MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **83** (2020) 357-372

ISSN 0340 - 6253

The Variations of Human miRNAs and Ising Like Base Pairing Models

Jyoti Prasad Banerjee*[1], Jayanta Kumar Das*†[2,3], Pabitra Pal Choudhury[2], Sayak Mukherjee[4], Sk. Sarif Hassan[5], Pallab Basu†[6]

[1]*National Centre for Biological Sciences, TIFR, Bangalore*-560065, *India*
[2]*Applied Statistics Unit, Indian Statistical Institute, Kolkata*-700108, *India*
[3]*Department of Pediatrics, School of Medicine, Johns Hopkins University, MD-21205, USA*
[4]*Institute of Bioinformatics and Applied Biotechnology, Bangalore*-560100, *India*
[5]*Department of Mathematics, Pingla Thana Mahavidyalaya, Paschim Medinipur-722140, West Bengal, India*
[6]*School of Physics, Witwatersrand University, Johannesburg, South Africa*

(Received May 13, 2019)

Abstract

miRNAs are small about 22-base pair long, RNA molecules are of significant biological importance. Like other longer RNA molecules, messages in miRNAs are words in an alphabet consists of only four nucleotide bases. However, just like words in any language, not all combinations of these alphabets are not meaningful. In fact, we find that the distributions of nucleotides bases in human miRNAs show significant deviation from randomness. First, a miRNA sequence containing four bases are mapped into a binary string with three kinds of classifications according to their chemical properties. Our analyses based on some statistical measurement clearly demonstrate that the purines-pyrimidines class of nucleotide bases resembles all the human miRNAs. Then, we propose a simple nearest neighbor model (Ising model) to understand the statistical variations in human miRNAs. One the way, we also discuss the limitations of a mean-field model.

*Joint first authors

†Corresponding authors: dasjayantakumar89@gmail.com (Jayanta Kumar Das), pallabbasu@gmail.com (Pallab Basu)

1 Introduction

Micro-RNAs (miRNAs) are small non-coding RNA molecules. They are made of the different permutations of nucleotide bases A, U, G and C. The number of bases in the miRNA is conserved across the species and they vary around 22 nucleotides (mostly). They play important role as the regulators of gene expression and they target some of messenger-RNAs (mRNAs) [2, 6, 9, 22, 32]. Mainly, the function of miRNAs is to down-regulate gene expression [10]. In recent years, the effects of miRNAs are also found in malignant cells; the miRNAs influence numerous cancer-relevant processes in a malignant cell [11, 12]. The set of miRNAs and their organization follow some kind of conservative patterns [13, 16, 17, 23], but their functional behaviour are very complex which is clearly evident in the studies of miRNAs in various diseases. Therefore, in-depth analysis of the scope and diversity of these rapidly regulatory molecules would definitely help the cutting-edge medical therapies of tomorrow.

There may be three possible grouping of nucleotide base pairing based on their chemical properties [13]. Each of the groups has important role that are observed over various datasets [7, 8, 18–21]. It will be worthy if the governing role of miRNAs could be apprehended from nucleotide bases and their chemical properties. The recent study have demonstrated how the miRNAs of five species are distributed on the basis of purine-pyrimidine bases utilizing different mathematical parameters [7, 8, 13]. To have a deeper understanding of the nature of miRNAs, here we study a few statistical properties of the miRNAs and propose some simple physics-inspired models to account for their ensemble properties. As we will discuss, our results may shade some light on how the miRNAs are produced in the cell.

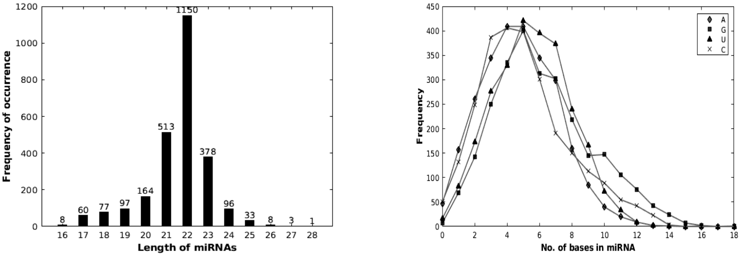
It is to be mentioned that models motivated by statistical physics have varied interdisciplinary applications. For an example, the Ising model has been used to model the long-range correlations, the kinetics of compaction and decompaction in DNA [24, 25], evolutionary processes and its relation to statistical biophysics and evolutionary genetics [26, 27], functional phylogeneies and gene regulation [28–31].

