From: ISMB-95 Proceedings. Copyright © 1995, AAAI (www.aaai.org). All rights reserved.

Graph-Theoretic Approach to RNA Modeling Using Comparative Data

Robert B. Cary and Gary D. Stormo
Dept. of Molecular, Cellular and Developmental Biology
University of Colorado, Boulder, CO 80309-0347
phone: 303-492-1476, FAX: 303-492-7744
rbcary@beagle.colorado.edu
stormo@beagle.colorado.edu

Abstract

We have examined the utility of a graph-theoretic algorithm for building comparative RNA models. The method uses a maximum weighted matching algorithm to find the optimal set of base-pairs given the mutual information for all pairs of alignment positions. In all cases examined, the technique generated models similar to those based on conventional comparative analysis. Any set of pairwise interactions can be suggested including pseudoknots. Here we describe the details of the method and demonstrate its implementation on tRNA where many secondary and tertiary base-pairs are accurately predicted. We also examine the usefulness of the method for the identification of shared structural features in families of RNAs isolated by artificial selection methods such as SELEX.

Introduction

Phylogenetic models of RNA secondary structure have provided a useful source of insight to the folding of structural RNAs. Indeed, for tRNA, the single case where crystalographic data is available, the phylogenetic model proved to be a highly accurate predictor of the secondary structure (Holley *et al.* 1965; Sussman & Kim 1976). The accuracy of comparative models arise in part from the evolutionary conservation of base interactions rather than absolute sequence (Noller 1984). Compensating base changes allow the evolutionary conservation of important interactions. Examining pairs of positions in a multiple sequence alignment reveals sites of compensating base changes. Sites that co-vary are likely to form interactions such as Watson-Crick base-pairs.

Many comparative models have been built by hand using manually aligned sequences and visual inspection (Holley *et al.* 1965; Fox & Woese 1975; Noller 1984). Recently, computational approaches have been applied to aid in both sequence alignment and the identification of consensus secondary structures (Altman 1993; Chiu & Kolodziejczak 1991; Eddy & Durbin 1994; Gutell *et al.* 1992; Klinger & Brutlag 1993; Sakakibara *et al.* 1994; Searls 1993). One quantitative approach is based on the calculation of the mutual information (MI) for two positions in a sequence alignment (Chiu & Kolodziejczak 1991; Gutell *et al.* 1992). The result is a measure of the covariance of two positions. MI analysis can be applied to an entire sequence, generating an exhaustive list of all possible position pairs and their associated MI.

Previous reports have described the use of filtered MI data sets to identify position pairs likely to interact. Filters are usually designed to reduce the data set size by removing positions for which MI values are below an arbitrary threshold or fail to be one of the few best scores for a given position. The filtered data is then used to manually build a model. This approach neglects a number of important considerations. Mutually exclusive helices may be indicated by the data. Manual data analysis leaves the resolution of helix conflicts up to the judgment of the researcher, neglecting the possible global consequences on the model. Additionally, long range interactions and complex structures such as pseudoknots may be difficult to decipher manually. The only reported example of automated covariance modeling depends on the use of a dynamic programming algorithm which is unable to describe pseudoknots (Eddy & Durbin 1994).

The limitations of current comparative analysis techniques leads to the question, does there exist an efficient method to find all favorable position pairs such that a globally optimum model will result? If so, does such a method yield a useful model? To address these questions, we have examined the feasibility of automating the process of building models from MI data. Automation is based on the assumption that the best model will be described by that subset of non-conflicting interactions which gives the greatest sum of MI. Finding those base-pairs that result in the greatest global sum of MI is a problem that can be solved in $O(N^3)$ time by the application of a well known graph-theoretic algorithm.

