MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **60** (2008) 277-290

ISSN 0340 - 6253

USING A NEURAL NETWORK TO IDENTIFY SECONDARY RNA STRUCTURES QUANTIFIED BY GRAPHICAL INVARIANTS

TERESA HAYNES, DEBRA KNISLEY, AND JEFF KNISLEY
INSTITUTE FOR QUANTITATIVE BIOLOGY
DEPARTMENT OF MATHEMATICS,
EAST TENNESSEE STATE UNIVERSITY,
JOHNSON CITY, TN 37614, USA

(Received December 14, 2007)

Abstract. Graphs have been used extensively in theoretical computer science to model various discrete structures, most notably data structures. Chemists have also utilized graphs to represent and quantify molecules and more recently, graphs have appeared in the literature as biomolecules such as RNA and protein structures. In this work, we quantify a graphical representation of secondary RNA structures using trees. By identifying a subset of the trees whose elements are known to model secondary RNA structure, we train a neural network to recognize the patterns of graphical invariants of trees that are RNA-like in structure. What is of particular interest is that these invariants are tools primarily from the field of theoretical computer science. We then identify additional trees that are potential representations of RNA secondary structure that may either occur naturally and have not been identified or may be considered a viable candidate for synthetically produced RNA.

1. Introduction

At one time scientists believed that the number of genes in the human genome would far exceed the number of genes found in less complex species. It is now known that this not the case, but that instead the genome of less complex species contain approximately the same number of genes as the human genome and in some cases even more. The gene regulatory network is now known to be much more complex than previously thought, and the paradigm of the DNA to RNA to proteins is under new scrutiny. It is now well established that RNA molecules in particular play multiple interacting roles and the class of non coding RNA's is rapidly expanding.[1] It has been shown, for example, that a large percentage of the mouse transcriptome is primarily composed of non coding RNAs[2] and evidence suggests that approximately half of the human RNAs are non coding.[3] In fact, the widespread conservation of secondary structure points to a very large number of functional RNAs in the human genome.[4, 5] Therefore it is apparent

that a comprehensive database of RNA motifs is an essential component for further research and a thorough examination of the structural properties of secondary RNA is merited.

Unlike the protein counterpart where sequence alignment methods have proven to be highly successful, many secondary RNA structures can exhibit highly similar features while having very different primary sequences.[6] Consequently, many classes of RNA molecules are characterized by highly conserved secondary structures having little detectable sequence similarity. Multiple sequence alignment methods alone cannot be considered reliable for determining RNA structural characteristics which implies that both sequential and structural information is required in order to expand the current RNA databases.[7] It is assumed that the natural tendency of the RNA molecule is to fold to its most stable conformation and this assumption is the basis for the lowest energy model. Combining sequential and structural information has led to the design of an efficient algorithm for local motif recognition by calculating a structure conservation index based on the minimum free energy paradigm.[4] However, algorithms that incorporate structural information based on the minimum free energy assumption are not always correct. A novel method based on Boltzmann-weighted structures showed marked improvement over structures based on the minimum free energy paradigm.[8] Thus, structural information that does not rely upon the assumption of minimum free energy alone deserves attention. In this work we demonstrate that when a secondary RNA structure is represented by a graph and quantified by graphical invariants, these numerical values are indicative of RNA structure. These graphical values are independent of the minimum free energy constraints.

Secondary RNA structures have frequently been represented by various modeling methods as graph-theoretic trees. RNA tree graphs utilized in this work were first developed by Le et al.[9] and Benededetti and Morosetti[10] to determine structural similarities in RNA. Although trees have been used previously to model secondary RNA structure, applying the graphical invariants of the corresponding graphs has been limited. This is unfortunate since using graph theory as modeling tool allows the vast resources of graphical invariants to be utilized. In previous work by Knisley et.al.[11] it was shown that graphical invariants commonly applied in fields such as computer network design or invariants studied in a purely mathematical setting are in fact indicative of secondary RNA structures. In this paper, we expand upon our findings and introduce additional graph-theoretic parameters. We then use parameters defined in terms

of graph invariants to train a neural network to recognize a tree that is RNA-like in structure. The neural network then identifies additional trees as candidates for novel secondary RNAs.

