Spectral Graph Partitioning Analysis of In Vitro Synthesized RNA Structural Folding

Stanley NG Kwang Loong[1,2] and Santosh K. Mishra[2,1]

[1] NUS Graduate School for Integrative Sciences & Engineering,
National University of Singapore
Block MD 11, Clinical Research Centre, #01-10, 10 Medical Drive, Singapore 117597
stanley@bii.a-star.edu.sg
[2] Bioinformatics Institute, Agency for Science, Technology and Research
Matrix, 30 Biopolis Street, #07-01, Singapore 138671
santosh@bii.a-star.edu.sg

Abstract. In this paper, we investigate the topological properties of synthetic RNAs (i.e., functional RNAs synthesized by *in vitro* selection technique), by applying the spectral graph partitioning technique. Our analysis shows that the majority of synthetic RNAs possess between two to six vertices and their second eigenvalues lie between one and two. In contrast, natural RNA structures mostly have nine or ten vertices and are less compact with the second eigenvalue below unity. Our statistical analysis (at 95 percentile) also reveals three criteria important for designing novel functional RNAs. Firstly, RNA sequences screened from a large random library, with length of 80 nucleotides and 32.31% paired bases, are very likely to fold into functional RNAs. Secondly, their predicted structures should possess two to six vertices inclusively. Thirdly, to minimize the number of false positives, a combination of filtering parameters should be included, the percentage G/C content of 65.95% and the normalized minimum free energy of -0.021 kcal/mol per nucleotide.

1 Introduction

Emerging experimental evidence demonstrates that many families of the naturally occurring non-coding protein RNA molecules (ncRNAs) found in prokaryotic and eukaryotic genomes, are actually integral players of the cellular machinery. The importance of functional ncRNAs participating at multiple regulatory layers and influencing a plethora of vital biological processes like transcriptional regulation, mRNA stability and localization, RNA processing and modification, and translation is becoming increasingly apparent [1-3].

Two recent and notable discoveries of regulatory ncRNAs, riboswitches [4;5] and microRNAs [6-8], further highlight their vital regulatory roles in many organisms. The former are highly conserved RNA regulatory elements embedded within the 5' untranslated region of biosynthesis genes or operons, and *cis*-modulate their expressions upon binding to metabolite (e.g., guanine and thiamine pyrophosphate), *without* involving protein cofactors. The latter are transcribed from the endogenous transcripts

J.C. Rajapakse, L. Wong, and R. Acharya (Eds.): PRIB 2006, LNBI 4146, pp. 81 – 92, 2006.

and subsequently processed by the Dicer enzyme to generate ~22 nucleotides mature microRNAs. These microRNAs bind to specific complementary sites on the mRNA transcripts to down-regulate expression of targeted genes either by inhibiting protein synthesis or causing degradation of their target mRNAs [9]. Comprehensive phenotypic and gene expression analysis have also suggested an intrinsic association between oncogensis and human microRNAs found in the tumor tissues [10-12].

"*In vitro* selection" or commonly known as "Systematic Evolution of Ligands by Exponential Enrichment (SELEX)", is an iterative combinatorial chemistry method [13]. Through this technique, many novel functional ncRNAs with specific physical or chemical properties can be isolated and preferentially enriched from a large $>10^{15}$ random RNA sequences library. Synthetic counterparts of natural functional ncRNAs include the target-binding RNA molecules (i.e., aptamers) that bind to a desired metabolite (such as ligands, antibiotics, peptides, and whole viruses) with nano-molar affinity, and the allosteric/catalytic RNAs (i.e., ribozymes) that *trans*-cleave the target mRNAs upon recognition of specific sequence patterns [14;15]. SELEX has been widely adopted as an important research and development tool in the fields of analytical, diagnostic, and therapeutic applications for synthesizing molecular switches and sensors [16;17], and therapeutic agents [18].

The functional activities of both synthetic and natural ncRNAs are pre-determined by their distinct local RNA secondary structures. Their capacity to perform ligand-binding, complementary base pairing, and catalytic reactions, are possible due to the conformational flexibility, modularity, and versatility of RNA molecules [19;20]. For instance, stem-loop structures present in the 5' untranslated regions of eukaryotic genomes may occlude association of the 40S ribosomal subunit with the mRNA to inhibit the translation initiation [21], while those present in bacteria can attenuate the mRNA degradation rate through inhibiting the nuclease activity [22]. Interestingly, hairpin structures discovered in the protein-coding regions of *E. coli*, *S. typhi*, and *S. cerevisiae* are also involved in the widespread regulation of RNA processing, mRNA stability, and translational control [23].

"Spectral graph partitioning" is a promising technique for analyzing RNA secondary structures, while providing an efficient organization of the structural folds using their topological properties [19;20]. The underlying principle is that the unique topology of RNA structural motifs (including the loops of hairpin, internal, bulge, and multi-branch, as well as stems) corresponds uniquely to a planar tree-graph representation, where the vertices are connected by incident edges. Applying the concept of spectral graph partitioning derived from the field of domain decomposition in parallel computing [24], the degree of connectivity of the tree-graph is represented by the Laplacian matrix. From which, the matrix's mathematical properties such as its associated eigenvalues and the number of vertices can be derived for characterizing and screening RNA secondary structures.

