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.

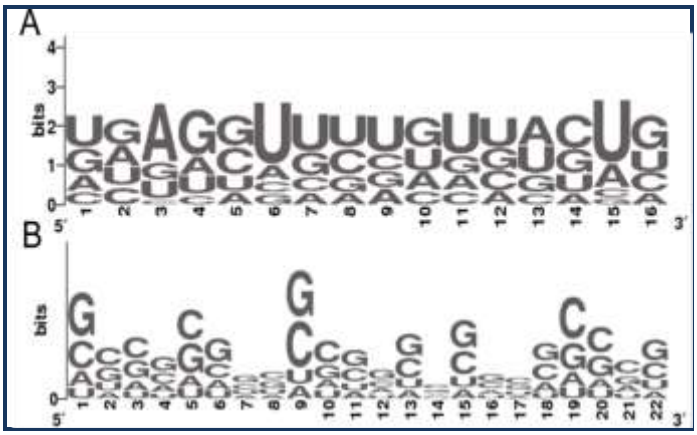

Figure 1: Sequence Conservation in miRs associated with cancer **(A)** and in the negative dataset **(B).**

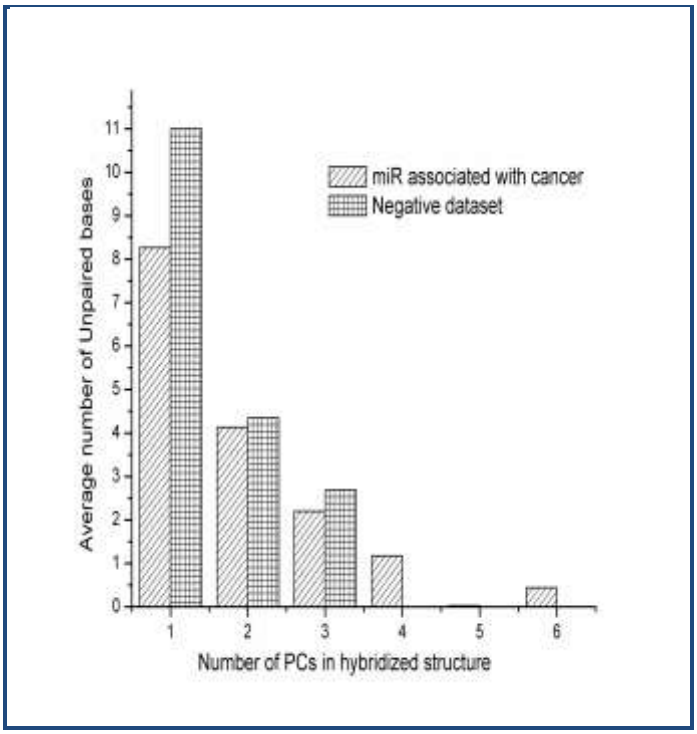

Figure 2: Variation in number of unpaired bases in miRs associated with cancer and the negative dataset. The first pair of bars stands for the variation in the hybrids having a single patch of complementarity (PC), the second for hybrids having two patches, and so on.

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.

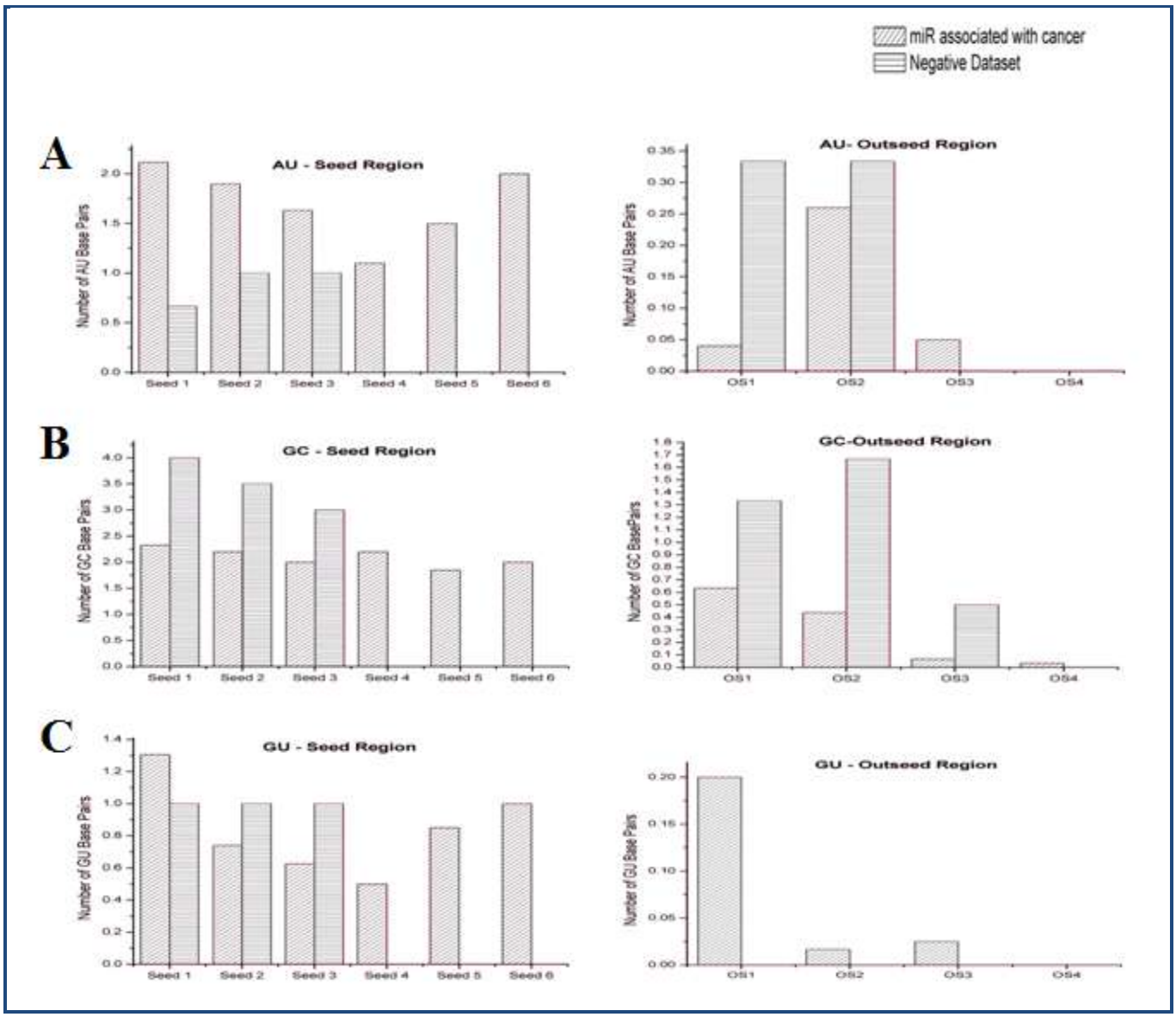

Figure 3: Distribution of the regular Watson – Crick (WC) and the non-WC base pairings between miR associated with cancer and the negative dataset. The panels on the left are for the pairings in the seed region, while the panels on the right are pairings in the regions outside the seed (OS).

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] Orom 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)