The rest of the paper is organized in the following manner. The Section 2 is the results and discussion that provides mathematical analysis of the datasets of human miRNA sequences in different subsections serially. Here we have discussed the variation of base pairing of nucleotide over miRNAs, their deviations from randomness, nearest neighbour model fittings and clustering probabilities. The Section 3 is the concluding remarks.

2 Results and Discussions

2.1 Dataset specification

The whole miRNA sequences (for *Human*, Hominidae family) are accessable through the website (a miRNA database: **http://www.mirbase.org/**). Their variation within mature miRNAs might be critical for normal miRNA regular activity and also important for understanding the structure and functions [23]. In human, 2588 mature miRNAs sequences are reported in release version 21. Their lengths vary from 16 to 28 nucleotide bases, but mostly within the range of 20-24 bases (Figure 1(a)).



(a) The frequency of different length miRNAs. **(b)** The frequency distribution of four nucleotides bases (A, G, U and C).

Figure 1. The length variations and average frequency distribution of four nucleotides of 2588 miRNAs.

2.2 Individual variation of nucleotides bases

Let L be the data set containing 2588 number of miRNAs i.e. $L = 2588$. Say, L_m is the average (or mean) length of each miRNA calculated over 2588, and we find $L_m = 21.588$. There are only four ribonucleotide bases which are distributed over the miRNAs. Here, we will see their distribution in terms of the probability measure.

Let, $Pr(X)$ is the probability of a nucleotide X ($X \in \{A, G, U, C\}$) to be present in the 2588 miRNAs, then

$$Pr(X) = \frac{f_X}{L_m} \tag{1}$$

where f_X is the average number of occurrence of the nucleotide X over 2588 miRNAs.

The average frequency distribution of all 2588 miRNAs are shown in Figure 1(b). From the data set, we find $f_A = 4.77$, $f_G = 6.21$, $f_U = 5.55$ and $f_C = 5.53$. So, $Pr(A) = 0.22$, $Pr(G) = 0.28$, $Pr(U) = 0.26$ and $Pr(C) = 0.24$. It is observed that $Pr(A) \neq Pr(G) \neq Pr(U) \neq Pr(C)$. Therefore, we find the individual variations of each nucleotide base over the 2588 miRNAs. And the order is $Pr(A) < Pr(C) < Pr(U) < Pr(G)$.

2.3 Grouping of nucleotide bases and transformations into binary

The four-letter alphabets (A/G/U/C) may be mapped onto two-letter (binary) alphabets in three distinct possible ways. As it turns out, these three classifications may be understood on the basis of chemical properties of nucleotide bases [1]. These are:

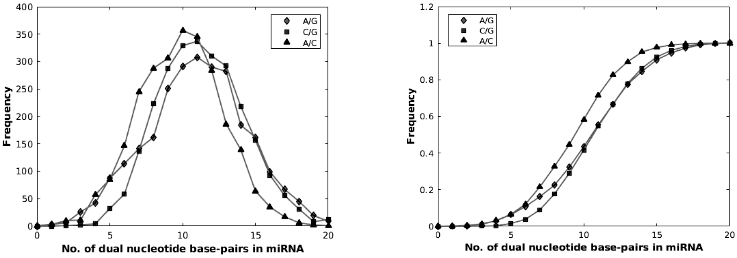
- **Purine-Pyrimidine classification:** The bases A/G and U/C are the purine and pyrimidine (Pu-Py) groups respectively.
- **Strong-Week H-bond classification:** The bases C/G and A/U are the strong H-bond and week H- bond (St-We) groups respectively.
- **Amino-Keto classification:** The bases A/C and G/U are the amino and keto (Am-Ke) groups respectively.

Among the three classifications, the purine-pyrimidine class is possibly the most important one [5, 7, 8]. Purines are the most widely occurring nitrogen-containing heterocyclic compounds in nature. In order to form DNA and RNA, both purines and pyrimidines are needed by the cell in approximately equal quantities. But, this may not hold true for miRNAs.

