Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 24, Issue Number 1, (2006)

Shape Similarity Measures for the Design of Small RNA Switches

http://www.jbsdonline.com

Assaf Avihoo[1]
Danny Barash[1,2,*]

[1]Department of Computer Science
Ben-Gurion University
84105 Beer-Sheva, Israel
[2]Genome Diversity Center
Institute of Evolution
University of Haifa
Haifa 31905, Israel

Abstract

Conformational switching in the secondary structure of RNAs has recently attracted considerable attention, fostered by the discovery of 'riboswitches' in living organisms. These are genetic control elements that were found in bacteria and offer a unique regulation mechanism based on switching between two highly stable states, separated by an energy barrier between them. In riboswitches, the energy barrier is crossed by direct metabolite binding, which facilitates regulation by allosteric means. However, other event triggers can cause switching to occur, such as single-point mutations and slight variations in temperature. Examples of switches with these event triggers have already been reported experimentally in the past. Here, the goal is to computationally design small RNA switches that rely on these triggers. Towards this end, our computer simulations utilize a variety of different similarity measures to assess the distances between an initial state and triggered states, based on the topology of the secondary structure itself. We describe these combined similarity measures that rely on both coarse-grained and fine-grained graph representations of the RNA secondary structure. As a result of our simulations, we provide some candidate sequences of approximately 30-50 nt, along with the exact triggers that drive the switching. The event triggers under consideration can be modelled by Zuker's mfold or the Vienna package. The proposed methodology that rely on shape measures can further be used to computationally generate more candidates by simulating various event triggers and calculating their effect on the shape.

Introduction

Recent discoveries have demonstrated the peculiar possibilities of an RNA molecule to control fundamental processes in living cells. Although the functional role of RNAs are often related to their three-dimensional structure, the RNA secondary structure is experimentally accessible and contains much information to shed light on the relationship between structure and function. In general, RNA folding is thought to be hierarchical in nature (1, 2), where a stable secondary structure forms first and subsequently there is a refinement to the tertiary fold. This phenomena that utilize the rugged energy landscape of an RNA molecule containing conformational traps can be understood at the level of RNA secondary structure (3). For example, in the recently discovered genetic control elements called 'riboswitches' (4, 5), a mechanism for prokaryotic gene regulation can be observed by examining the secondary structure alone. A switch between two highly stable states (*e.g.*, a transcription terminator and an anti-terminator in *Bacillus subtilis*) occurs as a consequence of direct metabolite binding that allows it to cross the energy barrier between these states. Conformational switching in the secondary structure can also be achieved by other event triggers. For example, it was noted in (6) that there is some probability that even a single mutation can substantially alter the RNA secondary structure. Experimentally, this was obeserved in the spliced leader of *Leptomonas collosoma* (7), among other biological systems. Reviews and valuable information about RNA secondary structure prediction (8) and RNA switches with event triggers other than metabolite binding are available in (9, 10). Before their discovery in living cells,

[*]Email: dbarash@cs.bgu.ac.il

Breaker and coworkers were experimentally designing artificial RNA switches that respond to metabolite binding [see (11), and a more recent review in (12)].

On the prediction side, given an RNA sequence, the paRNAss algorithm has been developed to evaluate the predictability of conformational switching in that sequence (13). However, paRNAss does not use temperature and mutations as variables in the prediction, unlike the simulations that are presented here. When attempting to computationally design new artificial switches, the seminal work of (14) can prove useful, since a multistable RNA is effectively a switch when the exact event triggers to cross the energy barrier between the stable states are found. Here, after developing four different similarity measures, we demonstratively construct small RNA switches by starting from numerous seed patterns represented by graphs or strings and simulating on them the specific event triggers that we would like to investigate. The seed patterns are mostly derived from cuts and pieces of ribosomal RNAs (rRNAs) and some natural RNAs that were discussed in (15). By further division and assembly of these patterns found in natural RNAs, we come up with small sequences of about 30-50 nt, consequently running a folding prediction on the corresponding sequences with *mfold* (16) or the Vienna package (17). The latest available energy parameters (18) are used. For each candidate sequence, we predict the wildtype secondary structure as well as all triggered combinations according to the drives we specify in advance. Here, we chose the event triggers to either be a difference of one or two degrees Celsius from the reference at room temperature, or single point mutations performed on the wildtype reference. Thus, the overall amount of calculation in this procedure is not expensive, noting that we are considering small RNAs. The challenging step is to automatically assess the differences between the secondary structure of the wildtype state and the triggered states as accurately as possible, to ensure we do not miss combinations of candidate switch constructs. Thus, we use a number of diverse similarity measures, all sharing in common that the distance estimation is based on the topology of the secondary structure itself. Next, we describe the pattern representations and similarity measures that are currently being used, where the representations consist of both coarse-grained and fine-grained graphs that correspond to the RNA secondary structure. Finally, we provide three small RNA switch candidates along with the temperature and mutation triggers that are causing the conformational switching, and suggest ways to expand the pool of candidate switch constructs in parallel with selecting a few for further biological testings.