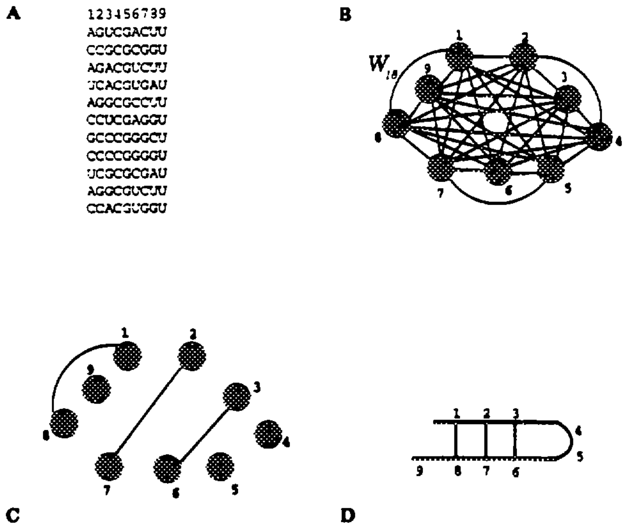
Graph-Theoretic RNA Modeling

By constructing a graph to represent a sequence alignment and associated pairwise MI, it is possible to find that set of non-conflicting base-pairs which has the highest summed MI. To understand the approach employed we must first consider some simple graph-theoretic concepts and their application to RNA modeling. A graph is composed of vertices and edges, or lines, connecting vertex pairs. We define a graph as $G(V, E)$ where V is the set of vertices and E the set of edges in graph G. We wish to represent a sequence alignment and pairwise MI as a graph. This representation is constructed by defining each position of a multiple sequence alignment as a vertex in V . An edge, $E(i, j)$, is defined as the edge incident to vertices i and j. Two vertices connected by an edge are said to be adjacent. In our graphical representation of sequence alignment and MI, there exists an edge for every vertex pair to represent the MI for every position pair. This results in a dense or complete graph where any two vertices are adjacent. We assign the MI associated with position pair (i, j) to edge $E(i, j)$ as weight W_{ij}. Figure 1 shows an example sequence alignment and representative graph.

The pairwise matching of bases in secondary or tertiary base-pairs disallows simultaneous interaction with other bases. This strict pairwise interaction correlates conveniently with the graph-theoretic concept of a matching. Using the same graph as described above, $G(V, E)$, a subset $M \subseteq E$ is a matching if no two edges in M are incident to the same vertex. Any RNA model that describes secondary and tertiary base-pairs is essentially a matching. In other words, just as no two base-pairs in an RNA model share a base, no two edges in a matching share a vertex. Thus, a matching on a graph that represents a RNA sequence alignment is a model of base-pairings.

We are now left with the finding that a matching can be used to describe a RNA model. However, the particular model that we require is the one described by the set of base-pairs with the greatest sum of MI. This model corresponds to the special matching on our sequence/MI graph known as the maximum weighted matching. A maximum weighted matching (MWM) is a matching for which the sum of the edge weights is maximum. It can easily be seen that the maximum weighted matching on the graph constructed as described, using MI values as edges weights, corresponds to a model composed of those base-pairs that give the greatest sum of MI. Finding a maximum weighted matching on a graph is a classic graph-theoretic problem that can be solved in $O(N^3)$ time by an algorithm due to Gabow (Gabow 1973; 1976).

Here we show that the application of Gabow's MWM algorithm to sequence alignment graphs builds RNA comparative models well suited for use in predicting secondary and tertiary base-pairings. Unlike approaches described previously which use dynamic programming or stochastic context-free grammars (Eddy & Durbin 1994; Nussinov & Jacobson 1980; Sakakibara *et al.* 1994; Searls 1993), the MWM approach does not disallow complex tertiary foldings such as pseudoknots. The method is demonstrated using an alignment of tRNAs where both secondary and tertiary base-pairs are accurately identified. We also demonstrate the approach on an alignment of synthetic RNAs isolated by *in vitro* selection for their shared ability to bind HIV-1 reverse transcriptase. The results indicate the technique is a useful means of identifying secondary and tertiary base-pairs when multiple sequences of conserved function are available.



```
123456739
AGUCGACUU
CCGCGCGGU
AGACGUCUU
UCACGUGAU
AGGCGCCUU
CCUCGAGGU
GCCCGGGCU
CCCCGGGGU
UCGCGCGAU
AGGCGUCUU
CCACGUGGU
```

