Published online 2014 November 25. **Research Article**

Relation Between RNA Sequences, Structures, and Shapes via Variation Networks

Javad Mohammadzadeh [2], Mohammad Ganjtabesh [1,*], Abbas Nowzari-Dalini [1]

[1]School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, I.R. IRAN
[2]Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, I.R. IRAN

[*]*Corresponding author*: Mohammad Ganjtabesh, School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, I.R. IRAN. Tel: +98-2166412178, E-Mail: mgtabesh@ut.ac.ir

Received: November 16, 2013; **Revised:** May 19, 2014

Background: RNA plays key role in many aspects of biological processes and its tertiary structure is critical for its biological function. RNA secondary structure represents various significant portions of RNA tertiary structure. Since the biological function of RNA is concluded indirectly from its primary structure, it would be important to analyze the relations between the RNA sequences and their structures. One important tool to perform this kind of analysis is the neutral network which is a collection of RNA sequences, all coding the same secondary structure, where each RNA sequence is distinguished from the others by no more than a single base mutation. Another high level and useful representation of an RNA secondary structure is the RNA shape, where it is holding the vicinity and nesting of structural components and reducing their lengths to one unit. This allows us to analyze the huge structural space corresponding to the larger RNA sequences.
Objectives: In this study, a new concept, entitled Variation Network, over the set of all RNA shapes is introduced. Based on this concept, the potential relations between random and natural RNA sequences, as well as their corresponding structures are analyzed.
Materials and Methods: To explore the relations between random and natural RNA sequences and their corresponding structures, different properties including frequency, normalized frequency, shape energy average, variation rate, normalized variation rate, neighborhood energy average, and stability were obtained and analyzed.
Results: The correlations among these properties of random and natural Variation Networks are presented. Base on the obtained correlations, all the employed datasets are highly correlated to each other from the frequency point of view, whereas they are not well correlated from the thermodynamic energy point of view.
Conclusions: Since the thermodynamic energy value of an RNA sequence over its secondary structure plays a key role in its function, this research conclude that the natural RNA sequences are not generated randomly.

Keywords: RNA Folding; RNA Inverse Folding; Single Mutation; Stability

1. Background

RNA plays key role in many aspects of biological processes. The function of RNA is related to its tertiary structure. Since dealing with tertiary structure of RNA is very complicated, RNA secondary structure has been studied in the literature (1-3). A secondary structure of an RNA sequence is basically a set of pairing connections among bases in the sequence, where each base can be paired with at most one another base. RNA secondary structure establishes various significant portions of RNA tertiary structure. Since the biological function of RNA is concluded indirectly from its primary structure, therefore it would be important to analyze the relationship between the RNA sequences and their structures. In this regard, the neutral network would be of great interest (4). Neutral network is a collection of RNA sequences, all coding the same secondary structure and each RNA sequence is distinct from other sequences by no more than a single base mutation (4). Neutral networks consequent to common structures saturate the space of RNA sequences (5, 6) and thus simplify the examination of a huge amount of alternative structures. This is achievable since diverse neutral networks are greatly meshed, i.e. all familiar structures can

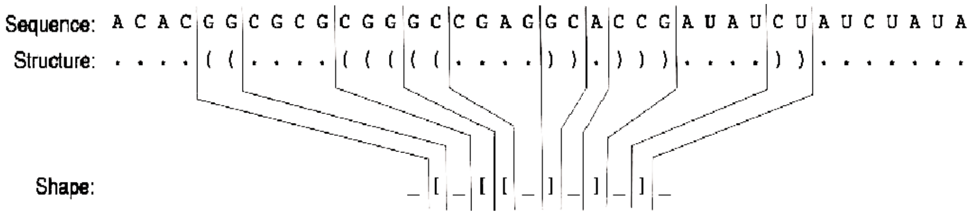
Sequence: A C A C G G C G C G C G G G C C G A G G C A C C G A U A U C U A U C U A U A
Structure: . . . . ((. . . . (((((. . . .)) .))) . . . .)) . . . .
Shape: _ [_ [[_] _] _] _

Figure 1. An example of RNA sequence, its corresponding secondary structure, and its shape (17)

be achieved within a few (mutational) walks starting from any arbitrary sequence (6).

Several structural properties of the RNA neutral networks have been studied previously (4-10), and the remarkably complex structures underlying it have been discovered. An upper bound $Sl=1.4848\times l^{-3/2}\times(1.8488)^{l}$ for the amount of distinct secondary structures for sequences of length l was obtained in (4). This indicates that the expected size of a neutral network develops as $4^{l} / S_{l} = 0.673\times l^{3/2} \times (2.1636)^{l}$ is an enormous quantity even for modest values of l. This expected amount is not illustrative for the real spreading of neutral network sizes, which is a very broad function (4, 11). The space of RNA sequences of length l, which is surrounded in a usual l dimension lattice, is directed by a reasonably small number of common structures which are tremendously plentiful and occur to be located as structural motifs in natural and functional RNA molecules (8, 12).

Among the considerable parts of the works that have been studied, Aguirre *et al*., (13) presented results of specific significance. They concentrated on the topology of RNA neutral networks and studied local and global parameters describing their structures. They have plotted neutral networks over all RNA sequences of length *12*, using the RNAfold as a folding method (14). Then they have obtained the topological properties of these neutral networks. Unfortunately, the information obtained in (13) cannot be generalized to the larger space (natural space) of RNA sequences and structures.

An additional useful representation of an RNA secondary structure is the RNA shape. RNA shape notion plots structure in a compact form, holding vicinity and nesting of structural components and reducing their lengths to one unit. It is motivated by the dot-bracket representation identified from the Vienna RNA package (14). Respect to the behind sequence and secondary structure from folding area in dot-bracket demonstration, the shape attitude proposes five deduction levels prepared in their level of abstraction. Also they shorten the loop and stack sizes, where unpaired areas are symbolized by an underline and stacking areas by a couple of formed brackets (15). Figure 1 represents the relationship between a small RNA sequence, its structure, and its shape. RNA shape can be well participated with dynamic programming algorithms, and consequently it can be employed through structure prediction rather than afterwards. This prevents exponential explosion and can still provide a non-heuristic and complete report of properties of the given RNA folding space (15).

The rest of this paper is organized as follows. In section 2, the short-term objectives to do this research are provided. In Section 3, the basic definitions as well as the Variation Network are presented. In section 4, the details of datasets construction and different measures are discussed. The obtained results and conclusions are presented in sections 5 and 6, respectively.

