. Annu Rev Plant Biol.* 2006 **57**: 19 [PMID: 16669754]
[6] Sevignani C *et al. Mamm Genome.* 2006 **17**: 189 [PMID: 16518686]
[7] Zhang R *et al. J Genet Genomics.* 2009 **36**:1 [PMID: 19161940]
[8] Orom UA *et al. Mol Cell.* 2008 **30**: 460 [PMID: 18498749]
[9] Hausser J *et al. Genome. Res.* 2013 **23**: 604 [PMID: 23335364]
[10] Bagga S *et al. Cell.* 2005 **122**: 553 [PMID: 16122423].
[11] Cannell IG *et al. Biochem Soc Trans.* 2008 **36**: 1224 [PMID: 19021530]
[12] Wu L *et al. Proc NIPR Symb.* 2006 **103**: 4043 [PMID: 16495412]
[13] Lewis BP *et al. Cell.* 2003 **115**: 787 [PMID: 14697198]
[14] Lekprasert P, *Plos One.* 2011 **6**:e20622 [PMID: 21674004]
[15] Baek *et al. Nature.* 2008 **455**: 64 [PMID: 18668037]
[16] Bartel DP, *Cell.* 2009 **136**: 215 [PMID: 19167326]
[17] Chi SW *et al. Nat Struc Mol Biol.* 2012 **19**: 321 [PMID: 22343717]
[18] Grimson *et al. Mol Cell.* 2007 **27**: 91 [PMID: 17612493].
[19] Schickel R *et al. Oncogene.* 2008 **27**: 5959 [PMID: 18836476]
[20] Croce CM, *Nat Rev Genet.* 2009 **10**: 704 [PMID: 19763153]
[21] Ørom UA & Lund AH, *Nature.* 2010 **451**: 1 [PMID: 19944134]
[22] Forbes SA *et al. Nucleic Acid Res.* 2009 **D38**: D652 [PMID: 19906727]
[23] Aragues R *et al. BMC Bioinformatics.* 2008 **9**: 172 [PMID: 18371197]
[24] Hsu SD *et al. Nucleic Acid Res.* 2011 **39**: D163 [PMID: 21071411]
[25] Griffiths-Jones S *et al. Nucleic Acid Res.* 2008 **36**: D154 [PMID: 17991681]
[26] Kinsella RJ *et al. Database.* 2011 [PMID: 21785142]
[27] Sharma S *et al. Bioinformation.* 2011 **6**: 364 [PMID: 21814397]
[28] Krüger J *et al. Nucleic Acid Res.* 2006 **34**: W451 [PMID: 16845047]
[29] Schneider TD & Stephens RM, *Sequences.* 1990 **18**: 6097 [PMID: 2172928]

Edited by P Kangueane
Citation: **Ram & Biswas,** Bioinformation 9(10): 524-527 (2013)

BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 11(1)

Handling class imbalance problem in miRNA dataset associated with cancer

Ram Kothandan

Department of Biological Sciences, BITS PILANI K K Birla Goa Campus, Zuarinagar, Vasco Da Gama, India; Ram Kothandan – Email: mailram1986@gmail.com

Received January 09, 2015; Accepted January 23, 2015; Published January 30, 2015

Abstract:
MiRNAs are small (~22nt long) non-coding RNA sequences; binds to the complementarity target sites in 3' Untranslated Region (UTR) of mRNA sequences but not restricted to other mRNA regions *viz.,* 5' UTR and Coding sequences (CDS). Complementarity binding of miRNA to mRNA target sites either results in complete degradation of the mRNA itself or it may regulate the mRNA as an oncogene or as a tumor suppressor gene. However, the exact mechanism involved in identifying a miRNA to be associated with cancer is still unclear. Further, with the outburst in the number of miRNAs sequences recorded every year in miRBase, the gap is still widening mainly due to the laborious and economically unfavorable experimental procedures associated with the functional annotation. Motivated by the fact, we constructed a two-step support vector machine-based predictive model - miRSEQ and miRINT. However, the major pitfall during the construction of the model is the class imbalance problem. Hence, in order to overcome class imbalance problem, in the present study we empirically compare the effectiveness of two different methods *viz.,* Synthetic Minority Oversampling Technique (SMOTE) and cost-senstive learning method. Performance measures were evaluated in terms of Precision and Recall. Based on our result, it was observed that for miRNA dataset with high class imbalance utilized for predicting association of cancer, cost-sensitive method outperformed the oversampling method.

