Quantification of miRNAs and Their Networks in the light of Integral Value Transformations

Sk. Sarif Hassan[1], Pabitra Pal Choudhury[1], Arunava Goswami[2], Navonil De Sarkar[3], Vrushali Fangal[1]

[1]Applied Statistics Unit, [2]Biological Sciences Division and [3]Human Genetics Unit,

Indian Statistical Institute, Calcutta-700108, India

Emails: sarimif@isical.ac.in, pabitra@isical.ac.in, agoswami@isical.ac.in & navonil.de@gmail.com, vrush2jlu@gmail.com

Abstract

MicroRNAs (miRNAs) which are on average only 21-25 nucleotides long are key post-transcriptional regulators of gene expression in metazoans and plants. *A proper quantitative understanding of miRNAs is required to comprehend their structures, functions, evolutions etc. In this paper, the nucleotide strings of miRNAs of three organisms namely Homo sapiens (hsa), Macaca mulatta (mml) and Pan troglodytes (ptr) have been quantified and classified based on some characterizing features. A network has been built up among the miRNAs for these three organisms through a class of discrete transformations namely Integral Value Transformations (IVTs), proposed by Sk. S. Hassan et al [1, 2]. Through this study we have been able to nullify or justify one given nucleotide string as a miRNA. This study will help us to recognize a given nucleotide string as a probable miRNA, without the requirement of any conventional biological experiment. This method can be amalgamated with the existing analysis pipelines, for small RNA sequencing data (designed for finding novel miRNA). This method would provide more confidence and would make the current analysis pipeline more efficient in predicting the probable candidates of miRNA for biological validation and filter out the improbable candidates.*

Keywords: *miRNAs, Integral Value Transformations (IVTs), Fractal dimension, Hurst Exponent, Mean Order.*

1. Introduction

Mature microRNAs (miRNAs) are a class of naturally occurring, small non-coding RNA molecules, of length about 21–25 nucleotides [3]. Mature microRNAs are partially or completely complementary to one or more messenger RNA (mRNA) molecules. MiRNA binds to target, mostly at 3’ UTR region in the metazoan although 2 exceptions have been reported recently [4]. In plants targets can be located in the 3' UTR but mostly in the coding region. MiRNA’s seed region’s perfect complimentarity with the target region is an essential issue for successful target binding, though a recent study has shown that alternative is also possible [5]. The primary function of miRNAs is to down-regulate gene expression in a variety of manners, including translational repression, mRNA cleavage, and deadenylation. They were first discovered in 1993 by Lee and colleagues, and the term microRNA was coined in 2001 [3]. Thousands of miRNAs genes found in intergenic regions or in anti-sense orientation to genes have since been identified in various organisms through random cloning and sequencing or computational prediction. Around 40% of miRNA genes lie in the introns of protein and non-protein coding genes or even in exons [6, 7, 8]. It is to be noted that several miRNAs have been reported to have links with certain types of cancer. Genetic signatures of miRNAs are required for individualized cancer treatment strategies as reported by C. Hatzis et al [9, 10]. A proper quantitative understanding of miRNAs is required to comprehend their structures, functions, evolutions etc. which will also help us to recognize a given string as a miRNA. In this paper nine such mathematical features of miRNA strings have been adumbrated and a primary classification is made for the miRNAs of three organisms viz. *hsa (Homo sapience), mml (Macaca mullata)* and *ptr (Pan troglodytes).* A network has been established among the miRNAs through Integral Value Transformations (IVTs) and validated through those nine features. Finally seven examples of strings are given, six of which are probable and one is non-probable as per quantitative studies.

Footnotes

- [1]To whom correspondence should be addressed. E-mail: sarimif@isical.ac.in
- *Author contributions*: Sk. Sarif Hassan and Navonil De Sarkar conceptualized the experiments and performed entire research with P. Pal Choudhury, A. Goswami and Vrushali Fangal made the necessary computer programs.
- Conflict of interest statement: The authors declare no conflict of interest.

1. Reviews and Fundamentals

In this section, some basics of *Fractal* and *Integral Value Transformations* (*IVTs*) have been discussed to warm up.

