



Second eigenvalue of the Laplacian matrix for predicting RNA conformational switch by mutation

Danny Barash

Genome Diversity Center, Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel

Received on October 6, 2003; revised on December 30, 2003; accepted on January 6, 2004

Advance Access publication February 26, 2004

ABSTRACT

Motivation: Conformational switching in RNAs is thought to be of fundamental importance in several biological processes, including translational regulation, regulation of self-cleavage in viruses, protein biosynthesis and mRNA splicing. Current methods for detecting bi-stable RNAs that can lead to structural switching when triggered by an outside event rely on kinetics, energetics and properties of the combinatorial structure space of RNAs. Based on these properties, tools have been developed to predict whether a given sequence folds to a structure characterized by a bi-stable conformation, or to design multi-stable RNAs by an iterative algorithm. A useful addition is in developing a local procedure to prescribe, given an initial sequence, the least amount of mutations needed to drive the system into an optimal bi-stable conformation.

Results: We introduce a local procedure for predicting mutations, by generating and analyzing eigenvalue tables, that are capable of transforming the wild-type sequence into a bi-stable conformation. The method is independent of the folding algorithms but relies on their success. It can be used in conjunction with existing tools, as well as being incorporated into more general RNA prediction packages. We apply this procedure on three well-studied structures. First, the method is validated on the mutation leading to a conformational switch in the spliced leader RNA from *Leptomonas collosoma*, a mutation that has already been confirmed by an experiment. Second, the method is used to predict a mutation that can lead to a novel conformational switch in the P5abc subdomain of the group I intron ribozyme in *Tetrahymena thermophila*. Third, the method is applied on Hepatitis delta virus to predict mutations that transform the wild-type into a bi-stable conformation, a configuration assessed by calculating the free energies using folding prediction algorithms. The predictions in the final examples need to be verified experimentally, whereas the mutation predicted in the first example complies with the experiment. This supports the use of our proposed method on other known structures, as well as genetically engineered ones.

Availability: An eigenvalue application will be available in the near future attached to one of the existing tools.

Contact: dbarash@research.haifa.ac.il

1 INTRODUCTION

The ability of certain RNA molecules to perform as conformational switches, alternating between two states, has been explored in a variety of systems and setups. Recently, an RNA sequence that can assume either of two different ribozyme folds, corresponding to two different functionalities, has been described in Schultes and Bartel (2000). Some systems can switch between two states by the intervention of an external factor: certain mRNA elements that are responsible for transcription termination and translation initiation in bacteria, called riboswitches, are known to alter their conformation between two forms in response to direct metabolite binding (Winkler *et al.*, 2002; Mironov *et al.*, 2002). Other systems may perform as self-induced switches: a metastable structure of the SV11 RNA that acts as a substrate for Q β replicase (Biebricher *et al.*, 1982; Biebricher and Luce, 1992) can switch to an inactive template, by refolding into a stable conformation in the course of about half an hour. Several other interesting examples of RNAs that alternate between two states have been discovered over the years, participating in diverse processes involving regulation, synthesis and splicing. Two recent review articles on RNA structural rearrangements (Nagel and Pleij, 2002; Micura and Höbartner, 2003) contain these examples that exploit the unique properties of conformational energy landscapes in RNA folding. Such properties were explored in several studies, including Chen and Dill (2000) and others referenced in the aforementioned reviews.

Consequently, analyses and tools have been developed to investigate, model and design structural RNA switches (Flamm *et al.*, 2001; Sczyrba *et al.*, 2003). Metastable states in viroid RNAs were simulated in Gulyaev *et al.* (1998) and Shapiro *et al.* (2001). An algorithm for designing multi-stable RNA sequences by combinatorial optimization was

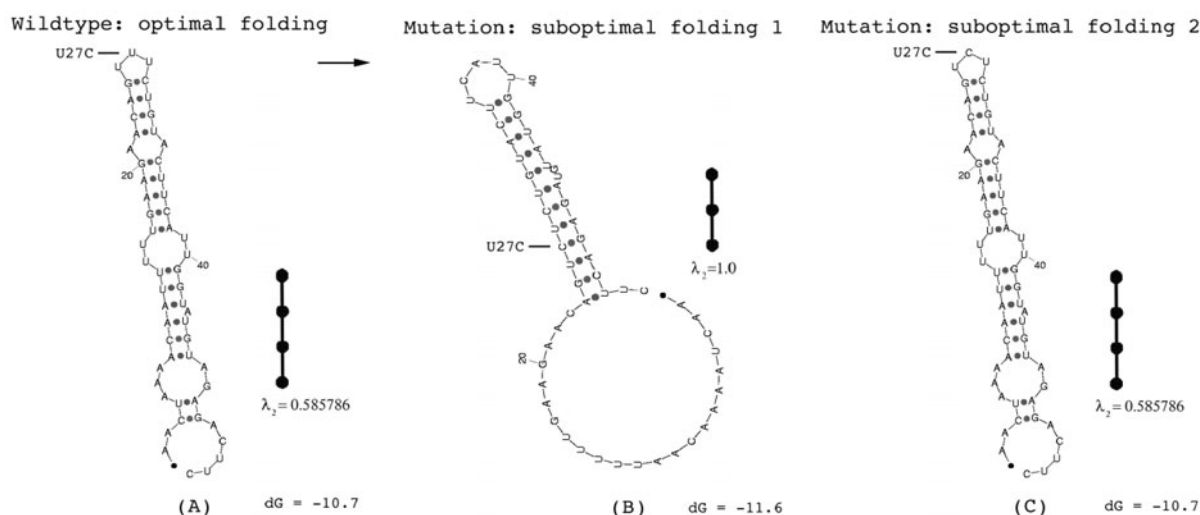


Fig. 1. A bi-stable conformation prediction for the secondary structure of the spliced leader RNA from *L.collosoma*. (A) Wild-type folded structure, along with its tree-graph representation and the corresponding algebraic connectivity $a(T) = 0.585786$ of the tree T . The computed mfold global minimum energy is $dG = -10.7$. The predicted single-point mutation U27C for a transition (see arrow) to a bi-stable configuration is pointed to by a line on the wild-type and mutant structures. (B) First suboptimal folded structure of the U27C mutant associated with $\lambda_2 = 1.0$. (C) Second suboptimal folded structure of the U27C mutant.