Pattern Representations

The secondary structure can be represented as graphs, strings, or matrices. In each, there is the possibility to represent the secondary structure at the nucleotide level (fine grain representation) or at the motif level (coarse grain representation). To obtain a matrix representation equivalent to a graph, let $T = (V,E)$ be a tree-graph with vertex set $V = v_1, v_2, \ldots, v_n$ 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 [also known to be the difference of the diagonal matrix of vertex degrees $D(T)$ and the adjacency matrix $A(T)$ (5, 14)] is $L(T) = (a_{ij})$, where
$L(T)$ is a symmetric, positive semidefinite and singular matrix. The lowest ei-

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

genvalue of $L(T)$ is always zero, since all rows and columns sum up to zero. Denote by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n = 0$ the eigenvalues of $L(T)$. The second smallest eigenvalue, λ_{n-1}, is called the algebraic connectivity (19) of T and labeled as $a(T)$. Some properties of $a(T)$ that are relevant to the application presented here will be mentioned in the next section.
To illustrate a coarse-grain and a fine-grain representation of RNAs using matri-

ces, let us examine the sequence "ACGGGGU". In this short sequence, simple Watson-Crick base pairing consideration (A connects with U, C connects with G) can be applied to yield the prediction "((…))" in dot bracket notation (20). Thus, in a matrix representation at the level of nucleotides, the corresponding Laplacian matrix reads [see (15) for other examples]:
and at the level of motifs, the stem-loop structure of "((…))" yields [see (2)

$$L = \begin{bmatrix} 2 & -1 & 0 & 0 & 0 & 0 & -1 \\ -1 & 3 & -1 & 0 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & -1 & 0 & 0 & -1 & 3 & -1 \\ -1 & 0 & 0 & 0 & 0 & -1 & 2 \end{bmatrix}$$

for other examples]:
when taking into account the 5′-3′ end, in addition to the hairpin denoted by

$$L = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},$$

"…", as nodes in the motif level tree-graph that the Laplacian matrix corresponds to. In addition, we chose not to assign nodes to internal loops and bulges consisting of a single nucleotide on either side of the loop. Next, we describe several similarity measures that are used to estimate how two pattern representations differ at both the fine-grain and coarse-grain levels, as well as a hybrid between the two. In Table I, all of the similarity measures are being calculated, in order to achieve a comprehensive analysis.

Table I
Similarity Measures: EV2 (second eigenvalue of the Laplacian matrix), Wiener Number, RNAdist (Vienna's RNADistance), HCT ('Hybrid Change Tracker').

Sequence Tag	EV2	Wiener	RNADist	HCT
A-WT -37 ºC	2	3297.9	0	1.00
A -35 ºC	1	1480.8	24	0.39
A -36 ºC	1	1480.8	24	0.39
A -38 ºC	2	3297.9	0	1.00
A -39 ºC	2	3297.9	0	1.00
A-U4G	1	1407.7	28	0.32
A-G21U	1	1446.1	24	0.38
A-C26A	2	3323.0	4	0.99
A-C26U	2	3615.9	4	0.91
B-WT -37 ºC	2	3472.5	0	1.00
B -35 ºC	1	2391.3	14	0.59
B -36 ºC	2	3472.5	0	1.00
B -38 ºC	2	3472.5	0	1.00
B -39 ºC	2	3472.5	0	1.00
B-A3G	2	3356.3	6	0.95
B-U5A	2	3472.5	0	1.00
B-A3U	1	2431.2	14	0.57
B-A9C	2	2532.6	16	0.79
C-WT -37 ºC	1	7009.6	0	1.00
C -35 ºC	1	7009.6	0	1.00
C -36 ºC	1	7009.6	0	1.00
C -38 ºC	1	7009.6	0	1.00
C -39 ºC	2	13793	48	0.16
C-A49C	1	7138.2	26	0.72
C-A16C	1	10135.7	40	0.23
C-U28C	1	8972.7	22	0.69
C-U32G	1	10436.8	26	0.70