2.1 Basics on Fractal and Fractal Dimension

Our artificial world can be described easily through Euclidean geometric shapes but there are many things in nature such as shape of cloud, geometry of lightening etc. could not be described through Euclidean geometry. Many mathematicians descended the challenge for a fair enough description of natural objects but after a long period in 1975, B. Mandelbrot took the challenge and gave the birth of a new geometry to describe nature which is known to us as 'Fractal Geometry' in short 'Fractal'. The precise definition of "Fractal" according to Benoit Mandelbrot is as a set for which the Hausdroff Besicovitch dimension strictly exceeds the topological dimension [11]. To gain a quantitative insight of Fractal, some fractal parameters namely Fractal dimension, Hurst exponent, succolarity, lacunarity etc. are also introduced in the literature. A brief discussion follows about one of the well-known methods of calculating fractal dimension namely '*Box-Counting method'*.

Box-Counting Method: This method computes the number of cells required to entirely cover an object, with grids of cells of varying size. Practically, this is performed by superimposing regular grids over an object and by counting the number of occupied cells. The logarithm of N(r), the number of occupied cells, versus the logarithm of 1/r, where r is the size of one cell, gives a line whose gradient corresponds to the box dimension [12].

2.2 Notion of Integral Value Transformations (IVTs)

Let us define the Integral Value Transformations (IVTs) [1, 2] in $\mathbb{N}_0^{\mathrm{K}}$as the following:

$IVT^{p,k}{}_j : \mathbb{N}_0^{\mathrm{K}} \rightarrow \mathbb{N}_0$
$IVT^{p,k}{}_j((n_1, n_2, \ldots . n_k)$= $(f_j(a_0{}^{n_1}, a_0{}^{n_2}, \ldots, a_0{}^{n_k}) f_j(a_1{}^{n_1}, a_1{}^{n_2}, \ldots, a_1{}^{n_k}) \ldots \ldots f_j(a_{l-1}{}^{n_1}, a_{l-1}{}^{n_2}, \ldots, a_{l-1}{}^{n_k}))_p = m$
where n_1= $(a_0{}^{n_1} a_1{}^{n_1} \ldots . a_{l-1}{}^{n_1})_p$, n_2= $(a_0{}^{n_2} a_1{}^{n_2} \ldots . a_{l-1}{}^{n_2})_p$,……$n_k$= $(a_0{}^{n_k} a_1{}^{n_k} \ldots . a_{l-1}{}^{n_k})_p$
f_j: $\{0, 1, 2, \ldots, p-1\}^k \rightarrow \{0,1,2, \ldots, p-1\}$.
m is the decimal conversion from the p adic number.

Let us fix the domain of IVTs as $\mathbb{N}_0$ (k=1) and thus the above definition boils down to the following:

$$\mathrm{IVT}^{\mathrm{p},1}{}_{\mathrm{j}}(\mathrm{x}) = \left(\mathrm{f_j}(\mathrm{x_n})\mathrm{f_j}(\mathrm{x_{n-1}}) \ldots \ldots \ldots \mathrm{f_j}(\mathrm{x_1})\right)_{\mathrm{p}} = \mathrm{m}$$

where m is the decimal conversion from the p adic number, and $\mathrm{x} = (\mathrm{x_n}\ \mathrm{x_{n-1}} \ldots \ldots \mathrm{x_1})_{\mathrm{p}}$.
Let us define precisely and particularly the IVT in $\mathbb{N}_0$ in 4-adic number systems. $\mathrm{IVT}^{4,1}{}_{\#}$is mapping a non-negative integer to a non-negative integer [3].

$$\mathrm{IVT}^{4,1}{}_{\#}(\mathrm{a}) = \left((\mathrm{f}_{\#}(\mathrm{a_n})\mathrm{f}_{\#}(\mathrm{a_{n-1}}) \ldots \mathrm{f}_{\#}(\mathrm{a_1})\right)_4 = b$$

Where 'a' is a non-negative integer and $a = (a_n a_{n-1} \ldots a_1)_4$ and 'b' is the decimal value corresponding to the 4-adic number.

For an example, let us consider $a = 225 = (3201)_4$ and $\# = 120$ so$\mathrm{f}_{\#}(0) = 0$; $\mathrm{f}_{\#}(1) = 2$; $\mathrm{f}_{\#}(2) = 3$ $and$$\mathrm{f}_{\#}(3) = 1$

Therefore, $\mathrm{IVT}^{4,1}{}_{120}(225) = (\mathrm{f}_{120}(3)(\mathrm{f}_{120}(2)(\mathrm{f}_{120}(0)(\mathrm{f}_{120}(1))_4 = (1302)_4 = 114$.