Since its introduction in 2003 [19;20], bioinformatics applications include the prediction of multiple mutations to disrupt motifs in riboswitches [25], the prediction of RNA conformational switch by mutation [26;27], the search and analysis of RNA secondary structures [28], the classification of RNA coarse-grained tree-graph structures [29;30], and lastly for systematically partitioning complex RNA structures into simpler fragments with maximal decoupling between them [20]. Together, they underscore the potential of spectral graph partitioning as an invaluable *in silico* tool to

elucidate the topological patterns hidden in genomic sequences and to offer a tremendous opportunity for an enhanced understanding of functional genomics.

To the best of authors' knowledge, the application of spectral graph partitioning on functional RNAs generated by SELEX, has *not* been systematically explored. Motivated by this challenge, our contribution in this paper, is a computational study based on the recently published spectral graph partitioning technique, for characterizing and analysis of *in vitro* synthesized RNAs. To date, the Aptamer Database [31] archives over 3,000 DNA/RNA sequences (across 300+ articles as of October 2005) of aptamers and synthetic ribozymes spanning diverse functions, that have been synthesized by *in vitro* selection method. This comprehensive and valuable knowledge-based resource, which is updated monthly, provides a unique opportunity and fertile ground for characterizing synthetic RNA's two-dimensional structures by applying new conceptual and mathematical approaches. A secondary motivation is, through this investigation, the prediction, screening, and engineering of novel functional ncRNA molecules *in silico* can be further advanced.

This paper is organized as follows. Section 2 describes the methods and materials. It introduces briefly the basic principles of predicting RNA secondary structure, the graph theoretic formalisms for RNA structures using tree-graph representation and Laplacian matrix, and the technique of spectral graph partitioning. In section 3, the results for 1,943 synthetic RNA sequences are presented and discussed. Finally, section 4 concludes with a discussion on the notable findings of the proposed methodology, and avenues for future research are outlined.

2 Methods and Materials

A typical experimental setup using spectral graph partitioning analysis is illustrated in **Fig. 1**. Briefly, through *in vitro* selection, a small RNA aptamer was isolated that bound with nano-molar affinity to human transcription factor *NF-kappa B* [32]. This transcription factor is a key regulator of inflammation, HIV-1 gene expression, and apoptosis. Experimental evidence showed binding of the small RNA aptamer effectively inhibited the ability of *NF-kappa B* to bind duplex DNA, performing as an anti-*NF-kappa B* [32].

Given the primary sequence of the anti-*NF-kappa B* described in FASTA format, (step **A**) its optimal secondary structure is predicted using RNAfold [33] (or mfold prediction server [34;35]). This program computes the lowest minimum free energy of folding (MFE) for the most favorable RNA conformation, based on the Zuker's energy minimization algorithm [34;35] and the experimentally derived Turner's energy rules [36]. Default parameters and temperature setting of 37°C were used throughout this study to ensure consistency. The output of RNAfold is a FASTA-like format appended with the optimal structure in Vienna dot-bracket notation with the base pairs and unpaired bases represented by brackets '( )' and dots '.' [33], respectively and the minimum free energy. In this example, the RNA secondary structure predicted by RNAfold has three hairpin loops, 5' and 3' termini, an internal loop, and two multi-branch loops – all of these stabilized by six stems. After applying the pair of vertex-edge rules (step **B**) [29;30;37], the structure in bracket notation is converted into a planar tree-graph consisting of six arbitrarily labeled vertices (•) connected by five unweighted edges (—). Finally, (step **C**) the six by six Laplacian matrix and its

corresponding eigenvalue spectrum are computed using the spectral graph partitioning technique. Of interest is the second eigenvalue $\lambda_2 = 0.438447$ (indicated by ←) that describes the compactness of the RNA structure [25;28]. This value increases monotonically when the tree-graph topology has greater degree of compactness.

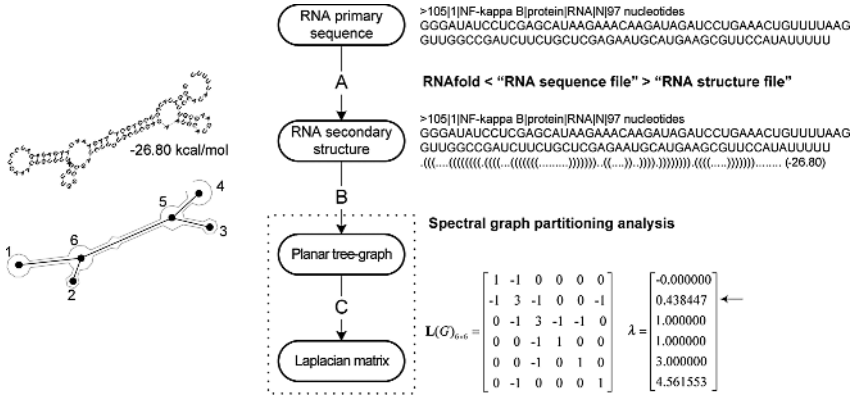

Fig. 1. Methodology for spectral graph partitioning analysis of *in vitro* synthesized RNA structural folding. Our predicted structure differs in shape from the one published [32], though both are predicted by Zuker's energy minimization algorithm. Possibly the cited work was using mulfold, an older Macintosh version of mfold with previous energy rules or different default settings. ←, *the second eigenvalue* λ_2.