Figure 1. An example of a sequence alignment represented as an undirected graph. The short nine nucleotide sequence alignment (A) can be represented as a graph with nine vertices (B), each vertex represents a position in the alignment. Each vertex is connected to all other vertices by lines known as edges. Each edge represents a possible position pair. The mutual information is calculated for every pair of positions and used as the weight associated with the edge representing the corresponding position pair. Here only a single representative weight (W_{18} in part B) is shown. A weighted matching on the graph gives the set of edges whose weights are maximized subject to the constraint that no two edges share a vertex (C). The matching is used to determine the structural model (D).

Methods

Data sets

The tRNA gene sequence alignment was obtained via anonymous ftp (Steinberg, Misch, & Sprinzl 1993). Part one of the trna.com file was edited to remove mitochondrial gene sequences. Mitochondrial tRNA structure is poorly understood and the proper alignment of these sequences with other tRNAs is difficult to determine. The tRNA alignment as used contained 807 sequences. The 16S rRNA multiple sequence alignment was obtained from The Ribosomal Database Project (Maidak *et al.* 1994). Alignment formats were modified to allow use with the programs described below. The synthetic RNA data set was taken from Tuerk, MacDougal, & Gold 1992 and aligned manually.

Calculation of MI values

The mutual information in two positions x and y of a multiple sequence alignment is a log-likelihood ratio normalized by the number of sequences in the alignment. A calculation of the MI for all base pairs in a sequence alignment generates a half matrix, since the measure is symmetric. To determine the MI values for the alignments, we made use of a modified version of the previously described MIXY (Mutual Information of X and Y) program written by A. Power (Gutell *et al.* 1992). Higher MIXY values indicate a higher probability that two bases do not vary independently. MI was calculated, as previously described (Gutell *et al.* 1992), using entropy measures:

$$M(x,y) = H(x) + H(y) - H(x,y) \qquad (1)$$

where $M(x,y)$ is the mutual information of positions x and y in the alignment, H is the entropy measure $H = -\sum_b f_b \ln f_b$, f_b is the frequency of base b (A, C, G, or U).

In some cases we examined the usefulness of subtracting a constant from the MI values. The constant used was based on the expected MI assuming independent variation. The calculations were made as follows:

$$M'(x,y) = M(x,y) - \nu/2N \qquad (2)$$

where $M'(x,y)$ is the corrected MI value for positions x and y, $M(x,y)$ is the MI value described in equation (1), ν is the degrees of freedom, and N is the number of sequences in the alignment.

Graph Construction and Matching

To find the set of position pairs that gives the greatest sum of MI values, we designed a program to convert MIXY output into a graph description. The resulting graph can be used to find a MWM in $O(N^3)$ time using an algorithm devised by Gabow (Gabow 1973; 1976). Graphs were defined such that each vertex represented a base in the RNA sequence. Edges were created for all vertex pairs (*i.e.* all position pairs) with the corresponding MI assigned as the edge weight (see Fig. 1). To find the MWM we used the wmatch program, an implementation of Gabow's maximum weighted matching algorithm, written by Ed Rothberg.

Results and Discussion

Initial tests of the MWM approach were carried out on an alignment of 807 tRNA gene sequences (Steinberg, Misch, & Sprinzl 1993). MI values were calculated for all possible position pairs using MIXY. MIXY output was converted into a graph description for finding the MWM. Weighted matching data was used to build a model such that each edge in the matching corresponded to a base-pair in the model. We compared the resulting model to the tRNA crystal structure to determine the extent to which a MWM is able to identify base-pairs given the MI comparative data (Fig. 2). The MWM predicted 25 of the 27 base-pairs present in the crystal structure (Sussman & Kim 1976; Holbrook *et al.* 1978).

