International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-4 Issue-6 December-2012

A Modified Algorithm for Quantifying of Pre-mature MiRNAs using Some Fractal Parameters

Joyshree Nath[1], Asoke Nath[2]
Department of Computer Science, Jogesh Chandra Chaudhuri College, Kolkata, India[1]
Department of Computer Science, St. Xavier's College (Autonomous), Kolkata, India[2]

Abstract

The prime research area now is to understand microRNA (miRNA) in a quantitative manner. The miRNAs are non-coding short ribonucleic acid (RNA) molecules, approximately ~25 nucleotides long. MicroRNAs have emerged as powerful regulators of diverse cellular processes with important roles in tissue remodeling. The scientists across the globe are working on quantitative estimation of microRNA (miRNA) from different angles. The authors in their earlier paper have already made a new addition in the present scenario. In the present paper the authors are appending few more fractal parameters to strengthen the algorithm. In the previous work the authors deciphered the inherent statistical behavior in pre-mature miRNA strings through few statistical parameters namely: Hurst Exponent values, Variance, Poly-String Mean and Poly-String Standard Deviation. Now, in the present work the authors have to tried to explore the quantification of the same miRNA dataset based on fractal and statistical parameters namely: (i) Fractal Dimension (FD) of 4 binary indicator matrices, (ii) FD of the DNA walk and (iii) Complexity of the miRNA strings, of the three organisms Homo sapiens (hsa), Macaca mulatta (mml) and Pan troglodytes (ptr).

Keywords

Fractal, microRNA,Homo sapiens,Macaca mulatta, Pan troglodytes,DNA Walk, Complexity, Fractal Dimension.

1. Introduction

Quantitative understanding of microRNA (miRNA) is now an emerging area of research. There is much to explore in the field of miRNAs for biological advancement. MiRNAs are a class of small, regulatory RNAs that play important role in biological processes like cell proliferation, cell death, cell development & differentiation, viral infection, hematopoiesis, oncogenesis and many more[6],[7]. It also helps to know the causes of lymphoma, leukemia, cancers and different cardiac problems [8]. MiRNAs help in apoptosis and fat metabolism [5]. For human and other vertebrate cell lines, miRNA genes are involved in tumor suppression, antiviral defense, adipocyte differentiation and susceptibility to cytotoxic T-cells [10]. In the present work we have quantified the nucleotide strings of pre-mature miRNAs of the three organisms (i) Homo sapiens (hsa), (ii) Macaca mulatta (mml) and (iii) Pan troglodytes (ptr) in the light of 3 statistical parameters in addition to our previous paper where we used another 4 statistical parameters. Of the 3 parameters used in this paper the first 2 are fractal parameters namely: (i) Fractal Dimension (FD) of 4 binary indicator matrices, (ii) FD of the DNA walk and the last one is a statistical parameter namely: (iii) Complexity of the miRNA strings . In our previous paper we have shown the Hurst exponent parameter is directly related to fractal dimension. These fractal parameters are used as fractals contribute hugely in the genre of bio-informatics and help in understanding and quantifying the biological parameters from mathematical point of view. We have shown that an unknown RNA sequence can be selected to be a probable premature miRNA. If the given sequence doesn't match according to the results of this work then it will be nullified. In section 2 we have discussed the biogenesis and functions of miRNA. In section 3 we have discussed our algorithms that have been implemented in the present work. In section 4 we have shown results that have been obtained from our proposed methods. In section 5 we discuss in the summary and ramifications of our method. Finally in section 6 we have given our conclusion and the future scope of the present study.

2. Algorithms used

Now we will show how we have used some statistical methods on pre-mature miRNAs.

A. Extracting the dataset and Generating individual sequence text files :-

For this work, we have extracted the pre-mature miRNA sequences of the three organisms Homo

sapiens (hsa), Macaca mulatta (mml) and Pan troglodytes (ptr). These miRNAs were extracted from the miRNA database, miRBase. Finally, 1424,479,599 pre-mature miRNA sequences were obtained for Homo sapiens (hsa), Macaca mulatta (mml) and Pan troglodytes (ptr) respectively. Before applying any method for the quantification of any of the pre-mature miRNAs, each of these strings of miRNAs for hsa, mml and ptr species were extracted and stored as text files. These text files were named as 1.txt, 2.txt…so on up till the last miRNA string for each of the species. The results we have shown in section 5. Here we will show how we have utilized some statistical and fractal features that have been utilized to work upon the pre-mature miRNA strings of hsa, mml and ptr.

