Sequence trademarks in oncogene associated microRNAs

Sumit Sharma & Sumit Biswas*

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

Abstract:
The last decade was taken by storm when the existence of a class of small (~22nt long) non – coding RNA species, known as microRNAs (miRNAs) came into light. MicroRNAs are one of the most abundant groups of regulatory genes in multicellular organisms and play fundamental roles in many cellular processes. Among these, miRNAs have been shown to prevent cell division and drive terminal differentiation, thus playing a causal role in the generation or maintenance of cancerous tumours. The unique expression profiles of different miRNAs in various types and stages of cancer suggest their performance as novel biomarkers. This discussion focuses on miRNAs implicated in cancer-associated events and strives to re-establish their sequential features which may classify them to be oncogenic.

Keywords: microRNA, oncogene, sequence trademarks, purines

Background:

MicroRNAs (miRNAs) are RNA molecules, on average only 22 nucleotides, which act as post-transcriptional regulators by binding to complementary sequences on target messenger RNA transcripts. First discovered while studying mutations that changed the timing of developmental events in *Caenorhabditis elegans* [1], they have now been identified in a variety of organisms [2] and can be found both in the plant and animal kingdoms. MicroRNAs are transcribed as parts of longer molecules (~ several kilobases) that are processed in the nucleus and form hairpin RNAs of 70–100 nt with the help of Drosha, the double-stranded RNA-specific ribonuclease [3, 4]. These hairpins are then transported to the cytoplasm, via an exportin 5-dependent mechanism, where they act as substrates for a second double-strand specific ribonuclease called Dicer [3, 5]. The single-stranded miRNA binds mostly to the 3' Untranslated Region (UTR) of the specific messenger RNA. Initial estimates put the number of microRNA genes in the mammalian genome in the range of 200 to 1000 (~1%–3% of known genes are represented by microRNAs). However, to this day, the number of microRNAs, including those electronically cloned, is over 1000 and still growing [6-8]. The majority of microRNAs (70%) are located in introns and/or exons, and approximately 30% are located in intergenic regions [9]. Analysis of the genomic location of human microRNA genes suggested that 50% of microRNA genes are located in cancer-associated genomic regions or in fragile sites, on chromosomes which are frequently deleted, amplified or rearranged in case of cancer development [10-12]. Altered expression pattern (over- or under-expression) of specific miRNAs has also been reported in tissues derived from various tumours [13, 14]. Changes in miRNA expression patterns have also been identified in abnormal cell proliferation, which is a hallmark of human cancers. The role of let-7 in lung cancer development, miR – 145 in breast cancer tissues, and the 28 different miRNAs in colonic adenocarcinoma, can all be cited as instances [14]. Some miRNAs have an increased expression in tumors (may be considered as oncogenes), others can also have decreased expression in cancerous cells (tumor suppressor genes). The oncogenic miRNAs are known as oncomirs and promote tumor development by negatively inhibiting tumor suppressor genes and/or genes that control cell differentiation or apoptosis. The present study was aimed at having a closer look into the sequential trademarks that may signify a miRNA to associate with cancer. Although there have been studies on prediction of oncogene regulation by miRNAs and other validation studies [12, 15, 16], this discussion and review focuses solely on the identification of sequential distinction present in the human oncogenically – involved miRNAs.

Approach and Findings:

A database of oncogenically involved miRNAs (oncomirs) was constructed using the entries listed in the methods of Koscianski *et al.* (2007) [12]. These miRNAs were selected from *Homo sapiens* sequences in the "miRNA Registry" Release 7.1 [17]. Koscianski *et al.* (2007) had listed these entries after scanning and validation against possible oncogenic sequences derived from the human 3'UTR database. We further culled these sequences using the identity criterion and non-redundant sequences were thus collected. The resulting sequences were sorted under two different categories, viz., hsa miRNA and hsa let. Thus, a total of 135 miRNAs could be sorted into two subgroups of 122 (hsa miRNA) and 13 (hsa let). While the longest sequence thus encountered was 24 for a hsa miRNA entry, the shortest one was of 18 nucleotides for a hsa miRNA. Since microRNAs have very short sequences, and sequence similarity is very high in miRNA families, we checked the rejected substitution score by using GERP conservation (miReval). Only three of the miRNA entries showed significant conservation – of which two were from hsa miRNA and one from hsa let. Hence, the final database was narrowed down to 132 miRNAs (120 hsa miRNA and 12 hsa let).
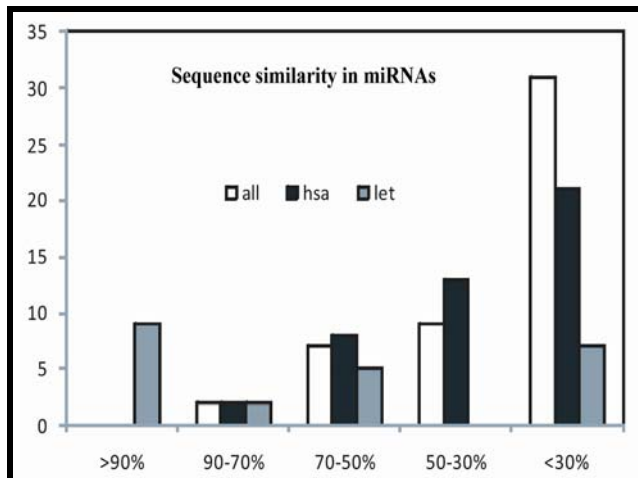
BIOINFORMATION