Similarity Measures

In our current simulation, we use four different similarity measures. The first is the second eigenvalue of the Laplacian matrix corresponding to the coarse-grained tree-graph representation of the RNA secondary structure. Despite its expected limitations when dealing with small RNAs due to the low amount of secondary structure motifs, in this particular case it happens to be a good switch indicator using simple intuition because the coarse-grained tree graphs typically contain either two or three nodes. The second similarity measure we describe is the Wiener number of the Laplacian matrix corresponding to the fine-grained graph representation of the secondary structure, as suggested by Merris (21). The third is the RNAdist in the Vienna package (17), relying on the fine-grained graph that is equivalent to the dot-bracket notation often used to represent the secondary structure (20). The fourth is an in-house 'Hybrid Change Tracker' (HCT), utilizing coarse-grained definitions from (22) as letters in a string that is being generated at the nucleotide level. Below, we elaborate more on these four similarity measures.

The second eigenvalue of the Laplacian matrix, as described in (15), starts from Shapiro's coarse-grain representation (23, 24) as calculated using the 'b2shapiro' routine available in (20). An equivalent representation of the coarse-grained tree-graphs is a Laplacian matrix, for which its second eigenvalue is a measure of the tree-graph compactness (19). When using the Laplacian matrix $L(T)$ to represent the tree T, the following mathematical properties become useful:

- The eigenvalues of $L(T)$ are nonnegative and the first eigenvalue is zero.
- The second smallest eigenvalue is the algebraic connectivity (19) of T, denoted by $a(T)$.
- $0 \le a(T) \le 1$.
- $a(T) = 2(1\text{-}cos(\pi/n))$ iff $T = P_n$ is a path on n vertices (19).
- $a(T) = 1$ iff $T = K_{1,n-1}$ is a star on n vertices (25, 26).

Interestingly, for the case of a star of any number of vertices N when N>2, the

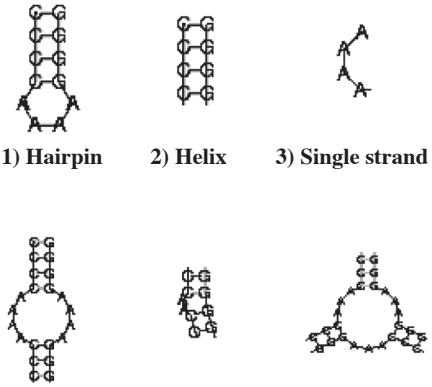

Figure 1: Coarse-Grain RNA Motifs Defined in (19).

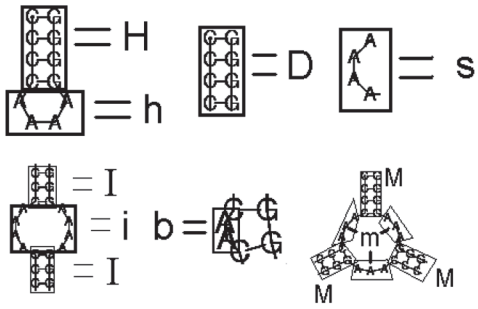

Figure 2: Illustration of Letter Tags that are Used in the 'Hybrid Change Tracker' (HCT).

second eigenvalue is 1.0 (26). However, for the exceptional case when N=2, the second eigenvalue is 2.0 [see (26, 15)] as occurs in the examples depicted in Figures 3-5. Thus, the difference between a star of two and three vertices can be easily traced by the Laplacian second eigenvalue, which serves as a rough indicator for conformational transitions in the cases presented in Figures 3-5. In other cases, such a similarity measure may not capture all the details when encountering finer conformational transitions. Therefore, as proposed to us by Merris (21), the reciprocals of all eigenvalues of the Laplacian matrix for the fine-grained graph (equivalent to the dot-bracket representation) can be added to extract the Wiener index (27). The Wiener index is a topological index defined as the sum of distances between all pairs of vertices in a tree. It has been used as a structural descriptor for molecular graphs. As proved in (21), the following corollary holds:

Let $T = (V,E)$ be a tree with vertex set $V = \{v_1, v_2, \ldots, v_3\}$ and Laplacian eigenvalues $\lambda_1 \geq \ldots \geq \lambda_{n-1} > \lambda_n = 0$. Then the Wiener index of T is
where $d(v_i, v_j)$ is a count of the edges in the unique path from v_i to v_j. This similarity

$$W(T) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d(v_i, v_j) = n \sum_{i=1}^{n-1} 1/\lambda_i,$$

measure, as well as the edit distance on strings provided by the RNAdist available in (17), both work at the level of fine-grained graph representations. We note that unlike information extracted from the Laplacian eigenvalues that often has an intuitive meaning but does not constitute a metric, the edit distance on strings is a metric. Finally, we describe an in-house 'Hybrid Change Tracker' (HCT), a heuristic procedure that we developed specifically for our purposes. It uses a hybrid strategy that combines coarse and fine grain representations. Coarse-graining can be mathematically defined in several ways and here we use the definitions found in (22) that distinguish between six different types of secondary structure motifs: Single strand, Double strand, Bulge, Hairpin, Interior loop, and Multi-branched loop as depicted in Figure 1. Next, we assign to each nucleotide in the given sequence a letter tag corresponding to one of the aforementioned motifs, further distinguishing between capital and small letters corresponding to single and double strands as illustrated in Figure 2. We arbitrarily choose to continue with the letter assignments in the old motif until we completely transform to the new motif. This results in strings at the nucleotide level (fine-grain representation), classified by coarse-graining, a hybrid between the two different levels of granularity. The HCT strings have the ability to change as a cause of slight variations at a resolution that can be useful when dealing with small RNAs. Next, the HCT strings (as observed in Figures 3-5) are aligned using simple dynamic programming. Both a gapped alignment and an ungapped alignment can be performed, with the latter being more sensitive to 'single-nucleotide slipping'. Scoring can be constructed in various ways. Here, considerations emanating from intuitive resemblance of secondary structure motifs are used. A possible extension is to use learning algorithms such as neural networks to construct a more sophisticated scoring scheme for future use with the HCT. The final output is a normalized distance measure between 0.0 and 1.0, where small numbers are robust indicators that a switch has been encountered.

Table I contains the results of applying the four similarity measures on three RNA switch candidates. In order to generate the switch candidates, we are looking for the largest distance measures between the triggered secondary structure and the wildtype secondary structure relative to the distances of all the other triggered structures. For example, let us examine sequence C in Table I. Using EV2, it is clear that only a change to 39 ºC may produce a conformational change from the predictive standpoint since in all other event triggers the second eigenvalue of the Laplacian matrix remains 1.0 (a star shape with three vertices or more). Furthermore, in this example, one also notices the benefit of using a non-metric measure such as EV2, since from the mathematical theorems about EV2 in the case of tree-graphs we can immediately deduce without a graphical picture that the resultant

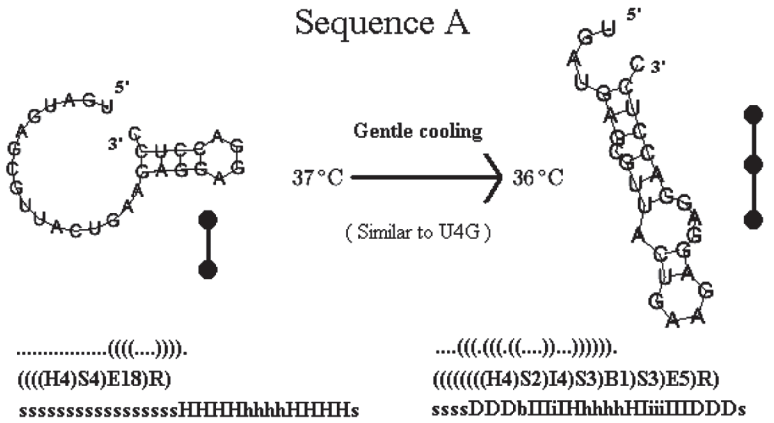

Figure 3: Small RNA Artificial Temperature Switch (A). Secondary Structure of Wildtype and Triggered State with Accompanied Representations: Coarse-Grain Tree-Graph, Dot-Bracket Notation, Coarse-Grain String, Hybrid (HCT) String.