described in Flamm *et al.* (2001). A package called ‘prediction of alternating RNA secondary structures’ (paRNAss) was developed to predict structural switching and visualize the transition between the predicted structures in a simulated animation (Giegerich *et al.*, 1999). In addition to the combinatorial structure space, it uses characteristics of the biophysical structure space in order to define a set of criteria for the energy distribution of the system that exists in a typical bi-stable conformation. Apart from RNA switches, the computational design of artificial RNAs is a growing research area that has recently been approached from various viewpoints (Cohen and Skiena, 2002). Here, attention is restricted to the design of RNA switches by attempting to predict bi-stable conformations. The ability to probe bi-stable secondary structures experimentally by comparative imino proton nuclear magnetic resonance (NMR) spectroscopy (Höbartner and Micura, 2003; Micura and Höbartner, 2003) further motivates the development of methods for the computational design of small RNA switches that can potentially lead to functional control.

The contribution of this paper is complementary to the aforementioned works that attempt to design multi-stable RNA sequences. Here, we focus on systems that initially reside in a stable state and concentrate on a local procedure to predict which mutation, given a stable wild-type structure as input, can be introduced in order to create the optimal bi-stable conformation assessed by standard folding prediction packages (Zuker, 2003; Hofacker, 2003). This concept, namely a transition by mutation from one form to the other, was demonstrated before on the spliced leader RNA from *Leptomonas*

collosoma by an experiment (LeCuyer and Crothers, 1994). In the course of applying our mutation prediction analysis on this structure, searching for switching mutations, we discover that the *L.collosoma* system exhibits a transition from a stable conformation to a bi-stable conformation by mutation as illustrated in Figure 1.

Here, a computational method for locally predicting selective mutations initially proposed in Barash and Comanicu (2003) is mathematically described and applied on several examples, with the purpose of locating bi-stable conformations. The method constructs eigenvalue tables, by calculating the second eigenvalue of the Laplacian matrix corresponding to the tree-graph representation of the RNA secondary structure for each mutant. A similar concept was used in Barash (2003) to predict RNA deleterious mutations for disrupting selective motifs in other examples. It also included the P5abc subdomain structure that is revisited here to predict a bi-stable configuration, by a mutation that was briefly mentioned in Barash (2003) but not analyzed since another predicted mutation was described that disrupts the stable hairpin without driving the system into a bi-stable configuration. We show that the same eigenvalue analysis can be used for identifying systems possessing a bi-stable conformation. Moreover, both the *L.collosoma* and the *Tetrahymena thermophila* sequences are 56 nt long, suggesting that other short sequences can potentially be designed by computational means possessing similar properties.

Our mutation prediction method is independent of the particular folding algorithms being used. For the results in this paper it relies on the success of mfold (Zuker, 2003) and the

Vienna package (Hofacker, 2003), both using the expanded energy rules by Mathews *et al.* (1999) to predict the foldings of small RNA sequences. After describing the method and algorithmic details, we apply it on three well-studied structures: the spliced leader RNA from *L.collosoma* (LeCuyer and Crothers, 1994), a predicted bi-stable conformation for the P5abc subdomain in the group I intron ribozyme of *T.thermophila* (Wu and Tinoco, 1998) and the Hepatitis delta virus (Lazinski and Taylor, 1995).

2 METHODS

2.1 RNA tree-graphs and algebraic connectivity

For the illustration of the method, we begin by examining the predicted secondary structure of the spliced leader RNA from *L.collosoma* taken from LeCuyer and Crothers (1994). The predicted wild-type secondary structure by mfold (Zuker, 2003), depicted in Figure 1, succeeds in capturing the exceptionally stable GUUUC loop with the non-canonical GC closing base pair (Shu and Bevilacqua, 1999) corresponding to the experiment (LeCuyer and Crothers, 1994). The problem we are concerned with, in general, is to predict the location of a mutation that will cause a structural rearrangement, such as disrupting the GUUUC hairpin, where the new folded structure as a consequence of introducing the mutation may assume a different shape than the wild-type secondary structure. In the course of looking for structural rearrangements, we will search for more specific mutations that will result in a bi-stable conformation.

In order to predict such mutations using the Laplacian second eigenvalue, as was first suggested in Barash and Comaniciu (2003) borrowing a concept that is used for spectral graph partitioning in parallel processing (Simon, 1991; Demmel, 1996, <http://www.cs.berkeley.edu/demmel/cs267/lecture20.html>), we will use the algebraic connectivity of a tree as an efficient similarity measure for comparing between the initial RNA fold and the folded structure of all possible mutants. The representation of RNA secondary structures as coarse grained tree-graphs was initially explored in Shapiro (1988), Le *et al.* (1989) and Hofacker *et al.* (1994) and the effect of single-point mutations using a combination of RNA tree-graph representation and string comparisons was addressed before in Margalit *et al.* (1989), without the reduction to eigenvalues by the methodology developed here. The second eigenvalues are fast to compute because the Laplacian matrices generated from the coarse grained tree-graphs of RNA sequences <100 nt in size are typically very small, corresponding to the number of loops in the secondary structure. As an example, in the test cases presented here, the size of the associated Laplacian matrices is always less than 10×10 . Other more expensive similarity measures for comparing between trees can be added (Shapiro and Zhang, 1990; Jiang *et al.*, 2002), which convey more information about the shape representation. However, for the purpose of detecting

conformational rearrangements, they are likely to contribute only as a refinement when analyzing sequences in the order of 100–200 nt or less. Shape similarity measures such as the second eigenvalue of the Laplacian matrix are real numbers that can be considered supplemental to the fitness of the RNA molecule (Stadler, 1999), which comprises energies and other biophysical quantities.