In the following section, we discuss the combinatorics of simple trees and then follow with a section that describes the modeling method by which simple trees are utilized to represent secondary RNA structure. We reference the RNA database RAG which catalogs all simple trees with eleven or fewer vertices.[12] In the RAG database, all trees of orders 2 through 11 are represented. For trees of orders 2 through 8, each tree has been classified as an RNA tree, an RNA-like tree or not RNA-like tree. For trees of order nine and above, trees that represent a known secondary RNA structure are identified as an RNA tree, but no trees are shown to be candidate structures, i.e. RNA-like. By finding graphical invariants of the trees of orders seven, eight and using the four additional trees of order nine as well, we train a neural network to identify new RNA-like structures of order nine.

2. **Graph-theoretic Trees**

2.1. **Combinatorial aspects of trees.** Trees have been highly studied in graph theory, both for application purposes and theoretical pursuits. A tree is frequently defined as a connected graph with the property that no two vertices lie on a cycle. These two properties of trees (connected and acyclic) completely characterize a tree since the removal of any edge will disconnect the graph and the addition of any edge will generate a cycle. This property of trees also implies that every tree with n vertices has exactly $n-1$ edges. In fact, given any positive integer n, the exact number of trees with n vertices is given by a formula derived by Harary and Prins.[13] Applying combinatorial enumeration techniques, they determined the counting polynomial for unlabeled trees. In particular, they determined the generating polynomial $\sum a_n x^n$ where the coefficient of x^n is the number of trees with n vertices. By the first 10 terms below, we see that there is exactly one tree with 1, 2, or 3 vertices, two trees with 4 vertices, three trees with 5 vertices, six trees with 6 vertices, eleven trees with 7 vertices, twenty three trees with 8 vertices and forty seven trees with 9 vertices and so forth.

$$p(x) = x + x^2 + x^3 + 2x^4 + 3x^5 + 6x^6 + 11x^7 + 23x^8 + 47x^9 + \cdots$$

Thus for unlabeled trees, the exact number of trees that can be drawn is known. Figure 1 contains the 11 trees of order 7.

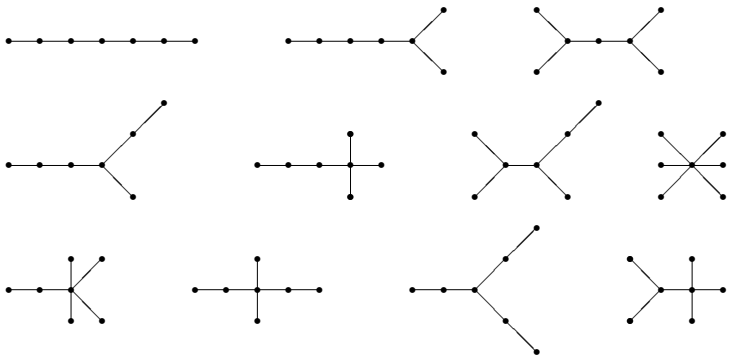

FIGURE 1. The 11 trees of order 7.

2.2. **Tree Representation of Secondary RNA Structures.** In the classic work of Waterman et.al.,[14] secondary RNA structure is defined as a graph where each vertex a_i represents a nucleotide base. If a_i pairs with a_j and a_k is paired with a_l where $i < k < j$, then $i < l < j$. Combinatorial techniques have been applied to enumerate small order graphs with these given constraints. The combinatorial counting results of Waterman are extended by enumerating a variety of sub-classes of secondary graphs by Hofacker et.al.[15] Their work was not concerned with the physical rules that govern the folding process, rather it was concerned with combining structural elements into a new valid structure using combinatorial techniques.

In this work we use the RNA database RAG and the tree model developed by Schlick et.al.[12] Unlike the classic model developed by Waterman et.al. where atoms are represented by vertices and bonds between the atoms by edges in the graph, this model represents stems as edges and breaks in the stems that result in bulges and loops as vertices. A nucleotide bulge, hairpin loop or internal loop are each represented by a vertex when there is more than one unmatched nucleotide or non-complementary base pair. This modeling method is illustrated in Figure 2.

