International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 6 (2014) pp. 184 - 195

Quantitative Deciphering of pre-mature and pattern recognition of mature miRNAs using some Statistical Parameters

Joyshree Nath[1], Asoke Nath[2]

[1] Machine Intelligence Unit(MIU), Indian Statistical Institute
203 B.T. Road, Kolkata-700 108, West Bengal, India
joyshreenath@gmail.com

[2] Department of Computer Science, St. Xavier's College(Autonomous)
30 Park Street, Kolkata, West Bengal, India
asokejoy1@gmal.com

Abstract: **Today the DNA and RNA structures and their underlying functions have a very significant role in every aspect of human life. Properly inferring on the gene functions leads us to be more affirmative about the functional system of every organism and thus to take necessary actions in recuperation from diseases. To understand the functional property of any gene sequence it is necessary to study the sequence structure. In the present work the authors have tried to assess the miRNA sequence structures using some mathematical and statistical parameters like calculating the range of hurst exponent or autocorrelation values of each miRNA string sequence, depicting the frequency distribution of each nucleotide in graphical pattern. The authors have also calculated the variance range of each miRNA string sequence. The last two decades the researchers are doing extensive work on quantitative estimation of miRNA. The miRNAs are non-coding short ribonucleic acid (RNA) molecules, approximately ~25 nucleotides long. MiRNAs help in understanding the entire scope of post-transcriptional gene regulation. MiRNAs regulate numerous cellular processes and have roles in cardiac activities, neural functions, psychological functions and lots more. The present work the authors tried to explore the quantification and classification based on statistical results on nucleotide strings of pre-mature miRNAs of the three organisms Homo sapiens (hsa), Gorilla gorilla (ggo) and Pongo pygmaeus (ppy). The present work is also extended to mature miRNAs of species like Pan troglodytes (ptr) to check the effect of linear binary rules on each of the miRNA sequence.**

Keywords**:** ***microRNA, Variance,Frequency Distribution, hsa, ggo, ppy*****.**

I. Introduction

Quantitative understanding of every kind of gene sequemces and especially microRNA (miRNA) is now an emerging area of research. The knowledge of miRNAs has accumulated rapidly in recent years but some of the important issues not yet explored which need to be explored. MiRNAs are a class of small, regulatory RNAs playing important roles in biological processes like cell proliferation, cell death, hematopoiesis, oncogenesis and many more [1]. MiRNAs also help to know the causes of lymphoma, leukemia, cancers and different cardiac problems [2]. In the present study the authors have assessed the nucleotide strings of pre-mature miRNAs of the three organisms (i) Homo sapiens (hsa), (ii) Gorilla gorilla (ggo) and (iii) Pongo pygmaeus (ppy) in the light of some statistical and mathematical parameters. The authors have shown that the results achieved through this method any unknown RNA sequence can be selected to be a probable premature miRNA, if it abides by the range of results obtained by this method. For validation of the sequence as a biological sequence the subsequent biological experiments would be needed. If however, the given sequence doesn't match according to the results of this work then it will be rejected. In section 2 the authors have given the algorithms that have been implemented in the present work. In section 3 the results are given which have been obtained from the proposed methods. In section 4 the effect of few linear rules on the mature miRNA sequences of Pan troglodytes species are shown. This effect will be studied in future for getting some intrinsic view of the sequence motifs of the mature miRNAs. In section 5 a summary is given. Finally in section 6 the conclusion and the future scope of the present study is given.

II. Algorithm Used :

In this section a brief overview is given on some statistical methods used on the pre-mature miRNAs of the three species organisms namely (i) Homo sapiens (hsa), (ii) Gorilla gorilla (ggo) and (iii) Pongo pygmaeus (ppy)

II.1 Extracting the experimental dataset:-

The data from the pre-mature miRNA sequences of the three organisms Homo sapiens (hsa), Gorilla gorilla (ggo) and Pongo pygmaeus (ppy) were extracted. These miRNAs were extracted from miRBase(version 19). The total number of

pre-mature miRNA obtained from miRBase version 19 were 322 for ggo, 1600 for hsa and 633 for ppy. Before applying any method for the quantification of any of the pre-mature miRNAs, each of these strings of miRNAs for hsa, ggo and ppy species were extracted and stored as text files. These text files were named as 1.txt, 2.txt, 3.txt …so on up till the last miRNA string for each of the species. The results we have shown in section 3. 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, ggo and ppy.

II.2. Calculation of Hurst Exponent of miRNA strings:-

The Hurst exponent occurs in several areas including biophysics, bioinformatics, etc.[3]

For calculating the Hurst exponent for each miRNA string we consider the following formula from [4]:

Let X be a string such that X={x_i},(i=1,2,3…n), [where n is the length of the string]

Therefore we calculate mean,

$$x_n = \frac{1}{n}\sum_{i=0}^{n} x_i$$

$$mi(i,n) = \sum_{j=1}^{i} \{Xj - Xn\},$$ [where x is the nucleotide at position j]

Now for deriving Hurst exponent for a miRNA string we calculate

R(n)=max of $m_i(i,n)$ – min of $m_i(i,n)$ 1<=i<=N[where N is the length of the miRNA string]

$$S_{n=}\frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_j - x_n)^2}$$

By Yu Zu Guo et al. [4] we applied the formula of hurst exponent H is defined which is $R(n)/S(n)\sim(\frac{n}{2})^H$, where n is the length of the miRNA string.

In the very beginning we consider the values of A=3, C=2, G=1, and U=0 and perform all our calculations for hurst exponent on the basis of these values using our own software developed in C-language. The values are just for storing the nucleotides and for the feasibility of mathematical calculations. The results obtained from all these operations are discussed in section 3.