Keywords: Cost-sensitive, SMOTE, miRNA-mRNA interaction, Support Vector Machines.

Background

A dataset is imbalanced if the classification categories are not equally represented **[1]**. Class imbalance or skewed dataset mainly arises when most of the instances are labeled as one class (majority class), while very few are labeled as the other class (minority class). Traditional classifiers utilizing the entire training set for prediction are not suitable to deal with imbalanced dataset because they show bias towards the majority class due to over-prevalence. Particularly in case of disease related dataset (like ours) - miRNA dataset associated with cancer, the number of experimentally validated miRNAs are much higher than the number of miRNAs not associated with cancer. The main problem in training a classifier with high imbalanced dataset is that the minority class is often considered as noisy dataset and hence overlooked by the majority class.

Performance of the classifier constructed with a certain level of class imbalance is always unpredictable or deteriorating in many cases. Hence, to overcome the problem of class imbalance, machine learning algorithms generally utilize two methods *viz.,* resampling at the data level *i.e.* either oversampling the minority class e.g. Synthetic Minority Oversampling Method (SMOTE) **[2]** or under sampling the majority class e.g. Easy Ensemble and Balancing Cascade method **[3]**. Utilizing a resampling method is entirely a data driven process. On the other hand, class imbalance is ignored at the algorithm level by adjusting the cost of the classes to counter imbalance, adjusting the probabilistic estimates (in case of decision trees) and adjusting the decision threshold. In certain cases, both cost and resampling methods are used in combination, i.e. individual models are adjusted with these methods and combined as an ensemble to provide better performance **[4]**.