Let $T = (V, E)$ be a tree with vertex set $V = \{v_1, v_2, \dots, v_n\}$ and edge set E . Denote by $d(v)$ the degree of v , where $v \in V$ is a vertex of T . The Laplacian matrix of T is $L(T) = (m_{ij})$, where

$$m_{ij} = \begin{cases} d(v_i), & \text{if } i = j, \\ -1, & \text{if } v_i, v_j \in E, \\ 0, & \text{otherwise.} \end{cases}$$

$L(T)$ is a symmetric, positive semidefinite and singular matrix. The lowest eigenvalue of $L(T)$ is always zero, since all rows and columns sum up to zero. Denote by $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n = 0$ the eigenvalues of $L(T)$. The second smallest eigenvalue, λ_{n-1} , is called the algebraic connectivity (Fiedler, 1973) of T and labeled as $a(T)$. Some properties of $a(T)$ that are relevant to the application presented here will be mentioned below, following the calculation of $a(T)$ for the RNA secondary structure example depicted in Figure 1.

2.2 Illustrative example

The eigenvalues of the Laplacian matrix are independent of the chosen labeling for the nodes in the tree-graph, which only amounts to interchanges of rows and columns. For an orderly labeling of the linear tree-graph example in Figure 1, containing four nodes as explained in the sequel, the Laplacian matrix $L(T)$ becomes:

$$L = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix},$$

which is clearly analogous, by the above definition of the matrix elements m_{ij} , to the Laplacian operator. The second smallest eigenvalue of the Laplacian matrix above, $a(T) = 0.585786$, is the algebraic connectivity corresponding to the tree T of the wild-type structure in Figure 1. It is the lower bound for the case of four vertices since the tree is linear. Note that by convention of the chosen tree-graph representation, loops with single isolated nucleotides are not accounted for as nodes, whereas the 5'-3' ends are counted as a node. For a star of four vertices, $a(T) = 1.0$, which is the upper bound. A star applies for a tree-graph with three vertices or more ($n \geq 3$) and its algebraic connectivity is always unity (Merris, 1987).

2.3 Properties of the algebraic connectivity

The algebraic connectivity $a(T)$ possesses special properties that are advantageous for the RNA secondary structure mutation prediction application presented here.

THEOREM 2.1. Let $T = (V, E)$ be a tree on n vertices with algebraic connectivity $a(T)$. Then:

- (1) $0 \leq a(T) \leq 1$.
- (2) $a(T) = 0$ iff T is not connected.
- (3) $a(T) = 1$ iff $T = K_{1,n-1}$ is a star on n vertices.

EXAMPLE 2.1. Because a tree T is a special case of a graph G , it follows that $a(T)$ is positive if and only if T is connected (Fiedler, 1973). Thus, in all our RNA test cases for each tree configuration representing the secondary structure, $a(T)$ is positive (e.g. Fig. 1) since loops, bulges and hairpins are connected through RNA stems.

EXAMPLE 2.2. Let $T = K_{1,n-1}$ be a star on n vertices, that is the tree T has n vertices: one vertex of degree $n - 1$ and $n - 1$ pendant (degree 1) vertices. Then the characteristic polynomial $q_T(x)$ of T becomes:

$$q_T(x) = x(x - n)(x - 1)^{n-2}.$$

This can be verified (Merris, 1987; Grone and Merris, 1990; Grone *et al.*, 1990; Mohar, 1991) by observing that $(1, 1, \dots, 1)$ is an eigenvector of $L(T)$ corresponding to the eigenvalue 0; $(n - 1, -1, \dots, -1)$ is an eigenvector corresponding to n and $\{(0, 1, -1, 0, \dots, 0), (0, 0, 1, -1, 0, \dots, 0) \dots, (0, 0, \dots, 0, 1, -1)\}$ is a set of $n - 2$ linearly independent eigenvectors corresponding to 1. An RNA structure example for the case of $n = 5$ exhibiting a star shape is the yeast phenylalanine tRNA. For $n = 5$, the spectrum of L is $\{0, 1, 1, 1, 5\}$ as a direct result of the characteristic polynomial above. Thus, the algebraic connectivity $a(T)$, or the second eigenvalue of $L(T)$, is smallest but positive when the RNA secondary structure assumes a linear shape and becomes identically 1 when the RNA secondary structure assumes a star shape.

EXAMPLE 2.3. Let T_{linear} be a linear tree on n vertices, $n > 2$, and T'_{linear} be a linear tree on $n - 1$ vertices. Then $a(T'_{\text{linear}}) > a(T_{\text{linear}})$, which is intuitive because a linear tree that is shortened becomes more compact. Therefore, the range of possible algebraic connectivities for a tree T on $n - 1$ vertices, between the lowest value $a(T'_{\text{linear}})$ and the highest value of 1, is smaller than the range of algebraic connectivities for a tree T on n vertices, with the lowest value $a(T_{\text{linear}})$ and the same highest value of 1. The third test case in Section 4, an eigenvalue analysis on the virusoid sequence, nicely illustrates this point since most of the mutant foldings are linear but associated with different number of vertices. We note that for the special case of $n = 3$, $a(T_{\text{linear}}) = 1$ since the tree can only assume a single shape which is a star, and $a(T'_{\text{linear}}) = 2$ corresponding to two vertices ($n - 1 = 2$) because of the mathematical properties of the Laplacian operator when the grid is reduced to two points.

3 ALGORITHM

We use the algebraic connectivity $a(T)$ of a tree T to construct a stepwise procedure that attempts to locate the least number of mutations needed to transform a stable RNA wild-type structure into a bi-stable conformation, specifying the mutation positions in the wild-type sequence as the final output. The prescribed procedure is slightly modified from a similar methodology to disrupt selected RNA motifs.