II.3. Variance of miRNA Strings:-

According to [5] we calculated the variance of the nucleotides in the miRNA sequence from their mean. For a given miRNA string sequence of N length {Y1, Y2… Y_N}, the variance at distance N-k (where k is 0,1,2…N-1) (N being the length of the miRNA sequence)is given as,

$$\sigma^2 \stackrel{\text{def}}{=} \frac{1}{N-k}\sum_{i=0}^{N-k-1} Y_i^2 - \left(\sum_{i=0}^{N-k-1} Y_i\right)^2$$

The results thus obtained are discussed in section 3.

II.4. MiRNA string sequence frequency distribution:-

In this section the nucleotide pattern and frequency distribution is checked via C programming and subsequent graph plotting. The frequency of every nucleotide A,C,G,U is calculated and the percentage of of the frequency of a single nucleotide(for example A,C,G,U each), frequency of dinucleotides(for example AA,CC,GG,UU each), frequency of trinucleotides(for example AAA,CCC,GGG,UUU each) and so on was calculated. The results obtained from all these operations are discussed in section 3.

III. Results and Discusion

Here we will show the results which we obtain from our proposed algorithm which we have stated in section 2.

III.1. Extracting the dataset and Generating individual sequence text files:-

There were 1600,322,633 pre-mature miRNA sequences for Homo sapiens (hsa), Gorilla gorilla (ggo) and Pongo pygmaeus (ppy) respectively from a well-known database called, miRBase version 19.These miRNA strings were extracted from each species were stored as as 1.txt, 2.txt…so on for each species.

For example, in Gorilla gorilla (ggo) the content of 1.txt came as the first sequence,i.e, the sequence for, ggo-let-7a MI0020621

AGACCGACUGCCCUUUGGGGUGAGGUAGUAGGUU
GUAUAGUUUGGGGCUCUGCCCUGCUAUGGGAUAAC
UAUACAAUCUACUGUCUUUCCUGAAGUGGCUGUAA
UAUCU

In Homo sapiens(hsa), the content of 2.txt came as the second sequence, i.e, the sequence for, hsa-let-7a-1 MI0000060

UGGGAUGAGGUAGUAGGUUGUAUAGUUUUAGGGU
CACACCCACCACUGGGAGAUAACUAUACAAUCUAC
UGUCUUUCCUA

In and Pongo pygmaeus (ppy), the content of 3.txt came as the third sequence, i.e, the sequence for, ppy-let-7a-3 MI0014781

GGGUGAGGUAGUAGGUUGUAUAGUUUGGGGCUCU
GCCCUGCUAUGGGAUAACUAUACAAUCUACUGUCU
UUCCU

III.2. Hurst Exponent of miRNA strings:-

The Hurst exponent values for hsa, mml, ptr were calculated using own developed program developed in C language.

The range of the Hurst exponents for the miRNA sequences for hsa, mml, ptr are:
HSA- (0.0112-0.0603), GGO-(0.0151-0.0316), PPY-(0.0124-0.0603)

The range of the Hurst exponent values came as $0 < H < 0.5$ which indicates a negative, auto-correlation in the case of these pre-mature miRNAs.

III.3. Variance of miRNA Strings:-

In this case all the variance measures for every nucleotide string converged to an approximated zero value when last but one nucleotide was left for the variance calculation. So for all these 3 species the lowest value for variance is 0.For ppy the highest value 2.25 comes for the 46th sequence which is for ppy-mir-29a MI0002660-AUGACUGAUUUCUUUUGGUGUUCAGAG UCAAUAUAAUUUUCUAGCACCAUCUGAAAUCGGUU AU. For ggo also the value came 2.25 for 20th sequence which is for `ggo-mir-18b MI0020857` UAUAAUGUGUCUCUUGUGUUAAGGUGCAUCUAGU GCAGUUAGUGAAGCAGCUUAGAAUCUACUGCCCUA AAUGCCCCUUCUGGCACAGGCUGCCUAAUAUACAG CAUUU. This value is same for hsa and for 17th sequence which is hsa-mir-9-1 MI0000466-GGGUUGGUUGUUAUCUUUGGUUAUCUA GCUGUAUGAGUGGUGUGGAGUCUUCAUAAAGCUA GAUAACCGAAAGUAAAAAUAACCCCA. In the below figure we have plotted values of k(where k=0,1,2…N-1) against the respectively variance or σ^2 values, for each miRNA sequence of each species.

