JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 11, Number 6, 2004

Pp. 1169–1174

Spectral Decomposition for the Search and Analysis of RNA Secondary Structure

DANNY BARASH

ABSTRACT

Scales in RNA, based on geometrical considerations, can be exploited for the analysis and prediction of RNA structures. By using spectral decomposition, geometric information that relates to a given RNA fold can be reduced to a single positive scalar number, the second eigenvalue of the Laplacian matrix corresponding to the tree-graph representation of the RNA secondary structure. Along with the free energy of the structure, being the most important scalar number in the prediction of RNA folding by energy minimization methods, the second eigenvalue of the Laplacian matrix can be used as an effective signature for locating a target folded structure given a set of RNA folds. Furthermore, the second eigenvector of the Laplacian matrix can be used to partition large RNA structures into smaller fragments. An illustrative example is given for the use of the second eigenvalue to predict mutations that may cause structural rearrangements, thereby disrupting stable motifs.

Key words: second eigenvalue of the Laplacian matrix, algebraic connectivity, RNA secondary structure, spectral bisection, deleterious mutations.

INTRODUCTION

METHODS FOR RNA STRUCTURE PREDICTION, a long-time companion to the field of protein structure prediction, have been developed since the late 1970's. In RNAs, because of a limited number of samples in databases as compared to proteins, homology modeling for predicting tertiary structure is currently not feasible. Nevertheless, the secondary structure of RNAs is experimentally accessible and contains important information relating to function. Moreover, RNA folding is known to be hierarchical (Tinoco and Bustamante, 1999), passing through its secondary structure before assuming a tertiary structure fold. Thus, computational prediction methods for RNA folding have concentrated mainly on examining its secondary structure, which is a representation of the complementary base pairings that are formed between nucleic acids given an initial RNA sequence.

The study of RNA secondary structure representation and folding prediction by optimization dates back to early works by Waterman (Waterman, 1978; Smith and Waterman, 1978) following several earlier attempts. Dynamic programming methods were applied by Nussinov *et al*. (1978) and by Nussinov and Jacobson (1980) for computing the maximum number of base pairings in an RNA sequence. Energy-minimization methods by dynamic programming (Zuker and Stiegler, 1981; Zuker and Sankoff, 1984) have gradually developed and resulted in sophisticated packages for RNA folding prediction. These packages, the mfold

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

prediction server (Zuker, 2003) and the Vienna package described by Hofacker (2003), rely on energy rules (Mathews *et al*., 1999) derived by experiments. They contain suboptimal folding solutions (Zuker, 1989), ranked by their percentage of optimality, which may correlate with the experimental result instead of the optimal solution as in the case of the tRNA example used in this paper. Other extensions and folding algorithm approaches (Rivas and Eddy, 1999; Chen *et al*., 2000; Shapiro *et al*., 2002; Gultyaev *et al*., 1995; Gutell *et al*., 2002) have been implemented; reviews are available (see Zuker, 2002, 2003).

RNA secondary structure consists of common elements, such as stems, loops, bulges, and hairpins. Their shape description (Schuster *et al*., 1994) has been the subject of many studies. Several simplified representations of the secondary structure have been devised, including a fine grained tree-graph representation (Fontana *et al*., 1993) that is advantageous for structure-comparison tasks (Hofacker *et al*., 1994) and a coarse grained tree-graph representation (Shapiro, 1988; Le *et al*., 1989) for which some possible applications are presented here by further simplifications using spectral decomposition. Although the most straightforward applications for the spectral decomposition method are achieved in using the coarse-grained tree representation, as proposed in this paper, other applications of this method can be thought of in using the fine-grained tree representation to assist with structure alignment and motif-based searches.