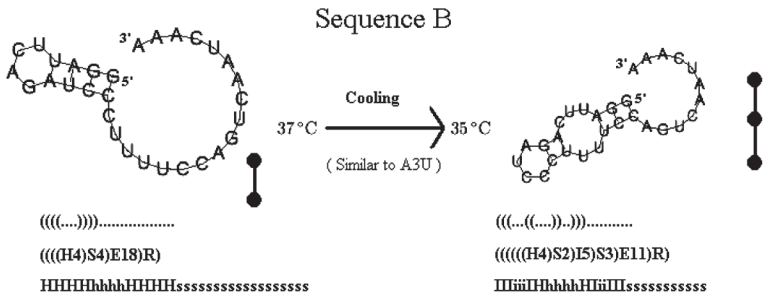

Figure 4: Small RNA Artificial Temperature Switch (B).

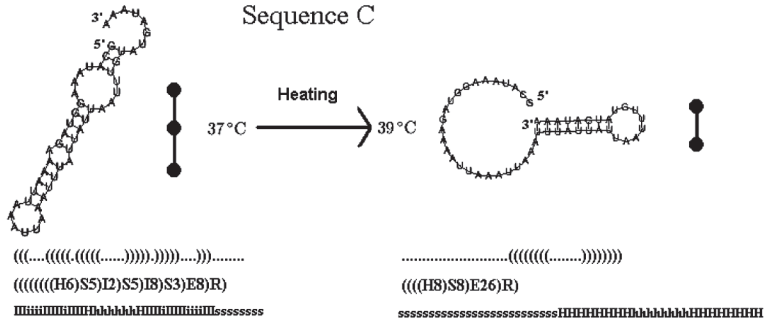

Figure 5: Small RNA Artificial Temperature Switch (C).

mutant possesses a tree-graph shape with two vertices (the only combination in which EV2 may possibly become 2.0). Such information about the overall configuration of the RNA secondary structure may not be possible to obtain by examining numbers that convey distances in a metric measure. Likewise, using RNAdist, we again observe a jump from 0 to the highest value of 48 at 39 ºC although this calculation that utilizes a fine-grain representation is independent from EV2 that uses a coarse-grain tree graph representation. We notice the same effect in the Wiener and HCT measures at 39 ºC: relative to the other event triggers, such as 38 ºC, a significantly large Wiener number of 13793 is reported with respect to the wildtype and a significantly small HCT value of 0.16 is reported relative to the wildtype. Indeed, inserting the 2 ºC increase in temperature to Vienna's RNAfold, we can observe that energy minimization predicts a conformational rearrangement (see Figure 5). Thus, most times the various distance measures that we use are in agreement with each other. When one of the distance measures is not in agreement with the other three because of an exceptional case that may report a false positive corresponding to this measure, it is possible to verify such a case using a direct simulation with RNAfold. The table reduces our searches substantially by discarding many unwanted possibilities when all distance measures are reporting that there is almost no change relative to the wildtype, such as in Table I (case C) for temperature values besides 39 ºC.

Finally, it is worthwhile noting that other similarity measures could be used [*e.g.*,

(28, 13)]. However, the simplest ones will not perform better because of the relatively small sequence sizes, because for each similarity measure that is being defined (*e.g.*, simply checking the number of base pairs in common) there are special cases that may generate false positives. For example, base pairing count may fail when generalizing our results to handle cases of varied lengths in sequences, which is in planning though not shown in this paper. Thus, by combining several independent similarity measures, we can detect special cases that may produce a false positive by direct comparison to the others.

Results

We provide three examples of candidate switches that were found using the approach outlined above. We note that temperature switches as a consequence of 1-2 degrees Celsius change are more scarce compared to single-point mutation switches, and the two are correlated; once we encounter a temperature switch candidate, it is most likely that there will be a single point mutation that results in approximately the same switching. We may deduce that such sequences can be triggered to cross the energy barrier by other means that are more difficult to simulate on a computer, such as radiation exposure and stress.

Conclusions

Our goal is to design small RNA temperature and mutation switches by a computer, and collect candidates that can further be tested experimentally. This line of research opens the door to genetically engineering RNA elements that can be used for detection and control purposes.