Consequently, $\mathrm{IVT}^{4,1}{}_{120}(225) = 114$.

It is worth nothing that there are 256 such transformations and out of those 4! (24)are bijective in$\mathfrak{T}^{4,1}{}_{\#}$. There are nine $\mathrm{IVT}^{4,1}{}_{\#}$ (for # values 108, 78, 180, 177, 228, 225, 201, 198, and 114) where $\mathrm{IVT}^{4,1}{}_{\#}(x) \neq x for x = 0, 1,2$ and 3. The set of these nine IVTs is denoted as $\mathfrak{T}^{4,1}{}_{\#}$ which is of our prime interest for this work.

3. Methods, Results & Its Analysis

For all the *hsa*, *mml* and *ptr* miRNAs, some features have been extracted as shown below.

3.1 *Generating Indicator Matrix and Its Quantification*

3.1.1 Method of Generating Indicator Matrix (IM)

The miRNA is composed of four basic nucleotides namely A=adenine, C=cytosine, U=uracil, G=guanine. Let $\mathcal{V} \stackrel{\text{def}}{=} \{A, U, G, C\}$be a finite set of nucleotides and $x \in \mathcal{V}$ be any member of the alphabet. A miRNA can be thought as a finite symbolic string$\mathcal{S} = \mathbb{N} \times \mathcal{V}$ so that $\mathcal{S} \stackrel{\text{def}}{=} x_i, i = 1, 2, \ldots, N$being $\mathrm{x_i} \stackrel{\text{def}}{=} (\mathrm{i}, \mathrm{x}) = \mathrm{x(i)}, (\mathrm{i} = 1, 2, \ldots, \mathrm{N}; \mathrm{x} \in \mathcal{V}$ the value of *x* at position *I and* N denote length of the string.

The notion of indicator matrix and its characterization through fractal dimension was proposed by C. Cattani in [13] as follows

$$f: \mathcal{S} \times \mathcal{S} \rightarrow \{0, 1\} \text{such that}$$

$$f(x_h, x_k) \stackrel{\text{def}}{=} \begin{cases} 1 \; if x_h = x_k \\ 0 \; if x_h \neq x_k \end{cases} x_h, x_k \in \mathcal{S}$$

Therefore, the indicator matrix of an N-length string can be easily described as $N \times N$ sparse symmetric, binary matrix which results from

$$M_{hk} = f_{x_h}(x_k) x_h, x_k \in \mathcal{S}, h, k = 1, 2, 3 \ldots, N$$

This definition of indicator matrix does not help us differentiate between zeros formed by distinct base pairs. A slight modified definition of f is as follows:

$$f: \mathcal{S} \times \mathcal{S} \rightarrow \{0, 1, 2, 3\}$$

$$f(x_h, x_k) \stackrel{\text{def}}{=} \begin{cases} 0 \; if x_h = x_k : x_h, x_k \in \mathcal{S} \\ 1 \; if x_h \neq x_k ; \; x_h, x_k \in \{G, U\} or \; \{A, C\} \\ 2 \; if x_h \neq x_k ; \; x_h, x_k \in \{U, C\} or \; \{A, G\} \\ 3 \; if x_h \neq x_k ; \; x_h, x_k \in \{C, G\} or \; \{A, U\} \end{cases}$$

Consequently, the matrix M_{hk} corresponding to a given miRNA is four threshold matrix, namely 0,1,2 and 3.Let us decompose the matrix M_{hk} into four binary matrices A1,A2,A3 and A4 as follows:

$$A1_{hk} = \begin{cases} 1 \; where x_h = x_k : x_h, x_k \in \mathcal{S} \\ 0 \; otherwise \end{cases}$$

$$A2_{hk} = \begin{cases} 1 \; where x_h \neq x_k ; \; x_h, x_k \in \{G, U\} or \; \{A, C\} \\ 0 \; otherwise \end{cases}$$

$$A3_{hk} = \begin{cases} 1 \; where x_h \neq x_k ; \; x_h, x_k \in \{U, C\} or \; \{A, G\} \\ 0 \; otherwise \end{cases}$$

And $$A4_{hk} = \begin{cases} 1 \; where x_h \neq x_k ; \; x_h, x_k \in \{C, G\} or \; \{A, U\} \\ 0 \; otherwise \end{cases}$$

From the indicator matrix we have an idea of fractal-like distribution of nucleotides in miRNAs. The fractal dimension for the graphical representation of indicator matrix can be computed as the average of number of p(n) of "1" in all the $n \times n$ minors in the $N \times N$ indicator matrix

$$D = \frac{1}{N} \sum_{n=2}^{N} \frac{log p(n)}{\log n}$$

Let us understand through an example considering the string MIMAT0018101,

CUCGUGGGCUCUGGCCACGGCC (3.1)

The indicator matrix can be disintegrated into four matrices A1, A2, A3, and A4.