For all of these three groupings, we can assign the bits 1's and 0's to the respective members of any of the aforementioned classifications. For example, for the purine-pyrimidine class we may have the following rule:

$$X = \begin{cases} 1, & \text{if } X \in \{A, G\} \\ 0, & \text{if } X \in \{U, C\} \end{cases} \tag{2}$$

We calculate mean occurrence for all the different classification. From the dataset, we get $M(Pu) = 10.99$ and $M(Py) = 10.59$ for purine-pyrimidine grouping, $M(St) = 11.27$



(a) The occurrence distribution. **(b)** The cumulative distribution function plot.

Figure 2. The occurrence distribution and cumulative distribution function (cdf) plot of Purine (A/G), Strong (C/G) and Amino (A/C) classes.

and $M(We) = 10.31$ for strong-week H-bond grouping, $M(Am) = 9.84$ and $M(Ke) = 11.74$ for amino-keto grouping. The occurrence distribution of Pu, St and Am is shown in Figure 2(a). The cumulative distribution function for these three groupings are plotted in Figure 2(b).

Interestingly, for all three categories, the distribution of occurrence of 0 (or 1) in the binary string representation of miRNAs fits well to a normal distribution. The group-specific distributions of the nucleotide base pairs vary significantly from each other. However, we observe that the mean occurrence of any member (Pu or Py) in the Purine-Pyrimidine classification is the closest to the occurrence averaged over all the members of all the three classifications (10.79), as compared to other two groupings. The mean and variance for each members of all the classifications are provided in the Table 1. The question that naturally arises is: which kind of model best describes the observed variances in the distribution of nucleotide bases?

2.4 Positional correlations in miRNAs and limitations of a binomial model

The simplest model is to start with is a binomial model. This is an iid (independent and identically distributed) model. In a binomial model, there is no correlation between locations in the string and probability of occurrence of a 0 (or 1) at each location is determined by an independent coin-toss with a probability p (or q) with $p+q=1$. Hence in a binomial model, Pu (Py) s may occur at each location independently and occurrence of a base in one location does not affect any other part of the micro-RNA.

The mean and the variance in a binomial distribution are related. The mean of an binomial distribution is given np, where 'n' is the number of trial (or coin toss). The binomial distribution has a simple property that variance is npq (Theoretical/Expected). Hence from the data we can estimate the mean and p, and check if the variance calculated from the data is close to the observed variance. (Here, $n = L_m = 21.588$). The details of Expected variance and observed variance for the bases of each classifications are shown in Table 1.

Table 1. Expected variance and observed variance in three classes.

Class	n	p	q	Mean (np)	Variance (npq)	Observed Variance
Pu-Py	21.59	0.509	0.491	10.988	5.40	11.7
St-We	21.59	0.522	0.478	11.268	5.39	8.48
Am-Ke	21.59	0.455	0.544	9.839	5.35	8.26

So, from the above results we find that the observed variance is greater than the variance obtained from the binomial model. However, the discrepancy in variance for the Pu-Py classification is much larger than that for the other classifications. For all other classes the expected variance is much closer to the observed variance. So, there are patterns of miRNAs where one group (Pu or Py) is more frequent than what is expected from chance alone. Therefore, it is interpreted that miRNAs have a stronger interdependency with respect to the purine-pyrimidine class. This means that the occurrence of Pu-Py group at different sites are correlated. To be noted is that, the existence of positional correlation also shows up in the direct computation of cluster probabilities, which we have performed in the Subsection 2.5.3.

To have a better statistical understanding of our claim that observed mean is different from the mean calculated from observed variance assuming binomial distribution, we performed one-sample t-test (h) with the null hypothesis that two observed mean are equal. We performed the t-test separately for all three cases and calculated the p-value. In all three cases we can reject the null hypothesis with high significance (Table 2).

2.5 Modeling the positional interactions in MiRNAs

In the previous subsection we have observed that the variance of the data is more than what is expected of the independent, binomially distributed binary strings. It is natural to guess that our assumption of independence was wrong and occurrence of elements at different positions in the strings are correlated. Physically, this discrepancy indicates that

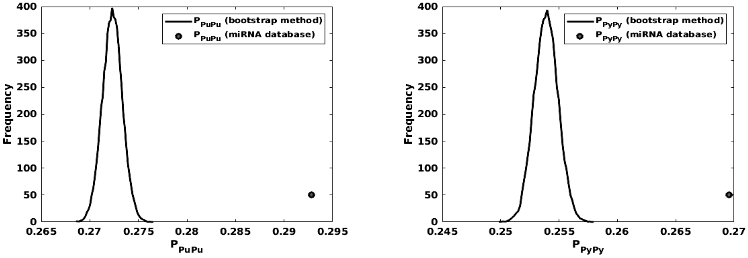
Table 2. Observed mean, Expected mean, *p*-value of each class for different cases.