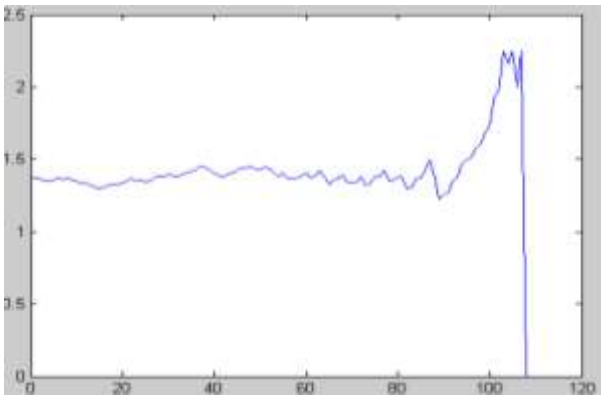

Fig-1: Variance for sequence 20.txt of ggo

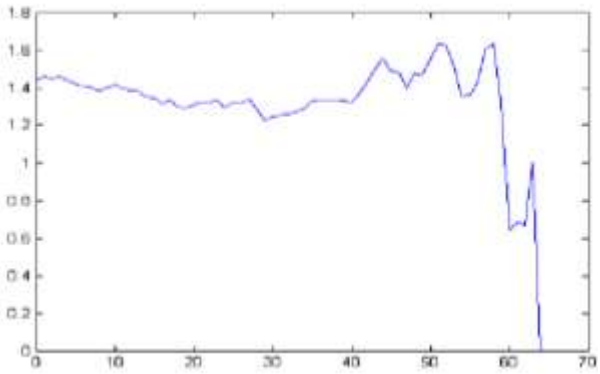

Fig-2: Variance for sequence 134.txt of hsa

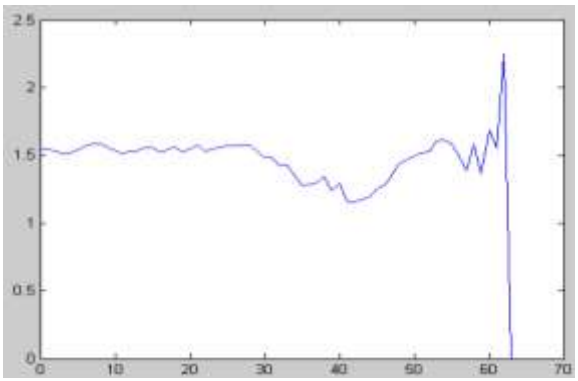

Fig-3: Variance for sequence 46.txt of ppy

III.3. Poly-String Mean & Standard Deviation of miRNA Strings:-

The frequency of every nucleotide A,C,G,U is calculated and the percentage of the frequency of a single nucleotide(for example A,C,G,U each), frequency of dinucleotides(for example AA,CC,GG,UU each), frequency of trinucleotides(for example AAA,CCC,GGG,UUU each) and so on was calculated. This was continued until any such pattern found in the entire nucloetide string sequence for each miRNA string of each species. Then it was plotted graphically combining the nucleotide distribution of every nucleotide subsequence (single, di, tri and so on) in the X axis and their respective percentages in that particular string sequence in Y axis.

The graphical plot of string 1.txt of ppy looks like-

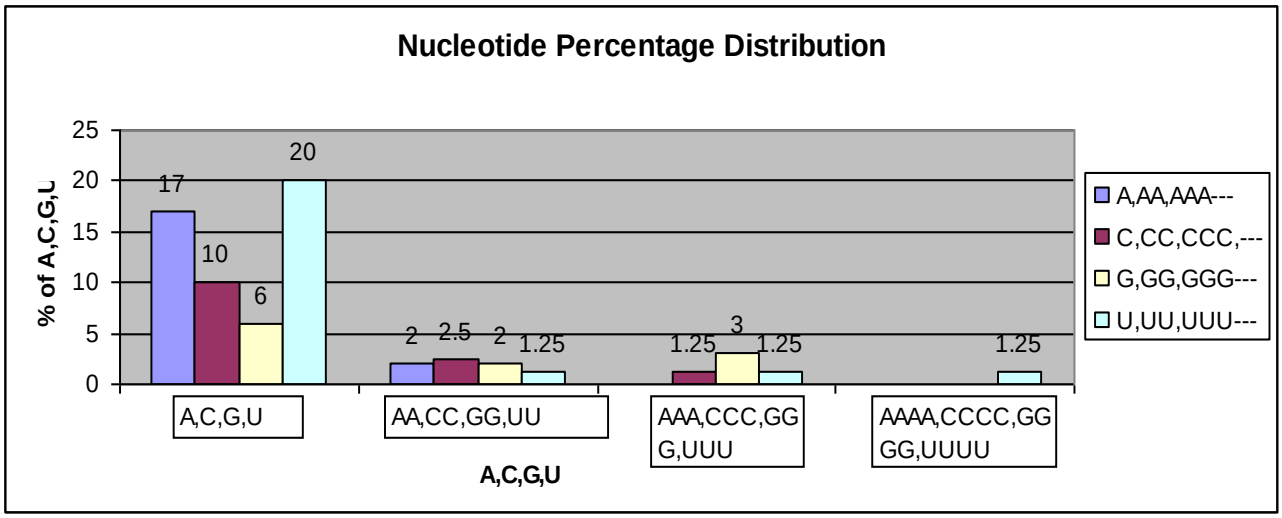

Figure 4. Percentage of different nucleotide subsequences in first microRNA string ppi(1.txt)

