Flanking region sequence information to refine microRNA target predictions

Russiachand Heikham[†] and Ravi Shankar[*†]

Department of Bioinformatics and Structural Biology, Indian Institute of Advanced Research, Gandhinagar 382 007, India

[†]*Contributed equally*

**Corresponding author (Email, ravish9@gmail.com; ravish@iiar.res.in)*

The non-coding elements of a genome, with many of them considered as junk earlier, have now started gaining long due respectability, with microRNAs as the best current example. MicroRNAs bind preferentially to the 3′ untranslated regions (UTRs) of the target genes and negatively regulate their expression most of the time. Several microRNA:target prediction softwares have been developed based upon various assumptions and the majority of them consider the free energy of binding of a target to its microRNA and seed conservation. However, the average concordance between the predictions made by these softwares is limited and compounded by a large number of false-positive results. In this study, we describe a methodology developed by us to refine microRNA:target prediction by target prediction softwares through observations made from a comprehensive study. We incorporated the information obtained from dinucleotide content variation patterns recorded for flanking regions around the target sites using support vector machines (SVMs) trained over two different major sources of experimental data, besides other sources. We assessed the performance of our methodology with rigorous tests over four different dataset models and also compared it with a recently published refinement tool, MirTif. Our methodology attained a higher average accuracy of 0.88, average sensitivity and specificity of 0.81 and 0.94, respectively, and areas under the curves (AUCs) for all the four models scored above 0.9, suggesting better performance by our methodology and a possible role of flanking regions in microRNA targeting control. We used our methodology over genes of three different pathways – toll-like receptor (TLR), apoptosis and insulin – to finally predict the most probable targets. We also investigated their possible regulatory associations, and identified a hsa-miR-23a regulatory module.

[Heikham R and Shankar R 2010 Flanking region sequence information to refine microRNA target predictions; *J. Biosci.* **35** 105–118]
DOI 10.1007/s12038-010-0013-7

1. Introduction

MicroRNAs are small non-coding RNAs which are formed by two-tiered processing of precursor gene products, resulting in mature oligomers with a size of ~23 nucleotides (Cullen 2004; Kim *et al.* 2006). These small but important components of the cellular system exist broadly in all multicellular organisms. More than 700 microRNA sequences have been reported for human and a total of 10 883 known microRNAs have been reported for various species in the Mirbase database. Mirbase is a database that archives updated information about microRNAs across various species with full annotation and classification (Ambros *et al.* 2003; Griffiths-Jones *et al.* 2006, 2008). About 30% of the total genes of higher eukaryotes are expected to be controlled by microRNAs (Griffiths-Jones

Keywords. Bioinformatics; genome; microRNA; non-coding; RNA

Abbreviations used: Ac, accuracy; AUC, area under the curve; FN, false negative; FP, false positive; MCC, Matthew correlation coefficient; ROC, receiver operating characteristic; Sn, sensitivity; Sp, specificity; SVM, support vector machine; TFBS, transcription factor-binding site; TLR, toll-like receptor; TN, true negative; TP, true positive; UTR, untranslated region; VDR, vitamin D receptor

Supplementary data pertaining to this article is available on the *Journal of Biosciences* Website at *http://www.ias.ac.in/jbiosci/mar2010/pp105-118-suppl.pdf*

2004). Additionally, in the case of vertebrates, the matter is further complicated by the fact that, unlike plants, there are very few cases with absolute complementarity between a microRNA and its target (Miller and Waterhouse 2005). A large part of abundance assessments is based upon predictions made by presently existing softwares which themselves agree with each other to a limited level and a maximum of 50% (Kiriakidou *et al.* 2004; Hammell *et al.* 2008). In general, these softwares predict large amounts of false-positive targets and leave ample space for improvement in target prediction methodologies. Most of them are primarily dependent on two major assumptions: (i) the free energy of interaction of the microRNA with its target, and (ii) the conservation and continuity of the seed region (Lai 2002, 2004).

An interesting study has been done by Kertesz *et al.* (2007) where information from partition function for the untranslated region (UTR) has been used to assess accessibility in terms of difference in free energy (ΔG). This approach has a predecessor in the microTAR program, which considers the difference between the energy required to open the structure of the target region and free energy of target formation (Thadani and Tammi 2005). The accuracy of such programs may decrease with an increase in length of the RNA as they are dependent upon single-sequence and thermodynamics-based RNA structure prediction methodologies. A detailed survey of the limitations of RNA structure prediction methodologies can be found elsewhere (Gardner and Giegerich 2004; Andronescu *et al.* 2005).

Some sofwares have successfully used phylogenetic conservation to improve predictions and have given strong support for assumptions that consider phylogenetic conservation of target sites as the basis for the targeting and prediction of valid targets (Lewis *et al.* 2005). A new generation of softwares has evolved now, which are based on a multivariate machine learning approach and consider multiple features for classification (Kim *et al.* 2006; Wang and El Naqa 2008). Due to the large number of false-positive results, the overall agreement between these softwares is very low. Recently, a support vector machine (SVM)-based target prediction refinement approach has been introduced through the MirTif software, which takes output from a target prediction tool and tries to identify the potential predicted targets based on k-mer information from a predicted Target:miR pair (Yang *et al.* 2008). Two recent studies on the relationship between microRNAs and targets have observed the significance of flanking region sequences around the target sites in a contextual manner and emphasized the need to look beyond seed pairing and miR:Target interactions alone (Grimson *et al.* 2007; Didiano and Hobert 2008). The motivation for the present work has largely been drawn from the observations made by these two landmark studies, along with the work done by Kertsez *et al.* (2007) on the role of flanking regions and their folding in target determination through target site accessibility. In the present study, we have considered the intrinsic sequence behaviour of the flanking regions around the target sites as a critical determinant, which could hold the signature for a possible target in the form of characteristic variations in the dinucleotide density profiles of the flanking regions.

2. Materials and methods

2.1 *Experimental dataset collection*

In the present study, we have used a newly released database of experimentally validated targets for human, miRecords version 1, in addition to TarBase version 4 (Sethupathy *et al.* 2006; Xiao *et al.* 2009). Many of the reported miR:Target interactions in MiRecords have been validated through site-specific mutational analysis, in addition to expression analysis. Both the databases differ from each other with just after overlapping entries, which have been removed in the present study. There were a total of 99 experimentally validated interactions for human in Tarbase version 4 and about 1072 experimentally validated entries in MiRecords. We considered only those entries where microRNA-binding site information was available and experimentally validated. We did not consider those entries in which reported binding sites were not found in the UTR sequences, sites that were poorly defined or where UTR sequences differed or were not available. Details of all the datasets used in this study are available in supplementary data 1. Besides these, a total of 38 experimentally validated negative interaction instances reported by various studies and used by Yang *et al.* (2008) were also collected. Details of other additional datasets used in model building and testing for SVMs have been given in the following related sections.