Figure 1: Plot depicting the sequence similarity of has miR and hsa let miRNAs against the frequency of their occurrences.

The key software used for this analysis was MATLAB. Multiple sequence alignment of the sequences was done using 'multialign' function in MATLAB with 'ExistingGapAdjust' option. It helps to control automatic adjustment based on existing gaps by introducing hyphens in between sequences so that there is maximum match at positions of sequences. From the resulting multiple sequence alignment, sequence trademarks and specific attributes were searched for. Finally the percentage match at all the positions in the sequences was calculated. A plot of the percentage sequence similarities versus the frequency in numbers **(Figure 1)** shows that the two classes of miRNAs more-or-less exhibit some sequence similarity but are not exactly so. While most of the hsa let family miRNAs have a very high percentage similarity (8 out of 12), the hsa miR family are only moderately similar in sequence with the biggest peak in the 30% similarity range.

What discriminates an oncogene-associated miRNA?
From the available results, some sequential features in miRNAs are evident. The first remarkable feature is the rarity of Cytosines among the four bases. On the contrary, Uracil seems to be the nucleotide of choice. In general, we could conclude that purines occur more frequently than pyrimidines in both families of oncogenic microRNAs. Considering the flexibility of RNA structure in general, the abundance of the bulkier purines is something which needs further investigation. Considering sequence similarities in the more similar hsa let family, it was found that a "gaugau" motif was common to all oncomirs of that family. In the more diverged hsa miRNA family, the common motif was altered at the second base to be represented as "ggu-ggu". To this effect, we concluded that human oncomirs can be identified by the presence of a sequence trademark of having a 'gg(/a)u-gg(/a)u motif. The available result could be used for further identification and validation of oncomirs, and the authors are in the process of developing an algorithm to that effect, based on the sequence trademarks.

References:
[1] Lee RC *et al*. *Cell* 1993 **75**: 843 [PMID: 8252621]
[2] Bartel DP. *Cell* 2004 **116**: 281 [PMID: 14744438]
[3] Lee Y *et al. Nature* 2003 **425**: 415 [PMID: 14508493]
[4] Cullen BR. *Mol Cell* 2004 **16**: 861 [PMID: 15610730]
[5] Bernstein E *et al. Nat Genet*. 2003 **35**: 215 [PMID: 14528307]
[6] Berezikov E *et al. Cell* 2005 **120**: 21 [PMID: 15652478]
[7] Friedman RC *et al. Genome Res.* 2009 **19**: 92 [PMID: 18955434]
[8] Xie X *et al. Nature* 2005 **434**: 338 [PMID: 15735639]
[9] Rodriguez A *et al. Genome Res*. 2004 **14**: 1902 [PMID: 15364901]
[10] Sevignani C *et al. Mamm Genome.* 2006 **17**: 189 [PMID: 16518686]
[11] Calin GA *et al. Proc Natl Acad Sci U S A*. 2004 **101**: 2999 [PMID: 14973191]
[12] Koscianska E *et al. BMC Mol Biol*. 2007 **8**: 79 [PMID: 17877811]
[13] He L *et al. Nature* 2005 **435**: 828 [PMID: 15944707]
[14] Lu J *et al. Nature* 2005 **435**: 834 [PMID: 15944708]
[15] Lindow M. *Methods Mol Biol.* 2011 **703**: 311 [PMID: 21125499]
[16] Oulas A *et al. Methods Mol Biol*. 2011 **676**: 23 [PMID: 20931387]
[17] Griffiths-Jones S. *Nucleic Acids Res*. 2004 **32**: D109 [PMID: 14681370]