RNA SCALES AND SPECTRAL DECOMPOSITION

In searching and analyzing a collection of RNA folded structures, such as perturbations from the wild-type structure as a consequence of introducing nucleotide mutations, the second eigenvalue and eigenvector of the Laplacian matrix corresponding to a tree-graph representation of the RNA secondary structure can play a potentially useful role in a number of predictive models relating to experimental data. The idea of using spectral decomposition in this context was first proposed by Barash and Comaniciu (2003), with analogy to computer vision scales and the digital total-variation image processing filter (Chan *et al*., 2001). The concept is borrowed from the field of domain decomposition in parallel computing, in which spectral-graph partitioning methods are used to subdivide a large domain and assign different processors to each subdomain in order to achieve load-balancing. Let us examine the notion of scales in RNA structures, with the various scales that are illustrated in Fig. 1. The *tertiary structure* can be reduced to a *secondary structure*, for which sophisticated computational structure prediction methods based on energy minimization exist: e.g., Zuker's mfold (Zuker, 2003) and the Vienna RNA package (Hofacker, 2003). The secondary structure can be further downscaled to a coarse-grained tree-graph (Shapiro, 1988; Le *et al*., 1989; Hofacker *et al*., 1994) and other types of coarse-grained graphs (Gan *et al*., 2003). Furthermore, for predicting clever nucleotide mutations that will perturb a given secondary structure, the idea of spectral decomposition is to represent these tree graphs by a Laplacian matrix and seek the second eigenvalue of the Laplacian matrix called the algebraic connectivity (Fiedler, 1973). Thus, at the coarsest scale, a single

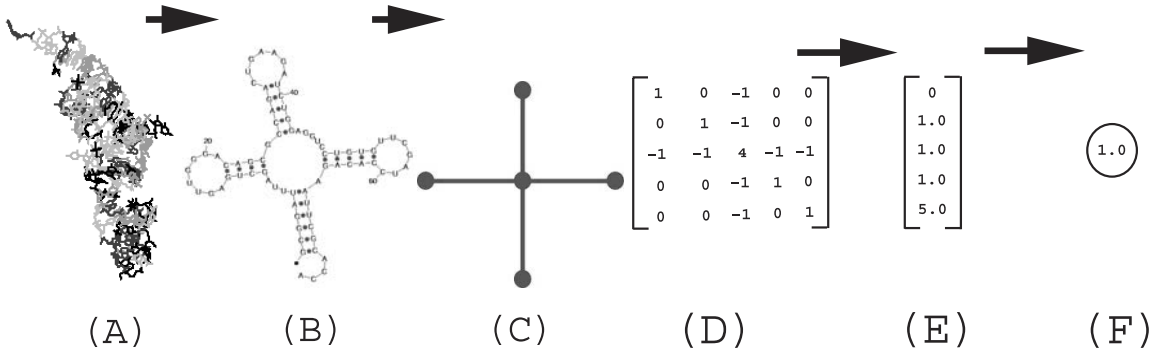

FIG. 1. Scales in the RNA biomolecule: (**A**) Tertiary structure of yeast phenylalanine tRNA. (**B**) Secondary structure, suboptimal prediction correlating with the experimental result. Prediction of suboptimal foldings when performing energy minimization in the secondary structure of RNAs (Zuker, 1989) is valuable since a suboptimal solution may correlate with the experiment instead of the optimal solution. (**C**) Tree-graph representation of the secondary structure. (**D**) Laplacian matrix corresponding to the tree-graph representation. (**E**) Spectrum of the Laplacian matrix. (**F**) Second eigenvalue of the Laplacian matrix.

real positive number—the second eigenvalue of the Laplacian matrix—exists that roughly indicates how the folding occurred as opposed to other folding possibilities. This reduced spatial information coexists with the energy of the folded structure, both being real scalar numbers that are calculated independently. Together, they form an effective signature for measuring similarities and discriminating against unwanted structures, given a target folded structure.

THE LAPLACIAN MATRIX AND ITS SECOND EIGENVALUE

The Laplacian matrix $L(T)$ corresponding to a tree graph T is a symmetric matrix, with one row and column for each node on the tree. It is constructed as follows: in the diagonal of L, the degree of the node (number of incident edges) is specified, while in the off-diagonals the value "-1" is inserted if there is an edge at that location, or "0" if there is no connecting edge. Thus, $L(T) = D(T) - A(T)$, where $D(T)$ is the diagonal matrix of vertex degrees and $A(T)$ is the adjacency matrix. As an example, for the tree-graph representation of an RNA secondary structure in Fig. 1(B), the corresponding Laplacian matrix is depicted in Figure 1(D).

Note that the rows and columns of the Laplacian matrix sum up to zero. The complete set of eigenvalues of the Laplacian matrix is called the spectrum of the graph. The spectrum is independent of how graph vertices are labeled. The following properties that characterize the Laplacian matrix eigenvalues are useful for their application in analyzing RNA structure:

- The eigenvalues of $L(T)$ are nonnegative and the first eigenvalue is zero.
- The second smallest eigenvalue is the algebraic connectivity of T, denoted by $a(T)$ (Fiedler, 1973).
- $0 \leq a(T) \leq 1$.
- $a(T) = 0$ iff T is not connected. This can occur, for example, if separate fragments of an RNA library are to be analyzed. Since loops, bulges, and hairpins are connected through stems, $a(T) > 0$ for each RNA molecule.
- $a(T) = 2(1 - \cos(\pi/n))$ iff $T = P_n$ is a path on n vertices (Fiedler, 1973).
- $a(T) = 1$ iff $T = K_{1,n-1}$ is a star on n vertices (Merris, 1987; Grone and Merris, 1987).