Class	Cases	Observed mean (O_m)	Expected mean (E_m)	p- value	$\frac{O_m}{E_m}$
Pu-Py	00c	5.54	4.96	0.0024	1.12
	01c	4.55	5.14	0.0000	0.89
	10c	4.47	5.14	0.0000	0.87
	11c	6.02	5.33	0.0000	1.13
St-We	00c	4.06	4.7	0.0000	0.86
	01c	5.75	5.14	0.0000	1.12
	10c	5.61	5.14	0.0000	1.09
	11c	5.71	5.61	0.0000	1.02
Am-Ke	00c	6.48	6.12	0.0000	1.06
	01c	4.68	5.12	0.0000	0.91
	10c	4.76	5.12	0.0000	0.93
	11c	4.66	4.26	0.0000	1.09

there could be interactions present within the elements of a string. These interactions may have a mean-field nature where every element of a binary string interacts with every other element within the same string with equal strength (all-to-all connected topology). In a mean field model, the occurrence of a Pu (or Py) at a position in the string increases the probability of occurrence of Pu (or Py) in other positions equally. A mean field model is blind to how 0 (or 1) is clustered in a string and probability that a string would contain N zeros (or ones) is only a function N. In contrast with the mean field model, there might be a strong localized interaction instead of a mean-field one. Accordingly, in a localized model, the occurrence of a Pu (or Py) at a position in the string affects the probability of occurrence of Pu (or Py) only in nearby positions . We would discuss the localized model in the subsection.

2.5.1 Testing Mean field models with miRNA data

First let us check how likely is the mean-field model given the observed data. We start with a null hypothesis (H0): data is well described by a mean-field model. . To test that hypothesis we implement of bootstrap method. We calculate and compare the probabilities of finding 2-clusters (i.e., like-pairs) of Purines and Pyrimidines in the binary strings obtained from the data (shown in Table 4) and the same obtained by the bootstrap method. In the bootstrap method we shuffle the elements of each binary string multiple times keeping the numbers of Purines and Pyrimidines constant. We then calculate the probabilities of a 2-cluster in the shuffled strings and obtain a distribution of them. The catch is that, probabilities calculated from a mean field model is invariant under shuffling. Hence, if the probabilities obtained from the data falls under the bootstrap distribution, preferably close to the pick of the distribution, then one can infer that shuffling does not change the cluster-probabilities, and a mean-field picture is enough to model the statistical properties of the miRNA sequences. The distributions and the values calculated from the data are shown in Figures 3(a) and 3(b).



(a) Purines.

(b) Pyrimidines.

Figure 3. Distribution of probability of 2-clusters of Purines and Pyrimidines from bootstrap method and from the data.

A statistically significant difference between the observed means and values obtained in the bootstrap analysis rules out our null hypothesis and indicates that there are strong position dependent correlations on how Pu (Py) are positioned in a miRNAs. Their special arrangements, therefore, suggest that they have a rather strong position dependent interaction.

2.5.2 Nearest Neighbor Model : Ising Model

Next we would explore models with localized interactions. The simplest among these class of models is a Nearest Neighbour (NN) model. An NN model assumes pair-wise interaction between any two nearest neighbour elements. There is no other interactions present in a NN model. In a NN model, the probability of occurrence of a particular string of length N with r zeros would be given by,

$$P \propto p^r q^{N-r} \exp(-J_{00}N_{00} - J_{11}N_{11} - J_{01}N_{01} - J_{10}N_{10}) \tag{3}$$

where ‘J’s are the strength of nearest neighbour interactions. The above definition of probabilities may be understood as a binomial model modified with nearest neighbour interaction. Now, the only the ratios of probabilities matter. Hence the parameters in our model is p/q and difference between various J' s. We would further assume $J_{00} = J_{11}$ and $J_{01} = J_{10}$. And define new parameters as $\mu = \frac{1}{2}\log(p - q)$ and $J = \frac{1}{2}(J_{00} - J_{01})$. In the physics literature a NN model is known as Ising model. Such models have already been used for various nucleotide sequencing data [4]. In principle one may consider a NN-model, where probability of occurence of each element at different position is different. In the context of Ising model, this is like turning on a position dependent magnetic field. As our miRNAs data is chiefly homogeneous, we won't take that complication into account.