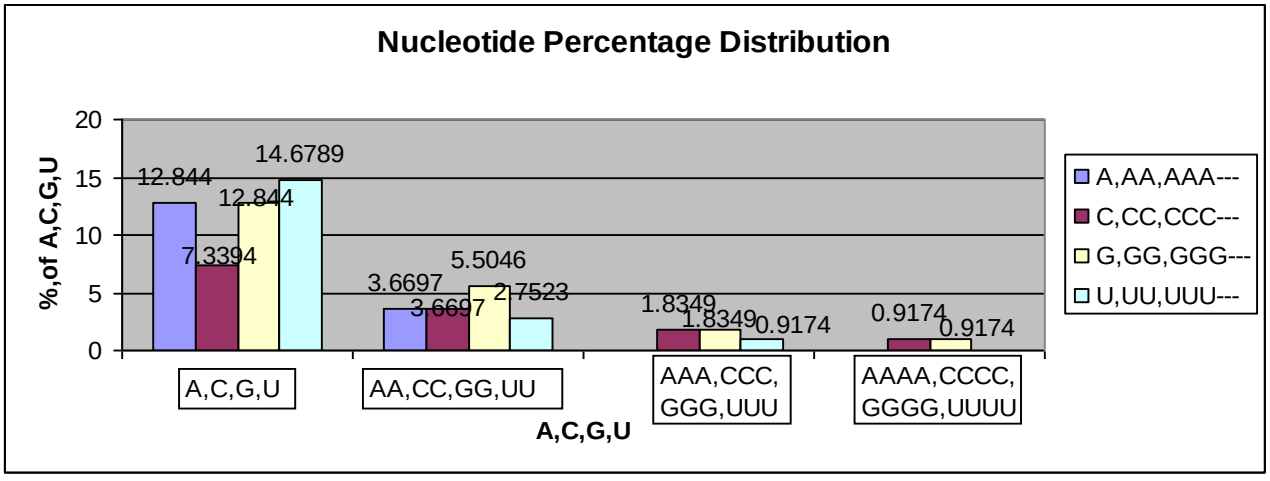

Figure 5. Percentage of different nucleotide subsequences in first microRNA string hsa(3.txt)

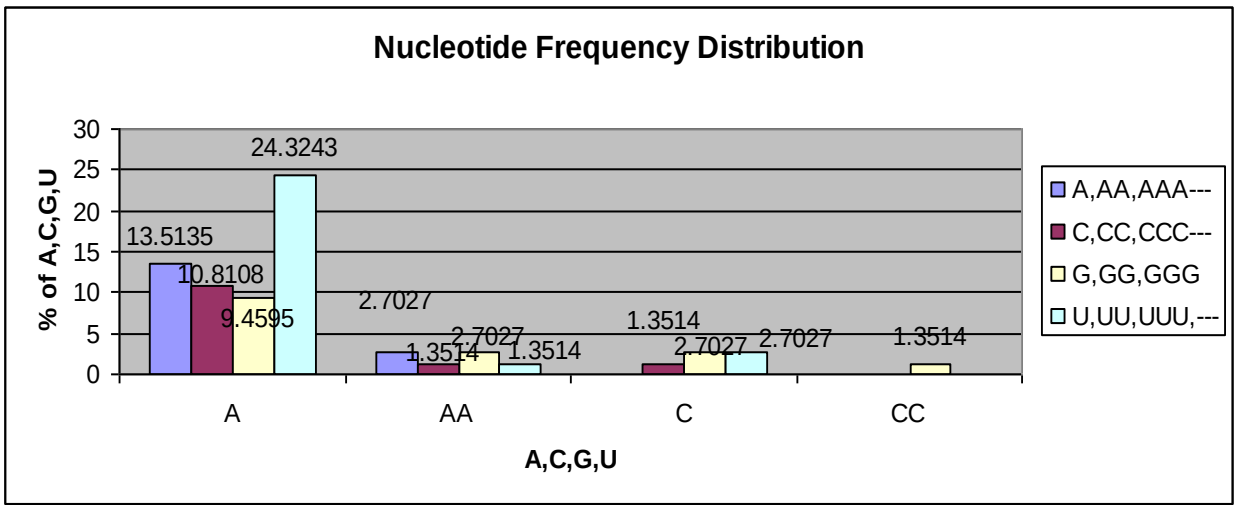

Figure 6. Percentage of different nucleotide subsequences in first microRNA string ggo(2.txt)

The numbers here show the percentage of every subsequence of nucleotide. The Y axis thus is the percentage of subsequences of nucleotides obtained in that particular sequence. The subsequence for which there is no corresponding bar graph, it means that there are no such subsequences in that particular sequence.

Also here is the over all chart for the maximum number of nucleotide subsequences found (single, di, tri etc. for each A,C,G,U) for the all the 1600 text sequences of miRNAs in hsa.

Total size of neucleotides=136756

Pattern	Frequency	Percentage
A	17	0.01243
AA	1125	0.82263
AAA	359	0.26251
AAAA	160	0.11700
AAAAA	47	0.03437
AAAAAA	11	0.00804
AAAAAAA	1	0.00073
AAAAAAAA	1	0.00073
AAAAAAAAA	2	0.00146
AAAAAAAAAA	1	0.00073
C	4835	3.53549
CC	1240	0.90672
CCC	390	0.28518
CCCC	87	0.06362
CCCCC	23	0.01682
CCCCCC	5	0.00366
G	5104	3.73219
GG	1491	1.09026
GGG	458	0.33490
GGGG	121	0.08848
GGGGG	39	0.02852
GGGGGG	5	0.00366
GGGGGGG	1	0.00073
U	5649	4.13071
UU	1303	0.95279
UUU	450	0.32905
UUUU	168	0.12285
UUUUU	66	0.04826
UUUUUU	11	0.00804
UUUUUUU	6	0.00439
UUUUUUUUU	2	0.00146

Likewise here is the over all chart for the maximum number of nucleotide subsequences found (single, di, tri etc. for each A,C,G,U) for the all the 322 text sequences of miRNAs in ggo.

Total size of neucleotides=34453

Pattern	Frequency	Percentage
A	10	0.02903
AA	318	0.92300
AAA	100	0.29025
AAAA	39	0.11320
AAAAA	7	0.02032
AAAAAA	1	0.00290
AAAAAAAA	1	0.00290
AAAAAAAAA	2	0.00581
C	1503	4.36246
CC	434	1.25969
CCC	133	0.38603
CCCC	39	0.11320
CCCCC	10	0.02903
CCCCCC	2	0.00581
G	1590	4.61498
GG	454	1.31774
GGG	136	0.39474
GGGG	54	0.15674
GGGGG	11	0.03193
GGGGGG	3	0.00871
U	1841	5.34351
UU	375	1.08844
UUU	131	0.38023
UUUU	51	0.14803
UUUUU	15	0.04354
UUUUUU	6	0.01742
UUUUUUU	1	0.00290

Likewise here is the over all chart for the maximum number of nucleotide subsequences found (single, di, tri etc. for each A,C,G,U) for the all the 633 text sequences of miRNAs in ppy.