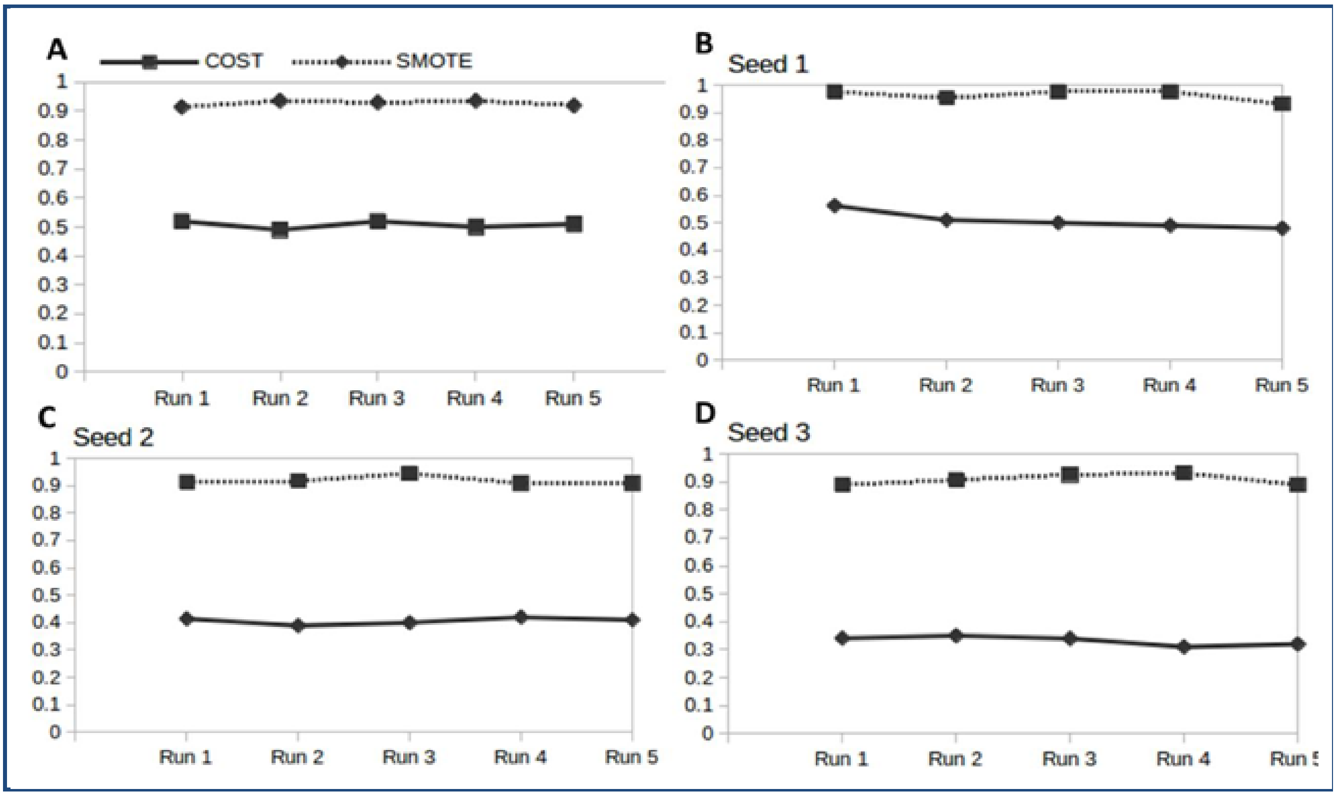


Figure 1: Comparison of SMOTE and cost-sensitive method to overcome class imbalance in the miRNA dataset associated with cancer: **A)** Comparison of SMOTE and cost-sensitive method with miRSEQ classifier trained with sequence based features only; **(B, C & D)** Comparison of SMOTE and cost-sensitive method with miRINT trained with miRNA-mRNA interaction based features. In both the classifier, SMOTE method tends to overfit the test dataset. SMOTE and cost-sensitive methods were tested with five independent test datasets (Run 1 to 5).

Generally in oversampling technique, class imbalance is overlooked by generating new instances with replacement from the minority class. But, generating similar instance at a specific region will overpopulate the minority class and results in bias during actual prediction **[5]**. Hence, in SMOTE, new synthetic samples are generated based on two parameters – the nearest neighbors (k) and the number of instance (n) required. In undersampling, multiple subset of majority class similar in size to the minority class is generated and trained. Since only a part of the dataset is utilized the computation cost and the time associated with this training is very less and efficient than the oversampling methods. However, undersampling methods ignore a large part of the training set making them vulnerable to miss many discriminative features present in them **[3].**

Most learning algorithms attempt to minimize the error rate in the classification by ignoring the difference between the types of misclassification errors. However, for real world problem this assumption wont hold true. Hence, to overcome the problem, cost-sensitive method is preferred generally over other class imbalance methods. Cost-sensitive method along with misclassification cost considers other cost like instance and attribute cost, active learning cost and computational cost. Among the cost, misclassification cost is more important in cost-sensitive learning and it can be either stationary (assigning a cost matrix) or dataset dependent. Thus, in the present study, we compare the effectiveness of two methods to overcome class imbalance in terms of precision and recall to construct an efficient classifier in predicting miRNAs associated with cancer.

Methodology:

Dataset Preparation

Dataset preparation was carried out for positive and negative set individually. For training purpose, 239 experimentally validated miRNAs obtained from our previous work would serve as positive dataset **[6]**. For negative dataset, precautionary steps were undertaken to avoid randomness in the dataset, i.e. randomly generated and predicted dataset were completely avoided. Only experimentally validated 32 miRNAs obtained from TargetMiner were considered as negative dataset **[7]**. For evaluating the effectiveness of the two methods compared in the study, we constructed an independent test dataset not utilized in training purpose. A 10-fold cross-validation method is used as a standard method for revalidation during training **[8].**

Feature Extraction

A list of 60 features were extracted from experimentally validated miRNA sequences, miRNA-mRNA interaction data and thermodynamics of miRNA-mRNA binding as obtained from RNAhybrid **[6, 9, 10]**. We utilized Pairfinder, a perl script to parse the various features from the miRNA-mRNA hybridized structure **[6].** In this present study, a two-step classifier (*viz.,* miRSEQ and miRINT) was constructed. MiRSEQ preliminarily predicts the miRNA associated with cancer based on 26 sequence-based features; whereas miRINT utilized 34 miRNA-mRNA interaction-based features to confirm the association of miRNA with cancer.