A total of 35 position pairs were suggested by the MWM. Of the 10 position pairs suggested by the MWM but absent from the crystal structure, most appeared to be well founded in the context of the accepted model. For example, the tertiary base-pairs formed by bases 18:55 and 19:56 are in close proximity to the suggested base-pairs 17:59, 20:60 and 21:57. Another MWM suggested pairing is an extension of the anti-codon stem (26:44 to 31:39) to a C:A wobble pair at 32:38. The presence of this particular base-pair has been described in some crystalographic analyses (see Sussman & Kim 1976).

It should be noted that compensating base changes could maintain critical aspects of tRNA structure unrelated to the formation of base-pairs and other direct interactions. The interpretation of comparative models must consider the array of constraints that might be represented by covariance relationships. Determining the exact nature of a covariance relationship is not possible based on comparative data alone. A number of the suggestions made by the MWM model could reflect indirect interactions. In particular the 36:37 interaction is likely to reflect anticodon-loop interactions that may be responsible for the effects which have led to the extended anticodon hypothesis (Yarus 1982). Other factors are likely at play for the interaction between the anticodon and the acceptor stem (35:73). This covariance probably reflects the involvement of position 73 and the anticodon in determining recognition by aminoacyl-tRNA synthetases (for review see (Saks, Sampson, & Abelson 1994)).

In some cases, the MWM appears to generate erroneous suggestions due to the constraints of the matching. For example, the three base triples identified in the crystal structure pose a problem for the MWM which allows only pairwise associations. The base triple 9:12:23 is identified by the MWM as base-pair 12:23 while position 9 is erroneously paired with position 45. Similar situations arise for the other two base triples (see Fig. 2).

The remaining unexpected position pairings result from more serious limitations of comparative analysis. For example, position 18 is unpaired in the MWM model while in the crystal structure 18 forms a tertiary base-pair with position 55. Similarly, position 58 is unpaired in the model though paired with position 54 in the crystal structure. The failure of the MWM to assign partners to positions 18 and 58 results from the nearly invariant nature of these positions. Invariant positions do not contain MI and are therefore ignored by the matching algorithm. In some instances positions are not strictly invariant but rather vary so infrequently that the MI is too low to be included by the matching. This is an unavoidable consequence of the comparative approach, those positions that are strictly conserved fail to provide any comparative information.

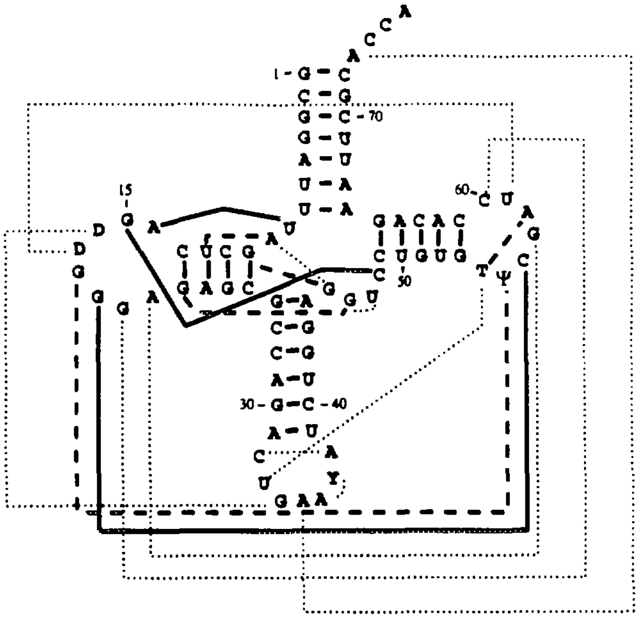


Figure 2. Results of a maximum weighted matching on a graph representing an alignment of 807 tRNA gene sequences. The matching was used to model the yeast phe-tRNA as shown here. Solid lines represent base interactions that are found by the matching and are also present in the crystal structure. Three of the five interactions found in the crystal structure but absent from the matching (dashed lines) are in base triples, an interaction disallowed by the definition of a matching. Dotted lines indicate interactions that are proposed by the matching but fail to be found in the crystal structure.