3. **Domination invariants**

There are a number of graphical invariants that are highly sensitive to even a slight change in the structure of a tree. The domination number of a graph is an example of such an invariant. The idea of domination is based on sets of vertices that are near (dominate) all the vertices of a graph. The domination number of a graph is the minimum

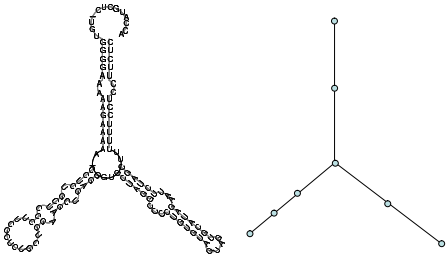

FIGURE 2. Illustration of the Modeling Method

number of vertices in a vertex set S with the property that all remaining vertices not in S are adjacent to at least one of the vertices in S. In [11] two parameters based on the domination number and variations on the domination number are defined by two classes of invariants and are subsequently shown to be indicative of secondary RNA structure. In one class, the domination numbers increase as the amount of branching in the tree increases while in the other class the opposite is true. That is, the numbers decrease as the branching increases. Hence we defined two distinct parameters, P_1 and P_2, grouping the invariants by their behavior. The domination number, total domination number and the global alliance number are used to define P_1 and differentiating domination number and the locating domination number are used for P_2.

$$\begin{aligned} P_1 &= \frac{\gamma + \gamma_t + \gamma_a}{n} \\ P_2 &= \frac{\gamma_L + \gamma_D}{n} \end{aligned}$$

In this work we also utilize the line graph of a tree and we calculate the diameter, the radius and the number of blocks of the line graph for each tree. These values are used to define P_3. To normalize the results, the sums were divided by the total number of vertices in the tree. Before proceeding, we formally define the invariants listed above. These definitions can be found in *Fundamentals of Domination in Graphs*,[16] *Chemical Graph Theory*,[17] or in *Graph Theory and its Applications*.[18]

3.1. **Definitions.** We denote the vertex set of a graph by $V(G)$, or simply V. The number of edges incident to a vertex v is the *degree* of the vertex $deg(v)$, and two vertices are *adjacent* if they are incident to the same edge. A vertex set S is a *dominating set* if for every vertex $u \in V - S$, u is adjacent to at least one vertex in S. The *domination number* $\gamma(G)$ is the minimum cardinality among all dominating sets in G. A set S is a *total dominating set* if for every vertex $u \in V$, u is adjacent to at least one vertex

in S (note here that even the vertices in S must be adjacent to a vertex in S). The *total domination number* $\gamma_t(G)$ is the minimum cardinality among all total dominating sets in G. The *neighborhood of a vertex* v, denoted by $N(v)$, is the set of all vertices adjacent to v and the *closed neighborhood of a vertex* u is $N[u] = N(u) \cup \{u\}$. A dominating set S is called a *locating-dominating set* if for any two vertices $v, w \in V - S$, $N(v) \cap S \neq N(w) \cap S$. Thus, in a locating dominating set, every vertex in $V - S$ is dominated by a distinct subset of the vertices of S. The *locating-domination number* of a graph G is the minimum cardinality among all locating dominating sets in G and is denoted by $\gamma_L(G)$. A dominating set S is called a *differentiating dominating set* if for any two vertices $v, w \in V$, $N[v] \cap S \neq N[w] \cap S$. The *differentiating domination number* of a graph G is the minimum cardinality among all differentiating dominating sets in G and is denoted by $\gamma_D(G)$. The *global alliance number* of a graph G is the minimum cardinality among all global alliances of G, where a set S is a global alliance if S is a dominating set and for each $u \in S$, the number of "allies" it has in S are at least as many as it has in $V - S$. In other words, S is a dominating set and for each vertex $u \in S$, it is true that $|N[u] \cap S| \geq |N(u) \cap (V - S)|$.