2. Objectives

In this paper, a new concept entitled Variation Network, which is based on the RNA shapes is introduced. Although the RNA shapes are obsolete in previous studies, here we have a special attitude. Based on the proposed Variation Network, different measures from frequency point of view, including frequency and variation rate, and thermodynamic energy point of view, including shape energy, neighborhood energy, as well as the stability are obtained in this paper. Also, the correlations among these measures are calculated for random and natural sequences and their corresponding shapes. Based on our analysis, we conclude that from the thermodynamic energy point of view, the natural RNA sequences are different from those generated randomly.

3. Basic Definitions

An RNA molecule is composed of a chain of nucleotides, namely Adenine (A), Cytosine (C), Guanine (G), and Uracil (U). An RNA sequence δ of length *l* can be considered as a string over Σ^l , where Σ is the set of alphabet (Σ = {A,C,G,U}). Let Δ denotes the set of all RNA sequences. An RNA sequence tends to fold to itself and forms pairs of bases by the creation of hydrogen bonds between Watson-Crick bases (A-U or C-G) and Wobble base (G-U). This set of base pairs is called the RNA secondary structure and it is defined as follows.

Definition 1. An RNA secondary structure corresponding to an RNA sequence d of length l is a set of pairs (i, j), where i, j ∈ $\{1,...,l\}$) and $i < j$, and for any two base pairs i_1 ,j_1 and i_2 ,j_2 form λ,, i_1+i_2 ⇔ $j_1=j_2$, and either $i_1< j_1 <i_2< j_2$ (disjoined) or $i_1< j_1 <i_2< j_2$ (nested) holds.

Let Λ denotes the set of all RNA secondary structure. Suppose that $\varphi:\Delta \rightarrow \Lambda$ maps any RNA sequence δ into its corresponding minimum free energy secondary structure $\lambda = \varphi(\delta)$ Considering ϕ as a relation, two sequences δ_1 and δ_2 are equivalence under ϕ if and only if $\varphi(\delta_1) = \varphi(\delta_2)$. Based on this equivalence relation, the induced equivalence class of any structure could be defined as follows.

Definition 2. The equivalence class of a structure λ under the mapping φ, denoted by $[\lambda]^{\varphi}$, is the set of RNA sequences having the same structure as λ, *i.e.* $[\lambda]^{\varphi} = \{\delta | \delta \in \Delta \text{ and } \varphi(\delta) = \lambda\}$.

Suppose that Γ represents the set of all shapes and $\psi:\Delta \rightarrow \Gamma$ maps any RNA secondary structure, say λ, to its corresponding shape γ, where $\gamma = \psi(\lambda)$. As a result, maps any RNA sequence to its corresponding shape. Similar to the equivalence class of structures, we can define the equivalence class of shapes under the mapping χ as follows.

Definition 3. The equivalence class of a shape γ under the mapping χ, denoted by $[\gamma]^{\chi}$, is the set of RNA sequences having the same shape as γ, i.e. $[\gamma]^{\chi} = \{\delta | \delta \in \Delta \textit{ and } \chi(\delta) = \lambda\}$.

Each equivalence class may contain different amount of RNA sequences. In order to measure this cardinality, the following two definitions are presented.

Definition 4. For any structure $\lambda \in \Lambda$, $f_{\varphi}(\lambda)$ is the cardinality of the equivalence class $[\lambda]^{\varphi}$.

Definition 5. For any shape $\gamma \in \Gamma$, $f_{\chi}(\gamma)$ is the cardinality of the equivalence class $[\gamma]^{\chi}$.

Now, based on our terminology, the classical neutral network is defined as follows.

Definition 6. For any structure λ, a graph $NN_{\lambda} = (V,E)$ is called a neutral network under the mapping ϕ where,

- $V = \{\delta | \delta \in \Sigma^l, \varphi(\delta) = \lambda\}$,
- $E = \{(\delta_1,\delta_2) \mid \delta_1,\delta_2 \in V, \text{dist}(\delta_1,\delta_2) = 1\}$

Figure 2 shows an example of the neutral network. The neutral network does not take into account the number of sequences that are transformed from $[\lambda_1]^{\varphi}$ to $[\lambda_2]^{\varphi}$ by performing a single mutation. Considering the equivalence classes of RNA shapes, we could measure how many sequences from $[\gamma_1]^{\chi}$ are transformed to $[\gamma_2]^{\chi}$ by performing a single mutation. To do this, the variation rate is defined as follows.

Definition 7. The variation rate between two shapes γ_1 and γ_2 , denoted by $\omega(\gamma_1, \gamma_2)$ is defined as follows (1):

$$\omega(\gamma_1 ,\gamma_2) = \sum_{\delta \in [\gamma 1]\chi} | N(\delta) \cap [\gamma_2]^{\chi}| = \sum_{\delta \in [\gamma 2]\chi} | N(\delta) \cap [\gamma_1]^{\chi}|$$

where N(γ) indicates all the sequences that are obtained from γ by performing a single mutation in different positions.

Figure 3 is a schematic representation of the previous definitions. Here, the solid lines between RNA sequences indicate the single base mutational neighborhoods, the dashed arcs represent the

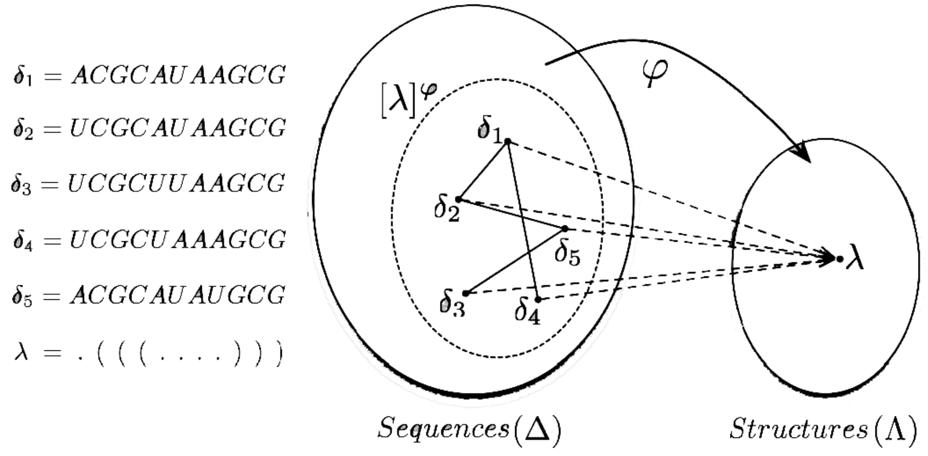

Figure 2. An example of neutral network (17)

mappings and the dashed eclipses show the equivalent classes for both structures and shapes. The variation rate between the shapes γ_1 and γ_2 is calculated as 3.

Based on the above definitions, the Variation Network can be defined over the set of all shapes as follows.