Total size of neucleotides=58027

Pattern	Frequency	Percentage
A	12	0.02068
AA	696	1.19944
AAA	212	0.36535
AAAA	96	0.16544
AAAAA	13	0.02240

AAAAAA	4	0.00689
AAAAAAAAA	1	0.00172
C	3005	5.17862
CC	797	1.37350
CCC	227	0.39120
CCCC	49	0.08444
CCCCC	13	0.02240
CCCCCC	7	0.01206
CCCCCCC	3	0.00517
G	3239	5.58188
GG	924	1.59236
GGG	303	0.52217
GGGG	66	0.11374
GGGGG	20	0.03447
GGGGGG	2	0.00345
GGGGGGG	1	0.00172
U	3573	6.15748
UU	832	1.43382
UUU	306	0.52734
UUUU	81	0.13959
UUUUU	23	0.03964
UUUUUU	5	0.00862
UUUUUUU	6	0.01034

So it can be said that if for any arbitrary miRNA sequence the nucleotide percentage for single or di or tri or tetra and so on sequence fall within the percentage range given here then it can be a potential choice for that particular species to which it's nucleotide frequency comes closest. From above tables we find out that in all these species frequency single nucleotide U is the maximum and that comes to be as 4.13071 for hsa, 5.34351 for ggo, 6.15748 for ppy.

IV. Research work on few mature miRNA sequences

We have continued our study on mature miRNA sequences.

IV.1 Extracting the experimental dataset:-

Firstly, the mature miRNA sequences of Pan troglodytes species(*ptr*)(580 in number) were downloaded from the famous database, miRBase(version 19, http://www.mirbase.org). Again like the pre-mature miRNAs the mature miRNA sequences were extracted and separated into separate text files like 1.txt, 2.txt, 3.txt and so on. This time the text sequences are much smaller than the corresponding pre-mature miRNA sequence of the same species(Pan troglodytes).

For example, 1.txt will be the file for the miRNA sequence

CAAGUCACUAGUGGUUCCGUUUA

2.txt will be the file for the miRNA sequence

AUAGGCACCAAAAAGCAACAA

3.txt will be the file for the miRNA sequence

UGAGGUAGUAGUUUGUGCUGUU and so on.

But this time the text files contained not the text sequences but instead the binary conversion of the text sequences. Already earlier it is mentioned that we are taking A=0, C=1, G=2, and U=3.
Now the binary format for A, C, G, U, that is, 00, 01, 10, 11 are used to represent the sequence and they are stored in text files. Few linear rules are applied on the entire sequence for few successive times. Thus all the rules are applied on each sequence with each linear rule applied on each sequence and applied for quite a number of times in succession. The text files have the main miRNA sequences and also the successive sequences portraying the effect of the linear rules on every bit of the miRNA sequence.

IV.2 Taking the graphical representations:-

Now after the linear rules are applied on the matrices of binary bits are fed into MATLAB programming language to check the graphical representation of the these bit patterns. These graphical representations are stored for further inference for entropy and fractal dimension values, that is shown in the next subsections. Here are few such graphical representations of the binary bit patterns which are all basically in a longitudinal fashion.

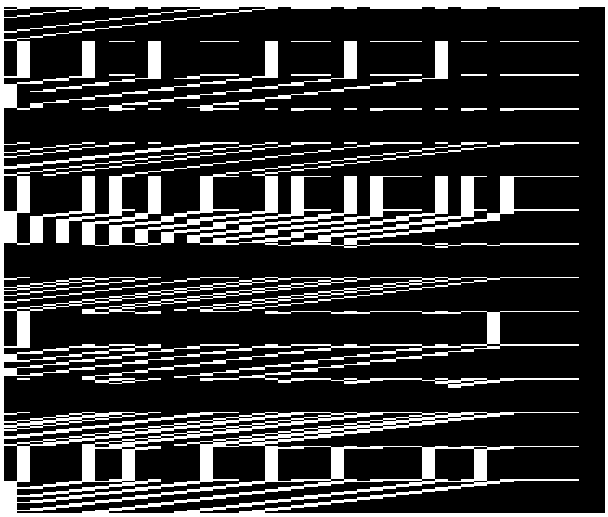

Figure 7: Graphical representation of sequence-1(1.txt)

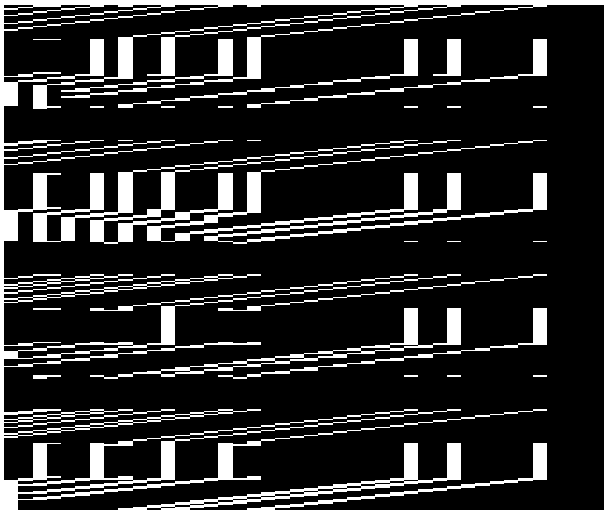

Figure 8: Graphical representation of sequence-2(2.txt)