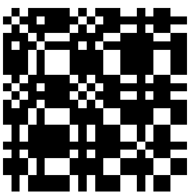

Fig.1:*A1*

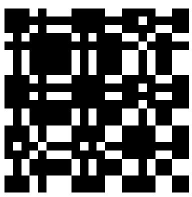

Fig.2:*A2*

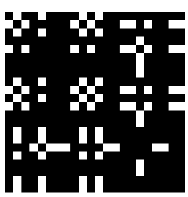

Fig.3:*A3*

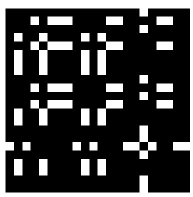

Fig.4:*A4*

The detailed calculations for fractal dimension of A1 are given below:

N	p(n)	log p(n)	log n	log p(n)/log n
2	1.2336	0.2099	0.6931	0.3029
3	2.8050	1.0314	1.0986	0.9388
4	4.9474	1.5989	1.3863	1.1533
5	7.6358	2.0328	1.6094	1.2631
6	11.0035	2.3982	1.7918	1.3385
7	15.0469	2.7112	1.9459	1.3933
8	19.5733	2.9742	2.0794	1.4303
9	24.4898	3.1983	2.1972	1.4556
10	29.6805	3.3905	2.3026	1.4725
11	35.6250	3.5730	2.3979	1.4901
12	42.3967	3.7471	2.4849	1.5079
13	50.1600	3.9152	2.5649	1.5264
14	59.2593	4.0819	2.6391	1.5467
15	68.8125	4.2314	2.7081	1.5625
16	78.6122	4.3645	2.7726	1.5742
17	88.3333	4.4811	2.8332	1.5816
18	98.9600	4.5947	2.8904	1.5897
19	111.625	4.7151	2.9444	1.6014
20	124.666	4.8256	2.9957	1.6108
21	136.00	4.9127	3.0445	1.6136
22	152.00	5.0239	3.0910	1.6253

*[**Table 1:** Calculation of fractal dimensions for A1]*

$d=\sum_{2}^{N}\frac{\log p(n)}{\log n}$=29.5784

D=d/N=1.3445

Similarly, the fractal dimensions for A2, A3 and A4 can be computed.

3.1.2 Results and Analysis

The intervals of fractal dimensions for *A1,A2, A3and A4* for the observed species are as follows:

Species /Intervals	***A1***	***A2***	***A3***	***A4***
hsa	(1.175, 1.667)	(0.173, 1.542)	(0.101, 1.531)	(0.068, 1.519)
mml	(1.211, 1.584)	(0.464, 1.495)	(0.627, 1.490)	(0.068, 1.420)
ptr	(1.211, 1.563)	(0.550, 1.448)	(0.763, 1.418)	(0.804, 1.420)

*[**Table 2**: Interval of fractal dimensions]*

The detailed table for fractal dimensions of A1, A2, A3 and A4 for all the miRNAs in the species as specified are given as a [*suppl. met-I, II and III*]. The interval of fractal dimension are listed above as resulted through detailed computation as given in [*Suppl. met-I, II and III*]. The followings are observed in case of human miRNAs.

I. The fractal dimensions for A2 and A3 of the miRNA sequence 229are -0.14045 and -0.23045 respectively [*Suppl. met-I*] of *Homo sapiens* due to fact that number of Us' and Gs' are more in number in the miRNA which leads to small and close values of p(n).

II. The fractal dimensions for A2 and A3of the miRNA sequence 684 [*Suppl. met-I*] of Homo sapiens species is undefined because the entire string of the miRNA contain only two bases G and U.

3.2 DNA Walk of miRNAs

3.2.1. Method

DNA walk is defines as a series $\sum Y_n, n = 1,2, \ldots\ldots, N \& Y_n \epsilon \{1,2,3,4\}$ which is the cumulative sum on the miRNA sequence representation$\{Y_1, Y_1 + Y_2, \ldots\ldots, \sum_{m=1}^{n-1} Y_m, \ldots\ldots, \sum_{m=1}^{N} Y_m\}$ [11].

