MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **72** (2014) 501-518

ISSN 0340 - 6253

Topological Properties of RNA Variation Networks over the Space of RNA Shapes

Javad Mohammadzadeh[a], Mohammad Ganjtabesh[*,b], Abbas Nowzari–Dalini[b]

[a]*Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran*
jmohammadzadeh@ibb.ut.ac.ir

[b]*School of Mathematics, Statistics, and Computer Science, University of Tehran, Tehran, Iran*
mgtabesh@ut.ac.ir, nowzari@ut.ac.ir

(Received October 24, 2013)

Abstract

The function of an RNA sequence is related to its tertiary structure. Since dealing with RNA tertiary structure is very complicated, the RNA secondary structure is used instead. RNA secondary structure denotes various considerable aspects of RNA tertiary structure, and the biological function of an RNA sequence is assumed to be related to its secondary structure. Another useful illustration of an RNA secondary structure is the RNA shape, where it is holding the vicinity and nesting of structural components and shrinking their lengths to one. It would be significant to analyze the relations between the RNA sequences and their structures. One of the useful methods to perform these kinds of analysis is the neutral network. A neutral network can be considered as a graph whose vertex set is a collection of RNA sequences, all coding the same secondary structure, in which two RNA sequences are connected if one of them can be obtained from the other by a single base mutation. In this paper, a novel concept, entitled *variation network*, over the set of all RNA shapes is proposed to analyze the relation between the RNA sequences and their shapes, as well as to discover different topological properties of the RNA shapes. Based on the variation network, several topological properties, such as clustering coefficient, topological coefficient, average shortest path distribution, and centrality are calculated for natural RNA sequences. Also, the correlations between power-law function and some distributions over the variation network are obtained. These correlations indicate that the variation network is a kind of complex biological network, having scale-free structure and small world property.

1. Introduction

In any biological organism, the information flows from DNA to RNA (transcription) and then from RNA to proteins (translation) [1, 2]. Hence, it seems that DNA, RNA, and proteins are the essential features of any biological organism [3].

The function of an RNA molecule is related to its structure which is represented in three levels: primary, secondary, and tertiary structures. Primary structure of an RNA is a sequence of nucleotides which is composed of Adenine (A), Cytosine(C), Guanine (G), and Uracil (U) [4]. A secondary structure of an RNA sequence is a set of pairing links between bases in the sequence, where each base can pair with at most one another base. A tertiary structure of an RNA is presented by the coordinates of its atoms in a three-dimensional space.

As dealing with tertiary structure of an RNA is extremely complicated, many efforts have been focused on RNA secondary structure in the literatures [5–7]. RNA secondary structure establishes various considerable aspects of RNA tertiary structure and the biological function of an RNA is assumed to be related to its secondary structure [8,9]. RNA shape is another valuable demonstration of an RNA secondary structure. RNA shape represents structure in a dense form, keeping vicinity and nesting of structural components and reducing their lengths to one. Figure 1 denotes a small RNA sequence, its secondary structure (in parenthesis representation), and its shape which is obtained by reducing the length of each structural component to one.

Since the biological function of an RNA is inferred indirectly from its sequence, consequently analyzing the relationships between the RNA sequences and their structures would be of great interest. The neutral network is a very useful method to perform these kinds of analysis [10–12]. A neutral network can be considered as a graph whose vertex set is a collection of RNA sequences, all coding the same secondary structure, in which two RNA sequences are connected if one of them can be obtained from the other by a single base mutation [10]. Neutral networks are generated by common structures which overfill the space of RNA sequences [11, 12] and consequently they simplify the investigations of an enormous quantity of alternative structures. Since various neutral networks are greatly meshed, all well-known structures can be reached within a few (mutational) walks, starting from any arbitrary or random sequence [12].

Dynamical processes of networks are related to its topological properties. The topolog-

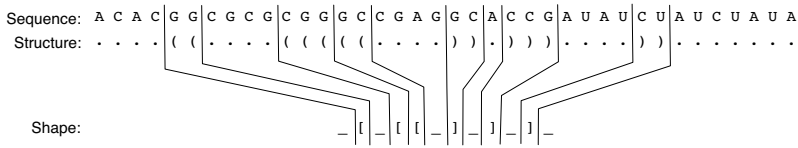


Fig. 1. An example of RNA sequence, its corresponding secondary structure, and its shape which is obtained by reducing the length of each structural component to one.

ical properties of biological networks have been studied in the literatures and reasonable results have been obtained [13, 14]. As a kind of biological network, metabolic networks, protein-protein interaction (PPI) networks, gene regulatory networks (GRN), and many others have been studied previously [23, 24]. Surprisingly in these biological networks, the degree distribution represents the scale-free connectivity [25, 26] as well as small-world structure (high clustering and short-path between nodes) [23, 27, 28]. Networks with scale-free connectivity are able to release any perturbation over the network in a few steps, and hence they are very robust against random failures [24]. Over the past decades, centrality has also became an acceptable strategy to deal with complex networks such as biological networks [29].