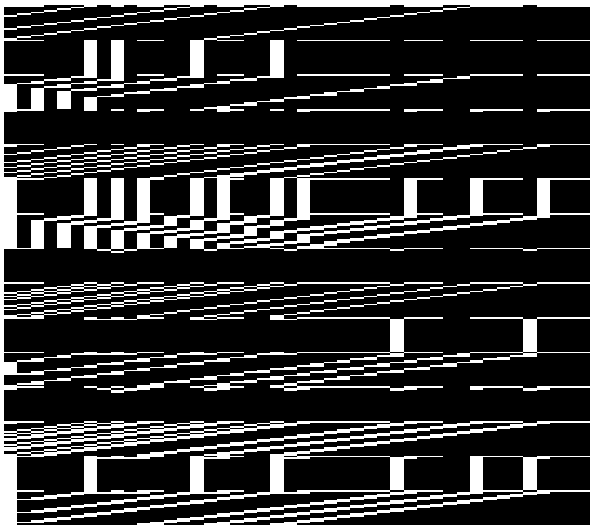

Figure 9: Graphical representation of sequence-3(3.txt)

IV.3 Taking the fractal dimensions:-

Fractal dimension as we see in the case of pre-mature miRNAs also, is hugely effective for studying biological organisms and quantifying them and making biologically important inferences. So the fractal dimension values were calculated for those graphical representations. The range of the fractal dimension values of those are given below in a chart format.

Fractal dimensions of Pan troglodytes miRNA sequences	
Highest value of fractal dimension obtained	Lowest value of fractal dimension obtained
1.09844	1.46946

Table-1: Fractal Dimension

IV.3 Taking the entropy values:-

Just like the statistical parameter fractal dimension, entropy value also gives indication on the biological aspects of the different sequences of the organisms. These are definitely very important making biologically important inferences. So the fractal dimension values were calculated for those graphical representations. The entropy values are given below for ptr species in a chart format. Since there were 580 sequences a … is given to show the continuation of the chart, skipping in between values.

Table-2: Entropy of different sequences

Sequence	Entropy	Sequence	Entropy
1	0.1893	38	0.3478
2	0.3484	39	0.1888
3	0.2707	40	0.4184
4	0.2784	41	0.3123
5	0.3206	42	0.2156
6	0.4016	43	0.2921
7	0.3716	44	0.2841
8	0.3887	45	0.2712
9	0.1914	46	0.2489
10	0.1565	47	0.3028
11	0.2738	48	0.3192
12	0.3298	49	0.1613
13	0.3245	50	0.3267
14	0.354	51	0.3983
15	0.3523	52	0.353
16	0.3782	53	0.2125
17	0.3057	54	0.3838
18	0.3065	55	0.3741
19	0.4042	56	0.2515
20	0.2387	57	0.2716
21	0.2656	48	0.4517
22	0.3712	59	0.2648
23	0.321	60	0.3339
24	0.2492	61	0.3047
25	0.457	62	0.3391
26	0.321	63	0.3283

27	0.2517	64	0.3773
28	0.3209	65	0.457
29	0.4101	66	0.286
30	0.3574	67	0.3386
31	0.2557	68	0.2448
32	0.2822	69	0.426
33	0.3279	70	0.3403
34	0.3107	71	0.3102
35	0.3658	72	0.3998
36	0.3478	73	0.4014
37	0.426	74	0.354
Sequence	Entropy	Sequence	Entropy
75	0.3142	111	0.2549
76	0.2862	112	0.2177
77	0.1959	113	0.3293
78	0.3351	114	0.3484
79	0.2801	115	0.3386
80	0.2784	116	0.2315
81	0.2428	117	0.1649
82	0.3028	118	0.3306
83	0.2762	119	0.2721
84	0.2333	120	0.2561
85	0.3095	121	0.2428
86	0.3281	122	0.321
87	0.4306	123	0.319
88	0.2721	124	0.2469
89	0.3047	125	0.2707
90	0.3641	126	0.3478
91	0.3903	127	0.3209
92	0.3678	128	0.4131
93	0.321	129	0.3208
94	0.1978	130	0.2917
95	0.3489	131	0.4146
96	0.3935	132	0.2145
97	0.3265	133	0.4146
98	0.3152	134	0.3806
99	0.3935	135	0.3933
100	0.3854	136	0.3072
101	0.2598	137	0.2499
102	0.2672	138	0.3736
103	0.2504	139	0.3445
104	0.4138	140	0.4399
105	0.2898	141	0.3523
106	0.2288	142	0.2648
107	0.2296	143	0.2325
108	0.3386	144	0.3028
109	0.2921	145	0.3412
110	0.2104	146	0.2701

Sequence	Entropy	Sequence	Entropy
147	0.3472	184	0.3591
148	0.4165	185	0.291
149	0.2549	186	0.2387
150	0.3766	187	0.3095
151	0.3691	188	0.3386
152	0.3228	189	0.3065
153	0.3887	190	0.3228
154	0.1981	191	0.1442
155	0.3393	192	0.2325
156	0.1543	193	0.356
157	0.2387	194	0.301
158	0.1756	195	0.3445
159	0.264	196	0.3193
160	0.264	197	0.4162
161	0.4146	198	0.3243
162	0.3658	199	0.3138
163	0.1966	200	0.3156
164	0.2249	201	0.2991
165	0.2241	202	0.4272
166	0.2762	203	0.1186
167	0.2529	204	0.2212
168	0.286	205	0.2648
169	0.2637	206	0.3083
170	0.4027	207	0.3138