Learning Algorithm

The choice of learning algorithm plays a critical role in overcoming class imbalance. In this present study, we employed Support Vector Machines (SVM) with Radial Basis Function (RBF) as kernel function for training the miRNA dataset **[11]**. In a binary classifier, SVM classifies two classes by constructing a hyperplane in three dimensional space separated by margins. We utilized LIBSVM package in Waikato Environment for Knowledge Analysis (WEKA) **[12]**. Random search method was employed to identify optimum algorithm parameter cost (c) and gamma (λ) rather than computationally expensive grid based method**.**

Both SMOTE and cost-sensitive method packages available within the WEKA environment were utilized to handle the class imbalance during the training process. For SMOTE, we considered the nearest neighbor to be five (k=5) and the percentage of instances generated (n) in each iteration to be 100. The number of iterations was limited till there is a shift in the class distribution. In a typical class imbalanced problem, cost-sensitive algorithms require a cost-matrix to represent costs for different misclassification types. The method tends to minimize the number of high cost error and then further generates a model with low misclassification cost. Misclassification cost can be assigned to both binary and multi-class classification problems. We constructed a 2x2 cost matrix for reweighing the data space. Cost for the correctly classified instances are assumed zero (*i.e.,* the cost associated with the True Positive (TP) and True Negative (TN) is zero) **[13]**. The main aim of utilizing cost-sensitive method is to construct a model with minimum misclassification cost and is given by the equation (1)

$$Cost = FNrate \times C(0,1) + FPrate \times C(1,0)$$

(Equation 1)

Where, C(0,1) and C(1,0) are the costs associated in prediction of False Positive (FP) and False Negative (FN) respectively.

Performance Evaluation

$$Precision = \frac{TP}{TP + FP} \qquad \text{(Equation 2)}$$

$$Recall = \frac{TP}{TP + FN} \qquad \text{(Equation 3)}$$

Results & Discussion:

The focus of the study is to obtain an efficient method for handling class imbalance in miRNA dataset associated with cancer. MiRSEQ and miRINT classifiers were constructed with both SMOTE and cost-sensitive method with SVM as the learning algorithm. Only experimentally validated miRNA were used for training purpose. Randomly generated, predicted miRNA sequences were neglected completely in order to avoid randomness in the dataset during the training process. Prior to training process dataset was normalized, since significant difference in the variance will dominate the RBF function and does not allow learning the dataset from other features. Utilizing mean value for missing attribute during the feature extraction was also avoided.

The performance of the constructed models were evaluated based on precision and recall. Usually in training machine learning algorithms, performance is evaluated using confusion matrix. However, for problems with high class imbalance, evaluating the performance of the classifier directly based on confusion matrix is not preferred. Alternatively, measures like precision and recall would reveal the actual predictive performance of the classifier. In disease related dataset, particularly miRNA dataset associated with cancer (like ours), precision would provide an exact measure of predictive performance of the constructed model since a single false prediction in disease related dataset would be catastrophic.

The predictive performance evaluated during the training process was marginally similar between the two methods being compared. However, when challenged with test dataset, cost-sensitive method performed better than the SMOTE. The underlying problem for poor predictive performance with SMOTE is due to overfitting (precision > 0.9 in all independent test runs are shown in **Table 1 See supplementary material).** One possible reason for overfitting with SMOTE is that the method centers more on the specific region in the feature space as the decision region for the minority class, than increasing the overall number of instances. Further, new instances are synthesized based on the number of the nearest neighbors chosen and also based on the number of new instances required per iteration. Thus SMOTE overpopulates a specific region rather than increasing the overall instances. Further, the classifier constructed with SMOTE method misclassified every instances as the minority class due to over-prevalence in the specific region during the independent test dataset prediction.

On the other hand, cost-sensitive method seamlessly performed better than SMOTE because it considers misclassification cost based on the dataset utilized in the training (precision 0.52 for miRSEQ and average precision 0.4 in all seed based models for miRINT) (**Table 1**). From (**Figure 1**)**,** it is evident that SMOTE method tends to overfit the dataset in both miRSEQ and miRINT classifier, whereas cost-sensitive showed significantly a steady performance in all test runs. Further, in order to boost the performance of classifier with SMOTE method, we reduced the number of instances generated per iteration. This will avoid over populating the minority class in a specific region. However, it was observed that there was no significant improvement in the performance measurement. For miRINT, the dataset was segregated based on the number of seed region formed in the hybridized structure. Similar to miRSEQ performance, the SMOTE method did not show much improvement in terms of precision, rather they tend to overfit (precision > 0.9) the dataset and thus left no room for further improvement.