- (1) Check, using mfold (Zuker, 2003) and the Vienna package (Hofacker, 2003), that the initial wild-type sequence contains a single global energy minimum relatively far away from suboptimal energies (e.g. a single energy solution is obtained from the folding prediction by using the default percentage of optimality parameter: 5% in mfold). This will mostly occur with short natural sequences, $\lesssim 100$ nt long, such as the ones dealt with in this paper or obtained by the subdivision procedure described in the next step.
- (2) Let N be the number of nucleotides in the given wild-type sequence. If $N > 100$, try subdividing the sequence into independently folded domains (i.e. the folding prediction of each subdomain by itself should be the same as the folding prediction of the whole sequence in the specific subdomain region). Note that such a partitioning scheme also requires the assumption that mutated subdomains will not interact with each other, i.e. each subdomain remains an independent folded entity even after mutations are introduced. Denote by N' the number of nucleotides in the artificial sequence, corresponding to the subdomain of interest.
- (3) Serially or in parallel, run a folding prediction calculation for each of the $N' \times 3$ single point mutants, since for each nucleotide there are three possible substitutions. Extract the tree T corresponding to the secondary structure of each mutant in the form of a Laplacian matrix $L(T)$. Calculate the algebraic connectivity $a(T)$, which is the second eigenvalue of $L(T)$. Derive the number of vertices in T to find how many mutants will assume the shape T (frequency of occurrence). Arrange the data in an eigenvalue table, as illustrated in Tables 1–3. Other shape comparison measures, as well as the energies, can be added to the table in separate columns for the purpose of refinements in the prediction.
- (4) If all $N' \times 3$ single-point mutants correspond to the same tree T of the wild-type, add additional layers of mutation by extracting the tree T and calculating the features in Step 3 for each one of the $(N' \times 3)^2$ double-point mutations, then $(N' \times 3)^3$ triple-point mutations, \dots , $(N' \times 3)^m$ m -point mutations, as necessary (see stopping criterion in next step).
- (5) Repeat the previous step until $m = m^*$, where m^* is the minimal number of mutations needed so that at least one

of the mutants will fold to a tree that is different from T of the wild-type, corresponding to a different Laplacian matrix. In most cases, the information conveyed in the columns of the eigenvalue table can be used to detect which mutations caused a change in the Laplacian matrix relative to that of the wild-type. Attempt to use prior information from step $i < j$ at step j , using data from the biology experiment if available, such that at step j only $(N' \times 3)^{m_j - m_i}$ folding calculations are needed instead of $(N' \times 3)^{m_j}$.

- (6) When $m = m^*$, analyze the final eigenvalue table and experiment with various eigenvalues. First, check the eigenvalues (i.e. visualize the predicted folded structure of mutants associated with this eigenvalue) that are furthest from the eigenvalue corresponding to the tree T of the wild-type. Second, check eigenvalues with different number of vertices than the wild-type, especially those with peculiarities (extreme number of vertices, low frequency of occurrence). When finding an eigenvalue inducing an interesting conformational rearrangement, check several mutations leading to this eigenvalue. Group together the ones that result in a single global energy that is far away from the single global energy of the wild-type structure in Step 1. Search for mutations that lead to two suboptimal solutions that are close in energy to each other whereas the other suboptimal energies are relatively far away from the first two. If such a mutation is found, label it as a candidate in the list of selective mutations corresponding to the bi-stable configuration example in Figure 1. Go back from the artificial sequence with $N' < N$ nucleotides to the original sequence with N nucleotides and report the positions of the nucleotide mutations within the wild-type sequence, leading to that transition.

At the completion of these steps, we obtain predicted mutations that lead from a stable conformation to a bi-stable conformation. These predictions are now ready to be tested in a laboratory experiment, attempting to locate mutations that lead to a change in function.

4 RESULTS

To begin our procedure, we verify that the three examples analyzed in this paper exhibit a wild-type sequence that contains a single global minimum energy. Using *mfold* with the default percentage of optimality, 5% in *mfold* 3.0, we observe that each of these sequences folds into an optimal energy solution, indicating its high degree of stability relative to randomly chosen sequences. In addition, manually experimenting with a few mutations on each of the example sequences confirms that these sequences are resilient to single-point mutations. Without applying a systematic procedure for locating selective mutations, the sequences tend to fold

Table 1. Eigenvalue table for the prediction of single-point deleterious mutation in the spliced leader RNA from *L.collosoma*

Second eigenvalue	Number of graph vertices	Wild-type vertices	Frequency
0.381966	5		6
0.585786	4	WT	Default
1.000000	3	*	61

The clustering to discrete eigenvalues enables to discriminate redundant folding possibilities and concentrate on examining only the most probable candidates for a deleterious mutation that may cause a structural rearrangement. Mutations associated with $\lambda_2 = 1.0$ are candidates for disrupting the stable GUUUC loop in the leftmost stem of Figure 1. In particular, one of these mutations (U27C) is predicted to cause a transition from a stable conformation to a bi-stable conformation, as illustrated in Figure 1. Eigenvalues signaling interesting structural rearrangements are labeled with an asterisk (*); eigenvalues that conform to the shape of the wild-type are labeled with ('WT').

to the same predicted shape with no apparent structural rearrangements as a response to introducing some random trial mutations.

Next, we generate an eigenvalue table for a single-point mutation in each of the three example sequences. The results are summarized in Tables 1–3. Because we succeed to transform from the eigenvalue of the wild-type to at least one different eigenvalue in all three examples, signaling a structural rearrangement, there is no need to introduce additional layers of mutation on top of the single-point mutation predictions. Therefore, we can analyze each of the three test cases listed in Tables 1–3, by performing the final step in the procedure outlined in Section 3. We successfully locate those mutations that lead to a bi-stable configuration. We describe how this is done for each of the test cases.