B. Generating Indicator Matrices and its Quantification:-

The miRNA consists of 4 basic nucleotides, i.e, A=Adenine, C=Cytosine, U=Uracil, G=Guanine. Let V , be a finite set of nucleotides and xЄVbe any member of the alphabet. A miRNA is a finite symbolic string S=N×V (N being a set of natural numbers) such that S Si, i=1,2,3…N(N here denotes the length of the miRNA string) being xi (i,x)= x (i), (i=1,2,3…N; xЄV the value of x at position i). The indicator matrix of an N-length string can be defined as an N×N sparse symmetric, binary matrix,

Mhk= fxh(xk) xh,xk ЄS,h,k=1,2,3…N

Here is the proposed definition of the indicator matrix. This is derived and modified from a predefined format (proposed by C. Cattani,[12]):

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

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

Consequently, the matrix Mhk corresponding to a given miRNA is a four threshold matrix, namely 0, 1, 2 and 3. The matrix Mhk, can be decomposed into four binary matrices A1, A2, A3 and A4 as follows [1]:

$$A1_{hk} = \begin{cases} 1, \text{ where } xh = xk : xh, xk \in S \\ 0, \text{ otherwise} \end{cases}$$

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

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

And

$$A4_{hk} = \begin{cases} 1, \text{ where } xh \neq xk ; xh, xk \in \{C,G\} or \{A,U\} \\ 0, \text{ otherwise} \end{cases}$$

C. DNA Walk of miRNAs:

The DNA digital representation is the N-length one-dimensional real signal {Yn}, n=1,2,…N. DNA walk is defined as a series ΣYn, Y{0,1,2,3} which is the cumulative sum on the miRNA sequence representation Y will be defined in details in the next section of this chapter. To calculate the DNA walk of a genome or more precisely, a miRNA string an initial table (Table-1) is considered. Then the summation of the matrix cell value for each nucleotide with the other nucleotides is calculated.

The definition follows as,

$$a_n \stackrel{def}{=} \sum_{i=1}^{n} f(A, x_i),$$
$$g_n \stackrel{def}{=} \sum_{i=1}^{n} f(G, x_i),$$
$$c_n \stackrel{def}{=} \sum_{i=1}^{n} f(C, x_i),$$
$$u_n \stackrel{def}{=} \sum_{i=1}^{n} f(U, x_i),$$

[where, i=1,2,3…N; x Є V the value of x at position i]

Table-1

f	A	U	C	G
A	0	3	1	2
U	3	0	2	1
C	1	2	0	3
G	2	1	3	0

Next the formula below is calculated,

$$W_n \stackrel{def}{=} sina_n^2 - sing_n^2 \text{ and } V_n \stackrel{def}{=} sinu_n^2 - sinc_n^2$$

[1]

The results thus obtained are discussed in chapter 5.

D. Complexity of miRNA strings

Non-repetitiveness of a string refers to its complexity unlike its periodicity and patchiness. The complexity of a string of length *n* is defined as [11]:

$$K = \frac{log\Omega}{n}, \text{ where} \Omega = \frac{n!}{a_n! u_n! c_n! g_n!} \& \; n = \{1, 2 \ldots N\}$$

The results thus obtained are discussed in chapter 5.

3. Algorithms used in the previous paper

A. Hurst Exponent of miRNA strings

The Hurst exponent occurs in several areas including biophysics, bioinformatics,etc.[4] For calculating the Hurst exponent for each miRNA string we used the following formula :

R(n)=maxY(i,n) minY(i,n) $1 \leq i \leq n$

S(n)= $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(xi-n)^2}$

Thus the Hurst exponent H is defined as :$(\frac{n}{2})H=\frac{R(n)}{S(n)}$, where n is the length of the string[1].
For details check our previous paper.

B. Variance of miRNA Strings:-

Another mathematical parameter that we calculated in our previous paper was the value of variance for each miRNA string and the variance is given as σ $\stackrel{def}{=} \frac{1}{N-k}\sum_{i=1}^{N-k} Y_i^2 \quad (\frac{1}{N-k}\sum_{i=1}^{N-k} Y_i)^2$ [1].
For details check our previous paper.

C. Poly-String Mean and StandardDeviation of miRNA Strings:-

These parameters were calculated for each miRNA string of each of the species, Homo sapiens (hsa), Macaca mulatta (mml) and Pan troglodytes (ptr). The poly-string mean (P_m^N) and poly-string standard deviation (P_{SD}^N) of N were defined as :

$$P_m{}^N = \frac{\sum_{i=1:n} m_i k_i}{\sum_{1:n} m_i} \text{ and } P_{SD}{}^N = \sqrt{\frac{1}{n}\sum_{i=1}^{n} m_i (k_i - P_m{}^N)^2}$$

[2]
For details check our previous paper.