2.2 *Encoded pattern generation and scan for equivalence*

In some experimental studies, it has been reported that the activity of microRNA is sensitive to the specific interactions in the entire length and even a single change at any specific position may alter the efficacy of inhibition (Doench and Sharp 2004; Didiano and Hobert 2008). In the present study, we have given due importance to this observation and considered the interaction patterns between the microRNA and its targets through the entire length. Encoded sequence patterns precisely retain the interactions between a microRNA and its target. For all the experimentally validated targets where mutational analysis was not performed in the experimental record, sequences were extracted with some extra sequences for both the flanks. They were aligned to their partner microRNA reverse complement using a standard alignment procedure and optimized scoring

matrix. The same was done for the microRNA:target pairs predicted by the RNAhybrid tool (Rehmsmeier *et al.* 2004). All alignments were converted into a single encoded sequence pattern represented by an alphabet of size five i.e. Σ = {M,X,B,b,W}; where M = match; X = mismatch, B = bulge in the target, b = bulge in the microRNA, W = wobble match. The encoded patterns generated for experimentally validated pairs were collected in a form of a library, giving precise information on interaction patterns observed so far and experimentally feasible for various microRNAs. In this manner, we generated 148 unique encoded patterns overall from the available experimental details in miRecords and TarBase for human, representing various interaction arrangements experimentally known in human. Further, we considered genes of the apoptosis, diabetes and insulin, and toll-like receptor (TLR) pathways for application of our methodology and associated studies from the findings. For every pathway, member genes were identified from the KEGG database (Kanehisa and Goto 2000). Using their Ensembl ID, the corresponding mRNA and UTR sequences were retrieved from the Ensembl Biomart server (Hubbard *et al.* 2002). We retrieved the sequences of 115 genes in the TLR, 88 genes in the apoptosis and 138 genes in the insulin pathways. For each of these sets of genes, we retrieved their UTR sequences, over which RNAhybrid was run separately with an increased energy cut-off parameter of –10 kcal.

For every experimentally validated encoded pattern, we looked for closely similar or identical matches across the enormous amount of predicted encoded patterns for every pathway. Such predicted patterns, exhibiting similarity to the experimental patterns, may point towards a biologically feasible miR:Target interaction. The similarity level can be categorized into various levels, depending upon the mismatches.

2.3 *Frequency of open nucleotides in target sites*

In a previous work by Long *et al*. (2007) with a limited amount of experimental data, it was argued that a continuous open block of four bases or more could be required by the target sites to find suitable interactions with the targeting microRNAs. With the presently available experimental data from both the databases, we looked into Ensembl for microRNA target genes and their sequences. All genes were searched for their respective UTRs and their respective UTRs were retrieved. The retrieved UTRs were subjected to secondary structure fold prediction using the Vienna Package RNAfold software (Hofacker and Stadler 2006). Using a python script, the target sites reported for each UTR sequence were searched and the corresponding regions from the secondary structure sequences were retrieved for further analysis.

In order to test the specificity of observations, we needed random data as the negative set for comparison and statistical significance. For negative data, we created two datasets: (i) randomly selected coding regions from the 5′ end, and (ii) randomized target UTRs retaining the same nucleotide composition, in order to avoid any possible bias. For contribution of openness, we considered three different features: (i) stretch of the longest continuous unpaired nucleotides present in the target site, (ii) total number of unpaired nucleotides present in the target region, and (iii) ratio of unpaired nucleotides versus the target region length. The R statistical package was used to perform the Wilcox test.

2.4 *Flanking region analysis*

To detect any overrepresented sequence motif in the flanking regions, we ran Gibbs motif sampler for DNA (Thompson *et al.* 2003*,* 2004). The inputs to this part of the analysis were the complete target region of 70-base flanking sequences around the microRNA target site. We also looked for any possible structural motif across the experimentally validated target sequences with 70 base flanks on both sides of the target region. We used tools such as RNAforester and mLocaRNA to get any possible common secondary structure motif present in the majority of these sequences (Höchsmann *et al.* 2003; Will *et al.* 2007). These tools utilize the RNA secondary structure information to develop multiple alignments as well as report the structural motifs.

2.5 *Dinucleotide density variation-based discrimination using support vector machines*

2.5.1 *Feature extraction, feature selection and window optimization:* In order to discriminate between the target UTRs and negative instances, we needed distance-specific dinucleotide density variation profiles for the flanking regions on both sides of the target sites, with some standardized window size for every positive and negative instance. We considered 70-base long flanks on both sides of the target sites. For every such instance, the target region was considered as the centre from which both terminals gave rise to two different subsequences in two opposite directions. On a trial and error basis, we tested various sliding windows with different sizes to scan the sequences in both the directions. Windows overlapping with a single base were made. For every training instance, we recorded the dinucleotide density in every window, which reflected the dinucleotide-based compositional variations at different distances and positions from the target region. In this way, the varying dinucleotide densities with respect to the distance from the target site estimated for each instance formed the feature vector set for the training instances. On the basis of various *F*-score values to estimate the discrimination

power of various features to separate the positive instances from the negative ones, the most discriminating features were found. The *F*-score values to measure every feature's discrimination capacity out of *n*- features were calculated using the following relationship:

$$F = \frac{(\bar{X}i^{(+)} - \bar{X}i)^2 + (\bar{X}i^{(-)} - \bar{X}i)^2}{1/(n+ - 1)\sum_{n+} (Xk,i^{(+)} - \bar{X}i^{(+)})^2 + 1/(n- - 1)\sum_{n-} (Xk,i^{(-)} - \bar{X}i^{(-)})^2}$$

Where,

$\bar{X}i^{(+)}$ = Mean value of i-th feature in experimentally validated positive instances

$\bar{X}i^{(-)}$ = Mean value of ith feature in negative instances

Xi = Overall mean value of the ith feature

$Xk,i^{(+)}$ = Feature value for k-th positive instance for ith feature

$Xk,i^{(-)}$ = Feature value for k-th negative instance for ith feature.

2.5.2 *About SVM and parameter optimization:* SVMs are kernel-based statistical learning machines, where a discriminant function is established for such a margin which is widest and successfully separates the positive sets from the negative sets using multidimensional feature space mapping (Drucker *et al.* 1997). The margin, which maximizes the separation between the two categories, is formed by support vectors, the training cases, which define the margin of separation around the discriminant line, i.e.

$$F(x) = \mathrm{sign}\left(\sum_{i=1}^{i} \mathrm{vi.} \cdot k\left\{x \cdot Xi\right\} + b\right)$$

Where

x = Input vector to be classified (in our case its sequence to be predicted as target)

xi = Support vector *i*

k(x.Xi) = Kernel function between the support vector and input vector

Vi = quadratic programming parameter

b = bias of the non-linear decision function

In the core of SVMs are kernel functions which actually measure the similarities between two objects defined over multiple features. Every object which is classified is defined over a number of features, and similarity between two such objects is estimated using various kernel functions. We opted for a Gaussian radial basis kernel. An SVM has two parts: training and testing. For training, one needs to feed known instances of positive and negative classes. Before performing the training, it is better to convert the data vectors into corresponding scaled values, which can be scaled on Gaussian parameters, as values coming in different ranges for different features can influence the estimate of similarity measurement and decision. In our case, we did not need scaling, as all the values taken were in the range of 0–1. The next step was a grid search to find the best training parameters for the radial basis kernel-based learner. This searches the best *C* and *γ* parameters, both of which influence the accuracy of the prediction. The *C* parameter specifically determines the optimal trade-off between the process of margin maximization and minimization of training error. We performed this by using a grid search script available in the LibSVM package, with cross validation value of 10 folds (Chang and Lin 2001).

2.5.3 *SVM models, training and testing:* We used LibSVM's training and prediction modules in our analysis. Since we had experimentally validated data from two different sources, we decided to prepare two separate SVM models from both the databases as the training sets. These two separate SVMs were formed with a total of 73 positive sequences from TarBase and 73 negative sequences derived from randomized sequence data, as well as randomly picked sequence data from non-coding regions (total 146 sequences), a total of 88 positive sequences from miRecords, 88 randomized sequences and randomly picked non-coding sequences as the negative set (total 176 sequences). Besides this, we also considered 32 experimentally validated negative interactions, which were predicted as microRNA targets but were found to be false positives when experimentally tested (Yang *et al.* 2008). Preparation of a training dataset is critical, as the wrong proportion of negative and positive instances as well as an unsuitable negative set can influence the accuracy through a class imbalance problem (Akbani *et al.* 2004). In this article, we have used terms such as 'Mirecords (based) dataset' for datasets where the positive instances have been derived from MiRecords, and 'Tarbase (based) dataset' for those sets that have their positive instances collected from the Tarbase database. For the TarBase-based SVM model, the complete miRecords-based model's negative and positive datasets worked as the testing set, and for the miRecords-based SVM model, the complete TarBase model's negative and positive datasets worked as the testing set. This way, the accuracies of both the SVMs were estimated over a relatively large amount of validated datasets, which has not been done before. In addition to this, other datasets were also prepared for further performance assessment using various combinations of positive and negative instances as described below.