2.1 Representing RNA Secondary Structure as Planar Tree-Graphs

The primary structure of a linear RNA chain molecule is the nucleotide sequence $\mathbf{s} = s_1 s_2 \ldots s_i \ldots s_L$, and runs in the direction 5' → 3' terminus. L is the number of nucleotides and $s_i \in \sum = (A, C, G, U)$ is the biochemical nucleotide at the i^{th} position. The RNA molecule **s** folds upon itself relatively rapid into a two-dimensional RNA secondary structure S [38]. The structure S is stabilized by the energetic interactions of Watson-Crick $G \equiv C$ and $A = U$, wobble $G = U$, and other non-canonical base pairings.

A planar RNA secondary structure S depicted in **Fig. 2**, is mathematically defined by a set of base pairings $(i, j) \in S$ connecting bases s_i and s_j, where $i < j$ [39]. Given (i, j) and $(k, l) \in S$, a nucleotide can base pair to at most one other nucleotide i.e., $i = k \Leftrightarrow j = l$. A set of $\Delta \in \mathbf{Z}^+$ consecutive base pairs defines a stem for stabilizing the structure against thermal fluctuations. In addition, the number of unpaired nucleotides between paired s_i and s_j should at most be $\theta \in \mathbf{Z}^+$ i.e., $i < j + \theta$. Otherwise, the structural motif is considered as an unpaired-loop of multi-branch, bulge, hairpin, or internal. Though unpaired-loops tend to destabilize the structures, they are generally necessary for the biological functions. For example, the cloverleaf structure of tRNA performs a critical role in transporting amino acids to the site of polypeptide synthesis [40]. A set of motifs containing nested base pairs i.e., $i < k < j \Leftrightarrow i < l < j$ are classified as pseudoknots. These pseudoknots are excluded in this spectral graph partitioning analysis, as both RNAfold and mfold assume the base pairs are *non-crossing* i.e.,

they disallow $i < k < j < l$ and $k < i < l < j$ for reasons of computational complexity. Illustrated in **Fig. 2**, the native RNA secondary structure consists of two hairpin loops, an internal loop, a bulge loop, a multi-branch loop, six stems, and a pseudoknot.

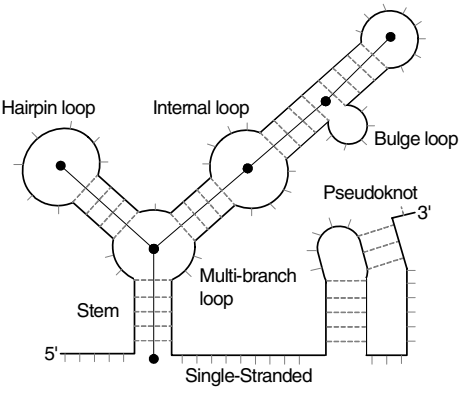

Fig. 2. Planar schematic of a RNA secondary structure and its embedded motifs. **Hairpin loop**, folds upon itself; **Internal loop**, an unpaired region between two stems due to mismatched (e.g., AG and CU) or unpaired bases; **Bulge loop**, an asymmetrical internal loop formed from one strand; **Multi-branch loop or junction**, more than two stems coincide with some unpaired bases; **Stem**, a base paired region; **Pseudoknot**, a long-range interaction between an upstream exposed loop (e.g., bulge loop) and downstream region. *Short and long dashed lines indicate unpaired nucleotides and paired bases, respectively.* (•) *and* (—) *represent vertex and edge, respectively after processing through the pair of vertex-edge rules* [29;30;37].

A RNA secondary structure S is represented as a RNA planar tree-graph $G = (V, E)$ after processing through the following pair of vertex-edge rules [29;30;37]. Referring to **Fig. 2**, the predicted RNA structure (ignoring the pseudoknot as it is not predicted by RNAfold and mfold) is represented as a planar tree-graph consisting of six vertices (•) and five edges (—).