The *eccentricity* of a vertex v is the maximum distance from v to any other vertex u in the graph where distance is defined to be the length of the shortest path and is denoted by $d(v, u)$. The *diameter* of G, $diam(G)$, is the maximum eccentricity where this maximum is taken over all eccentricity values in the graph G. That is

$$diam(G) = \max_{u,v \in V} \{d(v, u)\} \tag{1}$$

and the radius of a graph G, denoted by $rad(G)$ is given by

$$rad(G) = \min_{x \in V} \max_{y \in V} \{d(x, y\} \tag{2}$$

The line graph of a molecular graph is known to relay structural information about the corresponding molecule[19]. These values are frequently used as descriptors for QSAR models and properties of the second iterated line graphs also have been applied in computational geometry techniques for biomolecular conformations.[20] The *line graph of* G, denoted by $L(G)$, is a graph derived from G in such a way that the edges in G are replaced by vertices in $L(G)$. Two vertices in $L(G)$ are adjacent whenever the corresponding edges in G share a common vertex. See Figures 3 and 4 for illustrations.

The *block* of a graph as a maximal connected subgraph H such that no vertex of H can be removed resulting in a disconnected graph. Referring to Figure 3, T has 7 blocks and its line graph, $L(T)$ has 3 blocks.

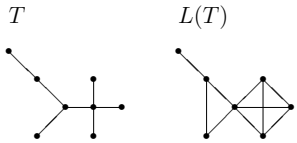

FIGURE 3. RNA-like tree T (8,15) and its associated line graph $L(T)$.

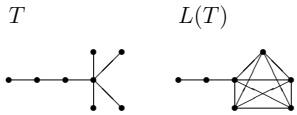

FIGURE 4. A Not RNA-like tree T and its associated line graph $L(T)$.

3.2. **Invariants as biomolecular quantifiers.** Parameters P_1 and P_2 are normalized sums of the graphical invariants defined in the preceding section:

The diameter and radius are indicative of the distances between the vertices and hence with respect to the model, the invariants in P_3 measure relative distance between the helical stems. We also include the clustering aspects of the stems by the number of blocks of the line graphs. Thus, P_3 measures the "average" of these invariants, normalized by n and where $L(T)$ represents the line graph of the tree and $|B|$ is the number of blocks in the line graph of the tree.

$$P_3 = \frac{diam(L(T)) + rad(L(T)) + |B|}{n}$$

4. An Artificial Neural Network

The graph theoretic parameters for the 26 trees of order 7, 8, or 9 that are either verified as RNA trees or classified as not RNA-like are used to predict the RNA-like status of the 55 remaining trees of order 7, 8, or 9. This method of classification yields results similar to the statistical analysis by the authors in a previous work and provided us with a predictive tool for the forty seven trees of order nine. These preliminary findings using graphical invariants as input values for the neural network are very promising and suggest that many other objectives in proteomics and genomics may be achieved by combining graph-theoretic models with neural network classifiers. In the following sections, we discuss the particulars of the training, the design of the algorithm, and the analysis of the training results.

4.1. **Algorithm and Implementation.** Our approach is to train a multi-layer perceptron (MLP) artificial neural network using a standard back-propagation algorithm.

Results from a back-propagation MLP can be reproduced independently by other researchers and can also provide information beyond simple predictions.

In an MLP, the i^{th} perceptron is an artificial neuron with an activation x_i that satisfies

$$x_i = \sigma\left(\sum_{j \neq i} w_{ij} x_j - \theta_i\right)$$

where w_{ij} is the weight of the connection between the i^{th} and j^{th} perceptrons, where x_j is the activation of the j^{th} perceptron, where θ_i is the threshold for the i^{th} node, and where

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

is the activation function.[21]

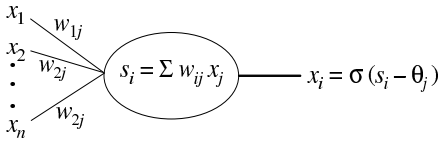

FIGURE 5. Node of an Artificial Neural Network.

A layer is a collection of perceptrons that are connected to other layers in the network but not to each other.

A 3-layer MLP is used to predict the RNA-like status of the trees. The first layer, or *input layer*, contains 3 perceptrons corresponding to P_1, P_2, and P_3. The last layer, or *output layer,* consists of 2 perceptrons with activations y_1 and y_2, where $y_1 = 1$ and $y_2 = 0$ if the tree is predicted to be an RNA tree and where $y_1 = 0$ and $y_2 = 1$ if the tree is not RNA-like. The middle layer, or *hidden layer*, is comprised of 12 perceptrons. The weights between the input and hidden layers will be denoted by w_{jk} and the weights between the hidden and output layers will be denoted by α_{ij}.