2.5.4 *Performance and SVM model testing:* Sensitivity (Sn), specificity (Sp) and accuracy (Ac) and Matthew

correlation coefficient (MCC) of the SVMs were estimated using the following equations:

$$Sn = TP/(TP + FN)$$
$$Sp = TN/(TN + FP)$$
$$Ac = TP + TN /(TP+TN+FP+FN)$$
$$MCC = \{\{TP*TN\} - \{FP*FN\}\} / \sqrt{\{TN+FN\}[TN+FP][TP+FN][TP+FP]}$$

Where TP = true positive, TN = true negative, FP = false positive and FN = false negative.

Performance was also measured by forming receiver operating characteristic (ROC) curves and calculation of the area under the curve (AUC) over four different dataset models, as accuracy measurement alone can be misleading sometimes. The ROC measures the performance behaviour of a TP with respect to an FN and provides information about the level of trade-off between these two. An AUC value of 0.5 or below suggests a random nature of the classifier and an inappropriate model. The higher the value of the AUC, the better the classifier model. Altogether, four different dataset models were formed to work as training and testing sets besides the above-mentioned test sets (Supplementary data 5). Three different tests, besides the above-mentioned tests, were carried out for performance assessment and comparison with a recently published refinement program MirTif (Yang *et al.* 2008). The positive instances were taken from experimentally validated data for human in the Mirecords and Tarbase databases, while the negative datasets were made by an equal number of randomized UTR sequences and randomly picked non-genic sequences. The randomized data from UTR sequences and randomly picked non-genic sequences as negative instances are different in both the datasets for model generation purposes. Besides this, 32 experimentally validated instances of negative predictions were considered from the complete negative dataset of a total of 38 sequences used by the recently published MicroRNA:target refinement tool, MiRTif (Yang *et al.* 2008). These 38 instances were experimentally validated psuedotargets, out of which we could consider 32 instances suitable for analysis due to sequence discrepancies and data format requirements. For all these models, the ROCs and AUCs were estimated for 10-fold cross-validation along with other performance measures.

2.6 *Regulatory and upstream region analysis*

For all of the predicted target genes, their 2 kb upstream promoter regions were extracted from Ensembl. The upstream promoter region sequences were scanned for putative transcription factor-binding sites (TFBS) using the Transfac database and Match software (Wingender *et al.* 2000; Kel *et al.* 2003). In order to get the optimum output from Match, we opted for settings that minimized the number of FPs and FNs.

To know which genes were similar to the target genes found on the basis of their expressions at different stages/tissues, we used the Gene Sorter module at USCS (*http://genome.ucsc.edu/cgi-bin/hgNear)*. We retrieved all the genes exhibiting expression proximity to every predicted target, and retrieved their corresponding 2 kb upstream promoter regions from Ensembl, which were analysed for TFBS along with the promoter region upstream sequences of the target genes. MicroRNA target predictions reported by TargetScan and miRDB were also considered in this analysis (Lewis *et al.* 2005; Wang and El Naqa 2008).

3. Results and discussion

3.1 *The optional: interaction pattern-based initial sorting*

For many experimentally validated microRNA:target pairs reported in databases such as Tarbase and miRecords, we noticed that the free energy of interaction was higher than –20 Kcal/mol, an interaction energy cut-off usually practised in target prediction. We also noticed different interactions for the same microRNA in many cases. The minimum free energy may not always ensure correct targets. This is supported by the fact that in human there are fewer cases of exact or very high complementarity ensuring lesser free energy. The same amount of minimum or lesser free energy can be obtained for a different interaction between the target and the microRNA. In the present study, for many cases, we observed different interaction patterns for the same microRNAs and many different interaction patterns for different microRNAs (supplementary data 2). Some recent studies have argued for high specificity of microRNA–target interactions and emphasized the need to review the rule of 5′ seed-based predictions as other parts of the miR:Target pairs were also found to be critical (Brennecke *et al*. 2005; Kim *et al*. 2006; Didiano and Hobert 2008; Hammell *et al*. 2008). Therefore, in the primary stage of this work, we assumed that local interaction geometry should be specific and helpful in the initial sorting of the potential targets from the predicted interaction. In total, we generated 148 unique encoded interaction patterns for human, which may be considered as the library of all possible experimentally validated interactions for various microRNAs in human.

For our analysis, we considered genes from the apoptosis, TLR and insulin pathways. The first-stage targets were predicted by RNAhybrid, which yielded a large number of predicted targets from which most potential targets had to be screened using our methodology. All predicted microRNA:target pairs were realigned for refined alignment through the same procedure as discussed above. This also

ensured a common format for both the target:microRNA interaction representations. The predicted patterns were scanned against the library of the experimental encoded patterns for various similarity levels. In every pathway, the least numbers were observed for a 100% match, which increased when the similarity levels were reduced. These numbers and candidates for various categories are available in supplementary data 3. We also assumed that the predicted interaction pattern matching some of the closest experimental encoded interaction patterns should have same microRNA partners, given that only the experimentally known interactions for a particular microRNA were considered. We observed that a smaller fraction of microRNAs shared 100% resemblance for their experimentally validated encoded interaction patterns with other microRNAs. Even for a single mismatch condition, a small number of microRNAs came in the same group (supplementary data 2). Through this step, we screened out a total of 17 potential target genes from the three pathways mentioned above and selected them for further study in the follow up of our SVM-based prediction of these.

3.2 *The transitional: measuring target region openness as a discriminating feature*

In an earlier study, it has been argued that the structural openness of the targets could be an important factor and the structure should be thermodynamically supportive in allowing a microRNA to bind to the target site (Robins *et al.* 2005). Robins *et al.* (2005) had considered the folded mRNA structure's role in predicting targets in *Drosophila* and hypothesized first about the possibility of considering the role of free base pairs as the targeting centre in a target mRNA. They used the predicted secondary structures of UTRs exclusively and estimated the single-stranded regions for at least 3 bases. A related study considering target site structure for openness was carried out by Long *et al.* (2007). They assumed that every potential target site should have four continuous bases with a high probability of being unpaired (Long *et al.* 2007). They experimentally tested it on a small dataset. We tried to look into this aspect of target sites in order to know if this particular aspect could be used as an efficient discriminating feature. We carried out a statistical test over predicted UTR structures considering UTRs with experimentally reported target sites, randomly generated UTR structures and non-UTR structures to test the strength of local target site openness as a possible discriminating feature as argued by Long *et al.* (2007). From our analysis, we observed that the mean continuous stretch of open regions for target sites was ~6 bases long, the average number of unpaired bases was ~9 bases and the average ratio of unpaired bases to target region length was ~0.4. In order to assess whether the observations reflected the openness properties of only microRNA target regions or not, we carried out the Wilcox Rank Sum tests for all the three classes of observations between all experimentally validated sites and two different representations of the random data. We did not find these features significant enough at a significance level of 5% to demarcate a target site away from a non-target site, as the *P* values were above 0.05 (*see* supplementary data 4). Therefore, parameters based on the openness assessment in the target sites alone did not appear sufficient enough to achieve a sensitive generalizable discrimination function for target site prediction, prompting us to look beyond the target regions and include the information and role of flanking regions around target sites.