The structure of the RNA neutral networks have been also considered previously [10–12, 15–22]. An upper bound $S_l = 1.4848 \times l^{-3/2} \times (1.8488)^l$ for the quantity of different secondary structures for sequences of length l is achieved [10]. This indicates that the anticipated volume of a neutral network is a huge quantity even for modest values of l. Among the substantial parts of the efforts that have been examined, the local characteristics of neutral networks are constrained in [30]. The analysis of the constraints allows to extract proper analytical estimations to some topological properties of neutral networks [30].

Using the RNAfold as a folding method, neutral networks over all RNA sequences of length 12 are constructed and the topological properties of these neutral networks are achieved [30]. Awkwardly, the obtained results for neutral networks in [30] can not be generalized to the superior space (real space), and we still need to examine the same specifications for longer sequences.

In this paper, the concept of RNA neutral network is extended over the space of RNA shapes. Since any RNA shape is representative for many RNA secondary structures (via reducing the length of components to 1 while preserving the vicinity and nestingness),

the weights over the nodes and edges are considered as extensions. This new version of neutral network, entitled variation network, is employed to analyze the relation between the RNA sequences and their shapes. Different topological properties are obtained for variation network of longer natural RNA sequences.

The rest of this paper is organized as follows. In Section 2., the basic definitions behind the variation network are presented. In Section 3., the datasets construction and different measures are discussed. The results and conclusions are presented in Sections 4. and 5., respectively.

2. Basic Definitions

As mentioned in Section 1., an RNA sequence is composed of four kinds of nucleotides, namely A, C, G, and U. Therefore an RNA sequence δ of length l can be considered as a string over Σ^l, where Σ is the set of nucleotides ($\Sigma = \{A, C, G, U\}$). An RNA sequence tends to fold to itself and forms pairs of bases by the formation of hydrogen bonds among Watson-Crick base pairs ($A-U$ and $C-G$) and Wobble base pair ($G-U$). This set of all base pairs is called the RNA secondary structure. Let Δ, Λ, and Γ, denote the collection of all RNA sequences, secondary structures, and shapes, respectively. The formal definition of RNA secondary structure is as follows.

Definition 1. *For an RNA sequence δ of length l, the RNA secondary structure λ is a collection of pairs (i, j), where $i, j \in \{1, ..., l\}$, $i < j$, and for any two base pairs i_1, j_1 and i_2, j_2 in λ, $i_1 = i_2 \iff j_1 = j_2$ and either $i_1 < i_2 < j_2 < j_1$ (nested) or $i_1 < j_1 < i_2 < j_2$ (disjoined) holds.*

Assume that $\varphi : \Delta \mapsto \Lambda$ maps any RNA sequence δ into its secondary structure $\lambda = \varphi(\delta)$. Also, assume that $\psi : \Lambda \mapsto \Gamma$ maps any RNA secondary structure λ into its corresponding shape $\gamma = \psi(\lambda)$. As a result, $\chi = \psi \circ \varphi : \Delta \mapsto \Gamma$ maps any RNA sequence into its corresponding shape. Considering φ as a relation, two sequences δ_1 and δ_2 are equivalence under φ, if and only if $\varphi(\delta_1) = \varphi(\delta_2)$. Similarly, two sequences δ_1 and δ_2 are equivalence under χ, if and only if $\chi(\delta_1) = \chi(\delta_2)$. Based on these equivalence relations, the generated equivalence classes of any structure and shape are defined as follows.

Definition 2. *The equivalence class of a structure λ under the mapping φ, indicated by $[\lambda]^\varphi$, is the collection of RNA sequences having the same structure as λ, i.e. $[\lambda]^\varphi =$*

$\{\delta \mid \delta \in \Delta \text{ and } \varphi(\delta) = \lambda\}$. *Similarly, the equivalence class of shape* γ *under the mapping* χ *is* $[\gamma]^{\chi} = \{\delta \mid \delta \in \Delta \text{ and } \chi(\delta) = \gamma\}$, *i.e. this class is the collection of RNA sequences having the same shape* γ.

The graphical representation of the equivalence structure and shape classes are presented in Fig. 2.

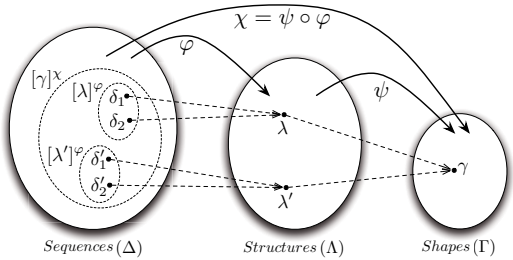


Fig. 2. Equivalence class of structures and shapes.

The formal definition of neutral network is defined as follows.