Definition 8. The Variation Network for the set of all shapes Γ is a weighted graph VN = (V,E,W), where

- $V = \{\gamma \mid \gamma \in \Gamma\}$,
- $E = \{(\gamma_1, \gamma_2) \mid \omega(\gamma_1, \gamma_2) > 0\}$,
- $\forall(\gamma_1, \gamma_2) \in E, W(\gamma_1, \gamma_2) = \omega(\gamma_1, \gamma_2)$

The Variation Network represents the imposed relations among the set of all shapes under the mapping χ as it is presented in Figure 4. With respect to the above definitions, we perform many experiments to explore the relations between RNA sequences, their structures and their shapes, both for frequency and thermodynamic energy points of view. The Variation Networks are created for random and natural RNA sequences and different measures, including frequency, shape energy, variation rate, neighborhood energy, and stability as well as the correlations coefficient between these measures are obtained. The details of our experiments as well as the results are presented in the following sections.

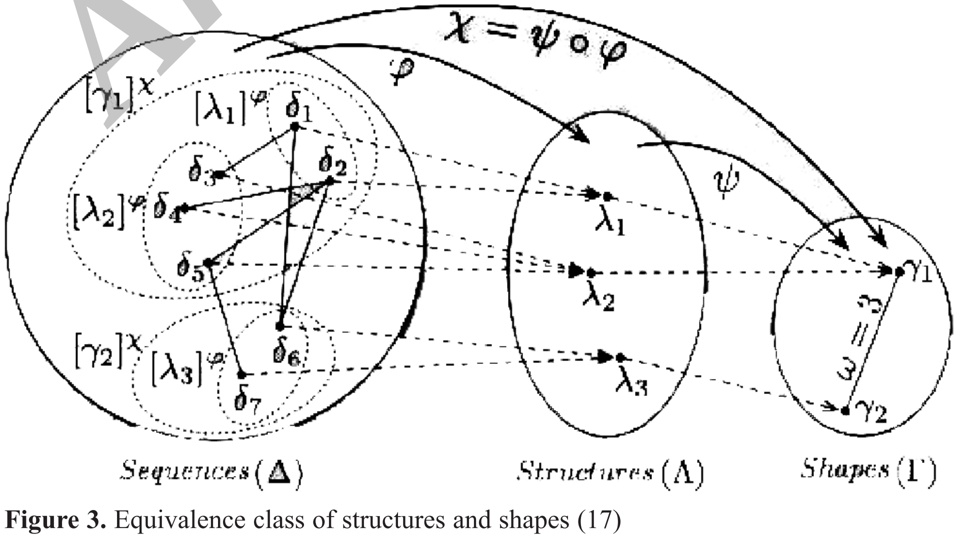

Figure 3. Equivalence class of structures and shapes (17)

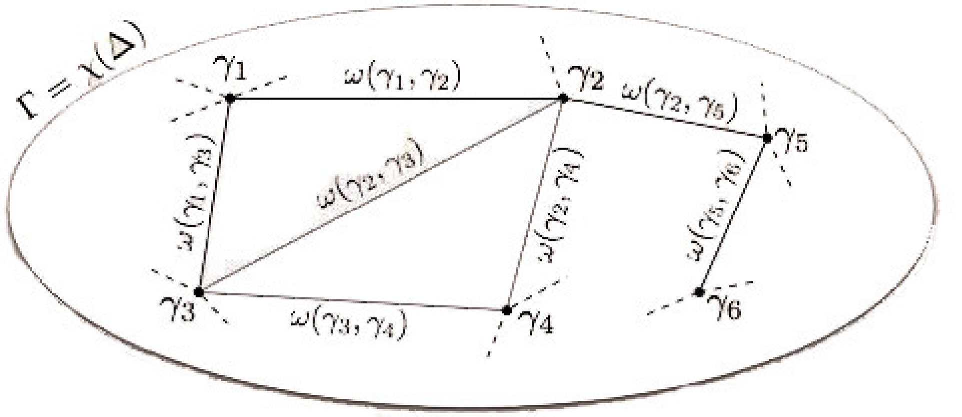
$\Gamma = \lambda(\Delta)$
γ_1
$\omega(\gamma_1, \gamma_2)$
γ_2
$\omega(\gamma_2, \gamma_5)$
γ_5
$\omega(\gamma_1, \gamma_3)$
$\omega(\gamma_2, \gamma_3)$
$\omega(\gamma_2, \gamma_4)$
$\omega(\gamma_5, \gamma_6)$
γ_3
$\omega(\gamma_3, \gamma_4)$
γ_4
γ_6

Figure 4. An example of Variation Network (17)

4. Materials and Methods

Since the number of RNA sequences grow exponentially with respect to their lengths, therefore analyzing the huge number of these sequences becomes difficult, especially for long sequences. The idea to tackle this difficulty is to use a small fraction of such sequences. On the other hand, using a small fraction of sequences may corrupt the accuracy of our analysis. But, our experiments show that the accuracy is not affected if sufficient number of sequences is employed. To justify this idea, two datasets of RNA sequences were constructed. The first one (RDS12A) contains all 4^{12} sequences of length 12 and the second one (RDS12R) contains only 111000 of such sequences (generated randomly with uniform distribution over the nucleotides). Our analysis over these two datasets indicates the high correlation rates among different properties of the corresponding Variation Networks. Therefore, employing small fraction of RNA sequences gives us the reasonable accuracy in our analysis.

After justifying the idea for sequences of length 12, we then employed sequences of length 50. To do this, seven different datasets of RNA sequences, each of length 50, were constructed. The first one (RDS50R) contains 20,000,101 sequences, which were generated randomly with uniform distribution over the nucleotides. Since the length of each RNA sequence is 50 and 150 sequences were generated by performing a single mutation in different positions, the smallest number greater than 20,000,000 which is a multiple of 151 is 20,000,101. The other six datasets were constructed by selecting all subsequences of length 50 from the natural RNA sequences (namely, Synthetic RNA, 5S Ribosomal RNA, Hammerhead Ribozyme, other Ribosomal RNA, other Ribozyme, Cis-regulatory element) taken from RNA STRAND server (16). After construction of the datasets, the RNAShape software (15) was employed to fold each RNA sequence in order to obtain its structure, its minimum free energy, and also its shape. Table 1 summarizes the construction details of the above mentioned datasets, as well as the number of distinct sequences, structures, and shapes in each one.

Since the neighborhood sequences of each RNA sequence play an important role in our analysis, therefore in construction process of random datasets (RDS12R and RDS50R), the small number of RNA sequences of desired length (l) were first generated and then all nucleotides appeared in each sequence weremutated (to three other nucleotides) to generate $3l$ more sequences (each of length l).