171	0.2283	208	0.3466
172	0.3887	209	0.3966
173	0.2687	210	0.3506
174	0.2387	211	0.3496
175	0.2954	212	0.2842
176	0.3799	213	0.3028
177	0.3421	214	0.2656
178	0.3114	215	0.301
179	0.2822	216	0.2609
180	0.3966	217	0.3196
181	0.3065	218	0.2999
182	0.435	219	0.3691
183	0.2749	220	0.3316
Sequence	Entropy	Sequence	Entropy
221	0.3008	257	0.2026
222	0.2935	258	0.4415
223	0.3849	259	0.3245
224	0.266	260	0.4248
225	0.3279	261	0.2759
226	0.3083	262	0.2189
227	0.2292	263	0.3316
228	0.1892	264	0.3369
229	0.3076	265	0.3298
230	0.2859	266	0.407
231	0.3281	267	0.3228
232	0.354	268	0.2802
233	0.3152	269	0.3286
234	0.2745	270	0.3935
235	0.1836	271	0.3138
236	0.4071	272	0.3997
237	0.2913	273	0.312
238	0.3506	274	0.2505
239	0.3391	275	0.2749
240	0.3174	276	0.3379
241	0.2099	277	0.3037
242	0.2954	278	0.3773
243	0.3228	279	0.2262
244	0.3695	280	0.3981
245	0.2366	281	0.3687
246	0.2648	282	0.2187
247	0.2543	283	0.4107
248	0.2879	284	0.2598
249	0.1981	285	0.3838
250	0.4169	286	0.321
251	0.3156	287	0.2895
252	0.2915	288	0.2584
253	0.2574	289	0.3192
254	0.1647	290	0.3658
255	0.301	291	0.1673
256	0.379	292	0.2536

Sequence	Entropy	Sequence	Entropy
293	0.2481	330	0.3362
294	0.3386	331	0.3165
295	0.2367	332	0.2656
296	0.4154	333	0.321
297	0.3174	334	0.3691
298	0.3822	335	0.1892
299	0.3377	336	0.3334
300	0.2509	337	0.3228
301	0.3369	338	0.3312
302	0.2598	339	0.3316
303	0.4081	340	0.3174
304	0.4262	341	0.3248
305	0.4522	342	0.3028
306	0.2991	343	0.2363
307	0.3047	344	0.2668
308	0.3382	345	0.3245
309	0.3069	346	0.2648
310	0.3028	347	0.2536
311	0.2407	348	0.3114
312	0.2926	349	0.2695
313	0.3083	350	0.2609
314	0.2589	351	0.3171
315	0.3114	352	0.3069

316	0.2954	353	0.343
317	0.3506	354	0.286
318	0.3107	355	0.3138
319	0.1467	356	0.2047
320	0.2365	357	0.2988
321	0.1573	358	0.3573
322	0.438	359	0.3065
323	0.3047	360	0.3837
324	0.2749	361	0.094
325	0.3039	362	0.1981
326	0.3608	363	0.3281
327	0.2648	364	0.3741
328	0.4042	365	0.2859
329	0.3016	366	0.3868
Sequence	Entropy	Sequence	Entropy
367	0.2241	403	0.3724
368	0.4476	404	0.3608
369	0.3687	405	0.2656
370	0.2569	406	0.4029
371	0.5102	407	0.3591
372	0.286	408	0.353
373	0.301	409	0.3065
374	0.2448	410	0.3832
375	0.2618	411	0.4795
376	0.3276	412	0.2256
377	0.2108	413	0.3838
378	0.4245	414	0.413
379	0.2347	415	0.5147
380	0.3192	416	0.2619
381	0.4275	417	0.3114
382	0.3592	418	0.3047
383	0.3557	419	0.3228
384	0.3369	420	0.354
385	0.3263	421	0.2406
386	0.4065	422	0.3935
387	0.321	423	0.2762
388	0.3806	424	0.3455

389	0.321	425	0.3502
390	0.3557	426	0.2628
391	0.296	427	0.2999
392	0.4291	428	0.3757
393	0.3038	429	0.3168
394	0.3228	430	0.4027
395	0.2637	431	0.243
396	0.3757	432	0.2786
397	0.3228	433	0.2509
398	0.3838	434	0.3699
399	0.2898	435	0.3806
400	0.39	436	0.4096
401	0.3142	437	0.2687
402	0.2026	438	0.2765

Sequence	Entropy	Sequence	Entropy
439	0.1661	476	0.3537
440	0.3379	477	0.2999
441	0.379	478	0.2652
442	0.3781	479	0.2648
443	0.4024	480	0.4582
444	0.2745	481	0.3047
445	0.2432	482	0.3838
446	0.4098	483	0.4032
447	0.291	484	0.2917
448	0.3019	485	0.2241
449	0.286	486	0.3245
450	0.3875	487	0.4016
451	0.3351	488	0.3138
452	0.3838	489	0.2716
453	0.3138	490	0.3649
454	0.145	491	0.4107
455	0.1034	492	0.3966
456	0.3002	493	0.189
457	0.3844	494	0.2609
458	0.2648	495	0.1592
459	0.1891	496	0.2898