There may be multiple approaches to fit NN model to the observed miRNA data. We would consider two. One is to calculate the mean and variance of the occurrence of 0 (or 1). This approach is equivalent to looking at the microRNA strings from a holistic point of view. The other approach is to calculate exactly the probabilities of n-clusters of a particular state and fit them to the observed values. However, first we would motivate our model by investigating occurrences of various two-clusters $(00, 01, 10, 11)$ in the data.

Table 3. Observed frequency and respective probability of each class for four cases.

Class	Cases	Frequency	Pr(00/01/10/11)
Pu-Py	00c	14341	0.27
	01c	11783	0.22
	10c	11575	0.22
	11c	15583	0.29
St-We	00c	10497	0.20
	01c	14881	0.28
	10c	14524	0.27
	11c	13380	0.25
Am-Ke	00c	16772	0.31
	01c	12112	0.23
	10c	12331	0.23
	11c	12067	0.23

From Table 3 we observe that $P_{00} \approx P_{11}$ and $P_{01} \approx P_{10}$ for all 3 types of classifications. This indicates that it might be sufficient to club the four different interactions to interactions between similar elements and those between the different elements. Also, we know that $P_0 \neq P_1$ (from Table 1) and $P_{00} \neq P_{01}$ (from Table 3) for all of the different classification schemes. These observations together suggest that a NN model can possibly be a good candidate to fit and explain the observed data.

As mentioned, in Physics literature the nearest neighbour model is called the Ising model and is relatively well studied. Historically, the Ising model was created/used to describe the paramagnetic to ferromagnetic transition in 1-dimension. It is a semi-classical model which, in simple terms, assumes that the magnetic moments (spins) of atoms in a 1-d atomistic chain can assume only $+\frac{1}{2}$ (↑) and $-\frac{1}{2}$ (↓) values in some suitable unit system along a particular (say, Z) axis. We would use results from physics literature to find out the cluster probabilities and other statistics of our NN model.

2.5.3 Cluster probabilities

As mentioned earlier, we can verify the exactness of our model by checking the probabilities for high order clusters (11 and 00, 111 and 000 and so on). This is shown in Table 4. (Pu, St, Am) for 1's and (Py, We, Ke) for 0's. And the corresponding 2-D line plot is shown in Figure 4 in two different styles for (Pu, St, Am) and (Py, We, Ke). The $Pr(r)$ values ($r > 4$) for Pu-Py class is more than the other classes.

Table 4. The calculated probability ($Pr(r)$) for different lengths strings, where r is the string of 1's or 0's.

	Pr(r)							
r=	1	11	111	1111	11111	111111	1111111	11111111
Pu	0.509	0.292	0.173	0.104	0.062	0.038	0.025	0.016
St	0.522	0.251	0.122	0.061	0.032	0.019	0.012	0.008
Am	0.455	0.226	0.112	0.056	0.028	0.015	0.008	0.005
r=	0	00	000	0000	00000	000000	0000000	00000000
Py	0.491	0.269	0.157	0.097	0.061	0.039	0.027	0.019
We	0.478	0.197	0.091	0.043	0.019	0.008	0.004	0.002
Ke	0.545	0.314	0.179	0.102	0.059	0.014	0.020	0.012

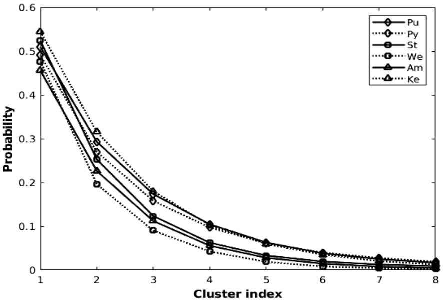


Figure 4. The probability distribution for the different lengths (1 to 8) strings of ones and zeros.