In order to perform the analysis and explore the potential relations between random and natural RNA sequences, as well as their corresponding structures, different measures have been employed in our analysis for each dataset. These measures are related to both frequency and thermodynamic energy as follow:

- *frequency*: For each shape γ, the cardinality of the equivalence shape class $[\gamma]^x$ is referred to as

Table 1. Summary of the constructed datasets and the number of sequences, structures, and shapes in each one.

Name	Length	# Sequences	Generation Method	Background Family	# Structure	# Shapes
RDS12A	12	16, 777, 216	All possible sequences	-	30	6
RDS12R	12	111, 000	Random sequences	-	29	6
RDS50R	50	20, 000, 101	Random sequences	-	4,704,322	7, 828
RDS50N1	50	252, 623	All Subsequence	Synthetic RNA	74,411	2, 059
RDS50N2	50	1, 248, 770	All Subsequence	5S Ribosomal RNA	331,611	3, 908
RDS50N3	50	214, 420	All Subsequence	Hammerhead Ribozyme	34,537	879
RDS50N4	50	292, 940	All Subsequence	Other Ribosomal RNA	89,316	2, 317
RDS50N5	50	219, 403	All Subsequence	Other Ribozyme	60,366	1, 444
RDS50N6	50	185, 428	All Subsequence	Cis-regulatory element	44,656	1, 626

frequency (f_χ (γ)).

normalized frequency (*nf*): This is the normalized value of frequency for each shape γ and it can be calculated as follows:

$$nf(\gamma) = \frac{f_\chi(y) - f_\chi^{\min}}{f_\chi^{\max} - f_\chi^{\min}}, \qquad (2)$$

where $f_\chi^{\min}$ and $f_\chi^{\max}$ denote the minimum and maximum frequency among the equivalence shape classes, respectively.

- *shape energy average* (*sea*): For each shape γ, consider the sequences in $[\gamma]^x$. For all these sequences, the minimum free energies over the corresponding structures were calculated and averaged, i.e.

$$sea(\gamma) = \frac{\sum_{\delta\in[\gamma]^\chi} energy(\varphi(\delta))}{f_\chi(y)}. \qquad (3)$$

- *variation rate* (*vr*): Although the variation rate is defined for any two shapes, here the variation rate for a shape γ is taken over all other shapes as follows:

$$vr(\gamma) = \sum_{\tau\in\Gamma,\tau\neq\gamma} \omega(y,\tau). \qquad (4)$$

The variation rate *vr*(γ) indicates that by performing a single mutation in different positions, how many sequences from $[\gamma]^x$ were transformed to the other equivalence shape classes.

normalized variation rate (*nvr*): This is the normalized value of variation rate and it can be calculated as follows:

$$nvr(\gamma) = \frac{vr(\gamma) - vr^{\min}}{vr^{\max} - vr^{\min}}, \qquad (5)$$

where $vr^{\min}$ and $vr^{\max}$ denote the minimum and maximum variation rates, respectively.

- *neighborhood energy average* (*nea*): For any shape γ, consider its neighborhood shapes in the Variation Network. The neighborhood energy average indicates the average minimum free energies over the sequences appeared in neighborhoods' equivalence shape classes as follows:

$$nea(\gamma) = \frac{\sum_{\tau\in N(\gamma)} \left(\sum_{\delta\in N([\gamma]^\chi)\cap[\tau]^\chi} energy(\varphi(\delta)) \right)}{vr(\gamma)}. \qquad (6)$$

- stability: This measure indicates the stability of a shape γ and it is calculated as follows:

$$\rho_{x,y} = \frac{\operatorname{cov}(X,Y)}{\sigma_x\sigma_y} = \frac{E\left[(X-\mu_x)(Y-\mu_y)\right]}{\sigma_x\sigma_y}, \qquad (7)$$

where ε is considered as 0.0001 to avoid the divide by zero exception.

All the above measures could be applied on a single shape. In order to compare the above mentioned measures obtained from different datasets, we employ a dimensionless metric as follows:

- ***Population Pearson correlation coefficient metric***

This metric is used to explore the linear dependency among two random variables. It is achieved by dividing the covariance of two random variables by the product of their standard deviations. Regarding the expected values (μ_x and μ_y) and the standard deviations (σ_x and σ_y) of two random variables (*X* and *Y*), the population Pearson correlation $p_{x,y}$ is calculated as:

$$stability(\gamma) = \frac{nf(\gamma) \times sea(\gamma)}{nvr(\gamma) \times nea(\gamma) + \varepsilon}, \qquad (8)$$

where *E* and *cov* denote the expected value and covariance, respectively. Also the corresponding *p-value*, which is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, is calculated. All the above mentioned measures were evaluated on different datasets and the obtained results are presented in the next section.

5. Results

As it is mentioned, the RNAshape software (15) is employed to fold each RNA sequence in order to obtain its structure, its minimum free energy, and also its shape.

For datasets RDS12A and RDS12R, six different shapes were obtained. Among them, two shapes have very low frequency and therefore they were not considered in our further analysis (they do not have much information). For each shape in these datasets, the corresponding measures were evaluated and presented in Table 2. Also, the correlation and p-value among different measures for these two datasets are presented in the last rows of Table 2. As it is understood from this table, all measures are correlated and therefore the Variation Network corresponding to the small fraction of sequences gives us a reasonable result about the properties of the Variation Network corresponding to the whole sequences.

The same evaluations have been done for other seven datasets of length 50 constructed from random and natural RNA sequences. The five most frequent shapes of each dataset, as well as their corresponding measures are presented in Table 3. To analyze the relationship between the Variation Networks of random and natural RNA sequences, the correlation and p-value of each measure is cal-

Table 2. Evaluation of the measures for datasets RDS12A and RDS12R

Datasets	Shapes	Frequencies (fγ(γ))	Normalized frequencies (nf)	Shape energy average (sea)	Variation rate (vr)	Normalized variation rate (nvr)	Neighborhood energy average (nea)	Stability
RDS12A	[_]_	855167	0.3687	-1.2542	2333959	0.2682	-1.2424	1.3878
	_[_]_	830439	0.358	-1.1362	3390606	0.3896	-1.857	0.5623
	_[_]	435209	0.1876	-1.0575	1794678	0.2062	-1.238	0.7772
	[_]	198277	0.0855	-1.5787	1177246	0.1353	-1.3585	0.7344
RDS12R	[_]_	5580	0.3593	-1.2792	1951	0.2763	-1.3938	1.1931
	_[_]_	5543	0.3569	-1.1599	2711	0.384	-1.8919	0.5698
	_[_]	2849	0.1834	-1.0916	1389	0.1967	-1.3045	0.7802
	[_]	1556	0.1002	-1.5972	1003	0.1421	-1.448	0.7778
Correlation	-	-	0.9988	0.9999	-	0.9969	0.991	0.9905
p-value	-	-	0.0012	0.0001	-	0.0031	0.009	0.0095

Table 3. Evaluation of the measures for the first five most frequent shapes of all dataset.