The results from the tRNA alignment were encouraging and suggested that MWM modeling may provide a means of rapidly extracting secondary and tertiary interactions from a sequence alignment. We were concerned however that the approach would be sensitive to the length of the sequence. To test the ability of a MWM to suggest a justifiable model given a sequence significantly longer than tRNA, we examined the performance of the method on the entire 16S-rRNA. To overcome memory limitations we removed those edges with zero weight. The matching suggested the vast majority of the interactions involved in forming the helices of the currently accepted secondary structure (Gutell 1994). In some helices, all base-pairs were not suggested, but in the majority of helices all of the base-pairs indicated in the proposed secondary structure were suggested by the MWM. In addition, a number of additional long-range interactions were suggested (to be described elsewhere). Results from 16S-rRNA indicated that the method was adequately robust to build a reasonable model from a data set of over 3.4 million possible base-pairs. It should be noted that no base-pairing rules were imposed, rather base-pairs were generated solely on the basis of MI content.

To examine the ability of the MWM algorithm to make useful predictions regarding the structure of small RNAs and where a limited number of sequences were available for alignment, we analyzed a set of oligonucleotides isolated using SELEX (Tuerk & Gold 1990). The short RNAs were isolated for their shared ability to bind to HIV-1 reverse transcriptase (Tuerk, MacDougal, & Gold 1992). The sequences have been described as sharing a pseudoknot structure. The consensus pseudoknot is formed by helices at alignment positions 10-13:32-35 and 21-25:46-50. The length of the helices varies between individual sequences but minimally contains 11-13:32-34 and 23-25:46-48. Eighteen sequences were aligned manually and processed to obtain a MWM. The data suggested two base-pairs of one pseudoknot stem (12-13:32-33) and six base-pairs of the second stem (20-25:46-51). These base-pair suggestions were the only set of interactions described by the model that formed consecutive base-pairs. The MWM failed to identify the position pair at 11:34, because it is invariant and therefore contains no MI. Although the base-pairs of the pseudoknot were the only suggestions describing helices, a large number of single base-pairs were suggested. The large number of base-pairs predicted by the MWM model led us to examine methods for reducing the incentives for the MWM algorithm to suggest relatively unlikely interactions.

The MWM algorithm attempts to maximize the sum of edge weights. This characteristic results in the addition of as many edges with positive weight as possible. Under the circumstances of our application to RNA modeling, it is favorable to avoid the addition of edges with relatively low MI. To address this issue we investigated the effects of negative edge weights on the MWM models. Negative edge weights were introduced by the subtraction of a constant from the MI data. The constant used was based on the expected MI of an independently varying position (2). We reasoned that half of the positions which varied randomly would fall below zero upon subtraction of the expected MI value. MI scores for the HIV-1 reverse transcriptase binding RNAs were adjusted by the subtraction of the expected MI of a randomly varying position. Reapplying the MWM algorithm to the corrected data resulted in a model with fewer base-pair predictions. The new model suggested 22 interactions versus 27 interactions suggested by the original MWM model. The effects on the pseudoknot prediction were minimal, indeed the only change was the elimination of a single base-pair at positions 20:51. The model derived from the modified data set predicted the pseudoknot while reducing the number of additional base-pair predictions. These findings suggest that the introduction of negative edge weights, by the subtraction of some reasonable constant, can have useful effects on the model proposed by the MWM. Of course the selection of the constant's value is somewhat arbitrary and only experimentation with different values and their effects on the model will reveal the true usefulness of such an approach in general.