1. Vertex, V (•) represents a set of $\theta \geq 1$ mismatched nucleotides or unmatched pairs of bases that of a bulge loop, hairpin loop, internal loop, the 5' and 3' unpaired termini, and the multi-branch loop. In general, vertices are arbitrarily labeled in the direction 5' $\rightarrow$ 3' terminus.
2. Edge, E (—) denotes a RNA stem having $\Delta \geq 2$ consecutive complementary pairs stabilized by the canonical Watson-Crick $G \equiv C$ and $A = U$, and wobble $G = U$. Edges may optionally be weighted by the frequency of paired bases, or the percentage G/C content, or the minimum free energy (MFE) of the helical duplex.

2.2 Spectral Graph Partitioning Theory of RNA Secondary Structures

A RNA planar tree-graph $G = (V, E)$ is a mathematical formalism composed of n vertices $v_i \in V$, $i = (1, 2, \ldots, |V|)$ connected by m incident undirected edges $(v_i, v_j) \in E$, each of which is assigned an edge weight E_{ij}. Without loss of generality, edges are

unweighted i.e., E_{ij} = 1 [25;28]. The tree-graph G is uniquely represented by the Laplacian matrix $\mathbf{L}(G)_{n\times n}$ as defined in Eq. (**1**).

$$G = (V, E) \leftrightarrow \mathbf{L}(G) = \mathbf{D}(G) - \mathbf{A}(G). \tag{1}$$

Where $\mathbf{D}(G)_{n\times n}$ and $\mathbf{A}(G)_{n\times n}$ are known as the degree and adjacency matrices of the tree-graph G, respectively. The diagonal elements d_{ij} of $\mathbf{D}(G)_{n\times n}$ specify the degree or the minimum number of incident edges that each vertex v_i connects with the other vertices $v_j \neq v_i$, denoted by deg(v_i). In relation to a RNA secondary structure, d_{ij} takes on values of deg(v_i) = 1 for hairpin loop, as well as 5' and 3' unpaired termini; deg(v_i) = 2 for internal and bulge loops; and deg(v_i) > 2 for multi-branch loop. The off-diagonal elements a_{ij} of $\mathbf{A}(G)_{n\times n}$ specify whether there exists an incident edge connecting the vertices v_i and v_j. If v_i and v_j are adjacent a_{ij} = 1, otherwise a_{ij} = 0.

$\mathbf{L}(G)_{n\times n}$ is a symmetric matrix having each of its rows and columns indexed by V, and individually total to zero. The value of element l_{ij} is given by the difference between d_{ij} and a_{ij}, as defined in Eq. (**2**). It specifies the degree of connectivity between the vertices v_i and v_j of the tree-graph G.

$$l_{ij} = \begin{cases} d_{ij} = \deg(v_i), \text{ if } i = j, \\ -a_{ij} = -1, \text{ if edge } (v_i, v_j) \in E \ \wedge \ i \neq j, \\ 0, \text{ if edge } (v_i, v_j) \notin E. \end{cases} \tag{2}$$

Applying the "Eigen-decomposition theorem" onto $\mathbf{L}(G)_{n\times n}$, as shown in Eq. (**3**),

$$\mathbf{L}(G)\mathbf{X} = \lambda\mathbf{X} \Leftrightarrow \left[\mathbf{L}(G) - \lambda\mathbf{I}\right]\mathbf{X} = \mathbf{O}. \tag{3}$$

Where eigenvalue λ is some scalar of $\mathbf{L}(G)_{n\times n}$ with its corresponding eigenvector $\mathbf{X} \in \Re^n \neq 0$. $\mathbf{I}$ and $\mathbf{O}$ are the identity and null matrices. Equation (**3**) has non-trivial solutions if and only if the condition given in Eq. (**4**) is satisfied,

$$\det\left|\mathbf{L}(G) - \lambda\mathbf{I}\right| = 0. \tag{4}$$

Solving the n^{th}-degree characteristic polynomial in Eq. (**4**) generates the entire set of ordered eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. This set is the matrix's eigenvalue spectrum quantifying the connectivity as well as characterizing the graph similarity. Generally, $\mathbf{L}(G)$ is always positive semi-definite such that the first eigenvalue λ_1 = 0 and those of higher orders $\lambda_{k>1} \in \Re^+$ [25;28]. Of interest is the second (also known as the Fiedler) eigenvalue λ_2, it represents mathematically the algebraic connectivity of the tree-graph G. In relation to the RNA secondary structure, λ_2 defines the compactness of the RNA topology at the coarsest scale [25;28]. RNA structures having similar values of λ_2 tend to be similar in topologies. Generally, the value of λ_2 increases monotonically with greater compactness in the RNA structure. Large values correspond to vertices of high degree that are in close proximity, while small values for more equally dispersed edge set. Maximum value of λ_2 is either 1 or 2 for an $n > 2$ perfectly connected star-shaped tree-graph or for n = 2 linear tree-graph, respectively [25;28].

3 Results and Discussion

We have retrieved 3,498 DNA/RNA sequences spanning across 308 articles from the Aptamer Database [31], as of October 2005. Among them, 1,943 RNA sequences gathered from 124 articles were used for investigation in the spectral graph partitioning analysis. **Fig. 3** summarizes the topological properties of novel functional RNAs generated by *in vitro* selection method (1,943 synthetic) and of naturally occurring ones (200 natural) [29;30].