Datasets	Shapes	Frequencies (fγ(γ))	Normalized frequencies (nf)	shape energy average (sea)	Variation rate (vr)	normalized variation rate (nvr)	Neighborhood energy average (nea)	Stability
RDS50R	_[_[_]_]_	1713382	1	-7.78	832555	0.93	-8	1.05
	_[_[_[_]_]_]_	1713251	1	-8.86	897861	1	-8.86	1
	_[_]_[_]_	883463	0.52	-8.41	452814	0.5	-8.48	1.01
	_[_]_[_[_]_]_	572828	0.33	-9.09	330260	0.37	-9.21	0.9
	_[_[_]_]_[_]_	571526	0.33	-9.01	321296	0.36	-9.12	0.92
RDS50N1	_[_[_[_]_]_]_	15078	1	-15.83	6375	1	-14.92	1.06
	_[_[_]_]_	12886	0.85	-14.9	5988	0.94	-12.9	1.05
	_[_]_[_]_	11724	0.78	-15.49	3746	0.59	-14.76	1.39
	_[_[_]_]_[_]_	8416	0.56	-14.69	3187	0.5	-13.54	1.21
	_[_]_[_[_]_]_	7218	0.48	-14.25	3743	0.59	-12.64	0.92
RDS50N2	_[_[_[_]_]_]_	74469	1	-11.23	39696	1	-10.31	1.09
	_[_[_]_]_	56031	0.75	-10.21	25875	0.65	-10.15	1.16
	_[_]_[_[_]_]_	41308	0.55	-11.19	16706	0.42	-10.83	1.36
	_[_[_[_[_]_]_]_]_	37712	0.51	-11.89	17816	0.45	-11.84	1.13
	_[_[_]_[_]_]_	35600	0.48	-11.35	14890	0.38	-12.07	1.2
RDS50N3	_[_[_]_]_	27054	1	-5.47	10737	1	-5.77	0.95
	_[_]_	26449	0.98	-3.6	5250	0.49	-4.23	1.7
	_[[_]_]_[_]_	15346	0.57	-14.34	1519	0.14	-13.06	4.4
	_[_[_[_]_]_]_	15012	0.55	-5.92	8134	0.76	-5.22	0.83
	_[_[_]]_	8757	0.32	-6.13	2394	0.22	-5.67	1.57
RDS50N4	_[_[_[_]_]_]_	19639	1	-14.34	7585	1	-13.64	1.05
	_[_[_]_]_	14842	0.76	-13.46	6572	0.87	-11.72	1
	_[_]_[_]_	10348	0.53	-13.99	4630	0.61	-14.04	0.86
	_[_[_]_]_[_]_	9523	0.48	-13.59	3982	0.52	-12.87	0.98
	_[_]_[_[_]_]_	9402	0.48	-13.16	4436	0.58	-13.06	0.83
RDS50N5	_[_[_]_]_	17757	1	-12.32	5416	0.71	-11.51	1.51
	_[_[_[_]_]_]_	15425	0.87	-13.13	7629	1	-12.34	0.92
	_[_]_[_]_	13061	0.74	-11.66	3915	0.51	-10.92	1.53
	_[_]_[_[_]_]_	7972	0.45	-12.78	3235	0.42	-11.72	1.15
	_[_[_]_]_[_]_	7906	0.45	-13.9	4527	0.59	-13.57	0.77
RDS50N6	_[_[_[_]_]_]_	9401	1	-9.31	5122	1	-8.84	1.05
	_[_[_]_]_	6556	0.7	-7.56	4280	0.84	-7.02	0.9
	_[_]_[_[_]_]_	6397	0.68	-9.69	2803	0.55	-9.71	1.24
	_[_]_[_]_	6015	0.64	-10.78	3294	0.64	-10.98	0.98
	_[_[_[[_]_]]]_	5553	0.59	-18.02	885	0.17	-16.99	3.63

culated separately. To do this, five different percentages of the most frequent shapes of each dataset, namely 100%, 20%, 10%, 5%, and 1%, were employed.

Then, for each measure, the correlation and p-value between any pairs of datasets with respect to the selected percentages of the most frequent shapes were calculated. For the normalized frequency measure, the correlations between different datasets, as well as the corresponding p-value, are presented in Table 4. The results for the other

Table 4. The correlations and p-values among random and natural datasets for normalized frequency measure.

Datasets	RDS50N1		RDS50N2		RDS50N3		RDS50N4		RDS50N5		RDS50N6	
	Corre-lation	p-value	Corre-lation	p-valu	Corre-lation	p-valu	Corre-lation	p-valu	Corre-lation	p-value	Corre-lation	p-value
	0.9347	0	0.9148	0	0.7769	0	0.9393	0	0.9395	0	0.8421	0
	0.9454	0	0.9225	0	0.7548	0	0.9532	0	0.939	0	0.8349	0
RDS50R	0.9437	0	0.9198	0	0.732	0	0.9516	0	0.9336	0	0.8251	0
	0.9417	0	0.9154	0	0.6817	0	0.9485	0	0.9239	0	0.8095	0
	0.9292	0	0.9185	0	0.4243	0.2948	0.9246	0	0.9252	0	0.8237	0.0002
			0.9137	0	0.7043	0	0.9626	0	0.9433	0	0.8554	0
			0.9027	0	0.6778	0	0.9594	0	0.9396	0	0.835	0
RDS50N1			0.8924	0	0.6375	0	0.9564	0	0.9346	0	0.8137	0
			0.8824	0	0.5509	0.0004	0.956	0	0.9278	0	0.8231	0
			0.8649	0	0.1875	0.6873	0.9422	0	0.9245	0	0.9009	0.0009
					0.5992	0	0.938	0	0.8702	0	0.8334	0
					0.5582	0	0.9314	0	0.8551	0	0.8081	0
RDS50N2					0.5145	0	0.9248	0	0.8414	0	0.7983	0
					0.4388	0.0084	0.9177	0	0.8171	0	0.8323	0
					0.6779	0.2085	0.8944	0	0.7373	0.0096	0.8702	0.0001
							0.6859	0	0.6984	0	0.5836	0
							0.6496	0	0.6621	0	0.604	0
RDS50N3							0.6196	0	0.6338	0	0.5583	0
							0.5212	0.0019	0.5278	0.0023	0.4612	0.0079
							0.2309	0.6184	0.0383	0.9512	0.4781	0.4154
									0.9435	0	0.8592	0
									0.9386	0	0.8457	0
RDS50N4									0.9345	0	0.8793	0
									0.9237	0	0.871	0
									0.8836	0	0.9473	0
											0.8241	0
											0.8586	0
RDS50N5											0.8474	0
											0.8651	0
											0.8525	0.0035

measures, namely *s*hape energy average, normalized variation rate, neighborhood energy average, and stability are presented in Tables 5, 6, 7, and 8, respectively.