Conclusion:

The work presented in this paper gives an empirical comparison of two methods to overcome class imbalance (*viz.,* SMOTE and cost-sensitive method) in prediction of miRNA associated with cancer. Among the two methods compared the SMOTE handles class imbalance at the data level and cost-sensitive method at the algorithm level. Handling class imbalance at the data level for disease related prediction (like ours) would induce several synthesized instances. Even though, oversampling method provide a good performance measure at the training step, when challenged with independent test datasets the performance of the classifier deteriorated completely. To further support the hypothesis, the prediction obtained from classifier constructed show overfitting of the test dataset.

On the other hand, cost-sensitve method provided a steady performance measure in each of the independent runs and thus acts as an effective method in handling class imbalance in miRNA dataset. The performance of cost-sensitive method can be further enhanced by utilizing appropriate feature selection method like Recursive Feature Elimination method (RFE) prior to the training process. Prioritizing most discriminative features would increase the performance of the classifier with cost-sensitive method. Further, utilizing different learning algorithm along with cost-sensitive method would boost the performance significantly and such a work is under progress in our group. Thus, we conclude that for prediction of miRNA associated with cancer with high class imbalance in dataset, cost-sensitive method performs better than the oversampling method.

Acknowledgment:
Ram Kothandan was supported by a scholarship from Council of Scientific Research and Industrial Research, Govt. of India.

References:

[1] Ding J *et al. BMC Bioinformatics* 2010 **14:** 11 [PMID: 21172046]

[2] Lertampaiporn S *et al. Nucleic Acids Res.* 2013 **41:** e21 [PMID: 23012261]

[3] Liu XY *et al. IEEE Trans Syst Man Cybern B Cybern*. 2009 **39:** 539 [PMID: 19095540]

[4] Yin HL & Leong TY, *Stud Health Technol Inform*. 2010 **160:** 856 [PMID: 20841807]

[5] Hao M *et al. Anal Chim Acta.* 2014 **806:** 117 [PMID: 24331047]

[6] Kothandan R & Biswas S, *Bioinformation* 2013 **9:** 524 [PMID: 23861569]

[7] Bandyopadhyay S & Mitra R, *Bioinformatics* 2009 **25:** 2625 [PMID: 19692556]

[8] Gamzon ER *et al. Plos One.* 2010 **5:** e13534 [PMID: 20975837]

[9] Batuwida R & Palade V, *Bioinformatics* 2009 **25:** 989 [PMID: 19233894]

[10] Sharma S & Biswas S, *Bioinformation* 2011 **6:** 364 [PMID: 21814397]

[11] Vapnik VN, *IEEE Trans Neural Netw*. 1999 **10:** 988 [PMID: 18252602].

[12] www.cs-waikato.ac.nz/ml/weka

[13] Zidelmal Z *et al. Comput Methods Programs Biomed.* 2013 **111**: 570 [PMID: 23849928]

Edited by P Kangueane
Citation: **Kothandan,** Bioinformation 11(1): 006-010 (2015)

Supplementary material:

Table 1: Comparison of SMOTE and cost-sensitive method in terms of Precision and Recall. Only average value of five independent runs are tabulated. For miRINT, miRNA-mRNA hybrid structures were segregated into seed 1, seed 2 and seed 3 models based on the number of seed region formed in their structures and trained individually.

MiRSEQ		**Precision**	**Recall**
Cost-sensitive		0.52	0.521
SMOTE		0.927	0.9345
MiRINT	**Number of Seeds**	**Precision**	**Recall**
Cost-Sensitive	**Seed 1**	0.562	0.426
	Seed 2	0.414	0.644
	Seed 3	0.341	0.584
SMOTE	**Seed 1**	0.9627	0.9042
	Seed 2	0.9181	0.931
	Seed 3	0.908	0.9141

* Models with Precision > 0.9 misclassified all instances as minority class in SMOTE