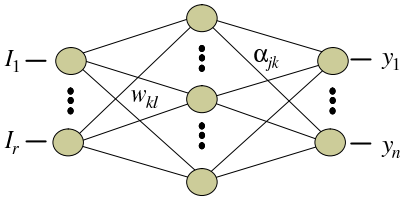

FIGURE 6. MLP for RNA Trees

The values of the 3 graphical parameters for the 26 trees that either are an RNA tree or not RNA-like determine a training set

$$TS = \left\{ \left(\mathbf{p}^i, \mathbf{q}^i \right) \right\}_{i=1}^{26}$$

where $\mathbf{p}^i = \langle p_1^i, p_2^i, p_3^i \rangle$ is the vector of normalized graphical parameters, $\mathbf{q}^i = \langle 1, 0 \rangle$ if the tree is known or predicted to be an RNA tree, and $\mathbf{q}^i = \langle 0, 1 \rangle$ if the tree is not RNA-like. The backpropagation algorithm is used to implement a gradient following minimization of the total squared error

$$E = \frac{1}{2} \sum_{i=1}^{26} \left\| \mathbf{y}\left(\mathbf{p}^i\right) - \mathbf{q}^i \right\|^2$$

where $\mathbf{y}\left(\mathbf{p}^i\right) = \langle y_1\left(\mathbf{p}^i\right), y_2\left(\mathbf{p}^i\right) \rangle$ is the output due to an input of $\mathbf{p}^i$ and the norm is generated by the corresponding dot product.

The weights are initially assigned random values close to 0. Then for each pair $\left(\mathbf{p}^i, \mathbf{q}^i\right)$, the weights α_{jk} are adjusted using

$$\alpha_{jk} \rightarrow \alpha_{jk} + \lambda \delta_j \xi_k$$

where $\xi_k = \sigma\left(\sum w_{kj} p_j^i - \theta_k\right)$, where $\lambda > 0$ is a fixed parameter called the *learning rate,* and where

$$\delta_j = y_j \left(1 - y_j\right) \quad \left(q_j^i - y_j\right)$$

The weights w_{kr} are adjusted using

$$w_{kl} \rightarrow w_{kl} + \lambda p_l^i \; \xi_k \left(1 - \xi_k\right) \sum_{j=1}^{2} \alpha_{jk} \delta_j$$

In each training session, the patterns should be randomly permuted to avoid bias, and training should continue until E is sufficiently close to 0.[21]

5. **Results**

The network was trained on the two classes of trees, RNA trees and not RNA-like trees. The 15 RNA trees of order 7, 8 and 9 and the 11 trees of order 7 and 8 classified as not RNA-like make up the training set of 26 trees. Figure 7 shows the 15 RNA trees that were used in the training set. Typically, the network converged to $E < 0.005$ in fewer than 2,000 training sessions, but for uniformity throughout testing and prediction with the MLP, all networks were trained for 10,000 training sessions.

The MLP artificial neural network was tested using *leave one out (LOO)* cross-validation in which each of the 26 known trees was omitted from the training set, the MLP was trained with the remaining 25, and then the trained network was used to

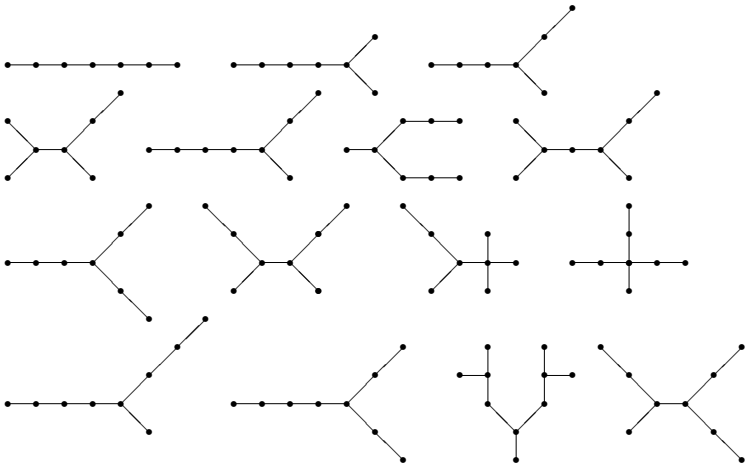

FIGURE 7. The 15 RNA Trees of the Training Set.