Also we define $a_n \stackrel{\text{def}}{=} \sum_{i=1}^{n} f(A, x_i), g_n \stackrel{\text{def}}{=} \sum_{i=1}^{n} f(G, x_i), c_n \stackrel{\text{def}}{=} \sum_{i=1}^{n} f(C, x_i) \& u_n \stackrel{\text{def}}{=} \sum_{i=1}^{n} f(U, x_i)$.
It has been resulted by plotting (W_n, V_n) as definedtwo functions: $W_n \stackrel{\text{def}}{=} \sin a_n^2 - \sin g_n^2$ and $V_n \stackrel{\text{def}}{=} sint_n^2 - \sin c_n^2$.
The DNA walk for the miRNA string (3.1) is given below and then box-counting dimension has been computed.

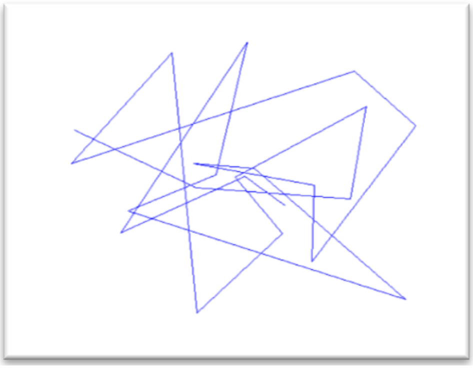

***Figure 1**: DNA Walk (V_n vs. W_n)*

The box-counting dimension for all such figures generated for all the miRNAs of three species, given as [*Suppli. met-I, II and III*], are computed by using well known software called BENOIT.

3.2.2 Results and Analysis

The interval of box-counting dimensions for *hsa*, *ptr* and *mml* DNA walks are given below:

Species/Interval	***hsa***	***Ptr***	***mml***
Fractal dimension	(1.94521, 1.94599)	(1.94505, 1.94627)	(1.94515, 1.94609)

[***Table 3:*** *Box-counting dimension for DNA walk*]

The box-counting dimension of DNA walk for all the miRNAs for all the specified species lie in very small interval, namely (1.945, 1.946).

3.3 Variance of miRNA Strings

3.3.1 Method

It is one of several descriptors, describing how far the values lie from the mean (expected value). For a given sequence $\{Y_1, Y_2 \ldots Y_N\}$, $\sigma^2 \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - (\frac{1}{N}\sum_{i=1}^{N} Y_i)^2$and the variance at distance N-k is given as
$\sigma^2 \stackrel{\text{def}}{=} \frac{1}{N-k}\sum_{i=1}^{N-k} Y_i^2 - (\frac{1}{N-k}\sum_{i=1}^{N-k} Y_i)^2$[11].
It is easily computable that the variance for the string (3.1) is 0.7430.

3.3.2 Results and Analysis:

The interval of variance for all the miRNA strings of inspected three species is given below:

Species/Interval	***hsa***	***Ptr***	***mml***
Variance	(0.4427, 1.9506)	(0.5871, 1.7284)	(0.4983, 1.7222)

***Table 4**: Variance*

From the results as shown in table 4, it has been observed that the variance of hsa-miR string 46 [*Suppl. met-I*] is minimum and the variance of hsa-miR string 210[*Suppl. met-I*] is maximum among all the miRNAs of the three species i.e. there are certain Human miRNAs which are adjacent to the mean and certain miRNAs which are distant from the mean. The intervals of variances of *ptr* and *mml* miRNAs are contained in the interval of variances of hs-miRNA string.

3.4 *Complexity of miRNA strings*

3.4.1 Method

Non-repetitiveness or singularity of a string signifies its complexity contrary to its normal behavior of periodicity and patchiness. The complexity of a string of length *n* is defined as follows [14]:
$K = \frac{log\Omega}{n}$, where$\Omega = \frac{n!}{a_n!u_n!c_n!g_n!}$& *n= {1, 2… N}*

In case of the miRNA string (3.1),$a_n = 69, u_n = 65, c_n = 41,\ g_n = 45$ and n = 22 and consequently the complexity of the string is -28.6464.