Because of these properties, the second eigenvalue of the Laplacian matrix becomes a useful positive scalar number for practical RNA structure calculations. Intuitively, it is monotonically increasing from its lowest value for a linear tree-graph structure to its highest value of 1.0 for a "star shaped" tree-graph structure, as illustrated in Fig. 2 for two possible formations in the case of a tree graph of 5 vertices. When transforming to a tree graph with a different number of vertices, as often occurs in RNA structures, it is useful to observe that the upper limit remains 1.0 for a star of any number of n vertices, $n > 2$ (Merris, 1987). The lower limit, being the second eigenvalue corresponding to a linear tree graph of n vertices, will increase as the number of vertices decreases since the tree graph becomes more compact. For example, transforming from a linear tree graph of four vertices to a linear tree graph of three vertices will increase the second eigenvalue to the upper limit, $a(T) = 1$, since T becomes a star for $n = 3$ (note that for the special case of $n = 2$, $a(T) = 2$).

Finally, the tRNA in Figure 1 corresponds to a known mathematical example (Merris, 1987). Let $T = K_{1,n-1}$ be a star on n vertices: one vertex of degree $n - 1$ and $n - 1$ pendant (degree 1) vertices. The characteristic polynomial becomes $q_T(x) = x(x - n)(x - 1)^{n-2}$, which can be verified by observing that $(1, 1, \ldots, 1)$ is an eigenvector of $L(T)$ corresponding to the eigenvalue 0; $(n - 1, -1, \ldots, -1)$ is an eigenvector corresponding to n; and $(0, 1, -1, 0, \ldots, 0)$, $(0, 0, 1, -1, 0, \ldots, 0)$, $\ldots$, $(0, 0, \ldots, 0, 1, -1)$ is a set of $n-2$ linearly independent eigenvectors corresponding to 1. Thus, for $n = 5$, the spectrum of $L(T)$ becomes $\{0, 1, 1, 1, 5\}$ as depicted in Fig. 1(E), which can be observed directly from the characteristic polynomial above.

EXAMPLE AND FUTURE WORK

Spectral decomposition in the context of RNA structure can be utilized in a number of different directions. First, as in the example of Fig. 2, it can be used to extend earlier work that addresses mutations at the