As it is understood from these tables, the frequency and variation rate measures in these datasets are highly correlated to each other. This indicates that the most frequent shapes and their

Table 5. The correlations and p-values among random and natural datasets for shape energy average measure

Datasets	RDS50N1		RDS50N2		RDS50N3		RDS50N4		RDS50N5		RDS50N6	
	Correlation	p-value	Correlation	p-valu	Correlation	p-valu	Correlation	p-valu	Correlation	p-value	Correlation	p-value
RDS50R	0.0757	0.0008	0.0864	0	0.2096	0	0.1122	0	0.0874	0.001	0.1773	0
	0.1557	0.0019	0.3225	0	0.5019	0	0.0951	0.0438	0.3832	0	0.4475	0
	0.0731	0.3013	0.4133	0	0.5912	0	-0.0526	0.4323	0.4927	0	0.6015	0
	-0.1415	0.1581	0.4369	0	0.6437	0	-0.0868	0.3562	0.4924	0	0.7132	0
	-0.2052	0.3856	0.5374	0.0007	0.6753	0.0661	-0.5452	0.0071	0.7918	0.0007	0.7281	0.0021
RDS50N1			0.099	0	0.0319	0.3858	0.2225	0	0.322	0	-0.0437	0.1392
			0.0976	0.0647	-0.1725	0.0325	0.3491	0	0.4073	0	0.1079	0.0926
			0.0717	0.3402	-0.381	0.0007	0.3574	0	0.2001	0.0247	0.1509	0.1075
			0.1162	0.2784	-0.4994	0.0017	0.4013	0.0004	0.1726	0.1872	0.1462	0.2779
			0.3918	0.1334	-0.6355	0.1251	0.58	0.0092	-0.4423	0.15	0.1767	0.6493
RDS50N2					0.0627	0.0693	0.0823	0.0004	0.0914	0.0009	-0.0176	0.5176
					0.044	0.5732	0.001	0.984	0.2475	0	0.0603	0.3233
					0.0765	0.4972	-0.047	0.5081	0.2664	0.0015	0.1033	0.2403
					0.1243	0.4768	0.0325	0.7433	0.2867	0.0196	0.2601	0.0396
					-0.0608	0.9226	-0.0815	0.7478	0.2196	0.5164	0.2724	0.3461
RDS50N3							0.0175	0.6278	0.1388	0.0003	0.1829	0
							-0.0534	0.5138	0.2875	0.0006	0.3169	0.0002
							-0.1007	0.3805	0.3506	0.0034	0.3691	0.004
							-0.161	0.3708	0.1602	0.3893	0.3159	0.0781
							-0.6803	0.0926	0.6839	0.2029	0.9247	0.0245
RDS50N4									-0.0235	0.4232	-0.055	0.0642
									-0.0319	0.6157	-0.0369	0.5735
									-0.1259	0.1569	-0.0193	0.8372
									-0.0133	0.9158	-0.2318	0.0827
									-0.4459	0.11	0.3896	0.2362
RDS50N5											-0.0032	0.9225
											0.095	0.1863
											0.2487	0.0188
											0.3389	0.0161
											0.1041	0.7899

Table 6. The correlations and p-values among random and natural datasets for normalized variation rate measure

Datasets	**RDS50N1**		**RDS50N2**		**RDS50N3**		**RDS50N4**		**RDS50N5**		**RDS50N6**	
	Correlation	p-value	Correlation	p-valu	Correlation	p-valu	Correlation	p-valu	Correlation	p-value	Correlation	p-value
RDS50R	0.9589	0	0.9457	0	0.9205	0	0.9486	0	0.9491	0	0.9181	0
	0.9679	0	0.9563	0	0.9184	0	0.966	0	0.9542	0	0.9179	0
	0.9668	0	0.9543	0	0.9163	0	0.9636	0	0.9493	0	0.9139	0
	0.9653	0	0.9514	0	0.9081	0	0.9595	0	0.9408	0	0.9032	0
	0.9582	0	0.9512	0	0.9105	0.0017	0.9334	0	0.9215	0	0.8849	0
RDS50N1			0.958	0	0.8749	0	0.975	0	0.9591	0	0.9292	0
			0.9521	0	0.8704	0	0.9719	0	0.9556	0	0.92	0
			0.9469	0	0.8676	0	0.9692	0	0.9532	0	0.912	0
			0.9409	0	0.8472	0	0.967	0	0.9471	0	0.9302	0
			0.943	0	0.8397	0.0181	0.963	0	0.9432	0	0.9693	0
RDS50N2					0.8143	0	0.9679	0	0.9511	0	0.9162	0
					0.8	0	0.9641	0	0.945	0	0.9043	0
					0.7912	0	0.9604	0	0.9413	0	0.8999	0
					0.7598	0	0.9565	0	0.9356	0	0.8923	0
					0.7239	0.1667	0.9336	0	0.9338	0	0.9204	0
RDS50N3							0.8663	0	0.8738	0	0.8303	0
							0.8568	0	0.861	0	0.8539	0
							0.8475	0	0.851	0	0.8615	0
							0.8077	0	0.8108	0	0.8434	0
							0.8072	0.0282	0.5591	0.3272	0.8908	0.0426
RDS50N4									0.9759	0	0.9216	0
									0.9747	0	0.9117	0
									0.9728	0	0.9084	0
									0.9696	0	0.8925	0
									0.9587	0	0.9412	0
RDS50N5											0.9093	0
											0.9075	0
											0.9037	0
											0.9162	0
											0.9378	00.0002

variation rates are almost identical in random and natural datasets. On the other hand, the shape energy average, neighborhood energy average, and stability measures are not well correlated. This indicates that, the natural RNA sequences are specialized to do a specific function inside the cell, where the random RNA sequences do not have such a special function (17).