If the length (N) of the string is large, the 1-cluster probability becomes (For the derivation and closed form expressions for higher-order clusters see App. A.2),

$$P_0 = 0.5 \left(\frac{e^{2J}\left(e^{2\mu} - 1\right)}{\sqrt{e^{4(J+\mu)} - 2e^{4J+2\mu} + e^{4J} + 4e^{2\mu}}} + 1 \right) \tag{4}$$

with

$$P_{11..} = \lim_{\mu \to -\mu} P_{00..}$$

Equations 4 and 18 (see App. A.2) are solved with the observed probabilities for Pu-Py classification which give, $\mu = -0.0143414$ and $J = 0.113675$. Feeding these values in eqn.s 19 and 20 we find, $P_{000} = 0.147573$ (6% deviation from the observed value) and $P_{0000} = 0.0809581$ (16.5% deviation from the observed value). However, doing the same thing with equations and probabilities for 1-clusters gives, $P_{111} = 0.167168$ (3.4% deviation from the observed value) and $P_{1111} = 0.0957027$ (8% deviation from the observed value).

2.5.4 Higher cluster probabilities

In their paper [15] Ivanytskyi & Chelnokov have calculated the exact expression for average occupancy number of *distinct* l-clusters in Ising model. The average occupancy number can be written as,

$$\langle n_{l \le N} \rangle = \frac{\delta_{N,l} + 2e^{-2J} + (N - l + 1)e^{-4J}}{2(1 + e^{-2J})^{l+1}} \tag{5}$$

Now, the number of all distinct and non-distinct q-clusters , $\tilde{n}_{q \le N}$, can be related to the number of all distinct clusters as,

$$\tilde{n}_{q \le N} = \sum_{l=q}^{N} (l - q + 1)\langle n_{l \le N} \rangle \tag{6}$$

Solving equation 6 for $q = 2$ and p_{00} from Table 4 we get, $J = 0.1728697$. Using this J in equations 5 and 6 we predict p_{000} from this model to be 0.1714543, which deviates from the observation (Table 4) by 1.05% only.

2.5.5 Variance and the limitations of NN model

In App. A.1 we calculate a frequently used physical quantity, known as the isothermal susceptibility, for the Ising chain for finite N. This quantity can be argued to be identical to the variance of difference between the population of two classes (i.e. $\chi_T \equiv \frac{\sigma^2_{n0-n1}}{N}$). At $\mu \to 0$ limit this variance is given as,

$$\chi_T = e^{2J} \tag{7}$$

Solving for J with $\sigma^2_{data}/N = 39.8470/22$ gives,

$$J = 0.297 \tag{8}$$

which is almost double the value obtained in subsection 2.5.4.

For the strong-weak H-bond grouping, $\sigma^2_{database}/N = 29.4017/22$ gives,

$$J = 0.145 \tag{9}$$

whereas, $J = 0.175$ for the amino-keto grouping. These higher than expected values of J shows that there are some long range interaction present in our data which is not captured well in NN model and NN model only captures a part of the observed variance. In short the data seems to have both NN and mean field like interactions present. As we demonstrated, a NN model of the captures the observed variation of small clusters. Our results indicates a more generalized hybrid model (NN+MF) is needed to take in to account the observed variance.

3 Concluding Remarks

All-inclusive, we observe that the Ising model with nearest neighbour interactions and chemical potential serves as a better descriptor of the miRNA data than the binomial model. We also observe that, within the scope of Ising model, the mean-field like quantities such as variance cannot lead us to a full description of the data. At least for the Pu-Py classification, we observe that there is hardly any difference between the finite and large N Ising models in describing the miRNA data. So the finite number of elements in the string does not play an important role in our approximated description [14].

References

[1] L. Shi, H. Huang, DNA sequences analysis based on classifications of nucleotide bases, in: S. D'Mello, A. Graesser, B. Schuller, J. C. Martin (Eds.), *Affective Computing and Intelligent Interaction*, Springer, Berlin, 2012, pp. 379–384.

[2] L. He, G. J. Hannon, MicroRNAs: small RNAs with a big role in gene regulation, *Nature Rev. Gen.* **5** (2004) 522–531.

[3] L. Flintoft, The wide reach of microRNAs, *Nature Rev. Gen.* **6** (2005) #164.