4. Results and Discussion

Here we will show the results which we obtain from our proposed methods stated in section IV.

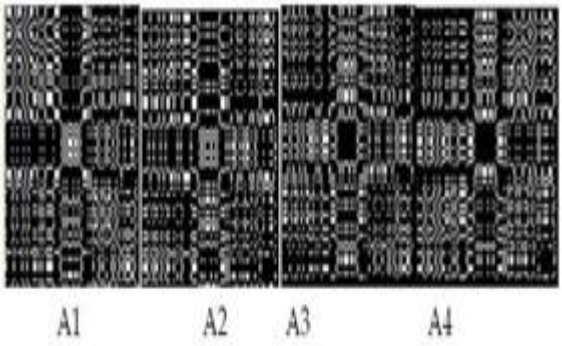
A. Extracting the dataset and Generating individual sequence text files:-

There were 1424,479,599 pre-mature miRNA sequences for Homo sapiens (hsa), Macaca mulatta (mml) and Pan troglodytes (ptr) respectively from a well-known database called, miRBase(version 17; http://www.mirbase.org/).
These miRNA strings were extracted from each species were stored as as 1.txt, 2.txt…so on for each species.
For example, in Macaca mulatta(mml), the content of 1.txt came as the first sequence,i.e, the sequence for, mml-let-7a -1 MI0007570
UGGGAUGAGGUAGUAGGUUGUAUAGUUUU
AGGGUCACACCCACCACUGGGAGAUAACUA
UACAAUCUACUGUCUUUCCUA
In Homo sapiens(hsa), the content of 2.txt came as the second sequence, i.e, the sequence for, hsa-let-7a-2 MI0000061
AGGUUGAGGUAGUAGGUUGUAUAGUUUAG
AAUUACAUCAAGGGAGAUAACUGUACAGCC
UCCUAGCUUUCCU
In Pan troglodytes (ptr), the content of 3.txt came as the third sequence, i.e, the sequence for, ptr-let-7a-3 MI0008399
GGUGAGGUAGUAGGUUGUAUAGUUUGGGG
CUCUGCCCUGCUAUGGGAUAACUAUACAAU
CUACUGUCUUUCCU

B. Generating Indicator Matrices and Its Quantification:

For each of the species a set of four binary matrices A1, A2, A3 and A4 for each of the text miRNA sequences were extracted using C programming. Each binary matrix was stored in a separate text file, namely, 1A1.txt, 1A2.txt, 1A3.txt, 1A4.txt, 2A1.txt, and so on. Here are excerpts of the binary indicator matrices of A1, A2, A3 and A4 for the miRNA strings of each species.

Now, mml sequence for mml-let-7a-1 MI0007570 gives the following graphical representations :



The range of the Fractal Dimensions (FDs) of thesemiRNA sequences taken from their respective graphical representations are:

Species	A1	A2	A3	A4
FDs of the miRNA of hsa	(1.46761, 1.68023)	(1.44106, 1.67951)	(1.44065, 1.66764)	(1.47664, 1.66853)
FDs of the miRNA of mml	(1.4328, 1.67459)	(1.41198, 1.67725)	(1.41653, 1.66758)	(1.43053, 1.67062)
FDs of the miRNA of ptr	(1.47253, 1.67775)	(1.42554, 1.67951)	(1.41653, 1.66758)	(1.43053, 1.66853)

C. **DNA Walk of miRNAs:**

Below are 3 graphical representations for Wn vs Vn for the string sequences, 1.txt, 2.txt and 3.txt miRNA sequences for the species mml, hsa, ptr respectively:

Graph for 1.txt sequence of *mml*
Graph for 2.txt sequence of *hsa*
Graph for 3.txt sequence of *ptr*

The range of the FDs for the miRNA sequences of *hsa, mml, ptr* are:

Species	MiRNA of hsa	MiRNA of mml	MiRNA of ptr
Fractal Dimension	(1.89608,1.94513)	(1.92149,1.94491)	(1.94083,1.94513)

D. **Complexity of miRNA strings:-**

Due to the constant increasing values of an,gn,un and cn the value of Ω gets decreased to 0 for every miRNA string sequence for every species. Subsequently the value of K goes from initial negative values to infinite values for every miRNA string sequence of every species.