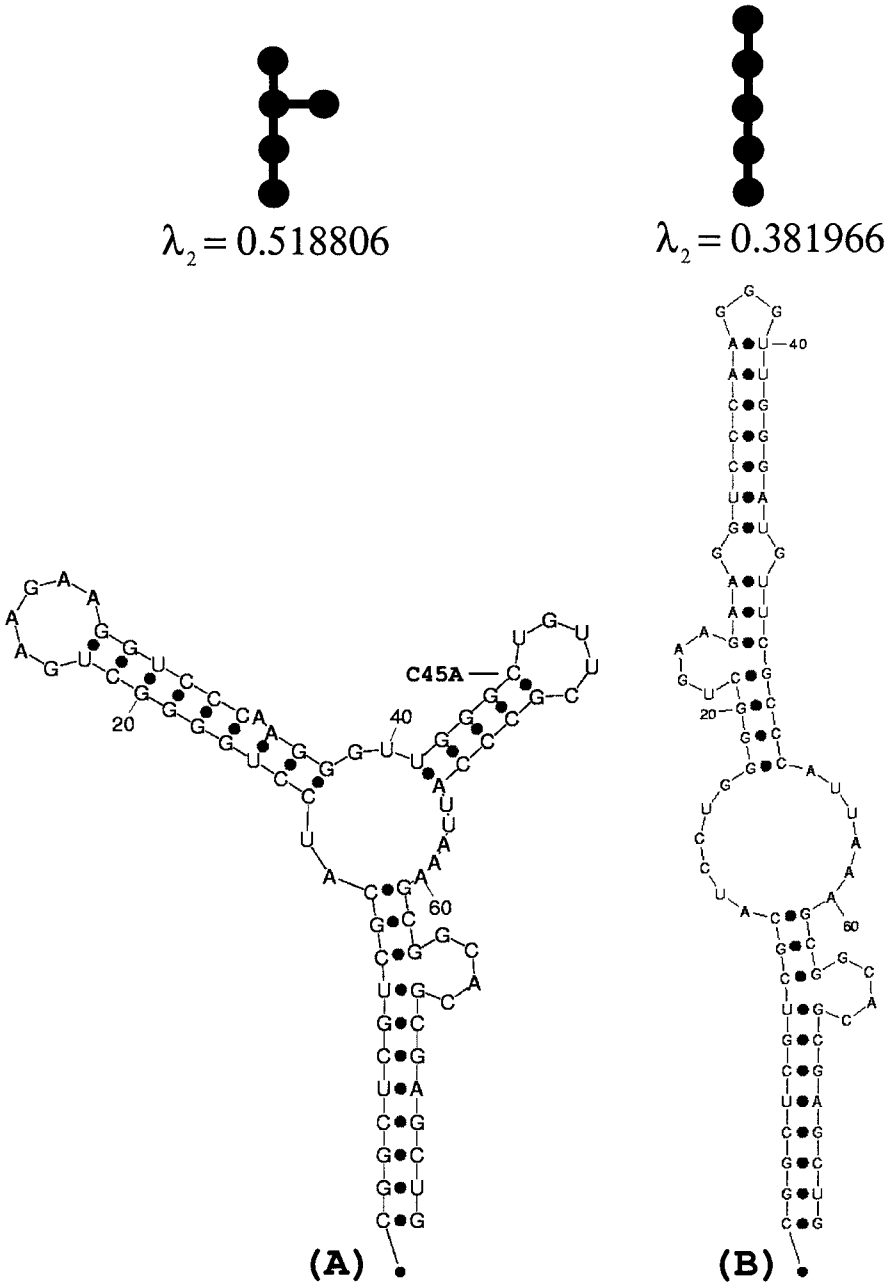

FIG. 2. Mutation prediction using the second eigenvalue of the Laplacian matrix. (**A**) The wild type structure of a domain *V* piece in the 5s rRNA of *T. thermophilus* from the x-ray crystallography experiment in Yusupov *et al.*, 2001 is well predicted by the optimal solution using default parameters in *mfold*, along with its tree-graph representation and second eigenvalue of the Laplacian matrix ($\lambda_2 = 0.518806$). Note that the coarse-grained tree-graph representation used here accounts for a "pseudo-node" that is based on the 5'-3' position in addition to the four nodes representing the bulge, internal loop, and a pair of hairpins. Calculating the second eigenvalue for each folded structure as a consequence of introducing a single point mutation, it is now possible to computationally predict a vulnerable spot in the sequence of the wild type structure (location labeled) whereby the single point mutation C45A leads to a change in the second eigenvalue. (**B**) The mutant structure after inserting the mutation C45A to the wild type structure in (A), causing a predicted conformational rearrangement, along with its linear tree-graph representation and second eigenvalue of the Laplacian matrix ($\lambda_2 = 0.381966$). The transition is from eigenvalue $\lambda_2 = 0.518806$ for the wild type to $\lambda_2 = 0.381966$ for the mutant. The example above serves as a proof of concept for more complicated structures, such as the riboswitches worked out by Barash (2003).

tree-graph scale (Margalit *et al*., 1989) by down-scaling to capture essential structural information, using the second eigenvalue of the Laplacian matrix as an efficient measure to assess perturbations from the tree graph of the wild type structure as a consequence of nucleotide mutations/deletions/insertions (Barash, 2003). For simple RNA structures, systematic mutation prediction using spectral decomposition can be tested in actual laboratory experiments and reveal interesting biological findings. Second, any procedure for searching or scanning a set of RNA secondary structures can utilize eigenvalue information. Third, it can be exploited in the classification of RNA coarse-grained tree-graph structures, by attaching the second eigenvalue as a quantitative catalogic number in the space of tree graphs. Fourth, it may perhaps be used for systematically partitioning large RNA structures into smaller fragments, such as in ribosomal structure calculations (Trifonov and Bolshoi, 1983), by attempting to use spectral bisection methods (Pothen *et al*., 1990).

ACKNOWLEDGMENTS

I am grateful to Drs. Alexander Bolshoy, Alexander Churkin, Eviatar Nevo, and Edward N. Trifonov for useful discussions and suggestions. The work was conducted at the Genome Diversity Center and supported by the Institute of Evolution, University of Haifa, Israel.