[4] A. Colliva, R. Pellegrini, A. Testori, M. Caselle, Ising-model description of long-range correlations in DNA sequences. *Phys. Rev. E* **91** (2015) #052703.

[5] Z. A. Shabarova, A. A. Bogdanov, *Advanced Organic Chemistry of Nucleic Acids*, Wiley, New York, 2008.

[6] J. Li, Z. Zhang, miRNA regulatory variation in human evolution, *Trends Gen.* **29** (2013) 116–124.

[7] B. A. Moffatt, H. Ashihara, Purine and pyrimidine nucleotide synthesis and metabolism, *Arabidopsis Book* (2002) #e0018.

[8] M. M. Daly, V. G. Allfrey, A. E. Mirsky, Purine and pyrimidine contents of some desoxypentose nucleic acids, *J. Gen. Phys.* **33** (1950) #497.

[9] V. Ambros, The functions of animal microRNAs, *Nature* **431** (2004) #350.

[10] L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, J. M. Johnson, Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs, *Nature* **433** (2005) #769.

[11] G. A. Calin, C. M. Croce, MicroRNA signatures in human cancers, *Nature Rev. Can.* **11** (2006) #857.

[12] L. A. MacFarlane, P. R Murphy, MicroRNA: biogenesis, function and role in cancer, *Curr. Gen.* **11** (2010) 537–561.

[13] J. K. Das, P. P. Choudhury, A. Chaudhuri, S. S. Hassan, P. Basu, Analysis of purines and pyrimidines distribution over miRNAs of human, gorilla, chimpanzee, mouse and rat, *Sci. Rep.* **8** (2018) #9974.

[14] R. K. Pathria, P. D. Beale, Statistical mechanics, *Butter Worth* (1996) #32.

[15] A. Ivanytskyi, V. Chelnokov, On bimodal size distribution of spin clusters in the onedimensional Ising model, *EPJ Web Conf.* **182** (2018) #03004.

[16] Y. Altuvia, P. Landgraf, G. Lithwick, N. Elefant, S. Pfeffer, A. Aravin, M. J. Brownstein, T. Tuschl, H. Margalit, Clustering and conservation patterns of human microRNAs, *Nucleic Acids Res.* **33** (2005) 2697–27706.

[17] E. Berezikov, V. Guryev, J. van de Belt, E. Wienholds, R. H. Plasterk, E. Cuppen, Phylogenetic shadowing and computational identification of human microRNA genes, *Cell* **120** (2005) 21–24.

[18] K. Tamura, M. Nei, Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees, *Mol. Biol. Evol.* **10** (1993) 512–526.

[19] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, A. L. Goldberger, Mosaic organization of DNA nucleotides, *Phys. Rev. E* **49** (1994) #1685.

[20] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H. E. Stanley, Long–range correlations in nucleotide sequences, *Nature* **356** (1992) #168.

[21] C. Fonseca Guerra, F. M. Bickelhaupt, Orbital interactions in strong and weak hydrogen bonds are essential for DNA replication, *Angew. Chem. Int. Ed.* **41** (2002) 2092–2095.

[22] A. Stepien, K. Knop, J. Dolata, M. Taube, M. Bajczyk, M. Barciszewska-Pacak, A. Pacak, A. Jarmolowski, Z. Szweykowska-Kulinska, Posttranscriptional coordination of splicing and miRNA biogenesis in plants, *WIREs: RNA* **8** (2017) #e1403.

[23] K. Rolle, M. Piwecka, A. Belter, D. Wawrzyniak, J. Jeleniewicz, M. Z. Barciszewska, J. Barciszewski, The sequence and structure determine the function of mature human miRNAs, *PloS One* **11** (2016) #e0151246.

[24] S. Ghosh, Ising model forB-Z transition in supercoiled DNA *Bull. Math. Biol.* **54** (1992) 727–732.

[25] N. N. Vtyurina, D. Dulin, M. W. Docter, A. S. Meyer, N. H. Dekker, E. A. Abbondanzieri, hysteresis in DNA compaction by Dps is described by an Ising model, *Proc. Natl. Acad. Sci.* **113** (2016) 4982–4987.

[26] I. Leuthäusser, An exact correspondence between Eigen's evolution model and a two-dimensional Ising system, *J. Chem. Phys.* **84** (1986) 1884-1885.