3.3 *The critical: flanking region analysis and SVM-based identification*

For identification of some structural and sequential motifs in the flanking regions of the target sites, we looked into the experimentally validated sequences using the approach described in the Methods section. For structural motif, we could not find any significant motif but at sequence level, we found three kinds of sequence motifs, overrepresented in a mutually exclusive manner – A-rich regions, U-rich regions and C-rich regions. The presence of AU-rich regions in the flanking regions has been observed to be associated with some of the target RNAs (Grimson *et al.* 2007; Didiano and Hobert 2008). Studies done by Didiano and Hobert (2008), Kertesz *et al.* (2007) and Long *et al.* (2007) suggest very clearly that microRNA:target interaction information alone may not be sufficient as a marker for successful target prediction, as information from the flanking regions, structure of the target UTR and site accessibility may also be critical for successful target prediction.

Considering the above-mentioned observations and limitations of sequence- and structural motif-based approaches, we looked for some sequence composition-based intrinsic features of the flanking regions around the target sites. Such features may retain the essence of structural information derived from the nearest neighbourhood principle and may also work like some signatures. The dinucleotide density variation profile as appeared to be a solution as it retains compositional information as well as the essence of the nearest neighbourhood principle used in RNA structure prediction. In a previous work, the dinucleotide densities of upstream regions have been successfully used to discriminate between genes of different functional categories as well as between species (Shankar *et al.* 2007). Recently, SVMs have been successfully used in microRNA:target prediction with multiple features, including k-mer features from target:miR interactions (Kim *et al.* 2006; Wang and El Naqa 2008). The same could be used for our purpose of discrimination based on position-specific variation of dinucleotide density.

The idea behind SVMs is generation of a maximum margin hyperplane to minimize error in the classification process. It is critical to select strong features which could generate maximum discrimination. We prepared the dinculeotide density variation profiles of the experimentally validated positive targets and false targets by considering the variation pattern of densities of 16 dinucleotides across the flanking regions around the target in a position-specific manner in both the directions. Every positional window and the dinucleotide densities for those positions were considered as possible discriminating features. For the above-mentioned dataset models, the most discriminating features were found through a feature selection procedure. The most discriminating position-specific dinucleotide densities were selected from the flanking regions of the various target sites from negative as well as positive instances, which were well able to discriminate between the positive and negative instances. The ten best common discriminating features among the top fifteen features obtained for Tarbase- and Mirecords-based datasets are listed in table 1, where the average validation accuracy for both the models was found to be above 90%.

We prepared sliding windows of various base lengths and recorded the performance of the SVMs to correctly classify targets and non-targets. For different window sizes, we tested the SVM models for their TPs, FPs, TNs, FNs, Sn, Sp, Ac and MCC. We also optimized the C and γ values for both the SVMs with 10-fold cross-validation. These values are listed in table 2. We obtained the best C and γ values of 8 and 2, respectively, for a positive set from the TarBase training data with a cross-validation accuracy of ~90%. For miRecords data as the training set, we observed the C and γ values of 2 and 2, respectively, with a cross-validation accuracy of ~95%. The best classification accuracy was observed for a window size of 20, where the best balance between specificities (0.98, 0.92) and sensitivities (0.92, 0.97) were also observed with the highest values for MCC (0.90, 0.797) for the miRecords-based model and TarBase-based model, respectively. For all the models, a window of 20 bases emerged as the one that could perform the best characterization based on position-specific variation in dinucleotide densities (figure 1). This study provides strong support for the intrinsic compositional role of flanking sequences around the target site. The dinucleotide profile variation patterns in the flanking regions could be useful in classifying the targets successfully. We extended our findings with these two separate SVM models over the targets predicted by the earlier steps and found positive predictions for a total of 11 genes out of 17 most likely potential targets screened during the previous steps. The results are detailed in table 3.

In order to assess the performance of our hypothesis further, we formed four different dataset models and carried out three different sets of performance tests. We also compared these with a recently published software to refine microRNA:target prediction, Mirtif (Yang *et al.* 2008). For every such model, we repeated the above-mentioned procedures for parameter estimation, cross-validation as well as formed ROCs and calculated their AUCs. MirTif uses the *k*-grams frequency information for nucleotide combinations from microRNA:target pairs for prediction. The authors claim ~82% accuracy and 0.86 AUC for their dataset. For model building, MirTif uses 195 positive instances from Tarbase and 38 negative instances. The authors have used the same data for the training set as well as the testing set. First, we tested the accuracy of MirTif against the positive instances in our Mirecords data, in which our Tarbase model had achieved 87% accuracy. On the positive instances of the MirRecords dataset, the accuracy of MirTif fell to 74% (table 4C). It should be noted that this dataset was unseen by MirTif as it was trained using Tarbase-positive data only. Thereafter, we assessed the Sn,

Table 1. The top 10 most significant features from the flanking regions of microRNA target sites found common in the two dataset models. The positive and negative instances were best discriminated by these common position-specific window densities of dinucleotides given here with their respective *F*-scores.

Feature# (Tarbase)	Dinucleotide	*F*-score	Feature# (MiRecords)	Dinucleotide	*F*-score
81	AA	0.73203	81	AA	0.25795
59	CG	0.21046	43	CG	0.20155
75	CG	0.19292	75	CG	0.15458
197	UA	0.19048	11	CG	0.15131
91	CG	0.13736	65	AA	0.13708
11	CG	0.12136	91	CG	0.13628
65	AA	0.10329	59	CG	0.11815
165	UA	0.09289	165	UA	0.11746
43	CG	0.08291	85	UA	0.10847
85	UA	0.07565	197	UA	0.09420

Table 2.

(A) Performance measure of the support vector machine (SVM) with the miRecords model for different window sizes. The best performance was observed for a window size of 20 bases.

Window size	C-value	Gamma	CV-fold	CV rate	Ac (%)	Sn	Sp	MCC
6	2	0.5	10	90.9	86.3	0.73	1	0.75
10	8	0.5	10	93.7	91	0.85	0.73	0.82
16	2	0.5	10	92.6	91.7	0.88	0.96	0.83
20	2	2	10	93.1	95.2	0.92	0.98	0.9
32	8	2	10	93.1	91.7	0.96	0.88	0.83

(B) Performance measure of SVM with TarBase model for different window sizes. Best performance was observed for window of size 20 bases.

Window size	C-value	Gamma	CV-fold	CV rate	Ac%	Sn	Sp	MCC
6	0.03	0.13	10	91.7	88.6	0.84	0.93	0.77
10	0.5	0.5	10	93.8	89.2	0.83	0.95	0.79
16	0.5	0.5	10	92.4	85.2	0.81	0.89	0.7
20	8	2	10	94.5	89.7	0.86	0.93	0.79
32	32	5	10	91.7	89.7	0.88	0.9	0.79

CV, cross validation; Ac accuracy; Sn sensitivity; Sp, specificity; MCC, Matthew correlation coefficient

Table 3. Predictions made by SVM on potential target genes selected through interaction pattern encoding match step over various pathways associated genes.