The results of the first test case taken from LeCuyer and Crothers (1994) and Giegerich *et al.* (1999) are found in Table 1 and Figure 1. Another indication for the high degree of stability in the wild-type structure of Figure 1 is the formation of a stable GUUUC hairpin, as reported in Shu and Bevilacqua (1999); the GUUUC hairpin is a highly stable loop with a non-canonical GC closing base pair that is not expected to be easily disrupted by single-point mutations. Table 1 contains three eigenvalues; it allows for our purposes to discard the majority of mutations, associated with the same eigenvalue 0.585786 of the wild-type, experimenting only with mutations associated with the other two eigenvalues. In addition, we observe that mutations leading to the eigenvalue 0.381966 increment the number of vertices in the folded shape but do not succeed to alter the GUUUC hairpin. Only the few mutations leading to the eigenvalue 1.0 result in a predicted folded shape that disrupts the GUUUC hairpin. Among these mutations, we find mutation U27C that disrupts the stable hairpin in its minimum energy solution and at the same time leads to two suboptimal solutions using *mfold* with default parameters, as illustrated in Figure 1. This selective

Table 2. Eigenvalue table for the prediction of single-point deleterious mutation in the P5abc subdomain of the *T.thermophila* group I intron ribozyme

Second eigenvalue	Number of graph vertices	Wildtype vertices	Frequency
0.381966	5	*	3
0.518806	5	WT	Default
1.000000	4		1

The clustering to discrete eigenvalues enables to discriminate redundant folding possibilities and concentrate on examining only the most probable candidates for a deleterious mutation that may cause a structural rearrangement. Mutations associated with $\lambda_2 = 0.381966$ are candidates for disrupting the stable GAAA tetraloop in the leftmost stem of Figure 2. In particular, one of these mutations (G15U) is predicted to cause a transition from a stable conformation to a bi-stable conformation, as illustrated in Figure 2. Eigenvalues signaling interesting structural rearrangements are labeled with an asterisk (**); eigenvalues that conform to the shape of the wild-type are labeled with ('WT').

Table 3. Eigenvalue table for the prediction of single-point deleterious mutation in a virusoid sequence from Hepatitis delta virus

Second eigenvalue	Number of graph vertices	Wildtype vertices	Frequency
0.198062	7		1
0.267949	6		11
0.324869	6	*	8
0.381966	5		101
0.585786	4	WT	Default
1.000000	3		6

The clustering to discrete eigenvalues enables to discriminate redundant folding possibilities and concentrate on examining only the most probable candidates for a deleterious mutation that may cause a structural rearrangement. Mutations associated with $\lambda_2 = 0.324869$ are candidates for disrupting the stable linear shape in the wild-type structure of Figure 3. In particular, one of these mutations (U40G) is predicted to cause a transition from a stable conformation to a bi-stable conformation, as illustrated in Figure 3. Eigenvalues signaling interesting structural rearrangements are labeled with an asterisk (**); eigenvalues that conform to the shape of the wild-type are labeled with ('WT').

mutation conforms with the experiment discussed in LeCuyer and Crothers (1994).

The results of the second test case taken from Wu and Tinoco (1999) are found in Table 2 and Figure 2. A further indication for the stability of the wild-type structure in Figure 2, besides the assumption that the P5abc subdomain of the ribozyme has settled into a robust configuration in the course of evolution and therefore its shape will not be easily altered as a response to single-point mutations, can be found by inspecting additional folding predictions concerning the highly stable L5b GAAA tetraloop. If we increase the temperature parameter in the folding prediction, a feature available using the Vienna package or mfold 2.3, the first motif to disappear is the bulge close to the 5'-3' end. By continuing to increase the temperature, the UGCAA hairpin breaks and its associated nucleotides get isolated. Only near the maximum allowed

temperature, the GAAA tetraloop finally disappears and all the 56 nt appear isolated with no predicted base pairings. Therefore, the GAAA tetraloop is not expected to be easily disrupted by single-point mutations. Examining Table 2, there are three possible eigenvalues corresponding to the different folded shapes, analogous to the previous test case in Table 1. Again, for our purpose the eigenvalue table allows us to discard almost all the $56 \times 3 = 168$ theoretically possible single-point mutations, since 164 mutations correspond to the eigenvalue 0.518806 of the wild-type. Moreover, the only mutation that leads to the eigenvalue 1.0 is A4C, decrementing the number of vertices by breaking the bulge close to the 5'-3' end but not affecting any of the P5abc hairpins. We are left with three mutations associated with the eigenvalue 0.381966: G15C, C22G and G15U. These mutations succeed to disrupt the GAAA tetraloop hairpin, causing a rearrangement of the part of the structure that is further away from the 5'-3' end. Examining the effect of these mutations by a detailed energy analysis available from mfold, we observe that mutation G15C illustrated in Barash (2003) causes a large energy gap between the wild-type folded structure (-25.6 kcal/mol) and the mutant folded structure (-19.5 kcal/mol). The mutation G15U, however, leads to a bi-stable configuration; only two suboptimal foldings with energies close to each other (-17.3 and -16.5 kcal/mol) are found for the mutant folded structure using mfold with default parameters, as illustrated in Figure 2. The mutation C22G also exhibits a bi-stable configuration with two suboptimal energies close to each other.