Definition 3. *For the mapping* φ *and structure* λ, *the neutral network is a graph* $NN_{\lambda} = (V, E)$ *where,*

- *The vertex set is a collection of RNA sequences, all coding the same secondary structure* λ, *i.e.* $V = \{\delta \mid \delta \in \Sigma^{l}, \varphi(\delta) = \lambda\}$,

- *Two RNA sequences* δ_1 *and* δ_2 *are connected if one of them can be obtained from the other by performing a single base mutation, i.e.* $E = \{(\delta_1, \delta_2) \mid \delta_1, \delta_2 \in V, dist(\delta_1, \delta_2) = 1\}$.

Fig. 3 demonstrates an example of neutral network for a set of five RNA sequences δ_i $(1 \leq i \leq 5)$, all coding the secondary structure λ. Let $N : \Sigma^{l} \mapsto 2^{\Sigma^{l}}$ be a neighborhood function, where for an RNA sequence δ, $N(\delta)$ denotes the set of all RNA sequences obtained from δ by performing a single mutation in different positions. This neighborhood function could be easily extended for the equivalence shape classes as $N^{*}([\gamma]^{\chi}) = \cup_{\delta \in [\gamma]^{\chi}} N(\delta)$. The neutral network does not take into account the amount of sequences that are converted from $[\lambda_1]^{\varphi}$ to $[\lambda_2]^{\varphi}$ by performing a single mutation. Considering the equivalence classes of RNA shapes, the transformation cardinality form $[\gamma_1]^{\chi}$ to $[\gamma_2]^{\chi}$ by performing a single mutation could be calculated via variation rate which is defined as follows.

Definition 4. *Variation rate for two equivalence shape classes* $[\gamma_1]^\chi$ *and* $[\gamma_2]^\chi$*, denoted by* $\omega(\gamma_1, \gamma_2)$*, is defined as:*

$$\omega(\gamma_1, \gamma_2) = |N^*([\gamma_1]^\chi) \cap [\gamma_2]^\chi| \ = \ |[\gamma_1]^\chi \cap N^*([\gamma_2]^\chi)|, \tag{1}$$

where $N^*(\cdot)$ *denotes the extended neighborhood function.*

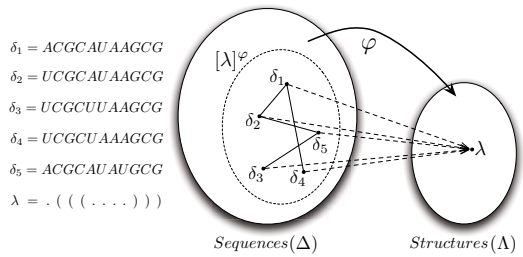


Fig. 3. An example of neutral network.

Based on the above definitions, the variation network can be expressed over the collection of all shapes as follows.

Definition 5. *For the collection of all shapes* Γ*, the variation network is a weighted graph* $VN = (V, E, W)$*, where*

- $V = \{\gamma \mid \gamma \in \Gamma\}$,
- $E = \{(\gamma_1, \gamma_2) \mid \gamma_1, \gamma_2 \in V \text{ and } \omega(\gamma_1, \gamma_2) > 0\}$,
- *For each* $(\gamma_1, \gamma_2) \in E, W(\gamma_1, \gamma_2) = \omega(\gamma_1, \gamma_2)$.

The variation network denotes the imposed relations between the collection of all shapes under the mapping χ as it is shown in Fig. 4. Regarding the above definitions, to illustrate the relations among RNA sequences and their shapes, many efforts have been done in this paper. The variation networks are constructed for natural RNA sequences and several topological measures are calculated for them. The details of our efforts as well as the results are described in the next sections.

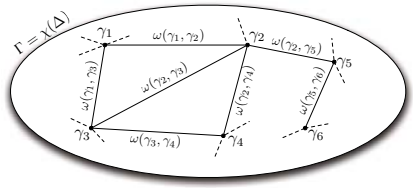


Fig. 4. An example of variation network.

3. Materials and Methods

3..1 Datasets

In order to analyze the different properties of variation networks, four datasets of RNA sequences, each of length between 1 to 50, 25 to 75, 50 to 100, and 75 to 125 nucleotides are selected from RNA STRAND server [1] [39]. Then all nucleotides appeared in different positions in each RNA sequence of length l are mutated (to three other nucleotides) to produce $3l$ more sequences (each of length l). After that the RNAShape software [31] is employed to determine their minimum free energy structures as well as their shapes. The length of sequences as well as the number of original and mutated sequences, structures, and shapes are presented in Table 1

Table 1. Constructed datasets and their characteristics.

dataset name	length	number of original sequences	number of sequences after mutation	number of structures	number of shapes
RDS01-50	1 to 50	530	48,014	9,200	520
RDS25-75	25 to 75	685	107,695	27,490	2,653
RDS50-100	50 to 100	901	190,966	56,928	6,961
RDS75-125	75 to 125	858	219,270	67,221	13,399

3..2 Measurements

In this subsection, we review the primary concepts and different available measures concerning the biological properties of networks.