Gene	MicroRNA	Pathway	Support vector machine (SVM) (miRecords model)	SVM (TarBase model)
PRKY	hsa-miR-196a	Insulin	Absent	Absent
SOCS4	hsa-miR-196a	Insulin	Absent	Absent
RAF1	hsa-miR-125b	Insulin	Present	Present
PIK3R3	hsa-miR-132	Insulin	Present	Absent
PPP1CB	hsa-miR-145	Insulin	Absent	Absent
SOS1	hsa-miR-23a	Insulin	Present	Present
AKT3	hsa-miR-34a	Insulin	Absent	Absent
PRKAB1	hsa-miR-1	Insulin	Present	Present
RELA	hsa-miR-20a	TLR	Present	Present
TMEM189	hsa-miR-132	TLR	Present	Present
BCL2	hsa-miR-221	Apoptosis	Absent	Present
APAF1	hsa-miR-132	Apoptosis	Present	Present
PIK3R3	hsa-miR-132	Apoptosis	Present	Absent
PRKY	hsa-miR-196a	Apoptosis	Absent	Absent
ATM	hsa-miR-132	Apoptosis	Present	Present
CASP7	hsa-miR-23a	Apoptosis	Present	Present
IKBKG	hsa-miR-20a	Apoptosis	Present	Present

Sp, Ac, MCC, ROC and AUC for various combinations of test and training data. While doing so, we built models and test sets in which we appended the negative dataset from the MirTif-negative dataset in models C (Tarbase-based data + 32 experimentally validated negative instances) and D (MiRecords-based dataset + 32 experimentally validated negative instances). In models A (Tarbase-based dataset only) and B (MiRecords-based dataset only), we directly used the MirTif-negative data as the negative test set without using them in the training set. Details of the Mirecords-based and Tarbase-based datasets have already been discussed in the previous sections. In all the tests, our methodology

performed better than MirTif, with most of the performance measures achieving better scores. All the AUC values scored were above 0.90 and higher than the MirTif AUC (0.86), all Ac values scored were higher than the MirTif Ac (81.9%), models A and B had better Sns and better Sps among all the models (table 4A, B and C). The complete details of the models and respective tests are available in supplementary data 5. The ROCs for these tests are given in figure 2.

The observed fall in the Ac level of MirTif, when tested with positive cases from our Mirecords dataset, could be due to the fact that MirTif uses the same dataset for training as well as testing, where all positive instances have been derived from Tarbase. MirTif considers a skewed dataset, having a total of 195 positive instances against only 38 negative instances. In our models, in all the above-mentioned tests, we never used same testing and training data completely. All predictions were largely in instances never seen before, whereas the Tarbase-derived dataset models have been tested over datasets containing all positive instances from miRecords and vice versa. With fully unseen test data, our models performed better. In addition, we maintained a reasonable ratio of positive and negative instances in order to avoid problems created by class imbalance. The basic work plan of the entire methodology is illustrated in figure 3. The related details of the datasets can be found in supplementary data 1.

3.5 *Application: regulatory module analysis over the predicted targets*

More than fifteen years ago, microRNAs came into prominence for their possible negative regulatory roles, with the first discovered microRNA, lin-4, exhibiting complementarity with the *lin-14* gene and its downregulatory impact observed in *C. elegans* (Lee *et al.* 1993). Later, they were found to downregulate a number of genes (Lim *et al.* 2005). Together with transcription factors, microRNAs present a complex regulatory system (Hobert 2004). Therefore, it was interesting to look into the regulatory roles of these microRNAs along with the transcription factor partners associated with them.

We selected SVM-supported predictions for further analysis. For every potential target, we carried out further analysis by looking into co-expressed genes using the gene sorter tool (Kent *et al.* 2005) and considering gene expression data for expression-based distance calculation. The gene sorter at the University of California Santa Cruz (UCSC) provides a module to look at related genes, based on the Euclidean distance computed over expression profiles of related genes across various tissues and stages. Our interest was to evaluate other co-expressed genes as well, which could be the targets for the same microRNA predicted to target the associated target gene. Further, we also looked into the promoter regions of these genes so that we could find some possible regulatory circuits associated with these genes, their associated microRNAs and transcription factors.

All these genes were compared with their respective predicted target genes with which their expression proximity had been observed. We retained only those genes that had the same common microRNA targeting them along with the predicted associated gene with which they were co-expressed (table 5). Besides this, for every such gene which exhibited a common expression and microRNA, we looked into the 5′ promoter regions up to 2 kb upstream for the presence of common TFBS. Every such target and its co-expressed and co-targeted genes shared some common set of TFBS. For detailed and pair-wise comparisons between the genes regarding TFBS sites, please see supplementary data 6. Some interesting findings which deserve to be mentioned here are about two genes of the apoptosis pathway, APAF1 and ATM. We found them to be targets sharing a common microRNA, hsa-miR-132, and standing close to each other for expression similarity, sharing about eleven transcription factors in common (table 5). Another set of interesting findings is associated with the microRNA hsa-miR-23a,

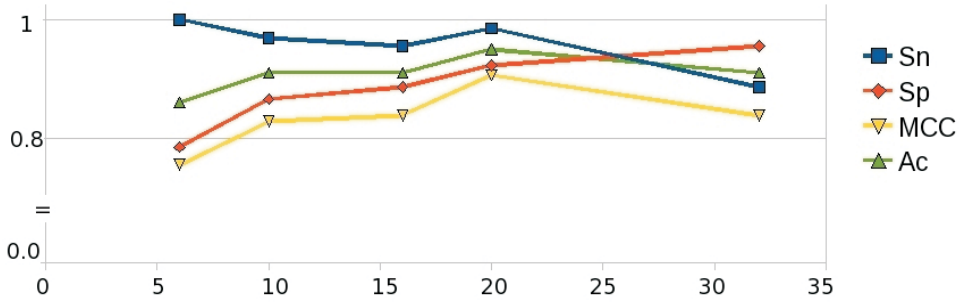

Figure 1. Performance measure of our support vector machine (SVM) model at different window sizes. The flanking regions around the target site were analysed for varying dinucleotide density profiles with varying distance as the SVM feature to discriminate between a true and a false target, with different window sizes. It was found that a window of 20 bases performs the best. Sp, specificity; Sn, sensitivity; MCC, Mathew correlation coefficient; Ac, accuracy.

Table 4. **(A)** Comprehensive performance test of our support vector machine (SVM) models. Details of these models and tests are available in supplementary data 5. Model A uses a 'Tarbase-based' dataset, B is 'MiRecords based', C is a Tarbase-based model where negative instances and positive instances are from Model A in addition to experimentally validated 32 negative instances. Similarly, Model D has an additional 32 experimentally validated negative instances added to the total negative instances in Model B

	CV rate	Ac%	TP	TN	FP	FN	Sn	Sp	MCC	AUC
Model A	94.52	88.46	76	108	12	12	0.86	0.91	0.78	0.94
Model B	93.18	89.33	64	95	10	9	0.88	0.90	0.75	0.98
Model C	93.26	86.54	65	115	5	23	0.74	0.96	0.73	0.9473
Model D	93.27	90.45	56	105	0	17	0.77	1	0.81	0.985

CV, cross validation; Ac, accuracy; Sn, sensitivity; Sp, specificity; TP, true positive; TN, true negative; FP, false positive; FN, false negative; MCC, Matthew correlation coefficient; AUC, area under the curve

(B) Average performance comparison between MirTif and our methodology. The average has been taken here after considering the values from the performances of all of the four support vector machine (SVM) models. The individual performances and respective accuracy, sensitivity, specificity and area under the curve (AUC) of our models are mostly higher as can been seen from (A)

	Average accuracy (Ac%)	Sensitivity (Sn)	Specificity (Sp)	Area under the curve (AUC)
MirTif	81.97	0.835	0.736	0.86
Our methodology	88.65	0.812	0.94	0.96

(C) Fall in accuracy observed in the MirTif prediction in positive instances from miRecords. MirTif has been trained and tested over the same set of data whose positive cases were taken from the Tarbase database

	Number of positive interactions	True positive	False positive	% Accuracy on miRecords positive data	% Accuracy reported in mirTif publication
miRecords	117	87	30	74	81.97