[27] K. Shekhar, C. F. Ruberman, A. L. Ferguson, J. P. Barton, M. Kardar, A. K. Chakraborty, Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes, *Phys. Rev. E* **88** (2013) #062705.

[28] G. Sella, A. E. Hirsh, The application of statistical physics to evolutionary biology, *Proc. Natl. Acad. Sci.* **102** (2005) 9541–9546.

[29] V. Mustonen, M. Lässig, Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies, *Proc. Natl. Acad. Sci.* **102** (2005) 15936–15941.

[30] M. Lässig, From biophysics to evolutionary genetics: statistical aspects of gene regulation, *BMC Bioinf.* **8** (2007) #S7.

[31] B. Obermayer, E. Levine, Exploring the miRNA regulatory network using evolutionary correlations, *PLOS Comput. Biol.* **10** (2014) #e1003860.

[32] L. Guo, T. Liang, J. Yu, Q. Zou, A comprehensive analysis of miRNA/isomiR expression with gender difference, *PloS One* **11** (2016) #e0154955.

A Appendix

A.1 Nearest neighbor model

Here we solve the finite N Ising chain using transfer matrix method [14]. The transfer matrix in this problem is given by,

$$T = \begin{pmatrix} e^{(J+\mu)} & e^{-J} \\ e^{-J} & e^{(J-\mu)} \end{pmatrix} \tag{10}$$

with the partition function, Z as,

$$Z = \lambda_+^N + \lambda_-^N \tag{11}$$

where

$$\lambda_\pm = e^J \cosh(\mu) \pm \sqrt{e^{2J}\sinh^2(\mu) + e^{-2J}} \tag{12}$$

At $N \to \infty$, the larger eigenvalue (λ_+) dominates and the free energy density is given as,

$$\begin{aligned} f &= -\ln(\lambda_+) \\ &= \ln(e^J \cosh(\mu) + \sqrt{e^{2J}\sinh^2(\mu) + e^{-2J}}) \end{aligned} \tag{13}$$

The isothermal susceptibility for this model is given by

$$\begin{aligned} \chi_T &= -\frac{\partial^2 f}{\partial h^2}\bigg|_T \\ &= \frac{e^J \cosh(\mu)}{\sqrt{e^{2J}\sinh^2(\mu) + e^{-2J}}} - \frac{e^{3J}\sinh^2(\mu)\cosh(\mu)}{[e^{2J}\sinh^2(\mu) + e^{-2J}]^{3/2}} \end{aligned} \tag{14}$$

A.2 Cluster Probabilities for large N

The cluster probabilities can be obtained using the transfer matrix method. For an n-cluster of 00.. the probability is obtained by using

$$P_{00..} = V_p^{-1} O V_p / \lambda_p^{n+1} \tag{15}$$

where λ_p is the largest eigenvector of the transfer matrix, T, and V_p is the corresponding eigenvector.

For $n = 2$,

$$O = \begin{pmatrix} e^{(3J+4\mu)} & e^{J+2\mu} \\ e^{J+2\mu} & e^{-J} \end{pmatrix} \tag{16}$$

for $n = 3$,

$$O = \begin{pmatrix} e^{4J+5\mu} & e^{2J+3\mu} \\ e^{2J+3\mu} & e^{\mu} \end{pmatrix} \tag{17}$$

Using these we arrive at the probabilities for n-clusters,

$$P_{00} = \frac{e^{2J+3\mu}AB}{ED^3} \tag{18}$$

where $E = \sqrt{-2\left(e^{4J}-2\right)e^{2\mu} + e^{4(J+\mu)} + e^{4J}}$

$A = \left(E - e^{2(J+\mu)} + e^{2J}\right)$,

$G = e^{2J+\mu}\left(e^{2J}\left(e^{2\mu}-1\right) + E\right)$,

$B = (G+2)^2$,

$D = E + e^{2(J+\mu)} + e^{2J}$

$$P_{000} = \frac{8e^{4J+5\mu}(E+F)}{ED^4} \tag{19}$$

Where $F = e^{2J}\left(e^{2\mu}\left(e^{\mu}(G+4)-1\right)+1\right)$,

$$P_{0000} = \frac{16e^{6J+7\mu}(E+F)}{ED^5} \tag{20}$$