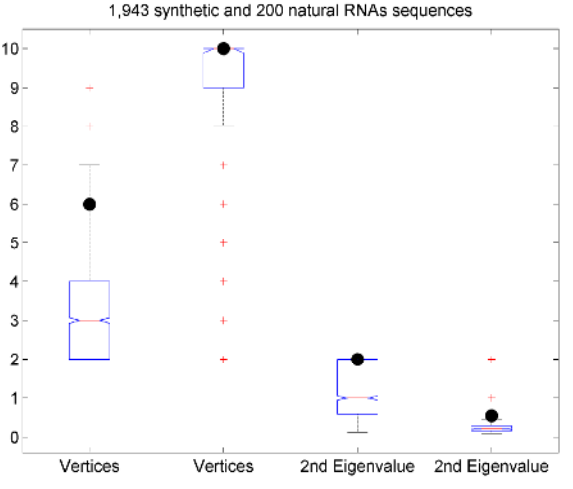

Fig. 3. Notched box-plots for the number of vertices V and second eigenvalue λ_2 of the 1,943 synthetic and 200 natural RNAs sequences [29;30]. *Non-overlapping notches indicate that their medians do differ at the 5% significance level. Box lines indicate the lower quartile, median, and upper quartile. Whisker lines extend to the most extreme data value or at most 1.5 times the box height. •, 95 percentile. +, outlier.*

At 95 percentile, 1,943 synthetic RNAs possess between two to six vertices, and their values of λ_2 lie between the interval one and two. In contrast, those of natural RNA structures have higher number of vertices ranging nine to ten and lower $\lambda_2 < 1$. Exceptions include the transfer RNA, which has a cloverleaf secondary structure consisting of either five or six vertices connected by four stems, but with $\lambda_2 = 0.5188$ [29;30]. According to the Mann-Whitney Rank Sum Test at $\alpha = 0.05$, there are statistically significant differences (p-values ≤ 0.001) in V and λ_2 between the two classes of RNAs. This finding suggests that the synthetic RNA structures have relatively stronger degree of connectivity per fewer numbers of vertices than natural RNA structures.

Table 1 summarizes the abbreviations and definitions of the seven measures investigated. **Fig. 4** depicts how the non-topological properties of the primary sequence (i.e., *L* and *G/C*) and of the secondary structure (i.e., *PB* and *MFE*) are correlated with

varying degrees to the topological measures (i.e., E, V, and λ_2) for the 1,943 synthetic functional RNAs generated by *in vitro* selection method.

Table 1. Abbreviations and definitions of seven measures for the 1,943 synthetic RNAs

Abbreviation	Definition
L	Length of RNA sequence **s**, nucleotides.
G/C	Percentage G/C content in the RNA sequence **s**.
PB	Percentage paired bases in the RNA structure S.
MFE	Normalized Minimum Free Energy (MFE) of S, kcal/mol per nucleotide.
E	Number of edges in the RNA tree-graph G.
V	Number of vertices in the RNA tree-graph G.
λ_2	Second (or the Fiedler) eigenvalue of $\mathbf{L}(G)$.

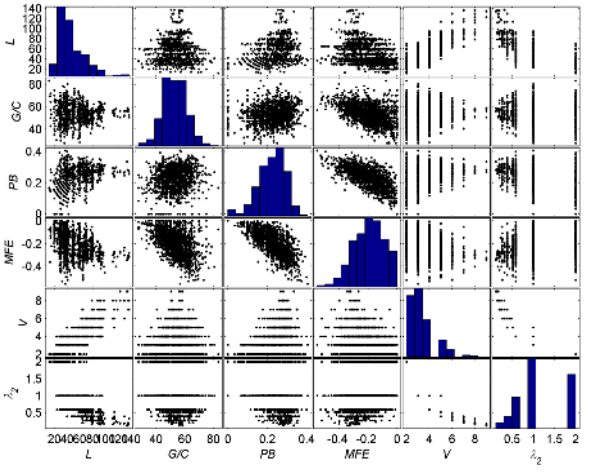

Fig. 4. Correlations between the seven measures for the 1,943 synthetic RNAs. Data for E is *not shown*, as graphs of E and V are identical. E correlates perfectly to V i.e., $E = V - 1$ in the planar tree-graph G, as pseudoknots are not considered in this work. (*Upper and lower diagonals*) *scatter plots*; (*Diagonal*) *histogram bar plot.*

To elucidate the correlation among the statistics, **Table 2** enumerates the values of Pearson correlation coefficient $C_p(f, g)$ defined in Eq. (**5**) between measures f and g in **Table 1**, 95th percentile, and p-values. We have cross-validated the trends of C_p matched similarly to those computed using spearman-rank C_s (ranks-based) and Kendall's C_k (relative ranks-based) correlation metrics (*data not shown*). Existing empirical studies have shown that C_p assumes a pseudo-Gaussian distribution of the dataset, and is not robust to outliers and to non-Gaussian distributions. In contrast, the non-parametric C_s and C_k are robust against non-guassian distributed outliers.