The results of the third test case taken from Lazinski and Taylor (1995) and Giegerich *et al.* (1999) are summarized in Table 2 and Figure 3. Note that unlike the first two examples applied on sequences that are both 56 nt long, this example contains a larger sequence 127 nt long; therefore, more eigenvalue possibilities exist. For the wild-type folded structure of the virusoid, we observe the typical linear secondary structure with very low energy values since most of the bases are paired. In this example, contrary to searching for unique mutations that disrupt a stable hairpin as in the previous two examples, our initial goal is to predict mutations that disrupt the linearity of the secondary structure. Examining Table 3, except for the eigenvalue $\lambda_2 = 0.324869$ all other eigenvalues correspond to linear tree structures with varying number of vertices, as discussed in Example 2.3 of Section 2. By discriminating the mutations belonging to these eigenvalues, we are left with eight mutations corresponding to $\lambda_2 = 0.324869$. Examining the effect of each of the eight mutations on the folding prediction, we find that most of the eight mutated sequences exhibit a bi-stable configuration in their secondary structure as illustrated in Figure 3 for the mutation U40G. This mutation prediction is clearly not intuitive to find; manually examining all $127 \times 3 = 381$ possible mutations is a formidable task, showing the usefulness of the algorithm outlined in the previous section for selective mutation prediction. The predicted selective mutations aim to disrupt stable configurations as

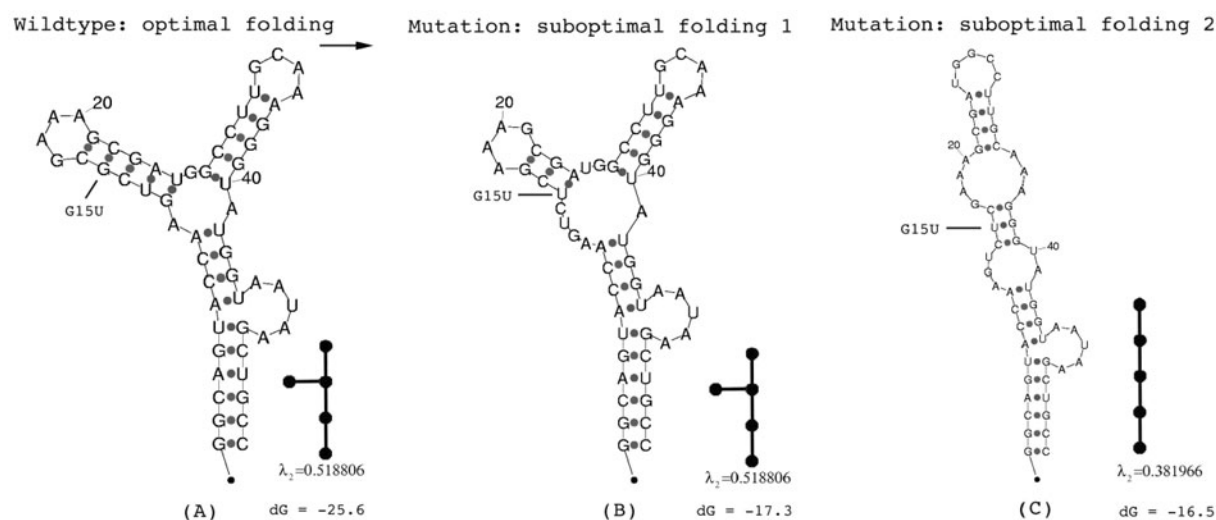


Fig. 2. A bi-stable conformation prediction for the secondary structure of the P5abc subdomain in the group I intron ribozyme of the *T. thermophila*. (A) Wild-type folded structure, along with its tree-graph representation and the corresponding algebraic connectivity $a(T) = 0.518806$ of the tree T . The computed mfold global minimum energy is $dG = -25.6$. The predicted single-point mutation G15U for a transition (see arrow) to a bi-stable configuration is pointed to by a line on the wild-type and mutant structures. (B) First suboptimal folded structure of the G15U mutant. (C) Second suboptimal folded structure of the G15U mutant associated with $\lambda_2 = 0.381966$.

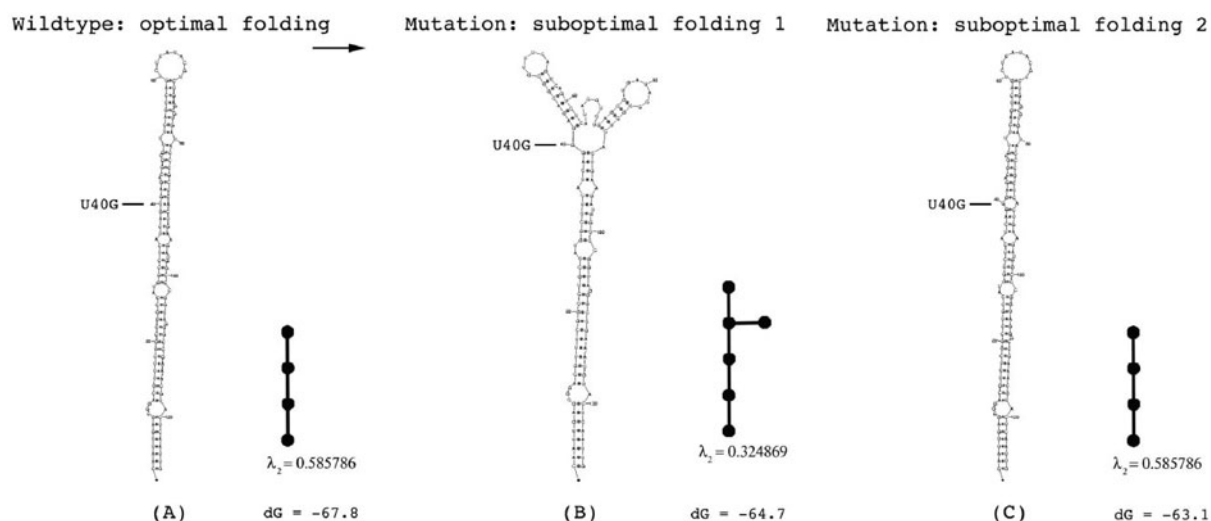


Fig. 3. A bi-stable conformation prediction for the secondary structure of the virusoid sequence from Hepatitis delta virus. (A) Wild-type folded structure, along with its tree-graph representation and the corresponding algebraic connectivity $a(T) = 0.585786$ of the tree T . The computed mfold global minimum energy is $dG = -67.8$. The predicted single-point mutation U40G for a transition (see arrow) to a bi-stable configuration is pointed to by a line on the wild-type and mutant structures. (B) First suboptimal folded structure of the U40G mutant associated with $\lambda_2 = 0.324869$. (C) Second suboptimal folded structure of the U40G mutant.

proposed in Barash (2003) but perhaps more importantly, they can potentially predict structural switches as examined here.