Expressly, an undirected network could be represented by a symmetric adjacency matrix $A = [a_{ij}]$, where $0 \leq a_{ij} \leq 1$. The following notations are used in the rest of this paper:

- n denotes the number of vertices in the network,

[1] http://www.rnasoft.ca/strand/

- u, v, and w denote vertices,
- e_v denotes the number of joined pairs between all neighbors of v,
- η denotes the degree of the network,
- $L(u, v)$ denotes the length of the shortest path between two vertices u and v,
- $J(u, v)$ denotes the number of neighbors shared between the vertices u and v.

In order to achieve the results, several topological measures have been employed in our analysis for each dataset as follows.

- **Average Shortest Paths Length** ($ASPL$) gives the average value of shortest paths length in the network.
- **Network Diameter** (D) of a network is marked out as the length of maximum shortest paths between any two nodes in the network.
- **Shortest Paths Distribution** gives the number of pairs (u, v) such that $L(u, v) = k$, for all $k \in \{1, 2, ..., D\}$.

  The shortest path length distribution and the network diameter might show small-world properties of the analyzed network [32]. Barabasi and Oltvai [33] hypothesized that the occurrence of small-world networks in biological systems may display an evolutionary advantage of such architecture.

- **Connectivity of Node** v (k_v) is the weighted sum over the neighbors of v, i.e.

$$k_v = \sum_{v \neq u} a_{vu}. \tag{2}$$

  The neighbors of node v is the collection of all nodes that are close to v involving v itself. In an unweighted network, the connectivity k_v of node v is the number of directly linked neighbors (degree of v).

- **Network Density** (N_d) is the normalized variety of average connectivity of nodes, i.e.,

$$N_d = \frac{\sum_i \sum_{j \neq i} a_{ij}}{\frac{n(n-1)}{2}}. \tag{3}$$

N_d displays how densely the network is populated with edges. A network which comprises of no edges has a density of 0, and in contrast, the density of a clique is 1.

- **Network Centralization (N_c)** is given by:

$$N_c = \frac{n}{n-2}(\frac{\eta}{n-1} - N_d), \tag{4}$$

where a network whose topology look likes a star, has N_c close to 1, whereas a decentralized network is characterized by N_c close to 0.

- **Network Heterogeneity (N_h)** is the coefficient of variation of the connectivity distribution, i.e.

$$N_h = \frac{\sqrt{variance(k)}}{mean(k)}. \tag{5}$$

This measure reflects the trend of a network to have highly connected nodes.

- **Clustering Coefficient of v (C_v)** is defined as

$$C_v = \frac{2e_v}{(k_v(k_v-1))}, \tag{6}$$

where $C_v \in [0,1]$ and it reflects the trend of nodes to cluster together. Evidence implies that in the majority of real-world networks, nodes tend to form closely knit groups characterized by a comparatively high density of ties [32,33].

- **Average Clustering Coefficient Distribution ($ACCD$)** gives the average value of the clustering coefficients for every node v with i neighbors ($i \in \{1,2,...,\eta\}$).

- **Degree Distribution (DD)** gives the number of nodes with degree i ($i \in \{1,2,...,\eta\}$).

Barabasi and Oltvai [33] used this property to differentiate between scale-free networks and random networks (as defined in [34,35]).

- **Topological Coefficients of v (T_v)** is computed as:

$$T_v = \frac{1}{k_v} \cdot average_u(J(v,u)) = \frac{1}{k_v} \cdot \frac{1}{n_J} \cdot \sum_{u \in V-\{v\}} J(v,u), \tag{7}$$

where n_J denotes the number of nodes, say u, that share at least one neighbor with v and $J(v,u)$ is defined for these kinds of nodes. T_v could be used to show the trend of the nodes to have shared neighbors.

- **Stress Centrality of** v **(**$C_s(v)$**)** is the number of shortest paths passing through the node v [13, 36].

- **Stress Centrality Distribution (**SCD**)** gives the number of nodes having the stress centrality i, for various values of i.

  A node has an extreme stress if it is traversed by a great number of shortest paths (vital vertex).

- **Betweenness Centrality of** v **(**$C_b(v)$**)** is calculated as:

$$C_b(v) = \sum_{u \neq w \neq v} \frac{\sigma_{uw}(v)}{\sigma_{uw}}, \tag{8}$$

  where u and w are nodes, σ_{uw} indicates the number of shortest paths from u to w, and $\sigma_{uw}(v)$ is the number shortest paths from u to w passing through v [41]. $C_b(v)$ denotes the amount of control that node v exerts over the connections of other nodes in the network.

- **Closeness Centrality of** v **(**$C_c(v)$**)** is calculated as:

$$C_c(v) = \frac{1}{average_u(L(v,u))} = \frac{n_L}{\sum_{u \in V} L(v,u)}, \tag{9}$$

  where n_L denotes the number of nodes that are accessible from v and $L(v,u)$ is defined for these kinds of nodes. Here, $C_c(v)$ is a metric that represents how fast the information is distributed from node v to other accessible nodes [37].