$$C_p(f,g) = \frac{(f-\overline{f})\cdot(g-\overline{g})}{\left\|f-\overline{f}\right\|\left\|g-\overline{g}\right\|}. \tag{5}$$

Table 2. Pearson correlation coefficients $C_p(f, g)$ between the measures f and g, 95th percentile, and p-values. (*Upper diagonal*) *-1.0 ≤ C_p ≤ 1.0, 1.0 for trend identical, -1.0 for perfect opposite, and 0.0 for complete independence. Bold, 0.5 ≤ |C_p| strongly correlated. Italic, 0.4 ≤ |C_p| < 0.5 weakly correlated;* (*Diagonal*) *95th percentile;* (*Lower diagonal*) *two-tailed p-values using the Student's t distribution. p-value < 0.05, statistically significant at 95% confidence.*

$C_p(f, g)$	*L*	*G/C*	*PB*	*MFE*	*E*	*V*	λ_2
L	80.0000	0.0181	0.3882	*-0.4258*	**0.8342**	**0.8342**	**-0.6779**
G/C	4.51E^{-01}	65.9505	0.1689	*-0.4530*	0.0763	0.0763	-0.0549
PB	1.02E^{-63}	1.33E^{-12}	0.3231	**-0.7178**	*0.4366*	*0.4366*	*-0.4761*
MFE	1.30E^{-77}	7.92E^{-89}	9.19E^{-276}	-0.0209	*-0.4061*	*-0.4061*	0.3897
E	0.00E^{+00}	1.44E^{-03}	5.90E^{-82}	4.00E^{-70}	5.0000	**1.0000**	**-0.8715**
V	0.00E^{+00}	1.44E^{-03}	5.90E^{-82}	4.00E^{-70}	0.00E^{+00}	6.0000	**-0.8715**
λ_2	1.25E^{-234}	2.21E^{-02}	3.57E^{-99}	3.14E^{-64}	0.00E^{+00}	0.00E^{+00}	2.0000

Notable observations on the relationships can be concluded as follows. Firstly, the measure *L* is weakly and inversely correlated to the *MFE* at -0.4258, as well as strongly correlated to the *E* and *V* at 0.8342, and to λ_2 at -0.6779. These correlations are statistically significant (p-values < 0.05) showing that a reduction in the length of synthetic RNA (i.e., its primary sequence has fewer nucleotides) weakens the thermo-stability of its structural folding, due to the fewer degree of freedom to form edges (i.e., stems) and vertices (i.e., loops). Consequently, this shifts λ_2 upwards for increased degree of compactness in the synthetic RNA structural topology at the coarsest scale.

Secondly, the statistic *G/C* is also weakly and negatively correlated to the *MFE* at -0.4530. This statistically significant observation (p-value < 0.05) is to be expected since each of the G≡C base pair has lower energy than the other possible base pairings, resulting in higher thermo-stability of the RNA structure. However, our analysis shows that the *G/C* measure is linearly independent to the others |C_p(*G/C*, $g \neq$ *MFE*)| < 0.2. Possibly, a primary sequence of synthetic RNA possessing higher percentage of G/C does not necessarily have greater number of base pairs, as the G/C nucleotides may be distributed in a manner that they do not interact with each other energetically. Consequently, these synthetic RNAs molecules fold into structures that may neither necessarily have more number of vertices *V* (i.e., loops) nor display a lesser degree of compactness λ_2 in their tree-graph topologies.

Thirdly, the structural measure *PB* is strongly and negatively correlated to the MFE at -0.7178, weakly related to the *E* and *V* at 0.4366, and to the λ_2 at -0.4761. These statistically significant findings (p-values < 0.05) are also to be expected since a synthetic RNA secondary structure having greater number of complementary base pairs, would structurally be more stable against thermal fluctuations, resulting in lower MFE than others. However, our statistical analysis suggests that more complementary pairing present in the structure does not necessarily give rise to increased likelihood of forming stems and vertices, and corresponding reduced degree of compactness λ_2.

Fourthly, the structural measure *MFE* is highly and negatively correlated to the *PB* at -0.7178 and weakly correlated to the rest except λ_2 at 0.3897. This statistically significant observation implies that a highly stable synthetic RNA structure does not

necessarily possess a compact topology, unless the degree of compactness provides the capacity to be biologically relevant to the RNA molecule's function.

Lastly, both the topological measures E and V are perfectly correlated, affirming the validity of the linear relationship $E = V - 1$ when pseudoknots are not considered. E and V are strongly and negatively correlated to λ_2 at -0.8715. This finding (p-value < 0.05) suggests that the degree of compactness of synthetic RNAs vary inversely to the number of vertices and stems. A possible explanation is that each vertex deletion tends to shift λ_2 upwards towards the maximum value of one or two as the linear tree-graph attains greater degree of compactness, and *vice-versa* [24].