5 DISCUSSION AND CONCLUSION

The common scheme in all three examples discussed in the previous section is their association with a natural short RNA sequence, exhibiting a single predicted optimal solution using

the default parameters in mfold and the Vienna package. Thus, unlike the riboswitch examples in Barash (2003), the eigenvalue tables are relatively simpler with fewer eigenvalue possibilities for grouping the mutations. The first two sequences (Figs 1 and 2) are 56 nt long, whereas the third sequence is 127 nt long; however, because of its exceptional linearity in the secondary structure, the folding predictions in

the third example are expected to be as accurate as in shorter sequences. All these structures tend to fold back to the same shape as a consequence of introducing single-point mutations. The initial goal in the first two examples was to find mutations that disrupt the stable GUUUC and GAAA stable hairpins, respectively, and in the third example to locate mutations disrupting the stable linear structure. After a unique eigenvalue corresponding to such mutations is found in each of the cases, we show that some of these mutations inserted in the wild-type structures depicted in Figures 1A, 2A and 3A lead to bi-stable conformations as in Figures 1B and C, 2B and C and 3B and C.

Some limitations of the proposed mutation prediction method concerning the presentation of alternative solutions should be noted. Folding predictions by energy minimization using dynamic programming produce a ranked list of suboptimal structures (Zuker, 1989; Wuchty *et al.*, 1999). However, the folding prediction packages may not necessarily find all suboptimal structures. In addition, a window parameter W controls the number of foldings that are computed in mfold. As W decreases, the number of predicted foldings increases. Initially, a default value for W is selected depending on the sequence length. In the examples presented here, it was verified that bi-stability can be observed regardless of window size modifications that only caused the number of suboptimal structures to change. In all these trials, the two lowest energy solutions remained the same and their energy difference was significantly smaller relative to their energy distance from other suboptimal structures. However, in each new example the sensitivity of a bi-stable conformation to the window parameter should be checked in order to validate the accuracy of the prediction.

Bi-stable secondary structures of small artificial RNAs were observed experimentally in Höbartner and Micura (2003) and Micura and Höbartner (2003) by using comparative imino proton NMR spectroscopy. The mutation prediction examples presented here can potentially assist in locating small natural RNAs possessing a bi-stable conformation. It is assumed that the two lowest energy structures in each of the present examples are prominent structures, showing two clearly distinct states, and that kinetic trapping may occur in longer sequences but is less likely to occur in short sequences like the ones presented here. It is also assumed that spontaneous transitions between the two prominent states are not expected since an energy barrier exists between the two, unless a switching is triggered by an outside event.

Predicting mutations that result in a bi-stable conformation, which is closely related to RNA switches as pointed out in Höbartner and Micura (2003), can be justified for their importance in several ways. First, it has been noticed in Flamm *et al.* (2001) that computationally, multi-stable RNA conformations can be found rather easily. These simulations suggest that a bi-stable configuration possessing properties of a switch may have been created in the course of evolution. Second, several

artificial bi-stable RNAs have already been detected experimentally by Höbartner and Micura (2003), raising the question of how many natural bi-stable RNAs can be detected and what is their functional role. Third, many variations in a given wild-type RNA sequence may result in functionally insignificant changes of produced structure, whereas variations that lead to a formation of an alternative structure will presumably have a deleterious effect on the functional role of the structurally affected RNA element. Therefore, a mutation prediction in which the mutant contains alternative foldings is interesting to explore experimentally for the investigation of functional RNA elements. The computational prediction of the first example in the previous section is indeed responsible for a conformational switch that has been shown to occur by an experiment (LeCuyer and Crothers, 1994). Other mutation predictions using the methodology that was described are awaiting experimental investigations.

The natural RNA sequences examined in this paper possess unique properties over randomly chosen ones. Only very few single-point mutations in certain vulnerable spots along the sequence can potentially disrupt stable structures. These mutations transform from a stable conformation to a bi-stable conformation in the majority of cases, and to a single stable conformation in the other cases. Such a combination may have evolved in the course of evolution for the purpose of regulation mediated by RNA structure using a switching mechanism. A bi-stable conformation contains two prominent states with an energy barrier between them, thus ensuring that switching between the two states does not occur spontaneously. Switching from one state to the other can be triggered by some outside event that is caused by the environment. Therefore, using the computational methodology described in this paper and expanding on it by generating eigenvalue tables repeatedly may assist in designing optimal artificial RNA sequences that mimic switching properties (Breaker, 2002; Höbartner and Micura, 2003), possessing a regulation mechanism behavior when interacting with the environment.

ACKNOWLEDGEMENTS

I am grateful to Drs Alexander Bolshoy, Ofer Peleg, Edward N. Trifonov and Eviatar Nevo for many useful discussions and suggestions. I would like to thank the anonymous referees for their helpful feedback and comments. The work was conducted at the Genome Diversity Center and supported by the Institute of Evolution, University of Haifa, Israel.

REFERENCES

- Barash,D. and Comaniciu,D. (2003) A common viewpoint on broad kernel filtering and nonlinear diffusion. In Griffin,L.D. and Lillholm,M. (eds), *ScaleSpace-03; Proceedings of the 4th International Conference on Scale-Space Theories in Computer Vision*, Lecture Notes in Computer Science, Vol. 2695. Springer-Verlag, Heidelberg, pp. 683–698.