460	0.3138	497	0.222
461	0.2633	498	0.3228
462	0.2966	499	0.2069
463	0.2387	500	0.2822
464	0.3316	501	0.2099
465	0.3357	502	0.3138
466	0.301	503	0.2991
467	0.2489	504	0.3246
468	0.3174	505	0.3076
469	0.3114	506	0.2888
470	0.3438	507	0.3742
471	0.3209	508	0.3627
472	0.3591	509	0.3412
473	0.3114	510	0.2234
474	0.3489	511	0.3159
475	0.366	512	0.3102
Sequence	Entropy	Sequence	Entropy
513	0.4065	549	0.6942
514	0.3325	550	0.7276
515	0.4242	551	0.7288
516	0.3903	552	0.7825
517	0.4306	553	0.7431
518	0.2366	554	0.6761
519	0.3458	555	0.707
520	0.3724	556	0.7318
521	0.1226	557	0.6956
522	0.2637	558	0.7039
523	0.2262	559	0.6669
524	0.4101	560	0.5788
525	0.3281	561	0.658
526	0.3997	562	0.6846
527	0.3741	563	0.719
528	0.2681	564	0.6263
529	0.379	565	0.7347
530	0.4076	566	0.6431
531	0.3083	567	0.5504
532	0.4086	568	0.6727
533	0.1165	569	0.737
534	0.2721	570	0.5844
535	0.1891	571	0.6434
536	0.321	572	0.6128
537	0.3643	673	0.7245
538	0.3028	574	0.671
539	0.2509	575	0.589
540	0.3981	576	0.7155
541	0.3591	577	0.7109
542	0.4092	578	0.6649
543	0.3724	579	0.5554
544	0.3047	580	0.7116
545	0.3478		
546	0.3455		
547	0.222		
548	0.3028		

V. Summary

Proper assessment of microRNA (miRNA) string sequence is evolving dynamically from research point of view. MiRNAs play important roles in cell proliferation, cell death, hematopoiesis, oncogenesis, cell differentiation, fat metabolism, growth control and much more. [7]. 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].

Now our present work will help in classifying and quantifying the nucleotide strings of pre-mature miRNAs of the three nearly biologically similar organisms (i) Homo sapiens (hsa), (ii) Gorilla gorilla (ggo) and (iii) Pongo pygmaeus (ppy) in the light of few 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 studying pre-mature miRNAs, to be exactly sure about the identification and classification of any unknown string of miRNA. So quantification of future unknown miRNAs under a biologically specific group of either hsa or ppy or ggo will be less time and money consuming, if the present biological methods has the backing of our mathematical method. Further the study and generation of computational mutation effects on the sequences of mature miRNAs is an added extension of the above statistical study. This will enrich the knowledge on mature miRNAs which are the ultimate goal of study and experiment for most biologists.

VI.Conclusion and Future scope :

In this work, the inherent mathematical behavior of pre-mature miRNAs were deciphered through some characterizing parameters. Three statistical parameters were used to quantify and classify these premature miRNA strings. This study would help in understanding the behavior of the strings of pre-mature miRNAs of the three organisms, (i) Homo sapiens (hsa), (ii) Gorilla gorilla (ggo) and (iii) Pongo pygmaeus (ppy). 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 reject those candidates to be a pre-mature miRNA without the requirement of any complex, expensive and time-consuming biological experiment will be necessary for that, if they do not fall in the range of values that the authors have obtained for the three organisms' pre-mature miRNA sequence. However, for assuring that candidate to be a pre-mature one would definitely need the help of biology. Also for the mature miRNA sequences the entropy value and fractal dimensions were calculated after applying some linear rules successively on the mature miRNA sequences. So as of now the future work is to extend this project work and include other 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. Also for mature miRNA sequences the calculated fractal dimensions and entropy values will be kept as standards to study the nature and characteristics of the sequences of the mature miRNAs already discovered and the ones that are yet to be discovered.

References:

[1] Got target? Computational methods for microRNA target prediction, Hyeyoung Min and SungrohYoon, ExpMol Med. April 30; 42(4): 233-244.

[2] MTar: a computational microRNA target prediction architecture for human transcriptome, Vinod Chandra, ReshmiGirijadevi, Achuthsankar S Nair, Sreenadhan S Pillai and Radhakrishna M Pillai,BMC Bioinformatics 2010, 11(Suppl 1):S2

[3]http://www.bearcave.com/misl/misl_tech/wavelets/hurst/

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

[5] C. Cattani (2010) Fractals and Hidden Symmetries in DNA. Mathematical Problem in Engineering, Vol-2010. 507056.

[6] Human MicroRNA Targets, Bino John, Anton J. Enright, Alexei Aravin, Thomas Tuschl, Chris Sander, Debora S. Marks,PLoSBiol 2(11): e363(2004).

[7] Advances in microRNAs: implications for gene therapists., Marquez RT, McCaffrey AP., Hum Gene Ther. 2008;19:27–38.

[8] MicroRNA target prediction by expression analysis of host genes, Vincenzo Alessandro Gennarino, Marco Sardiello, Raffaella Avellino, Nicola Meola, VincenzaMaselli, SantoshAnand, Luisa Cutillo, Andrea Ballabio, and SandroBanfi,Genome Res, v.19(3); Mar 2009.

Author Biographies

Joyshree Nath passed M.Tech(IT) from C.U. in 2012. Now she is working as Junior Research Fellow at Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. She is involved in research work in microRNA genre of Bioinformatics. She has presented papers in International conferences in India and in abroad.

Asoke Nath is the Associate Professor in Department of Computer Science, St. Xavier's College(Autonomous), Kolkata. Apart from his teaching assignment he is involved with various research work in Cryptography, Steganography, Green Computing, E-learning. He has presented papers in International conferences in India and abroad