4 Conclusion

The main contribution of this paper is, we have characterized existing functional RNAs generated by *in vitro* selection method (SELEX) using a recently published spectral graph partitioning technique. This computational approach operates independently of the folding algorithms, but relies on and is limited by their predictions. For example, pseudoknots were not analyzed in this work, as RNAfold does not predict such motifs.

Our topological and in-depth statistical analysis reveals three criteria important for the functionality of synthetic RNAs generated by SELEX, which can be adopted as part of an *in silico* design methodology. Firstly, the length of RNA sequence L and the percentage paired bases *PB* are two critical non-topological parameters for screening likely functional RNA sequences from a large random RNA sequence libraries. At the 95 percentile, RNA sequences with $L = 80$ nucleotides and can fold with approximately 32.31% base pairings, are probable targets to begin with. Secondly, the predicted RNA structure should possess two to six vertices inclusively. Putative functional RNAs with second eigenvalue λ_2 deviating furthest from the wild-type λ_2 requires further inspection, especially those that occur rarely and possess unique traits. λ_2 can be regarded as an effective similarity measure between a library of RNA structures and for discriminating them against a particular fold. Thirdly, to minimize the number of false positives, a combination of filtering parameters should be included, the percentage G/C content of 65.95% and the normalized minimum free energy *MFE* of -0.021 kcal/mol per nucleotide, at the 95 percentile. This last criterion is highly applicable in rare cases of multiple RNA structures possessing identical values of λ_2. To resolve the ambiguity, their sequence compositions (e.g., the percentage G/C content) and structural properties (e.g., the *MFE*) can be used to distinguish them.

Spectral graph partitioning is a promising technique for extracting the topological properties of the detailed RNA structures. This reduced spatial information along with existing characteristics at both the structural and sequence levels are suitable for high-throughput prediction, screening, and engineering of novel functional RNA molecules. Part of our ongoing research is to investigate their topological properties, in order to gain important insights into several open questions. Firstly, whether *in vitro* synthesized RNA sequences belonging to different functional classes are well conserved structurally and topologically. Secondly, whether they possess unique topological characteristics in comparison to randomized RNA sequences.

References

1. N. K. Gray and M. Wickens, "Control of Translation Initiation in Animals," *Annual Review of Cell and Developmental Biology*, vol. 14, no. 1, pp. 399-458, 1998.
2. S. R. Eddy, "Non-coding RNA genes and the modern RNA world," *Nat. Rev. Genet.*, vol. 2, no. 12, pp. 919-929, Dec 2001.
3. G. Storz, "An Expanding Universe of Noncoding RNAs," *Science*, vol. 296, no. 5571, pp. 1260-1263, May 2002.
4. M. Mandal and R. R. Breaker, "Gene regulation by riboswitches," *Nat. Rev. Mol. Cell Biol.*, vol. 5, no. 6, pp. 451-463, Jun 2004.
5. J. R. Hesselberth and A. D. Ellington, "A (ribo) switch in the paradigms of genetic regulation," *Nat. Struct. Biol*, vol. 9, no. 12, pp. 891-893, Dec 2002.
6. M. Lagos-Quintana et al., "Identification of Novel Genes Coding for Small Expressed RNAs," *Science*, vol. 294, no. 5543, pp. 853-858, Oct 2001.
7. N. C. Lau et al., "An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans," *Science*, vol. 294, no. 5543, pp. 858-862, Oct 2001.
8. R. C. Lee and V. Ambros, "An Extensive Class of Small RNAs in Caenorhabditis elegans," *Science*, vol. 294, no. 5543, pp. 862-864, Oct 2001.
9. L. He and G. J. Hannon, "MicroRNAs: small RNAs with a big role in gene regulation," *Nat Rev. Genet*, vol. 5, no. 7, pp. 522-531, Jul 2004.
10. J. Lu et al., "MicroRNA expression profiles classify human cancers," *Nature*, vol. 435, no. 7043, pp. 834-838, Jun 2005.
11. J. Jiang et al., "Real-time expression profiling of microRNA precursors in human cancer cell lines," *Nucl. Acids. Res.*, vol. 33, no. 17, pp. 5394-5403, Sep 2005.
12. M. V. Iorio et al., "MicroRNA Gene Expression Deregulation in Human Breast Cancer," *Cancer Res*, vol. 65, no. 16, pp. 7065-7070, Aug 2005.
13. D. S. Wilson and J. W. Szostak, "In vitro selection of functional nucleic acids," *Annual Review of Biochemistry*, vol. 68, no. 1, pp. 611-647, 1999.
14. J. B. Thomson et al., "In vitro selection of hammerhead ribozymes containing a bulged nucleotide in stem II," *Nucl. Acids. Res.*, vol. 24, no. 22, pp. 4401-4406, Nov 1996.
15. N. K. Vaish, P. A. Heaton, and F. Eckstein, "Isolation of hammerhead ribozymes with altered core sequences by in vitro selection," *Biochemistry*, vol. 36, no. 21, pp. 6495-6501, May 1997.
16. S. Tombelli, M. Minunni, and M. Mascini, "Analytical applications of aptamers," *Biosensors and Bioelectronics*, vol. 20, no. 12, pp. 2424-2434, Jun 2005.
17. M. Rajendran and A. D. Ellington, "In vitro selection of molecular beacons," *Nucleic Acids Res.*, vol. 31, no. 19, pp. 5700-5713, Oct 2003.
18. D. Proske et al., "Aptamers-basic research, drug development, and clinical applications," *Applied Microbiology and Biotechnology*, vol. 69, no. 4, pp. 367-374, Dec 2005.
19. D. Barash, "Spectral decomposition of the Laplacian matrix applied to RNA folding prediction," Bioinformatics Conference 2003, 2003, pp. 602-603.
20. H. H. Gan, S. Pasquali, and T. Schlick, "Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design," *Nucl. Acids. Res.*, vol. 31, no. 11, pp. 2926-2943, Jun 2003.
21. H. A. Meijer and A. A. Thomas, "Control of eukaryotic protein synthesis by upstream open reading frames in the 5'-untranslated region of an mRNA," *Biochem J*, vol. 367, no. Pt 1, pp. 1-11, Oct 2002.