- **power-law (**$f(x)$**)** is a function which has exponentially relation with x. i.e.,

$$f(x) \sim ax^b. \tag{10}$$

  This function generally explains a system where the larger events are more infrequent than smaller events.

All the above mentioned measures are evaluated on the variation networks corresponding to different datasets and the obtained results are presented in the next section.

4. Results and Discussions

As mentioned, the RNAshape software [31] is employed to determine the minimum free energy structures of all RNA sequences appeared in the constructed datasets and their

corresponding shapes. Then the variation networks over the resulting shapes of each dataset is constructed separately. After that, the clustering coefficient distribution, node degree distribution, shortest path length distribution, topological coefficient, stress centrality distribution, betweenness centrality distribution, and closeness centrality of these variation networks are calculated and presented in Figs. 5 to 11, respectively. All these figures are provided by the cytoscape software [38, 42].

For each constructed variation network, the related properties are shown in Table 2. Since the N_d (network density) of each variation network is very low (between 0.0002 and 0.011), the C_v, D and number of components are considerable. Except the last variation network, the low number of connected components proposes a stronger connectivity in the constructed variation networks. By ignoring the small clusters in all variation networks, they are mostly connected. In the worse case, each node in these variation networks is accessible from other nodes by a path of length between seven to seventeen ($D \in \{7, 8, 11, 17\}$).

Although N_d is low, the robustness of these networks does not corrupt by deleting a random node. It is happened because of the relatively high value of C_v. By omitting a random node from a network, it is not divided into two separate subnetworks. As presented in Table 2, N_c is lower than 0.25, and therefore the variation network is a kind of a decentralized network.

Variation network as a kind of biological network attends to be very heterogeneous. In other words, while some nodes are highly connected, the popular nodes attend to have very few connections. As indicated in Table 2, N_h reflects the trend of a network to have highly connected nodes [40].

The clustering coefficient for every node in each variation network is also computed. The power-law function is calculated (by fitting the curve over the data points) for this property and it is highly correlated with it as presented in Fig. 5 and Table 3.

For each variation network, the node degree distribution is calculated and presented in Fig. 6. As it is understood, the variation network is a kind of scale-free network. Evidence for this claim is the dense path length distribution and its highly correlation with power-law function as presented in Fig. 7 and Table 3.

According to Table 3 and Fig. 7, $ASPL$ of each variation network is also correlated with the power-law function. Considering this correlation, removal of an accidental node

hardly ever causes a dramatic increase in $ASPL$ (or a dramatic decrease in the C_v). Because of a huge number of shortest paths run through highly connected nodes, if an unimportant node is omitted, it is doubtful to affect the other paths between the remaining nodes.

T_v is also calculated for each node in these variation network and planned against the number of neighbors (Fig. 8). As indicated in Table 3, this measure is also correlated with the power-law function. As illustrated in Fig. 8, T_v is reduced with respect to the number of neighbors, and hence the nodes with highest degree are not artificially clustered together. Furthermore, it verifies the modular network organization specified by the C_v.

For each variation network, stress centrality distribution is calculated and presented in Fig. 9. To calculate the stress centrality, the obtained values are grouped into bins whose sizes grow exponentially by a factor of 10. In all variation networks, the majority of nodes (about a thousand nodes) have an extreme stress, i.e. $10^3 \leq C_s(v) \leq 10^7$. Hence, they are traversed by a great number of shortest paths in the variation networks.

C_b is also calculated and presented in Fig. 10 for each variation network. Although C_b raises against the number of neighbors, it is low in most of the cases. As mentioned in Subsection 3..2, $C_b(v)$ denotes the ratio of the number of shortest path passing through node v to the number of all shortest paths, i.e., C_b is the amount of control that this node exerts over the connections of other nodes in the network. As presented in Fig. 10, the control of each variation network is distributed in all nodes. As network $RDS01-50$ is composed of one component, $C_b(v) < 0.15$ holds for all its nodes . Therefore, all the nodes lie in a community (dense subnetworks), rather than join communities. By omitting tiny components, this fact could be generalized to the other three variation networks.

The closeness centrality of all nodes is also calculated and drawn against the amount of neighbors (Fig. 11). C_c is a metric that show how fast information distributed from a given node to other accessible nodes [37]. As shown in Fig. 11, for majority of nodes, C_c varies between 0.1 and 0.3.

5. Conclusion

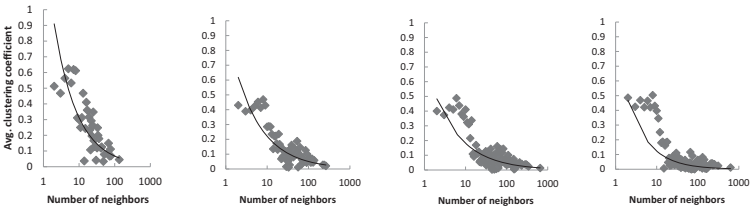
In this paper, a novel concept, entitled variation network, is introduced to analyze the relationships between RNA sequences and their shapes. Although the function of an RNA molecule is related to its tertiary structure, the RNA shape shows the higher order of its

Table 2. Summery of measurements corresponding to the constructed variation networks. C_v, D, N_h, N_c, $aspl$, and N_d denote the clustering coefficient, network diameter, network heterogeneity, network centralization, average shortest path length, and network density, respectively.