REFERENCES

Barash, D. 2003. Deleterious mutation prediction in the secondary structure of RNAs. *Nucl. Acids Res.* 31(22), 6578–6584.

Barash, D., and Comaniciu, D. 2003. A common viewpoint on broad kernel filtering and nonlinear diffusion. *Proc. 4th Int. Conf. on Scale-Space Theories in Computer Vision*, 683–698, Springer Verlag, Berlin.

Chan, T.F., Osher, S., and Shen, J. 2001. The digital TV filter and nonlinear denoising. *IEEE Trans. Imag. Proc.* 10(2), 231–241.

Chen, J.H., Le, S.Y., and Maizel, J.V. 2000. Prediction of common secondary structures of RNAs: A genetic algorithm approach. *Nucl. Acids Res.* 28(4), 991–999.

Fiedler, M. 1973. Algebraic connectivity of graphs. *Czechoslovak Math. J.* 23, 298–305.

Fonatana, W., Konings, D.A.M., Stadler, P.F., and Schuster, P. 1993. Statistics of RNA secondary structures. *Biopolymers* 33(9), 1389–1404.

Gan, H.H., Pasquali, S., and Schlick, T. 2003. Exploring the repertoire of RNA secondary motifs using graph theory with implications for RNA design. *Nucl. Acids Res.* 31, 2926–2943.

Grone, R., and Merris, R. 1987. Algebraic connectivity of trees. *Czechoslovak Math. J.* 37, 660–670.

Gultyaev, A.P., van Batenburg, F.H.D., and Pleij, C.W.A. 1995. The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.* 250(1), 37–51.

Gutell, R.R., Lee, J.C., and Cannone, J.J. 2002. The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.* 12(3), 301–310.

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.

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.

Margalit, H., Shapiro, B.A., Oppenheim, A.B., and Maizel, J.V. 1989. Detection of common motifs in RNA secondary structure. *Nucl. Acids Res.* 17(12), 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(5), 911–940.

Nussinov, R., and Jacobson, A.B. 1980. Fast algorithm for predicting the secondary structure of single stranded RNA. *Proc. Natl. Acad. Sci.* 77(11), 6309–6313.

Nussinov, R., Pieczenik, G., Griggs, J.R., and Kleitman, D.J. 1978. Algorithm for loop matchings. *SIAM J. Appl. Math.* 35, 68–82.

Pothen, A., Simon, H., and Liou, K.P. 1990. Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal.* 11, 430–452.

Rivas, E., and Eddy, S.E. 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* 285(5), 2053–2068.

Schuster, P., Fontana, W., Stadler, P.F., and Hofacker, I. 1994. From sequences to shapes and back: A case study in RNA secondary structures. *Proc. R. Soc. Lond. B.* 255, 279–284.

Shapiro, B.A. 1988. An algorithm for comparing multiple RNA secondary structures. *Comput. Appl. Biosci.* 4(3), 387–393.

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(1), 27–44.

Smith, T.F., and Waterman, M.S. 1978. RNA secondary structure: A complete mathematical analysis. *Math. Biosci.* 42, 257–266.

Tinoco, I. Jr., and Bustamante, C. 1999. How RNA folds. *J. Mol. Biol.* 293(2), 271–281.

Trifonov, E.N., and Bolshoi, G. 1983. Open and closed 5S ribosomal RNA, the only two universal structures encoded in the nucleotide sequences. *J. Mol. Biol.* 169(1), 1–13.

Waterman, M.S. 1978. Secondary structure of single-stranded nucleic acids. *Adv. Math. Suppl. Stud.* I, 167–212.

Yusupov, M.M., Yusupova, G.Z., Baucom, A., Liberman, K., Earnest, T.N., Cate, J.H., and Noller, H.F. 2001. Crystal structure of the ribosome at 5.5Å resolution. *Science* 292, 883–896.

Zuker, M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science* 244, 48–52.

Zuker, M. 2002. Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.* 10(3), 303–310.

Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucl. Acids Res.* 31(13), 1–10.

Zuker, M., and Sankoff, D. 1984. RNA secondary structures and their predictions. *Bull. Math. Biol.* 46, 591–621.

Zuker, M., and Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. *Nucl. Acids Res.* 9, 133–148.

Address correspondence to:
Danny Barash
Genome Diversity Center
Institute of Evolution
University of Haifa
Mount Carmel, Haifa 31905
Israel

E-mail: dbarash@research.haifa.ac.il