22. A. Diwa et al., "An evolutionarily conserved RNA stem-loop functions as a sensor that directs feedback regulation of RNase E gene expression," *Genes Dev.*, vol. 14, no. 10, pp. 1249-1260, May 2000.
23. L. Katz and C. B. Burge, "Widespread Selection for Local RNA Secondary Structure in Coding Regions of Bacterial Genes," *Genome Res.*, vol. 13, no. 9, pp. 2042-2051, Sep 2003.
24. P. Alex, D. S. Horst, and L. Kan-Pu, "Partitioning sparse matrices with eigenvectors of graphs," *SIAM J. Matrix Anal. Appl.*, vol. 11, no. 3, pp. 430-452, 1990.
25. D. Barash, "Deleterious mutation prediction in the secondary structure of RNAs," *Nucl. Acids. Res.*, vol. 31, no. 22, pp. 6578-6584, Nov 2003.
26. D. Barash, "Second eigenvalue of the Laplacian matrix for predicting RNA conformational switch by mutation," *Comput. Appl. Biosci.*, vol. 20, no. 12, pp. 1861-1869, Aug 2004.
27. D. Barash, "Second eigenvalue of the Laplacian matrix for large predicting RNA conformational switch by mutation," *Comput. Appl. Biosci.*, p. bth157, Feb 2004.
28. D. Barash, "Spectral Decomposition for the Search and Analysis of RNA Secondary Structure," *Journal of Computational Biology*, vol. 11, no. 6, pp. 1169-1174, 2004.
29. D. Fera et al., "RAG: RNA-As-Graphs web resource," *BMC Bioinformatics*, vol. 5, no. 1, p. 88, 2004.
30. H. H. Gan et al., "RAG: RNA-As-Graphs database--concepts, analysis, and features," *Comput. Appl. Biosci.*, vol. 20, no. 8, pp. 1285-1291, May 2004.
31. J. F. Lee et al., "Aptamer Database," *Nucl. Acids. Res.*, vol. 32, no. 90001, pp. D95-100, Jan 2004.
32. L. L. Lebruska and L. J. Maher, III, "Selection and characterization of an RNA decoy for transcription factor NF-kappa B," *Biochemistry.*, vol. 38, no. 10, pp. 3168-3174, Mar 1999.
33. I. L. Hofacker, "Vienna RNA secondary structure server," *Nucl. Acids. Res.*, vol. 31, no. 13, pp. 3429-3431, Jul 2003.
34. M. Zuker and P. Stiegler, "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information," *Nucleic Acids Res.*, vol. 9, no. 1, pp. 133-148, Jan 1981.
35. M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucl. Acids. Res.*, vol. 31, no. 13, pp. 3406-3415, Jul 2003.
36. D. H. Mathews et al., "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure," *Journal of Molecular Biology*, vol. 288, no. 5, pp. 911-940, May 1999.
37. Z. Weinberg and W. L. Ruzzo, "Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy," *Comput. Appl. Biosci.*, vol. 20, no. suppl_1, p. i334-i341, Aug 2004.
38. J. Tinoco, I and C. Bustamante, "How RNA folds," *Journal of Molecular Biology*, vol. 293, no. 2, pp. 271-281, Oct 1999.
39. V. Moulton et al., "Metrics on RNA Secondary Structures," *Journal of Computational Biology*, vol. 7, no. 1-2, pp. 277-292, 2000.
40. M. Sprinzl and K. S. Vassilenko, "Compilation of tRNA sequences and sequences of tRNA genes," *Nucl. Acids. Res.*, vol. 33, no. suppl_1, p. D139-D140, Jan 2005.