predict the classification of the omitted tree. Leave one out cross-validation is a reliable measure of the generalization error of the network when the training set is not too large.[22]

The RMS error across all 26 LOO trials is 0.1456. The individual errors are shown in Table 1. The RNA tree in Figure 3 is the only tree that the MLP had difficulty

TABLE 1. Error in Predicting the Class of the Given Tree using Leave One Out Cross Validation

RAG[a]	Class[b]	Error[c]	RAG	Class	Error
7.1	1	3.8268E-07	8.14	0	1.7051E-05
7.2	1	8.4021E-02	8.15	1	7.3690E-01
7.3	1	7.9215E-08	8.17	0	2.2156E-02
7.6	1	2.1239E-05	8.18	0	3.2810E-08
7.9	0	9.9406E-08	8.19	0	6.5133E-08
7.10	0	9.5272E-08	8.20	1	3.1600E-07
7.11	0	1.8141E-08	8.21	0	2.2388E-02
8.3	1	2.2135E-08	8.22	0	2.3965E-08
8.5	1	5.3006E-08	8.23	0	3.1796E-09
8.7	1	1.2076E-06	9.6	1	2.8982E-07
8.9	0	2.6632E-05	9.11	1	5.4533E-08
8.10	1	2.8906E-08	9.13	1	8.7496E-05
8.11	1	2.8228E-08	9.27	1	7.3754E-08

[a] Labels from the RAG RNA database [12].
[b] Class = 1 if an RNA tree, Class = 0 if not RNA-Like.
[c] Average deviation from predicted class.

classifying during the LOO analysis. Performance of the MLP was improved somewhat by designating trees 8.17 and 8.21 as unclassified rather than as not RNA-like. In particular, when LOO analysis is performed on the remaining 24 trees, the RMS error reduces to 0.0605, and the error in predicting 8.15 reduces to 0.2954. However, the overfitting of data–an issue considered in depth for this project–is a greater concern for this reduced data set.

More generally, the MLP was also tested by *predicting complements*, in which the 26 known trees are randomly partitioned into a training set and a complement (also known as leave-v-out cross-validation). Once the network was trained, the RMS error in predicting the complement was calculated. Predicting complements was performed over 10 trials for each of 6, 13, and 20 trees, respectively, in the complement, with the results shown in Table 2. Again, the results in Table 2 can be improved slightly by designating trees 8.17 and 8.21 as unclassified rather than as not RNA-like.

TABLE 2. $|Comp| =$ Number of Trees in Complement

	$\|Comp\| = 6$	$\|Comp\| = 13$	$\|Comp\| = 20$
Average Error	0.084964905	0.161629391	0.305193489
Standard Deviation	0.125919698	0.127051425	0.188008046

The MLP was subsequently re-trained over the entire 26 tree training set, and then the MLP was used to predict whether or not unclassified trees of orders 7, 8, or 9 could be predicted to be an RNA tree or to be not RNA-like. The results are shown in Table 3. Predictions produced by the reduced training set formed by designating trees 8.17 and 8.21 as unclassified rather than as not RNA-like differs only slightly from Table 3. Specifically, the two trees of orders 7 and 8 predicted to be not RNA-like in Table 3 are predicted to be RNA trees in the reduced training set. For trees of order 9, only the prediction for tree 9.9 differs for the reduced training set.

Finally, it appears that the relationships between P_1, P_2, and P_3 vary between the RNA and not RNA-like trees, as is illustrated in Table 4. Indeed, Table 4 seems to reinforce the discussion above about the significance of the graphical invariants. Averages for the reduced training set formed by designating 8.17 and 8.21 as unclassified rather than as not RNA-like are similar.