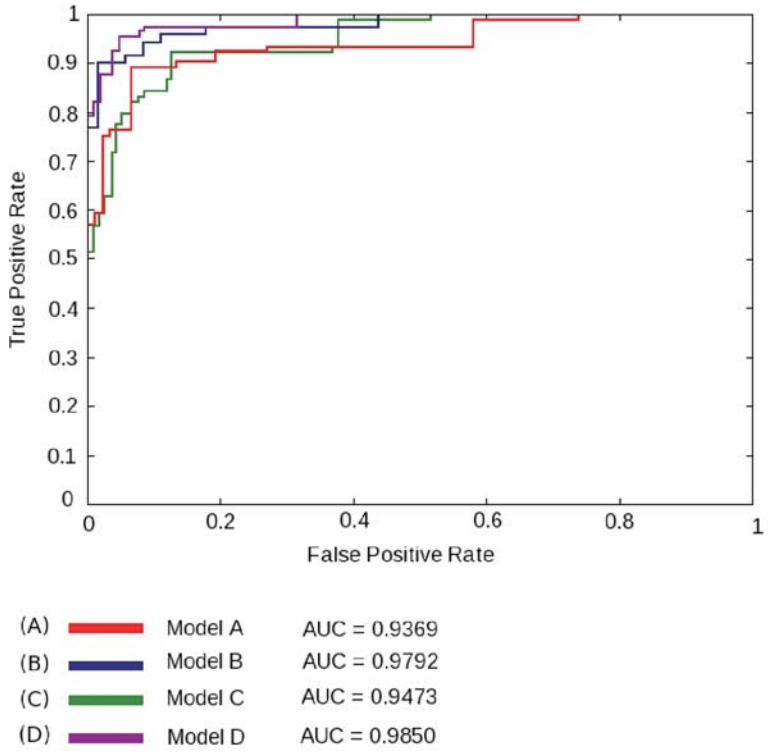

Figure 2. The receiver operating characteristic (ROC) plots for performance measure. Four different combinations of datasets were used to measure the performance of our support vector machine (SVM) models. Details about the models are available in supplementary data 5. For all the models the measured performance was significantly higher than MirTif (Yang *et al*. 2006). AUC, area under the curve.

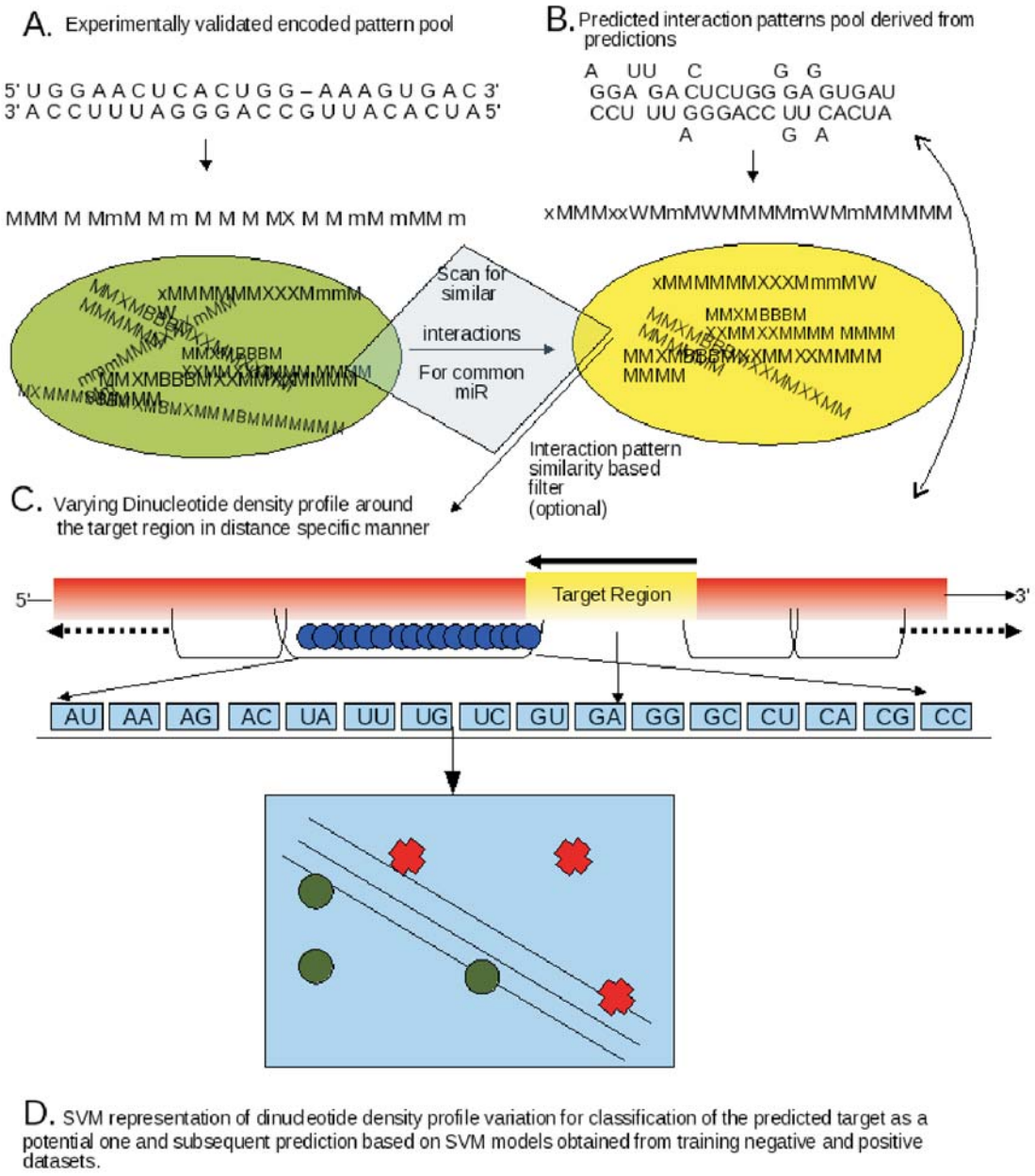

Figure 3. The basic work plan. The flow of the entire methodology practised here to predict most probable targets out of predictions made by target-finding softwares. The initial step is sorting based on similarity with experimentally known interactions. The sorted candidates are subjected to flank analysis and support vector machine (SVM)-based final prediction which utilizes the distance-based dinucleotide density variation profile in the flanking regions to characterize a possible target.

which targets the CASP7 gene and transcription factor vitamin D receptor (VDR), whose site is present in CASP7 as well as in all of its five co-expressed and co-targeted genes (HMGN2, NFX1, KIAA0323, JMJD2A and TWA1). This is a significant observation as CASP7 is a critical component of apoptosis and T-cell signalling (Korfali *et al.* 2004). Since VDR is common to all these genes, which are co-expressed and co-targeted with CASP7 and hsa-mir23a, respectively, a complete regulatory circuit involving hsa-miR-23a, VDR, CASP7 and these co-expressed genes emerges. HMGN2 is involved in chromatin structure and transcription, NFX-1 has transcription factor activity (Song *et al.* 1994), KIAA0323 is a hypothetical protein, JMJD2A is a histone demethylase transcription repressor (Zhang *et al.* 2005), while TWA1 is a nuclear protein, which is supposed to be critical in microtubule formation and cell division (Umeda *et al.* 2003).