- Barash,D. (2003) Deleterious mutation prediction in the secondary structure of RNAs. *Nucleic Acids Res.*, **31**, 6578–6584.
- Biebricher,C.K., Diekmann,S. and Luce,R. (1982) Structural analysis of self-replicating RNA synthesis by Q β replicase. *J. Mol. Biol.*, **154**, 629–648.
- Biebricher,C.K. and Luce,R. (1992) *In vitro* recombination and terminal elongation of RNA by Q β replicase. *EMBO J.*, **11**, 5129–5135.
- Breaker,R.R. (2002) Engineering allosteric ribozymes as biosensor components. *Curr. Opin. Biotechnol.*, **13**, 31–39.
- Chen,S.-J. and Dill,K.A. (2000) RNA folding energy landscapes. *Proc. Natl Acad. Sci., USA*, **97**, 646–651.
- Cohen,B. and Skiena,S. (2002) Natural selection and algorithmic design of mRNA. *J. Comput. Biol.*, **10**, 419–432.
- Demmel,J. (1996) Graph partitioning. Lecture Notes, University of California at Berkeley, Berkeley.
- Fiedler,M. (1973) Algebraic connectivity of graphs. *Czechoslovak Math. J.*, **23**, 298–305.
- Flamm,C., Hofacker,I.L., Maurer-Stroh,S., Stadler,P.F. and Zehl,M. (2001) Design of multistable RNA molecules. *RNA*, **7**, 254–265.
- Giegerich,R., Haase,D. and Rehmsmeier,M. (1999) Prediction and visualization of structural switches in RNA. *Pac. Symp. Biocomp.*, **4**, 126–137.
- Grone,R. and Merris,R. (1987) Algebraic connectivity of trees. *Czechoslovak Math. J.*, **37**, 660–670.
- Grone,R. and Merris,R. (1990) Ordering trees by algebraic connectivity. *Graphs and Combinatorics*, **6**, 229–237.
- Grone,R., Merris,R. and Sunder,V.S. (1990) The Laplacian spectrum of a graph. *SIAM J. Matrix Anal. Appl.*, **11**, 218–238.
- Gulyaev,A.P., van Batenburg,F.H.D. and Pleij,C.W.A. (1998) Dynamic competition between alternative structures in viroid RNAs simulated by an RNA folding algorithm. *J. Mol. Biol.*, **276**, 43–55.
- Höbartner,C. and Micura,R. (2003) Bistable secondary structures of small RNAs and their structural probing by comparative imino proton NMR spectroscopy. *J. Mol. Biol.*, **325**, 421–431.
- Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
- Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
- Jiang,T., Lin,G.-H., Ma,B. and Zhang,K. (2002) A general edit distance between RNA structures. *J. Comput. Biol.*, **9**, 371–388.
- Lazinski,D.W. and Taylor,J.M. (1995) Regulation of hepatitis delta virus ribozymes: to cleave or not to cleave? *RNA*, **1**, 225–233.
- Le,S.Y., Nussinov,R. and Maizel,J.V. (1989) Tree graphs of RNA secondary structures and their comparisons. *Comput. Biomed. Res.*, **22**, 461–473.
- LeCuyer,K.A. and Crothers,D.M. (1994) Kinetics of an RNA molecular switch. *Proc. Natl Acad. Sci., USA*, **91**, 3373–3377.
- Margalit,H., Shapiro,B.A., Oppenheim,A.B. and Maizel,J.V. (1989) Detection of common motifs in RNA secondary structures. *Nucleic Acids Res.*, **17**, 4829–4845.
- Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
- Merris,R. (1987) Characteristic vertices of trees. *Lin. Multi. Alg.*, **22**, 115–131.
- Micura,R. and Höbartner,C. (2003) On secondary structure rearrangements and equilibria of small RNAs. *ChemBioChem*, **4**, 984–990.
- Mironov,A.S., Gusarov,I., Rafikov,R., Lopez,L.E., Shatalin,K., Kreneva,R.A., Perumov,D.A. and Nudler,E. (2002) Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell*, **111**, 747–756.
- Mohar,B. (1991) The Laplacian spectrum of graphs. In Alavi,Y., Chartrand,G., Oellermann,O.R. and Schwenk,A.J. (eds), *Graph Theory, Combinatorics, and Applications*. Wiley, Vol. 2, pp. 871–898.
- Nagel,J.H.A. and Pleij,C.W.A. (2002) Self-induced structural switches in RNA. *Biochimie*, **84**, 913–923.
- Schultes,E.A. and Bartel,D. (2000) One sequence, two ribozymes. *Science*, **289**, 448–452.
- Sczyrba,A., Kruger,J., Mersch,H., Kurtz,S. and Giegerich,R. (2003) RNA-related tools on the Bielefeld Bioinformatics Server. *Nucleic Acids Res.*, **31**, 3767–3770.
- Shapiro,B.A. (1988) An algorithm for comparing multiple RNA secondary structures. *Comput. Appl. Biosci.*, **4**, 387–393.
- Shapiro,B.A. and Zhang,K. (1990) Comparing multiple RNA secondary structures using tree comparisons. *Comput. Appl. Biosci.*, **6**, 309–318.
- Shapiro,B.A., Bengali,D., Kasprzak,W. and Wu,J.C. (2001) RNA folding pathway functional intermediates: their prediction and analysis. *J. Mol. Biol.*, **312**, 27–44.
- Shu,Z. and Bevilacqua,P.C. (1999) Isolation and characterization of thermodynamically stable and unstable RNA hairpins from a triloop combinatorial library. *Biochemistry*, **38**, 15369–15379.
- Simon,H.D. (1991) Partitioning of unstructured problems for parallel processing. *Comput. Syst. Eng.*, **2**, 135–148.
- Stadler,P.F. (1999) Fitness landscapes arising from the sequence-structure maps of biopolymers. *J. Mol. Struct.*, **463**, 7–19.
- Winkler,W., Nahvi,A. and Breaker,R.R. (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, **419**, 952–956.
- Wu,M., Tinoco,I., Jr (1998) RNA folding causes secondary structure rearrangement. *Proc. Natl Acad. Sci., USA*, **95**, 11555–11560.
- Wuchty,S., Fontana,W., Hofacker,I.L. and Schuster,P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.
- Zuker,M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.
- Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.