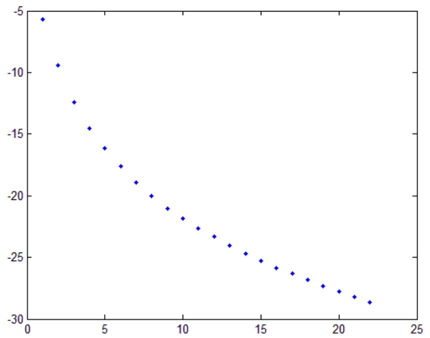

Figure 1: *Complexity for sequence (3.1)*

For the given string the complexity lies between -*5* and -*30.*

3.4.2 Results and Analysis:
As per the observation of complexities of miRNA strings, the following results were inferred:
I. The complexity of miRNA strings of three species lies between -*25* and -*30.*
II. The complexity of miRNA strings containing only three nucleotides lies between -*15&* -*20.*
III. The miRNA string 684 [*suppl.met-1*] of human is a two nucleotides (*G, U)* string has complexity of -12.3361.
IV. Complexities of certain miRNA strings are undefined due to Ω values tending to zero.

3.5 *Hurst Exponent of miRNA strings*

3.5.1 Method:

Hurst exponent is referred to as the "index of dependence," and is the relative tendency of a time series either to regress strongly to the mean or to cluster in a direction. It is a measure of long range correlation of one dimensional time series [12].
Let us consider a string $X = \{x_i\}, i = 1,2, \dots, n$

$$m_{x,n} = \frac{1}{n}\sum_{i=1}^{n} x_i \; Y(i,x) = \sum_{j=1}^{i}\{x_{j-}m_{x,n}\}$$
$$R(n) = \max Y(i,n) - \min Y(i,n) \;\; 1 \le i \le n$$
$$S(n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - m_{x,n})^2}$$

The Hurst exponent H is defined as :$(\frac{n}{2})^H = \frac{R(n)}{S(n)}$, where n is the length of the string. The range for which the Hurst exponent, H indicates negative, positive auto-correlation are $0 < H < 0.5$ and $0.5 < H < 1$ respectively. A value of $H=0.5$ indicates a true random walk, where it is equally likely that a decrease or an increase will follow from any particular value.
Binary string conversion has been made for all the miRNA strings through the transformation:
A=00, C=01, G=10 and U=11.
It is noted that the Hurst exponent of the binary string (3.1) is 0.6167 signifying a smooth trend.

3.5.2 Results and Analysis

The interval of Hurst exponent for examined three species is given below:

Species/Interval	***hsa***	***ptr***	***mml***
Hurst exponent	(0.420,0.893)	(0.431,0.862)	(0.397,0.872)

Table-5: *Hurst exponent*

As per the data available in [*supple. met. I, II, III*], it is observed that all the miRNA binary strings have almost similar range of Hurst exponent, with most of the miRNA binary strings showing positive autocorrelation.

3.6 Mean Ordering of miRNA strings

3.6.1 Method

A miRNA is a string constituting of different permutations of the base pairs *A, C, U &G* where repetition of a base pair is allowed. We can classify the miRNA sequences based on the ordering of poly-string mean of *A, C, U, & G* in the string. Given a string *X*, we calculate the mean of poly-strings consisting only of *A, C, U & G* separately [15, 16].

Mean $N_u= 2(N_{u1}+N_{u2}+N_{u3}+\ldots +N_{un})/n.(n+1)$ where $N_{ui}\in \{A, U, C, G\}, i = 1,2, \ldots, n and n$ is the length of the longest poly-string over the string.
According to the non-decreasing order of mean, we have classified all the miRNAs for different species into different classes. The mean order of sequence (3.1) is AUGC i.e. mean of poly-string of A is less than the same of U and so on.

3.6.2 Results and Analysis

There are 256 possible poly-strings of length four using four nucleotides A, U, C and G. It is worth noting that all the *hsa-miRs*, *mml-miRs* and *ptr-miRs* have been classified into only 73, 69 and 71 classes respectively according to the order of poly-string mean. Mean ordering of the three species enlightens that the mean of all the four nucleotides are never equal. The classification of miRNAs based on mean ordering is given as a [*Suppl. met-V*].

So far *hsa-miR*, *mml-miR*, and *ptr-miRs'* quantifications a re done. In the subsequent section, a mathematical network among the miRNAs has been deciphered in the light of IVTs.