Name	C_v	*#Connected Components*	D	N_h	N_c	*#Shortest Path*	$ASPL$	*#Nodes*	N_d
$RDS01-50$	0.243	1	7	2.072	0.249	269,880	3.378	520	0.011
$RDS25-75$	0.137	6	8	2.898	0.098	6,706,324	4.081	2,653	0.002
$RDS50-100$	0.111	16	11	3.941	0.094	42,781,838	4.595	6,961	0.001
$RDS75-125$	0.078	43	17	4.100	0.047	129,759,980	6.730	13,399	0.0002

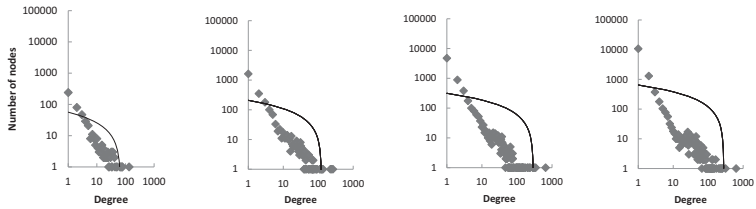
Table 3. The parameter of power-law function $f(x) \sim ax^b$ obtained by fitting the curve over the data points. The correlation among the specified data points and the related points on the fitted curve is represented by *corr*. Additionally, the R-squared value is reported as rs.

Dataset	Node Degree Distribution				Clustering Coefficient Distribution				Topological Coefficient			
Name	a	b	$corr$	rs	a	b	$corr$	rs	a	b	$corr$	rs
$RDS01-50$	100.84	−1.171	0.987	0.874	1.441	−0.664	0.761	0.574	0.902	−0.7	0.978	0.946
$RDS25-75$	385.44	−1.318	0.972	0.876	0.964	−0.64	0.848	0.553	0.934	−0.81	0.987	0.977
$RDS50-100$	653.58	−1.331	0.96	0.859	0.737	−0.608	0.838	0.43	0.947	−0.852	0.99	0.983
$RDS75-125$	650.62	−1.287	0.929	0.84	0.83	−0.828	0.851	0.342	0.998	−0.902	0.985	0.984



(a) RDS01-50 (b) RDS25-75 (c) RDS50-100 (d) RDS75-125

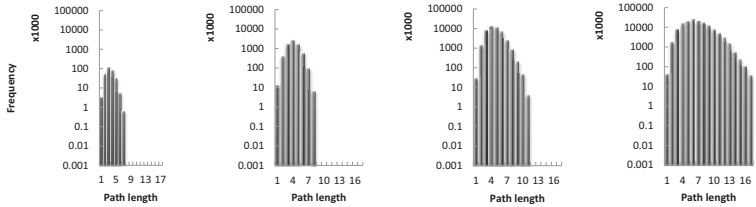
Fig. 5. The clustering coefficient distribution of the constructed variation networks.



(a) RDS01-50 (b) RDS25-75 (c) RDS50-100 (d) RDS75-125

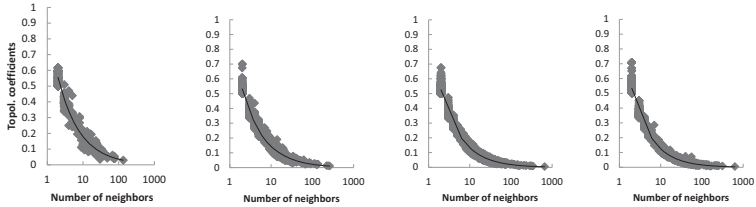
Fig. 6. The node degree distribution of the constructed variation networks.

functionality inside the cell. Based on natural RNA sequences, variation networks are constructed and some topological properties are evaluated. Correlation between degree distribution and power-law function makes evidence that the obtained variation network is a kind of scale free networks. As diameters of obtained networks are lower than seventeen, each RNA shape equivalence classes is accessible with in most seventeen mutational steps.



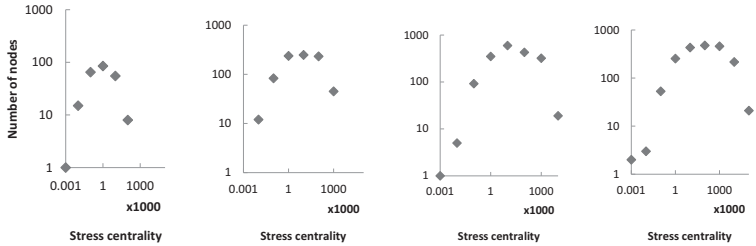
(a) RDS01-50 (b) RDS25-75 (c) RDS50-100 (d) RDS75-125

Fig. 7. The shortest path length distribution of the constructed variation networks.