Concluding Remarks

The computation of an RNA model from MI data using the MWM algorithm is an efficient approach to comparative analysis that builds models in $O(N^3)$ time and space. The technique generates accurate suggestions for tRNA and rapidly finds all major helices of the 16S-rRNA model. The approach should be of substantial value for the analysis of both naturally occurring structural RNAs and families of synthetic RNAs artificially selected for shared characteristics.

Modifications to the approach could allow the detection of interactions disallowed by the constraints imposed by pairwise matchings. In the MWM tRNA model discussed above, a source of erroneous base-pair suggestions appeared to be the failure of the MWM to recognize non-pairwise interactions in the form of base-triples. By altering the matching algorithm to allow two edges adjacent to each vertex in the matching, base-triples would no longer be disallowed. So called 2-matchings could provide interesting suggestions for sites likely to participate in base-triples. We have developed a b-matching program and preliminary results indicate that 2-matchings can identify some base-triples as well as the secondary and tertiary base-pairs suggested by the MWM approach described above.

In some cases, a large number of interactions are proposed by the MWM algorithm. A statistical approach could be used to reduce the set of interactions to those most likely to be of interest. MI is a log-likelihood ratio normalized by the number of sequences in the alignment. This fact can be exploited to obtain a statistical measure of the significance two positions do not vary independently. Multiplication of the log-likelihood ratio by two conforms to a χ^2 distribution from which significance may be assessed (see Gutell *et al.* 1992). Eliminating interactions falling below a user specified significance threshold would reduce the number of spurious suggestions made by the MWM algorithm.

One of the major short comings of the comparative approach is the failure to provide information regarding invariant positions. This is a true disadvantage for the analysis of highly conserved regions in molecules otherwise well suited for comparative analysis. Thermodynamic RNA modeling techniques, on the other hand, are unaffected by conserved regions. A significant difficulty with thermodynamic modeling, however, is the vast number of alternative structures that exist within 5-10% of the calculated free-energy minimum (Zuker 1989). Biologically relevant structures need not exist at the free-energy minimum but rather maybe present in the large number of near optimal structures. The weaknesses of comparative analysis and thermodynamic modeling are largely complementary. A method that efficiently integrates comparative and thermodynamic modeling techniques should result in a more robust approach capable of modeling structures problematic for either approach alone. The general nature of the MWM algorithm may present an opportunity for the integration of the two disparate modeling approaches. Using edge weights containing comparative and thermodynamic information, the MWM algorithm could be employed to find the optimum structure based on the two forms of data. The use of comparative data should reduce the number of sites where thermodynamic information is employed to conserved regions. This would impose constraints determined by the more biologically relevant comparative information.

The MWM modeling method is a rapid and efficient approach to building comparative RNA models. It should be of use for both naturally occurring families of RNAs with conserved functions as well as for sets of synthetically generated RNAs isolated by *in vitro* selection procedures. Unlike methods based on dynamic programming and stochastic context-free grammars, the MWM method does not disallow pseudoknots, a class of structures recognized as playing a role in defining important RNA structures (Gold *et al.* 1993). Extensions of the MWM approach to address current limitations of comparative analysis may provide more robust approaches to RNA modeling.

Acknowledgments

We would like to thank Sean Eddy for suggesting this idea, and Hal Gabow for very helpful discussions about the approach and for advice and pointers to the liturature. John Kececioglu also provided useful suggestions and provided us with the wmatch program, written by Ed Rothberg. This work was supported by DOE grant ER61606.

References

Altman, R. B. 1993. Probabalistic structure calculations: A three-dimensional tRNA structure from sequence correlation data. In *Proceedings of the First International Conference on Intelligent Systems in Molecular Biology*, 12–20. Menlo Park, CA: AAAI Press.

Chiu, D. K. Y., and Kolodziejczak, T. 1991. Inferring consensus structure from nucleic acid sequences. *Computer Appl. Biol. Sci.* 7:347–352.