4. miRNA Network in the light of IVTs

4.1 Method & Results

4.1.1 Use of IVT for evolution of miRNAs:

In this section, a network has been established through 4-adic IVTs. We have used nine IVTs from the set $\mathfrak{T}^{4,1}{}_{\#}$ to evolve each of the mml-miRs (without loss of generality). So corresponding to each of those miRNAs, there are 9 distinct strings which have been blasted in the miRNA database (miRBase) to have significant similarities Also all the features for the blasted strings have been enumerated (*Suppl. met-IV*). For example, let us consider mmu-mir-874 and the blast result for it is given below:

$IVT_{\#}^{4,1}$ (#)	***ID of blast results***
108	bmo-miR-2745
78	cin-miR-4049-5p
180	No match
177	sko-miR-4824
228	bta-miR-2887
225	mml-miR-664
201	ebv-miR-BFRH1-3
198	tca-miR-3883-5p
114	cin-miR-4023-3p

Table-6: *IVT results*

Through these nine IVTs it is possible to have a network among all the miRNAs over all the species with closest features explained in section 3 [*Suppl. met-IV*].

Through this ample study of quantifications of miRNAs, is it possible to justify or nullify a given string as probable miRNA? The answer is affirmative. The result follows in the next section.

5. Inference about Probable miRNAs

Here we have randomly chosen a stretch of around 90 nucleotides from *Macaca mulatta* genome and used random shifting window to generate these 7 candidates of miRNA .

String-I: 3’ AUUCCUGCUAAGCACGUACGAUGG 5’

String-II: 3’ GAGGUAGUAGGUUGUAUAGUUU 5’

String-III: 3’ GUAGGUUGUAUAGUUUUAGGGU 5’

String-IV: 3’ AGGUUGUAUAGUUUUAGGGUCA 5’

String-V: 3’ GAUGAGGUAGUAGGUUGUAUAG 5’

String-VI: 3’ UGAGGUAGUAGGUUGUAUAGUU 5’

String-VII: 3’ GGUAGUAGGUUGUAUAGUUUUAGGGU 5’

Now problem is to nullify or justify them as a probable miRNA. The features extracted for the seven strings are shown below in the table-7. The *string-I* have same poly-string mean for all nucleotides and also variance of the string does not fall in the interval of interest. It is to be noted that there is no *hsa, mml and ptr* miRNA having the same poly-string mean. Therefore it is certified that the *string-I* is not a *hsa, mml* and *ptr* miRNA.

Features	***I***	***II***	***III***	***IV***	***V***	***VI***	***VII***
FD for A1	1.3325	1.4074	1.4411	1.4155	1.4006	1.4045	1.4660
FD for A2	**1.1521**	1.1162	1.1228	1.1440	1.0817	1.0967	1.1748
FD for A3	1.1567	1.0646	0.9264	0.9884	1.1565	1.0978	1.0430
FD for A4	1.1516	1.3223	1.3435	1.3326	1.2911	1.3180	1.3847
FD for DNA walk	1.9452	1.94537	1.94536	1.94558	1.94525	1.9453	1.9451
Variance	**2.1875**	0.5723	0.5123	0.6942	0.6632	0.5723	0.5399

Complexity	Undefined	-19.3307	-19.5140	-28.4624	-19.3175	-19.3306	-20.5409
Hurst exponent	0.7514	0.6264	0.5822	0.6631	0.5597	0.6264	0.5810
Mean Order	**CCCC**	CAGT	CAGT	CAGT	CATG	CAGT	CAGT

Table-7: *Features for given candidates*

The values of the nine features, as shown in the above table, lie in the required intervals indicating that *string-II* to string-VII are all probable miRNA candidate for *mml* as per studies.

6. Conclusion and future research efforts:

We have developed the analysis protocol, using 9 mathematical parameters in characterizing miRNA strings, which is a novel approach of quantifying the strings pattern. Here we are reporting the pattern of all reported mature miRNA strings from miRBase of three different organisms, *viz. Homo sapiens*, *Macaca mulatta* and *Pan troglodytes*. On the basis of these observed features, we have established a network among different miRNAs of a given organism through implementation of IVTs and determined the bounded thresholds for all of the above mentioned mathematical features for miRNAs of the studied three organisms. On the basis of this extracted information this prescribed pipeline can now convincingly characterize a given string and define it whether as a probable mature miRNA or a non probable mature miRNA sequence, at least for the 3 studied organisms, without any prior conventional biological experiment(s). This adds to one prominent dimension to this study. Since efficient and successful prediction of miRNA is still one of the major issues in bioinformatics, and our developed protocol would add efficiency to already existing mature miRNA prediction tools, we believe our work is appropriate in present time frame.