5. Summary

Quantitative understanding of microRNA (miRNA) is evolving dynamically from research point of view. MiRNAs play important roles in biological processes like cell proliferation, cell death, fat metabolism, growth control etc. [7].It also helps to know the causes of lymphoma, leukemia, cancers and different cardiac problems [8]. Now our previous work classified and quantified the nucleotide strings of pre-mature miRNAs of the three organisms namely: (i) Homo sapiens (hsa), (ii) Macaca mulatta (mml) and (iii) Pan troglodytes (ptr) in the light of 4 statistical parameters, which were, Hurst Exponent values, Variance, Poly-String Mean and Poly-String Standard Deviation. In the present paper we append 3 more fractal and statistical parameters.

By our method any unknown RNA sequence can be selected to be a probable premature miRNA candidate. If the given sequence doesn't match the results of this work then it will be straightway nullified with the help of our method. This can be clubbed with existing biological works on quantifying pre-mature miRNAs, to be exactly sure about quantifying and classifying any unknown string of miRNA. So classification of future unknown miRNAs under a biologically specific group of either *hsa* or *mml* or *ptr* will be less time and money consuming if the present biological methods has the backing of our mathematical method.

6. Conclusion and Future scope

In this work, the inherent mathematical behavior of pre-mature miRNAs were deciphered through some fractal parameters, as the behavior of miRNA string sequences become more meas urable in the light of fractals. Four statistical parameters were used to quantify and classify these premature miRNA strings in our earlier paper. Along with that our present paper appends 3 more fractal and statistical parameter. This study would help in understanding the behavior of the strings of pre-mature miRNAs of the three organisms, Homo sapiens, Macaca mulatta and Pan troglodytes in the light of fractals(Fractals refer to structure within structure) and Statistical parameters .

This work will help ultimately in concluding on the probable candidates of premature miRNAs, from some set of unknown RNA sequences. The present method will nullify certain candidate to be a pre-mature miRNA and no complex, expensive and time-consuming biological experiment will be necessary for that. However, for assuring that candidate to be a

pre-mature one would definitely need the help of biology. So as of now the future work is to extend this project work and include other fractal and morphological parameters in it to conclude if a miRNA in question is a probable candidate for a pre-mature one or not. These parameters can help one to strike out a candidate from being a pre-mature one in a different light of statistical analysis.

References

[1] Hassan, Sk Sarif, et al. "Quantification of miRNAs and Their Networks in the light of Integral Value Transformations." Nature Precedings: npre 2 (2011).
[2] Nath, Joyshree, and Asoke Nath. "A Comprehensive Study of Target Prediction Algorithms for Animal MicroRNAs (miRNAs)." International Journal of Computer Applications 40.15 (2012): 8-11.
[3] www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_089374.pdf
[4] http://www.bearcave.com/misl/misl_tech/wavelets/hurst/
[5] John, Bino, et al. "Human microRNA targets." PLoS biology 2.11 (2004): e363.
[6] Wang, Xiaowei, and Issam M. El Naqa. "Prediction of both conserved and nonconserved microRNA targets in animals." Bioinformatics 24.3 (2008): 325-332.
[7] Min, Hyeyoung, and Sungroh Yoon. "Got target?: computational methods for microRNA target prediction and their extension." Experimental & molecular medicine 42.4 (2010): 233-244.
[8] Chandra, Vinod, et al. "MTar: a computational microRNA target prediction architecture for human transcriptome." BMC bioinformatics 11.Suppl 1 (2010): S2.
[9] Gennarino, Vincenzo Alessandro, et al. "MicroRNA target prediction by expression analysis of host genes." Genome research 19.3 (2009): 481-490.
[10] Marquez, Rebecca T., and Anton P. McCaffrey. "Advances in microRNAs: implications for gene therapists." Human gene therapy 19.1 (2008): 27-38.
[11] Zu-Guo, Yu, et al. "Fractals in DNA sequence analysis." Chinese Physics 11.12 (2002): 1313.
[12] Cattani, Carlo. "Fractals and hidden symmetries in DNA." Mathematical Problems in Engineering 2010 (2010).

Joyshree Nath Born on 17-th August 1986. After finishing M .Tech(IT)from C.U. now she is taking classes of Computer Science in B.Sc. in a part - time basis. She is also approved academic counsellor of IGNOU M CA programme. She is involved in research work in Quantum computing, Bio-informatics and symmetric key cryptography.

Asoke Nath is the Associate Professor in Department of Computer Science. Apart from his teaching assignment he is involved with various research work in Cryptography, Steganography, Green Computing, E-learning. He has presented papers and invited tutorials in different International and National conferences in India and in abroad.