TABLE 3. Predictions for the 55 unclassified trees

RAG[a]	Class[b]	Error[c]	RAG	Class	Error	RAG	Class	Error
7.4	0	0.00947	9.9	0	0.0554	9.31	1	0.0247
7.5	1	0.0245	9.10	1	2.65E-06	9.32	0	1.99E-06
7.7	1	7.45E-05	9.12	1	5.28E-07	9.33	1	0.0462
7.8	1	1.64E-07	9.14	1	2.32E-07	9.34	1	0.00280
8.1	1	1.05E-06	9.15	0	1.82E-04	9.35	0	2.46E-06
8.2	1	1.24E-06	9.16	1	5.35E-04	9.36	0	7.41E-05
8.4	1	0.0138	9.17	1	6.24E-06	9.37	0	7.41E-05
8.6	1	0.0138	9.18	1	4.87E-07	9.38	1	4.86E-05
8.8	1	5.43E-05	9.19	1	6.06E-07	9.39	0	2.46E-06
8.12	1	3.59E-06	9.20	1	0.0247	9.40	0	4.79E-08
8.13	0	0.0157	9.21	1	6.38E-05	9.41	0	4.79E-08
8.16	1	8.81E-06	9.22	1	0.0247	9.42	1	2.51E-07
9.1	1	1.48E-07	9.23	0	7.41E-05	9.43	1	4.86E-05
9.2	1	0.0151	9.24	1	1.47E-05	9.44	1	0.0247
9.3	1	0.0121	9.25	0	3.85E-07	9.45	0	7.41E-05
9.4	1	4.05E-07	9.26	1	1.48E-04	9.46	0	4.79E-08
9.5	1	5.24E-05	9.28	0	7.41E-05	9.47	0	2.33E-08
9.7	1	6.38E-05	9.29	1	3.61E-07			
9.8	1	6.38E-05	9.30	1	1.47E-05			

[a]Labels from the RAG RNA database [12].
[b]Class = 1 if predicted to be an RNA tree, Class = 0 if not RNA-Like
[c]Average deviation from predicted class.

TABLE 4. Averages for P_1, P_2, and P_3 over subsets of trees of orders 7, 8, and 9.

	P_1	P_2	P_3
Overall	1.2418	1.2262	1.1440
Classified Trees (26)	1.2453	1.2399	1.1023
RNA trees Trees (15)	1.4085	1.0776	1.3701
not RNA-like Trees (11)	1.0227	1.4610	0.7370
Unclassified Trees (55)	1.2403	1.2198	1.1637
Predicted RNA trees (38)	1.3261	1.1374	1.3063
Predicted not RNA-like (17)	1.0484	1.4040	0.8450

6. **Conclusion**

Using graphical invariants of RNA trees of orders 7, 8, and 9 together with trees of orders 7, 8, and 9 classified as not RNA-like, we train an artificial neural network to recognize a tree as an RNA tree or as not RNA-like in structure. We then predict RNA tree or not RNA-like for all remaining trees of orders seven, eight and nine. The results for the trees of orders seven and eight agree with the RNA database RAG classification

with the exception of two trees out of the possible 34 tree structures. We also classify the forty four trees of order nine which are not classified in the database. This approach has many implications for the emerging area of RNA as a tool for drug development since RNA-based drug discovery requires general structural information to guide rational drug design.[23, 24, 25] It is clear that structural information on secondary RNA molecules is needed from a variety of standpoints. Functional clusters of RNAs are a valuable source of diagnostic targets for methods of managing disease. RNA-based drug discovery requires general approaches for detecting and quantifying RNA-protein interactions.

In this work we demonstrate that graphical invariants from the field of mathematical graph theory can be used to numerically identify characteristics of secondary RNA structures sufficiently well that an artificial neural network can be trained to recognize the difference. These invariants do not rely upon the minimum free energy paradigm, but rather they measure the structural characteristics of the molecule when represented as a graph. This does not imply that the minimum free energy supposition is invalid, but rather that it is inherent in the structure. The field of graph theory is a rich source of invariants that can be utilized as biomolecular descriptors. Given that a biomolecule may be represented as a graph, the wealth of graphical invariants and the implications of their values with respect to the corresponding structural characteristics of the graph provides an extensive, unexplored means for quantification and interpretation.

Acknowledgments

This work was supported in part by the National Science Foundation grant, DMS-0527311 and by the Institute for Mathematics and its Applications with funds provided by the National Science Foundation. The authors also thank the Schlick group for the creation and maintenance of the web resource RAG: RNA-As-Graphs.

References

[1] S. Washietl, I. Hofacker, M. Lukasser, A. Huttenhofer, P. Stadler, Mapping on conserved RNA secondary structures predicts thousands of functional non coding RNAs in the human genome, *Nature Biotech.* **23** (2005) 1383–1390.