Table 7. The correlations and p-values among random and natural datasets for neighborhood energy average measure

Datasets	RDS50N1		RDS50N2		RDS50N3		RDS50N4		RDS50N5		RDS50N6	
	Correlation	p-value	Correlation	p-valu	Correlation	p-valu	Correlation	p-valu	Correlation	p-value	Correlation	p-value
	0.1496	0	0.2845	0	0.033	0.3303	0.1896	0	0.1114	0	0.099	0.0001
	-0.1049	0.0375	-0.0869	0.0179	-0.0337	0.6583	-0.1573	0.0008	-0.0517	0.3849	0.0834	0.1497
RDS50R	-0.1055	0.1352	0.0485	0.3507	0.2318	0.0318	-0.2092	0.0016	0.175	0.0366	0.2011	0.0136
	-0.2868	0.0036	0.2058	0.0041	0.499	0.0008	-0.258	0.0054	0.1893	0.1113	0.4011	0.0004
	-0.287	0.2198	0.3249	0.0532	0.6127	0.1063	-0.3477	0.1041	0.7413	0.0024	0.6552	0.008
			0.2941	0	0.2996	0	0.4253	0	0.4886	0	0.2673	0
			-0.0208	0.694	0.1033	0.2024	0.3136	0	0.3645	0	0.0446	0.488
RDS50N1			-0.1001	0.1826	0.0509	0.6647	0.1631	0.0358	0.1948	0.0288	0.0651	0.4896
			-0.0979	0.3613	-0.4002	0.0141	0.1578	0.1764	0.1521	0.2461	-0.1734	0.197
			0.1209	0.6557	-0.525	0.2263	0.5026	0.0283	-0.5491	0.0644	0.3628	0.3372
					0.1525	0	0.3197	0	0.2913	0	0.2049	0
					0.0315	0.6874	0.0347	0.4848	0.0708	0.2436	-0.0056	0.9271
RDS50N2					0.0709	0.5295	0.03	0.6721	0.0869	0.3092	0.129	0.1418
					0.2519	0.1443	0.0107	0.9143	-0.1026	0.4121	0.0297	0.8172
					0.3848	0.5224	-0.0929	0.7139	-0.1227	0.7193	0.2264	0.4363
							0.2402	0	0.2451	0	0.1832	0
							0.0832	0.3081	-0.1179	0.1683	0.0694	0.4275
RDS50N3							0.1353	0.2377	-0.1528	0.2134	0.0582	0.6613
							-0.205	0.2525	-0.1972	0.2876	-0.1503	0.4117
							-0.5885	0.1645	0.6167	0.2679	0.7367	0.1556
									0.3525	0	0.2492	0
									0.1866	0.0031	0.0352	0.5913
RDS50N4									0.1293	0.1459	0.0111	0.9057
									0.212	0.0875	-0.1138	0.3994
									-0.306	0.2874	0.1996	0.5563
											0.2079	0
											0.0237	0.7423
RDS50N5											0.0278	0.7957
											0.2072	0.1488
											-0.3654	0.3335

Table 8. The correlations and p-values among random and natural datasets for stability measure

Datasets	RDS50N1		RDS50N2		RDS50N3		RDS50N4		RDS50N5		RDS50N6	
	Correlation	p-value	Correlation	p-valu	Correlation	p-valu	Correlation	p-valu	Correlation	p-value	Correlation	p-value
RDS50R	-0.0164	0.4679	-0.0108	0.5206	-0.0031	0.9269	-0.0128	0.5491	-0.0860	0.0012	-0.0953	0.0002
	-0.0417	0.4096	-0.0170	0.6440	-0.0579	0.4462	0.0115	0.8085	-0.1126	0.0577	-0.0217	0.7078
	-0.0407	0.5653	-0.0186	0.7203	-0.0829	0.4481	-0.0004	0.9955	-0.0273	0.7458	0.0533	0.5174
	-0.0889	0.3765	0.0806	0.2652	0.4337	0.0041	0.1583	0.0911	-0.0423	0.7240	-0.0191	0.8707
	-0.1576	0.5071	0.4320	0.0085	-0.0489	0.9085	-0.0479	0.8283	0.3011	0.2954	-0.1072	0.7036
RDS50N1			0.0350	0.1448	0.0958	0.0092	0.1378	0	0.2034	0	0.1584	0
			0.0242	0.6477	0.1321	0.1024	-0.0070	0.9012	0.0785	0.2249	0.0594	0.3555
			-0.0113	0.8803	-0.0214	0.8554	-0.0455	0.5603	-0.0371	0.6799	-0.0134	0.8873
			-0.0223	0.8359	-0.0465	0.7845	-0.0246	0.8340	-0.0335	0.7996	-0.0366	0.7869
			-0.1223	0.6518	-0.1662	0.7217	-0.0556	0.8211	-0.0917	0.7769	0.2466	0.5224
RDS50N2					-0.0145	0.6755	0.0741	0.0013	0.0294	0.2867	0.0421	0.1208
					0.0073	0.9252	0.0255	0.6078	0.0408	0.5018	0.0582	0.3409
					-0.0264	0.8148	-0.0058	0.9354	-0.0067	0.9377	0.0137	0.8766
					0.4470	0.0071	-0.0112	0.9105	0.1130	0.3663	0.1317	0.3035
					0.8028	0.1020	-0.5036	0.0331	0.0313	0.9271	1	0
RDS50N3							0.1042	0.0037	0.0581	0.1297	0.0377	0.3283
							0.0867	0.2884	0.0023	0.9787	-0.0495	0.5714
							-0.0066	0.9540	-0.0523	0.6716	-0.0471	0.7231
							-0.0400	0.8249	0.2487	0.1773	-0.0425	0.8173
							-0.1673	0.7199	-0.6759	0.2104	0.8841	0.0465
RDS50N4									0.3141	0	0.1390	0
									0.2993	0	0.0602	0.3585
									0.3541	0	-0.0095	0.9194
									0.1581	0.2048	-0.0314	0.8165
									-0.0773	0.7927	0.4152	0.2041
RDS50N5											0.2234	0
											0.2040	0.0042
											0.3222	0.0021
											0.5781	0
											-0.1250	0.7486

Also, the correlation between the random dataset (RDS50R) and six other natural datasets are higher than the correlation between natural datasets. This indicates that, the natural RNA sequences do not follow the uniform distribution because of their special functionality inside the cell. On the other hand, the big amount of employed random RNA sequences gives us a large number of shapes which covers almost many natural produced shapes (from natural sequences).

The correlations among natural datasets become higher if the small percentages of most frequent shapes were employed in our analysis. This indicates that the most frequent shapes appeared in different families of natural RNA sequences.

6. Conclusions

In this paper, the Variation Network concept has been introduced to analyze the relationship between RNA sequences, structures, and shapes. Although the function of an RNA sequence is related to its structure, the RNA shape indicates the higher level of representation of its functionality inside the cell. Bases on the Variation Networks corresponding to the random and natural RNA sequences, different measures were calculated and the correlation among them are presented in this study. The obtained results indicate that from the frequency point of view, all the employed datasets are highly correlated to each other, but from the thermodynamic energy point of view they are not well correlated. These conclude that the natural RNA sequences are not generated randomly.