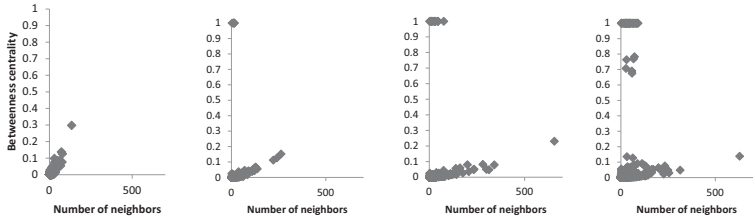
(a) RDS01-50 (b) RDS25-75 (c) RDS50-100 (d) RDS75-125

Fig. 8. The topological coefficient of the constructed variation networks.



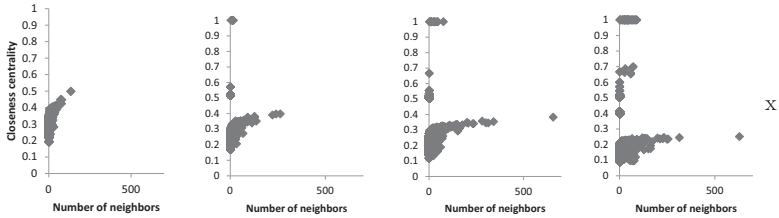
(a) RDS01-50 (b) RDS25-75 (c) RDS50-100 (d) RDS75-125

Fig. 9. Stress centrality distribution of the constructed variation networks.



(a) RDS01-50 (b) RDS25-75 (c) RDS50-100 (d) RDS75-125

Fig. 10. Betweenness centrality distribution of the constructed variation networks.



(a) RDS01-50 (b) RDS25-75 (c) RDS50-100 (d) RDS75-125

Fig. 11. Closeness centrality of the constructed variation networks.

It means that the variation network is a small-word network. Also, topological coefficient shows that the nodes with many neighbors are not artificially clustered together in variation network. Furthermore, it demonstrates the modular network organization defined by the clustering coefficient. Finally, some centrality measures are calculated for each RNA shape equivalence class in each network. Although betweenness centrality raises against the number of neighbors, this measure is low in most of the results. Obtained closeness centrality is lower than 0.4 and it raises slowly by increasing the number of neighbors. With respect to the obtained results, we could decide that the variation network is a complex network and it has some dynamic information for further researches.

References

[1] F. H. C. Crick, On protein synthesis, *Symp. Soc. Exp. Biol.* **12** (1958) 63–138.

[2] F. H. C. Crick, Central dogma of molecular biology, *Nature* **227** (1970) 561–563.

[3] N. Takeuchia, P. Hogewegb, Evolutionary dynamics of RNA-like replicator systems: A bioinformatic approach to the origin of life, *Phys. Life. Rev.* **3** (2012) 219–263.

[4] M. Ganjtabesh, J. M. Steyaert, Enumerating RNA structures, including pseudoknots of any topology, *MATCH Commun. Math. Comput. Chem.* **66** (2011) 399–414.

[5] A. Nakayaa, A. Yonezawaa, K. Yamamotob, Classification of RNA secondary structures Using the techniques of cluster analysis, *J. Theor. Biol.* **183** (1996) 105–117.

[6] M. A. Huynen, D. A. Konings, P. Hogeweg, Multiple coding and the evolutionary properties of RNA secondary structure, *J. Theor. Biol.* **165** (1993) 251–267.

[7] R. Aguirre–Hernandez, H. H. Hoos, A. Condon, Computational RNA secondary structure design: empirical complexity and improved methods, *BMC Bioinformatics* (2007) 8–34.

[8] F. Zare–Mirakabad, M. Sadeghi, H. Ahrabian, A. Nowzari-Dalini, RNA Comp: A new method for RNA secondary structure alignment, *MATCH Commun. Math. Comput. Chem.* **61** (2009) 789–816.

[9] S. Bernhart, I. Hofacker, P. Stadler, PM-Match - A new way to align RNA structures *Abstracts of Math-Chem-Comp*, 2004, http://mcc.irb.hr/mcc04/abs04.html.

[10] P. Schuster, W. Fontana, P. F. Stadler, I. L. Hofacker, From sequences to shapes and back: a case study in RNA secondary structures, *Proc. Biol. Sci.* **255** (1994) 279–284.

[11] W. Gruner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, P. Schuster, Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structures of neutral networks and shape space covering, *Monatsh. Chem.* **127** (1996) 375–389.

[12] C. Reidys, P. Stadler, P. Schuster, Generic properties of combinatory maps: neutral networks of RNA secondary structures, *Bull. Math. Biol.* **59** (1997) 339–397.