[2] M. Suzuki,Y. Hayashizaki, Mouse-centric comparative transcriptomics of protein coding and non coding RNAs, *BioEssays* **26** (2004) 833–843.

[3] S. Cawley, S. Bekiranov, H. Ng, P. Kapranov, E. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl, T. Gingeras, Unbiased mapping of transcription factor binding site along human chromosomes 21 and 22 points to widespread regulation of non coding RNAs, *Cell* **116** (2004) 499–509.

[4] S. Washietl, I. Hofacker, P. Stadler, Fast and reliable prediction of non coding RNAs *Proc. Natl. Acad. Sci. USA* **101** (2005) 2454–2459.

[5] S. Washietl, I. Hofacker, P. Stadler, Genome-wide mapping of conserved RNA Secondary Structures Reveals Evidence for Thousands of functional Non-Coding RNAs in Human (preprint).

[6] S. Bernhart, I. Hofacker, P. Stadler, PM-Match - A New Way to Align RNA Structures *Abstracts of Math-Chem-Comp 2004* http://mcc.irb.hr/mcc04/abs04.html, 2004.

[7] R. Backofen, S. Will, Local Sequence-Structure Motifs in RNA, *J. Bio. Comp. Biol.* **2** (2004) 681–698.

[8] Y. Ding, C. Chan, C. Lawrence, RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble, *RNA* **11**(2005) 1157–1166.

[9] S. Le, R. Nussinov, J. Maziel, Tree graphs of RNA secondary structures and their comparison,*Comp. Biomed. Res.* **22** (1889) 461–473.

[10] G. Benedetti, S. Morosetti, A graph-topological approach to recognition of pattern and similarity in RNA secondary structures, *Biol. Chem.* **22** (1996) 179–184.

[11] T. Haynes, D. Knisley, E. Seier, Y. Zoe, A Quantitative Analysis of Secondary RNA Structure Using Domination Based Parameters on Trees, *BMC Bioinformatics* **7** (2006) 108.

[12] D. Fera, N. Kim, N. Shiffeldrim, J. Zorn, U. Laserson, H. Gan, T. Schlick, RAG; RNA-As-Graphs web resource, *BMC Bioinformatics* **5** (2004) 88.

[13] F. Harary, G. Prins, The number of homeomorphically irreducible trees and other species, *Acta Math* **101** (1959) 141–162.

[14] M. Waterman, *An Introduction to Computational Biology: Maps, Sequences and Genomes* Chapman Hall/CRC; 2000.

[15] I. Hofacker, P. Schuster, P. Stadler, Combinatorics of RNA secondary structures *Discrete Appl. Math.* **88** (1998) 207–237.

[16] T. Haynes, S. Hedetniemi, P. Slater, *Fundamentals of Domination in Graphs* Marcel Dekker, 1998.

[17] N. Trinajstic, *Chemical Graph Theory* CRC Press, 1992.

[18] J. Yellen, J. Gross, *Graph Theory and Its Applications* CRC Press, 1998.

[19] E. Estrada, N. Guevara, L. Gutman, L. Rodriguez, Molecular connectivity indices of iterated line graphs. A new source of descriptors for QSAR and QSPR studies, *SAR QSAR Environ. Res.* **9** (1998) 229–240.

[20] D. Dix, An Application of Iterated Line Graphs to Biomolecular Conformation (preprint).

[21] N. K. Bose and P. Liang, *Neural Network Fundamentals with Graphs, Algorithms, and Applications*, McGraw-Hill, New York, 1996.

[22] J. Shao, Linear model selection by cross-validation, *J. Am. Statistical Association*, **88** (1993) 486–494.

[23] J. Zorn, H. Gan, N. Shiffeldrim, T. Schlick, Structural Motifs in Ribosomal RNAs: Implications for RNA Design and Genomics, *Biopolymers* **73** (2004) 340–347.

[24] H. Gan, S. Pasquali, T. Schlick, Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design, *Nucleic Acids Research* **31** (2003) 2926–2943.

[25] Kim N, Shiffeldrim N, Gan H, Schlick T, Candidates for novel RNA topologies, *J. Mol. Biol.* **341** (2004) 1129–1144.