References

1. Huynen MA, Konings DA, Hogeweg P. Multiple Coding and the Evolutionary Properties of RNA Secondary Structure. *J Theor Biol*. 1993;**165**(2): 251-267.doi:10.1006/jtbi.1993.1188
2. Nakayaa A, Yonezawaa A, Yamamotob K. Classification of RNA Secondary Structures Using the Techniques of Cluster Analysis. *J Theor Biol.* 1996;**183**:105-17.
3. Aguirre-Hernandez R, Hoos HH, Condon A. Computational RNA secondary structure design: empirical complexity and improved methods. *BMC Bioinformatics* 2007;8-34.
4. Schuster P, Fontana W, Stadler PF, Hofacker IL. From sequences to shapes and back: a case study in RNA secondary structures. *Proc Biol Sci*. 1994;**255**:279-84.doi: 10.1098/rspb.1994.0040
5. Gruner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Schuster P, Stadler PF. Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structures of neutral networks and shape space covering. *Monatsh Chem*.1996;**127**:375-389.doi:10.1007/BF00810882
6. Reidys C, Stadler P, Schuster P. Generic properties of combinatory maps: neutral networks of RNA secondary structures. *Bull Math Biol*.1997;**59**:339-397.doi:10.1007/BF02462007
7. Cowperthwaite MC, Economo EP, Harcombe WR, Miller EL, Ancel Meyers L. The ascent of the abundant: How mutational networks constrain evolution. *PLoS Comput Biol*.2008;**4**(7):1-11.doi: 10.1371/journal.pcbi.1000110.
8. Fontana W, Konings DAM, Stadler PF, Schuster P. Statistics of RNA secondary structures. *Biopolymers* 1993;**33**:1389-1404.PMID:7691201[PubMed - indexed for MEDLINE]
9. Jorg T, Martin OC, Wagner A. Neutral network sizes of biological RNA molecules can becomputed and are not atypically small. *BMC Bioinformatics*.2008;**9**:464-476. doi:10.1186/1471-2105-9-464.
10. Reidys C, Forst CV, Schuster P. Replication and mutation on neutral networks. *Bull Math Biol*. 2001;**63**:57-94.
11. Stich M, Briones C, Manrubia SC. On the structural repertoire of pools of short, random RNA sequences. *J Theor Biol*. 2008;**252**:750-763.doi:10.1016/j.jtbi.2008.02.018
12. Gan HH, Pasquali S, Schlick T. Exploring the repertoire of RNA secondary motifs using graph theory with implications for RNA design. *Nucl Acids Res*. 2003;**31**:2926-2943.doi:10.1093/nar/gkg365
13. Aguirre J, Buldu JM, Stich M, Manrubia SC. Topological Structure of the Space of Phenotypes: The Case of RNA Neutral Networks. *PLoS One*. 2011;doi: 10.1371/journal.pone.0026324.
14. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. *Monatsh Chem*.1994; **125**:167-188.doi:10.1007/BF00818163
15. Steffen P, Vo B, Rehmsmeier M, Reeder J, Giegerich R. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* 2006;**22**:500-503.doi:10.1093/bioinformatics/btk010
16. Andronescu M, Bereg V, Hoos HH, Condon A. RNA STRAND: The RNA Secondary Structure And Statistical Analysis Database. *BMC Bioinformatics*. 2008;**9**:340.doi:10.1186/1471-2105-9-340.
17. Mohammadzadeh J, Ganjtabesh M, Nowzari-Dalini A. Topological Properties of RNA Variation Networks over the Space of RNA Shapes. *MATCH Commun Math Comput Chem*. 2014;**72**:501-18.

Iranian Journal of Biotechnology

شبکه تغییر ابزاری برای ارتباط بین توالیها، ساختارها و شکل های RNA

جواد محمدزاده[2]، محمد گنج تابش[1,*]، عباس نوذری دالینی[1]

[1]دانشکده ریاضی، آمار و علوم کامپیوتر، پردیس علوم، دانشگاه تهران، تهران، ایران
[2]مرکز تحقیقات بیوشیمی-بیوفیزیک (I)، دانشگاه تهران، تهران، ایران

اطلاعات مقاله

نوع مقاله
مقاله پژوهشی

کلمات کلیدی:
پایداری
مسأله معکوس پیشگویی ساختار دوم RNA
مسأله پیشگویی ساختار دوم RNA
جهش

خلاصه مقاله

مقدمه: مولکولهای RNA نقش کلیدی را در بسیاری از فرآیندهای زیست شناختی ایفا می کنند. فعالیت RNA وابسته به ساختار سوم آن بوده و ساختار دوم RNA تا حدودی ویژگیهای ساختار سوم آن را نشان می‌دهد. از آنجا که فعالیت زیستی RNA به صورت غیر مستقیم از ساختار اولیه آن نتیجه می شود، ارتباط بین توالیها و ساختارها اهمیت ویژه ای پیدا می کند. یکی از ابزارهایی که برای تحلیل این نوع ارتباطات استفاده می شود، شبکه خنثی است. در شبکه خنثی توالیها، گره ها را تشکیل می‌دهند و دو گره زمانی به هم متصل می‌شوند که فقط در یک باز متفاوت باشند و در عین حال دارای ساختار یکسانی باشند. شیپ (شکل) RNA یکی دیگر از روشهای نمایش ساختار دوم RNA است. این نحوه نمایش یک نمایش مختصر شده است که با حفظ پیچیدگی ساختار، طول مولفه های آن را به یک واحد کاهش می‌دهد. با استفاده از این نمایش، ساختارهای با طول بلندتر به راحتی در آنالیز توالیها شرکت می کنند.

اهداف: در این مقاله، یک مفهوم جدید به نام شبکه تغییر معرفی شده است. این شبکه بر روی مجموعه شکل های RNA ایجاد شده است. به وسیله این شبکه ارتباطات بالقوه ای که بین RNA های تصادفی و طبیعی وجود دارد به همراه تحلیل ساختار آنها به‌دست می آید.

مواد و روشها: بعضی از معیارها شامل بسامد، بسامد نرمال شده، میانگین انرژی شکل، نرخ تغییر، میانگین انرژی همسایگی، و پایداری محاسبه و تحلیل شده است.

نتایج: همبستگی بین شبکه های تغییر تصادفی و طبیعی محاسبه شده است و بر اساس همبستگی محاسبه شده، همه ی معیارها به غیر از معیار انرژی ترمودینامیکی همبستگی بالایی داشتند.

نتیجه گیری: با توجه به اینکه مقدار انرژی ترمودینامیکی توالی RNA بر روی ساختار آن، نقش کلیدی در فعالیت RNA دارد، این تحقیق نشان می‌دهد که توالیهای طبیعی به صورت تصادفی به وجود نیامده اند.

[*] نویسنده مسئول