[13] U. Brandes, A faster algorithm for betweenness centrality, *J. Math. Sociol.* **25** (2001) 163–177.

[14] S. Wuchty, E. Ravasz, A. L. Barabasi, The architecture of biological networks, in: T. S. Deisboeck, J. Y. Kresh (Eds.), *Complex Systems Science in Biomedicine*, Springer, New York, 2006, pp. 165–182.

[15] W. Fontana, D. A. M. Konings, P. F. Stadler, P. Schuster, Statistics of RNA secondary structures, *Biopolymers* **33** (1993) 1389–1404.

[16] W. Gruner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, P. Schuster, Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks, *Monatsh. Chem.* **127** (1996) 355–374.

[17] M. C. Cowperthwaite, E. P. Economo, W. R. Harcombe, E. L. Miller, L. A. Meyers, The ascent of the abundant: How mutational networks constrain evolution, *PLoS Comput. Biol.* **4** (2008) e1000110.

[18] T. Jorg, O. C. Martin, A. Wagner, Neutral network sizes of biological RNA molecules can be computed and are not atypically small, *BMC Bioinformatics* **9** (2008).

[19] M. Stich, S. C. Manrubia, Motif frequency and evolutionary search times in RNA populations, *J. Theor. Biol.* **280** (2011) 117–126.

[20] M. Huynen, Exploring phenotype space through neutral evolution, *J. Mol. Evol.* **43** (1996) 165–169.

[21] M. Huynen, P. Stadler, W. Fontana, Smoothness within ruggedness: The role of neutrality in adaptation, *Proc. Natl. Acad. Sci. U.S.A.* **93** (1996) 397–401.

[22] E. van Nimwegen, J. Crutchfield, M. Huynen, Neutral evolution of mutational robustness, *Proc. Natl. Acad. Sci. U.S.A.* **96** (1999) 9716–9720.

[23] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, M. Gerstein, Genomic analysis of regulatory network dynamics reveals large topological changes, *Nature* **431** (2004) 308–312.

[24] M. E. J. Newman, The structure and function of complex networks, *SIAM Review* **45** (2002) 167–256.

[25] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A. L. Barabasi, The large-scale organization of metabolic networks, *Nature* **407** (2000) 651–654.

[26] R. Tanaka, Scale–rich metabolic networks, *Phys. Rev. Lett.* **94** (2005).

[27] S. H. Yook, Z. N. Oltvai, A. L. Barabasi, Functional and topological characterization of protein interaction networks, *Proteomics* **4** (2004) 928–942.

[28] A. Wagner, D. A. Fell, The small world inside large metabolic networks, *Proc. R. Soc. Lond. B.* **268** (2001) 1803–1810.

[29] J. Wang, M. Huihui, F. Wang, F. Jin, Exploring the network structure and nodal centrality of China's air transport network: A complex network approach, *J. Transport Geog.* **19** (2011) 712–721.

[30] J. Aguirre, J. M. Buldu, M. Stich, S. C. Manrubia, Topological structure of the space of phenotypes: The case of RNA neutral networks, *PLoS One* **6** (2012) 10.1371.

[31] P. Steffen, B. Vo, M. Rehmsmeier, J. Reeder, R. Giegerich, RNAshapes: an integrated RNA analysis package based on abstract shapes, *Bioinformatics* **22** (2005) 500–503.

[32] D. J. Watts, S. H. Strogatz, Collective dynamics of 'small–world' networks, *Nature* **393** (1998) 440–442.

[33] A. L. Barabasi, Z. N. Oltvai, Network biology: understanding the cell's functional organization, *Nat. Rev. Genet.* **5** (2004) 101–113.

[34] P. Erdos, A. Renyi, On the evolution of random graphs, *Publ. Math. Inst. Hung. Acad. Sci.* **5** (1960) 17–61.

[35] B. Bollobás, *Random Graphs*, Cambridge Univ. Press, Cambridge, 2001.

[36] A. Shimbel, Structural parameters of communication networks, *Bull. Math. Biophys.* **15** (1953) 501-507.

[37] M. E. J. Newman, A measure of betweenness centrality based on random walks, *Social Networks* **27** (2005) 1–15.

[38] Y. Assenov, F. Ramirez, S. E. Schelhorn, T. Lengauer, M. Albrecht, Computing topological parameters of biological networks, *Bioinformatics* **242** (2008) 282–284.

[39] M. Andronescu, V. Bereg, H. H. Hoos, A. Condon, RNA STRAND: The RNA secondary structure and statistical analysis database, *BMC Bioinformatics* **9** (2008).

[40] J. Dong, S. Horvath, Understanding network concepts in modules, *BMC. Syst. Biol.* **24** (2007).

[41] E. Estrada, Characterization of topological keystone species local, global and meso–scale centralities in food webs, *Ecological Complexity* **4** (2007) 48–57.

[42] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A software environment for integrated models of biomolecular interaction networks, *Genome. Res.* **13** (2003) 2498–2504.