Owing to increasing popularity of massively parallel sequencing techniques for RNA, a rapid increment in number of validated miRNA is occurring. Sequencing protocol for small RNA has been standardized both in SoliD platform (Invitrogen) and Hi-Seq platform (illumina). First step of post sequencing analysis is to search for sequence match with the databases like miRBase for already validated miRNA sequences. The left out small RNA fragments are then filtered through an analysis pipeline to find out small number of most probable miRNA candidates for novel miRNA biological validation experiment(s). More and more efficient pipeline is required for more efficient prediction of probable miRNA candidate(s) to reduce the load of costly, cumbersome and may be redundant biological experiments in a greater way. Amalgamation of existing analysis pipeline with the proposed protocol will make the prediction of probable candidate(s) and exclusion of non-probable candidate(s), far more efficient. Above all, the proposed method can draw inference just reading around 26nt long character strings only. So, this method would be applicable for very short read lengths, which are as small as 26 nt in length.

In near future, we would be extending our study for rest of organisms having quite a large number of already reported miRNA in miRBase.

Appendix-I

Supplementary materials:

1. *hsa Databse.xls*
2. *ptr Database.xls*
3. *mml Databse.xls*
4. *Blast.xls*
5. *Mean classification.xls*

Acknowledgements

The authors express their gratitude to Dr. B. S. Daya Sagar for his thoughtful research suggestions. Also, Authors extends note of thanks to CSIR for providing fellowship of NDS.

References

1. Hassan, S. Sk. Et al. Collatz Function like Integral Value Transformations, Alexandria Journal of Mathematics, Vol. 1- Num.2 – Nov. 2010, pp 30-35.

2. Hassan, S. Sk., et al. DNA Sequence Evolution through Integral Value Transformations. (In press of The journal Interdisciplinary Sciences--Computational Life Sciences, Springer).

3. Lee R. C. et al. The C. elegansheterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 1993 *75*,843–854.

4. Wang WX et al. Individual microRNAs (miRNAs) display distinct mRNA targeting "rules". RNA Biol. 2010; 7(3): 373–380.
5. M E Peter et al. Targeting of mRNAs by multiple miRNAs: the next step, *Oncogene* (2010) **29,** 2161–2164; doi:10.1038/onc.2010.59; published online 1 March 2010.

6. Ruvkun, G. Molecular Biology: Glimpses of a Tiny RNA World. *Science* 2001, *294*, 797–799.

7. Lau NC, Lim LP, Weinstein EG, Bartel DP (October 2001). "An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditiselegans". *Science* 294 (5543): 858–62.
8. Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A (October 2004). "Identification of mammalian microRNA host genes and transcription units". *Genome Res.* 14 (10A): 1902–10.
9. He L, Thomson JM, Hemann MT, *et al.* (June 2005). "A microRNA polycistron as a potential human oncogene". *Nature* 435 (7043): 828–33.
10. A Genomic Predictor of Response and Survival Following Taxane-Anthracycline Chemotherapy for Invasive Breast Cancer, 2011;305(18):1873-1881.
11. B. Mandelbrot, The Fractal Geometry of Nature. W.H. Freeman and Company..ISBN 0-7167-1186-9. Briggs, John (1992).

12. M. Bransley, Fractals Everywhere. Boston: Academic Press Professional, 1993. ISBN 0-12-079061-0.

13. Carlo Cattani, Fractals and Hidden Symmetries in DNA, Mathematical Problems in Engineering Volume 2010 (2010), Article ID 507056

14. Yu Zu-Guo *et al* Fractals in DNA sequence analysis, 2002 *Chinese Phys.* Vol 11 Num. 12.

15. Hassan, S. Sk. et al. Underlying Mathematics in Diversification of Human Olfactory Receptors in Different Loci. Available from Nature Precedings <http://hdl.handle.net/10101/npre.2010.5475.1> (2010)

16. Hassan, S. Sk. et al. Understanding Genomic Evolution of Olfactory Receptors through Fractal and Mathematical Morphology. Available from Nature Precedings <http://hdl.handle.net/10101/npre.2011.5674.1> (2011).