In this paper, following the formulation and discussion of four different shape similarity measures, we have exemplified these methods with a procedure to computationally generate small RNA secondary structure switch candidates. The similarity measures may become less reliable for larger RNAs. Thus, no large RNA examples are given. As a consequence of applying our procedure, we have obtained a list of such candidates, among them we have chosen three for illustration. After investigating these switches in laboratory experiments, when more is known about their biochemical and physical properties, it can be useful to include many more sequence variations that were not simulated in the present procedure because the initial population was not randomly chosen. Our shape similarity measures could then be applied to generate additional small RNA candidate switches in a systematic manner.

Acknowledgements

We thank Russell Merris for the suggestion to examine the use of the Wiener index as a similarity measure, and the anonymous reviewers for their helpful comments. The research was partially supported by a grant from the Israel USA binational science foundation BSF 2003291.

References and Footnotes

1. I. Tinoco and C. Bustamante. *J. Mol. Biol. 293*, 271-281 (1999).
2. P. Brion and E. Westhof. *Annu. Rev. Biophys. Biomol. Struct. 26*, 113-137 (1997).
3. C. Flamm, W. Fontana, I. L. Hofacker, and P. Schuster. *RNA 6*, 325-338 (2000).
4. W. Winkler, A. Nahvi, and R. R. Breaker. *Nature 419*, 952-962 (2002).
5. A. S. Mironov, I. Gusarov, R. Rafikov, L. E. Lopez, K. Shatalin, R. A. Kreneva, D. A. Perumov, and E. Nudler. *Cell 111*, 747-756 (2002).
6. P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. *Proc. R. Soc. Lond. B. Biol. Sci. 255*, 279-284.
7. K. A. LeCuyer and D. M. Crothers. *Proc. Natl. Acad. Sci. 91*, 3373-3377 (1994).
8. M. Zuker. *Curr. Opin. Struct. Biol. 10*, 303-310 (2000).
9. R. Micura and C. Höbartner. *ChemBioChem 4*, 984-990 (2003).
10. J. H. Nagel and C. W. Pleij. *Biochimie 84*, 913-923 (2002).
11. G. A. Soukup and R. R. Breaker. *Trends. Biotechnol. 17*, 469-476 (1999).

12. R. R. Breaker. *Nature 432*, 838-845 (2004).
13. B. Voss, C. Meyer, and R. Giegerich. *Bioinformatics 20*, 1573-1582 (2004).
14. C. Flamm, I. L. Hofacker, S. Maurer-Stroh, P. F. Stadler, and M. Zehl. *RNA 7*, 254-265 (2001).
15. D. Barash. *Bioinformatics 20*, 1861-1869 (2004).
16. M. Zuker. *Nucleic Acids Res. 31*, 3406-3415 (2003).
17. I. L. Hofacker. *Nucleic Acids Res. 31*, 3429-3431 (2003).
18. D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. *J. Mol. Biol. 288*, 911-940 (1999).
19. M. Fiedler. *Czechoslovak Math. J. 23*, 298-305 (1973).
20. I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. *Monatsh.Chem. 125*, 167-188 (1994).
21. R. Merris. *Lin. Multi. Alg. 25*, 291-296 (1989).
22. D. Sankoff, J. B. Kruskal, S. Mainville, and R. J. Cedergren. Fast Algorithms to Determine RNA Secondary Structures Containing Multiple Loops, In *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, pp. 913-923. Eds., D. Sankoff and J. B. Kruskal. Addison-Wesley Reading, MA (1983).
23. B. A. Shapiro. *Comput. Appl. Biosci. 4*, 387-393 (1988).
24. H. Margalit, B. A. Shapiro, A. B. Oppenheim, and J. V. Maizel. *Nucleic Acids Res. 17*, 4829-4845 (1989).
25. R. Merris. *Lin. Multi. Alg. 22*, 115-131 (1987).
26. R. Grone and R. Merris. *Czechoslovak Math. J. 37*, 660-670 (1987).
27. H. Wiener. *J. Am. Chem. Soc. 69*, 17-20 (1947).
28. B. A. Shapiro and K. Z. Zhang. *Comput Appl Biosci. 6*, 309-318 (1990).

Date Received: September 30, 2005

Communicated by the Editor Ramaswamy H. Sarma