Table 5. Regulatory analysis of co-expressed, co-targeted and co-regulated genes. The predicted target genes were analysed for co-expressed genes which share same microRNA targets and were studied for their enriched transcription factor-binding sites

Gene	Associated genes	MicroRNA	Associated transcription factor-binding sites
IKBKG	*ARHGEF18*	hsa-miR-20a	CACD; DBP; Pax-8; RFX; Tal-1beta:E47
PIK3R3	*CAMSAP1L1*	hsa-miR-132	AhR; CACD; C/EBPalpha; c-Ets-1(p54); Ets; GATA-X; Oct-1; OG-2; RFX1; Spz1
ATM	*APAF1*	hsa-miR-132	c-Ets-1; c-Ets-1(p54); DBP; Ets; Oct-1; Pax-8; Spz1; Tel-2; v-Myb; ZF5
CASP7	*HMGN2,NFX1,KIAA0323, JMJD2A, C20orf11*	hsa-miR-23a	BRCA1:USF2, C/EBP, CKROX, CREB, ETF, KROX, LRF, Spz1, TBX5,ZF5 (3); AP-2, CACD, DBP, MAZ, Oct-1, Pax-8,RFX,RFX1,Sp1,YY1(4); c-Ets-1, Ets; GARP, v-Myb(5); VDR3(6);
APAF1	*ATM, CNOT2*	hsa-miR-132	ATF6, GABP, Oct-1, Octamer, Pax-8, v-Myb
PRKAB1	*IMPACT, ASH2L,TBC1D15*	hsa-miR-1	AP-2alpha, C/EBPalpha, FAC1, GATA-X, Hand1:E47, MYB, Nkx2-5, Oct-1, OG-2, RUSH-1alpha, ZF5 (2); BRCA:USF2, C/EBP, RFX, TBX5, v-Myb (3); CACD, c-Ets-1, Ets, Pax-8 (4);
RELA	*MAP3K14, RNH1*	hsa-miR-20a	AhR:Arnt; AP-2; c-Ets-1; ETF; KROX; LRF; Nkx2-5; Oct-1; OG-2; RFX; Sp1; TBX5 (2); CACD; C/EBP; C/EBPalpha; Hand1:E47; kid3; Sp1(3)
TMEM189	*VAPA*	hsa-miR-132	AP-2; C/EBP; Kid3; NF-Y; Pax-8; SREBP; v-Myb
BCL2	*TET2, YTHDC1, VEZF1*	hsa-miR-221	AP-2; CdxA; C/EBPalpha; ER; FAC1; Oct-1; Sp1; TBX5(4); AP-2alpha; CACD; c-Ets-1; CREB; Ebox; KROX; LRF; OG-2; Pax-8; RFX; SREBP; ZF5(3); Hand1:E47(5)
SOS1	*WHAM,MCM3AP, NCOA6, WDR37*	hsa-miR-23a	AP-2; BRCA1:USF2; CKROX; FAC1; Hand1:E47; Oct-1 CACD; Kid3; MAZ; v-Myb CdxA; c-Ets-1; RFX; RFX1; Spz1; VDR; WT1

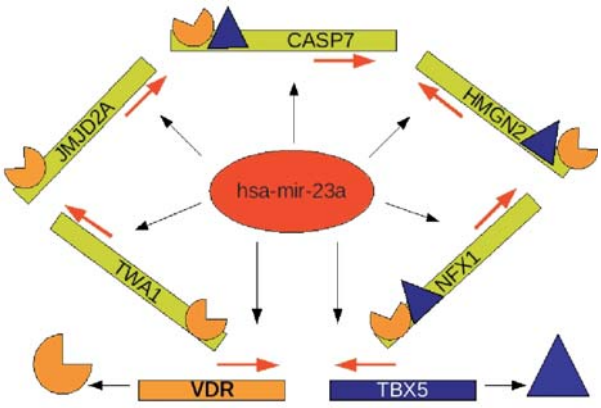

Figure 4. The proposed hsa-miR-23a regulatory module. It was found that hsa-miR-23a could be critical in regulating apoptosis by regulating these five genes which are co-expressed together as well as co-targeted by hsa-miR-23a and share many transcription factor-binding sites (TFBS) in common, of which *VDR* and *TBX5* appear to be the most critical as they too were found to have an hsa-miR-23a target site.

Interestingly, most of these genes are nuclear factors and in some way associated with cell growth and death. Another transcription factor, TBX5, was found to be common with two co-expressed and co-targeted genes (NFX1 and HMGN2) of CASP7 as well as a target for the microRNA hsa-miR-23a. TBX5 was found to be common in many of the predicted target genes associated with apoptosis reported in this study. TBX5 protein is a T-box transcription factor with a repressor role in transcription control of the cell cycle and plays a very important role in cell growth (He *et al.* 2002).

What we have reported here is supported by an experimental study with hsa-miR-23a, where it was found that after an antagomir inhibition of hsa-miR-23 in lung carcinoma cells, A549, downregulation of cell growth was noticed (Cheng *et al.* 2005). With the CASP7 system, these observations strongly indicate that hsa-miR-23a is a critical component in cell growth and death (figure 4). In this study as well, hsa-miR-23a was found to target SOS1 and associated genes (WHAM, MCM3AP, NCOA6, WDR37).

4. Conclusion

We have presented a novel methodology to refine microRNA target prediction, using dinucleotide density profile variation in the flanking regions of target sites. As observed in some recent work, the role of the target flanking regions in controlling microRNA targeting could be significant (Grimson *et al*. 2007; Kertesz *et al*. 2007;

Didiano and Hobert 2008). We found that the sequence-based intrinsic features derived from these regions, when represented as the varying dinculeotide density profiles of the flanking regions around the target sites with the help of the SVM learning approach, were able to successfully discriminate between the positive and negative instances. MicroRNAs are important components of the regulatory system, which work in a concerted fashion along with transcription factors. Finding the correct targets for these components of the regulatory system may shed light on the behaviour of various pathways and genes at different levels. We concluded this study by looking into the regulatory aspects of our findings in association with co-expressed and co-targeted genes sharing common transcription factor sites, and identified some regulatory modules. The procedure described here could be useful in narrowing the predicted targets and can be easily extended to other species. Based on the methodology presented in this article, some novel software can also be developed.

Acknowledgements

We thank Mr Amit Chaurasia and Dr Mitali Mukerji of Institute of Genomics and Integrative Biology, Delhi, for sharing TFBS data on human sequences using the TP database. We thank Ms Shivani Pareek and Mr Bhavesh Kataria for helping with this study.

References

Akbani R, Kwek S and Japkowicz N 2004 Applying support vector machines to imbalanced datasets; in *Proceedings of the 15th ECML* (Italy: Springer)

Ambros V, Bartel B, Bartel D P, Burge C B, Carrington J C, Chen X, Dreyfuss G, Eddy S R *et al*. 2003 A uniform system for microRNA annotation; *RNA* **9** 277–279

Andronescu M, Zhang Z C and Condon A 2005 Secondary structure prediction of interacting RNA molecules; *J. Mol. Biol*. **4** 987–1001

Brennecke J, Stark A, Russell RB and Cohen S M 2005 Principles of microRNA-target recognition; *PLoS Biol*. **3** e85

Chang C and Lin C 2001 LIBSVM: a library for support vector machines *http://www.csie.ntu.edu.tw/~cjlin/libsvm*

Cheng A M, Byrom M W, Shelton J and Ford L P 2005 Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis; *Nucleic Acids Res*. **33** 1290–1297

Cullen B R 2004 Transcription and processing of human microRNA precursors; *Mol. Cell* **16** 861–865

Didiano D and Hobert O 2008 Molecular architecture of a miRNA-regulated 3′ UTR; *RNA* **14** 1297–1317

Doench J G and Sharp P A 2004 Specificity of microRNA target selection in translational repression; *Genes Dev*. **18** 504–511

Drucker H, Burges C, Kaufman L, Smola A and Vapnik V 1997 Support vector regression machines; *Adv. Neural Inf. Processing Syst.* **9** 155–161

Gardner P P and Giegerich R 2004 A comprehensive comparison of comparative RNA structure prediction approaches; *BMC Bioinformatics* **5** 140

Griffiths-Jones S 2004 The microRNA registry; *Nucleic Acids Res*. **32** D109–D111

Griffiths-Jones S, Grocock R J, Van D S, Bateman A and Enright A J 2006 miRBase: microRNA sequences, targets and gene nomenclature; *Nucleic Acids Res*. **34** D140–D144