Eddy, S., and Durbin, R. 1994. RNA sequence analysis using covariance models. *Nucl. Acids Res.* 22:2079–2088.

Fox, G. E., and Woese, C. R. 1975. 5S secondary structure. *Nature* 256:505–507.

Gabow, H. 1973. *Implementation of algorithms for maximum matching on non-bipartite graphs.* Ph.D. Dissertation, Stanford University Dept. Computer Science.

Gabow, H. N. 1976. An efficient implementation of Edmonds' algorithm for maximum matching on graphs. *J. ACM* 23:221–234.

Gold, L.; Tuerk, C.; Allen, P.; Binkley, J.; Brown, D.; Green, L.; MacDougal, S.; Schneider, D.; Tasset, D.; and Eddy, S. R. 1993. RNA: The shape of things to come. In Gesteland, R. F., and Atkins, J. F., eds., *The RNA World.* Plainview, NY: Cold Spring Harbor Laboratory Press.

Gutell, R. R.; Power, A.; Hertz, G. Z.; Putz, E. J.; and Stormo, G. D. 1992. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucl. Acids Res.* 20:5785–5795.

Gutell, R. R. 1994. Collection of small subunit (16S- and 16S-like) ribosomal RNA structures: 1994. *Nucl. Acids Res.* 22:3502–3507.

Holbrook, S. R.; Sussman, J. L.; Warrant, R. W.; and Kim, S. 1978. Crystal structure of yeast phenylalanine transfer RNA. ii. Structural features and functional implications. *J. Mol. Biol.* 123:631–660.

Holley, R. W.; Apgar, J.; Everett, G. A.; Madison, J. T.; Marquisee, M.; Merrill, S. H.; Penswick, J. R.; and Zamir, A. 1965. Structure of a ribonucleic acid. *Science* 147:1462–1465.

Klinger, T. M., and Brutlag, D. L. 1993. Detection of correlations in tRNA sequences with structural implications. In *Proceedings of the First International Conference on Intelligent Systems in Molecular Biology*, 225–233. Menlo Park, CA: AAAI Press.

Maidak, B. L.; Larsen, N.; McCaughey, M. J.; Overbeek, R.; Olsen, G. J.; Fogel, K.; Blandy, J.; and Woes, C. R. 1994. The Ribosomal Database Project. *Nucl. Acids Res.* 22:3485–3487.

Noller, H. F. 1984. Structure of ribosomal RNA. *Ann. Rev. Biochem.* 53:119–162.

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

Sakakibara, Y.; Brown, M.; Hughey, R.; Mian, I. S.; Sjolander, K.; Underwood, R.; and Haussler, D. 1994. Stochastic context-free grammers for modeling RNA. In *Proceedings of the Hawaii International Conference on Systems Sciences: Biotechnology Computing, Vol. V*, 284–293. Los Alamitos, CA: IEEE Computer Society Press.

Saks, M. E.; Sampson, J. R.; and Abelson, J. N. 1994. The transfer RNA identity problem: A search for rules. *Science* 263:191–197.

Searls, D. B. 1993. The computational linguistics of biological sequences. In *Artificial Intelligence and Molecular Biology.* AAAI Press. 47–120.

Steinberg, S.; Misch, A.; and Sprinzl, M. 1993. Compilation of tRNA sequences and sequences of tRNA genes. *Nucl. Acids Res.* 21:3011–3015.

Sussman, J. L., and Kim, S. 1976. Three-dimensional structure of a transfer RNA in two crystal forms. *Science* 192:853–858.

Tuerk, C., and Gold, L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249:505–510.

Tuerk, C.; MacDougal, S.; and Gold, L. 1992. RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proc. Natl. Acad. Sci. USA* 89:6988–6992.

Yarus, M. 1982. Translational efficiency of transfer RNA's: Uses of an extended anticodon. *Science* 218:646–652.

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