Griffiths-Jones S, Saini H K, Dongen S and Enright A J 2008 miRBase: tools for microRNA genomics; *Nucleic Acids Res*. **36** D154–D158

Grimson A, Farh K K, Johnston W K, Garrett-Engele P, Lim L P and Bartel D P 2007 MicroRNA targeting specificity in mammals: determinants beyond seed pairing; *Mol. Cell* **6** 91–105

Hammell M, Long D, Zhang L, Lee A, Carmack C S, Han M, Ding Y and Ambros V 2008 mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts; *Nat. Methods* **5** 813–819

He M L, Chen Y, Peng Y, Jin D, Du D, Wu J, Lu P and Lin M C 2002 Induction of apoptosis and inhibition of cell growth by developmental regulator hTBX5; *Biochem. Biophys. Res. Commun*. **297** 185–192

Hobert O 2004 Common logic of transcription factor and microRNA action; *Trends Biochem. Sci.* **29** 462–468

Höchsmann M, Toller T, Giegerich R and Kurtz S 2003 Local similarity in RNA secondary structures; *Proceedings of the IEEE Bioinformatics Conference CSB-2003* (California, USA: Stanford) pp 159–168

Hofacker I L and Stadler P F 2006 Memory efficient folding algorithms for circular RNA secondary structures; *Bioinformatics* **22** 1172–1176

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T and Cuff J 2002 The Ensembl genome database project; *Nucleic Acids Res*. **30** 38–41

Kanehisa M and Goto S 2000 KEGG: Kyoto encyclopedia of genes and genomes; *Nucleic Acids Res.* **28** 27–30

Kel A E, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis O V and Wingender E 2003 MATCH: a tool for searching transcription factor binding sites in DNA sequences; *Nucleic Acids Res.* **31** 3576–3579

Kent W J, Hsu F, Karolchik D, Kuhn R M, Clawson H, Trumbower H and Haussler D 2005 Exploring relationships and mining data with the UCSC gene sorter; *Genome Res*. **15** 737–741

Kertesz M, Iovino N, Unnerstall U, Gaul U and Segal E 2007 The role of site accessibility in microRNA target recognition; *Nat. Genet*. **39** 1278–1284

Kim S K, Nam J W, Rhee J K, Lee W J and Zhang B T 2006 miTarget: microRNA target gene prediction using a support vector machine; *BMC Bioinformatics* **7** 411

Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z and Hatzigeorgiou A 2004 A combined computational-experimental approach predicts human microRNA targets; *Genes Dev*. **18** 1165–1178

Korfali N, Ruchaud S, Loegering D, Bernard D, Dingwall C, Kaufmann S H and Earnshaw W C 2004 Caspase-7 gene disruption reveals an involvement of the enzyme during the early stages of apoptosis; *J. Biol. Chem*. **279** 1030–1039

Lai E C 2002 Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation; *Nat. Genet*. **30** 363–374

Lai E C 2004 Predicting and validating microRNA targets; *Genome Biol*. **5** 115

Lee R C, Feinbaum R L and Ambros V 1993 The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14; *Cell* **75** 843–854

Lewis B P, Burge C B and Bartel D P 2005 Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets; *Cell* **120** 15–20

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

Lim L P, Lau N C, Garrett-Engele P, Grimson A, Schelter J M, Castle J, Bartel D P and Linsley P S *et al.* 2005 Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs; *Nature (London)* **433** 769–773

Long D, Lee R, Williams P, Chan C Y, Ambros V and Ding Y 2007 Potent effect of target structure on microRNA function; *Nat. Struct. Mol. Biol*. **14** 287–294

Miller A A and Waterhouse P 2005 Plant and animal microRNAs: similarities and differences; *Funct. Integr. Genomics* **5** 129–135

Rehmsmeier M, Steffen P, Hochsmann M and Giegerich R 2004 Fast and effective prediction of microRNA/target duplexes; *RNA* **10** 1507–1517

Robins H, Li Y and Padgett R 2005 Incorporating structure to predict microRNA targets; *Proc. Natl. Acad. Sci. USA* **102** 4006–4009

Sethupathy P, Corda B and Hatzigeorgiou A G 2006 TarBase: a comprehensive database of experimentally supported animal microRNA targets; *RNA* **12** 192–197

Shankar R, Chaurasia A, Ghosh B, Chekmenev D, Cheremushkin E, Kel A and Mukerji M 2007 Non-random genomic divergence in repetitive sequences of human and chimpanzee in genes of different functional categories; *Mol. Genet. Genomics* **277** 441–455

Song Z, Krishna S, Thanos D, Strominger J L and Ono S J 1994 A novel cysteine-rich sequence-specific DNA-binding protein interacts with the conserved X-box motif of the human major histocompatibility complex class II genes via a repeated Cys-His domain and functions as a transcriptional repressor; *J. Exp. Med*. **180** 1763–1774

Thadani R and Tammi M T 2006 MicroTar: predicting microRNA targets from RNA duplexes; *BMC Bioinformatics* **18** 7

Thompson W, Rouchka E C and Lawrence C E 2003 Gibbs recursive sampler: finding transcription factor binding sites; *Nucleic Acids Res*. **31** 3580–3585

Thompson W, Palumbo M J, Wasserman W W, Liu J S and Lawrence C E 2004 Decoding human regulatory circuits; *Genome Res*. **14** 1967–1974

Umeda M, Nishitani H and Nishimoto T 2003 A novel nuclear protein, Twa1, and Muskelin comprise a complex with RanBPM; *Gene* **303** 47–54

Wang X and El Naqa I M 2008 Prediction of both conserved and nonconserved microRNA targets in animals; *Bioinformatics* **24** 325–332

Will S, Reiche K, Hofacker I L, Stadler P F and Backofen R 2007 Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering; *PLoS Comput. Biol*. **3** 4

Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prüss M *et al.* 2000 TRANSFAC: an integrated system for gene expression regulation; *Nucleic Acids Res*. **28** 316–319

Xiao F, Zuo Z, Cai G, Kang S, Gao X and Li T 2009 miRecords: an integrated resource for microRNA-target interactions; *Nucleic Acids Res*. **37** D105–D110

Yang Y, Wang Y and Li K 2008 MirTif: a support vector machine-based microRNA target interaction filter; *BMC Bioinformatics* **9** S4

Zhang D, Yoon H G and Wong J 2005 JMJD2A is a novel N-CoR-interacting protein and is involved in repression of the human transcription factor achaete scute-like homologue 2 (ASCL2/Hash2); *Mol. Cell. Biol*. **25** 6404–6414

MS received 30 September 2009; accepted 14 December 2009

*e*Publication: 23 February 2010

Corresponding editor: Seyed E Hasnain

Flanking region sequence information to refine microRNA target predictions

Russiachand Heikham[†] and Ravi Shankar[*†]

J. Biosci. **35(1), March 2010, 105–118 © Indian Academy of Sciences**

Supplementary Material

The complete supplementary data have been made available online in zipped file format at: *http://ravishankar.150m.com/ShankarSupplementaryData_zip*. While downloading, please save it as *ShankarSupplementaryData.zip* file in order to successfully open it.

Supplementary data 1	Various datasets used in this analysis
Supplementary data 2	Encoded patterns of various microRNAs used in this study and their groupings based on similarities with the maximum allowed difference of two mismatches
Supplementary data 3	Results of a similarity search between predicted encoded patterns and experimental encoded patterns for genes of the insulin, apoptosis and toll-like receptor (TLR) pathways
Supplementary data 4	Openness estimation data from the target site secondary structure and random structures. Information on target sites being affected due to splice variant transcripts is also included.
Supplementary data 5	Performance testing of our methodology and comparison with the performance of MirTif
Supplementary data 6	Regulatory analysis of predicted target genes and their co-expressed and co-targeted partner genes