Nucleic Acids Research, Vol. 20, No. 21 5785–5795

Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods

R.R.Gutell, A.Power, G.Z.Hertz, E.J.Putz and G.D.Stormo
MCD Biology, Campus Box 347, University of Colorado, Boulder, CO 80309, USA

Received June 25, 1992; Revised and Accepted September 17, 1992

ABSTRACT

Comparative sequence analysis addresses the problem of RNA folding and RNA structural diversity, and is responsible for determining the folding of many RNA molecules, including 5S, 16S, and 23S rRNAs, tRNA, RNAse P RNA, and Group I and II introns. Initially this method was utilized to fold these sequences into their secondary structures. More recently, this method has revealed numerous tertiary correlations, elucidating novel RNA structural motifs, several of which have been experimentally tested and verified, substantiating the general application of this approach. As successful as the comparative methods have been in elucidating higher-order structure, it is clear that additional structure constraints remain to be found. Deciphering such constraints requires more sensitive and rigorous protocols, in addition to RNA sequence datasets that contain additional phylogenetic diversity and an overall increase in the number of sequences. Various RNA databases, including the tRNA and rRNA sequence datasets, continue to grow in number as well as diversity. Described herein is the development of more rigorous comparative analysis protocols. Our initial development and applications on different RNA datasets have been very encouraging. Such analyses on tRNA, 16S and 23S rRNA are substantiating previously proposed associations and are now beginning to reveal additional constraints on these molecules. A subset of these involve several positions that correlate simulataneously with one another, implying units larger than a basepair can be under a phylogenetic constraint.

INTRODUCTION

Within the past few years, our perception of RNA has undergone a significant and rewarding change. Functionally, this molecule was perceived in a subordinate position, playing a secondary role to proteins and DNA. It is now appreciated that RNA can act on other macromolecules in a variety of interesting ways. A short list of such functions includes RNA cleavage/ligation (Group I introns (1) and RNase P (2)) and direct involvement of ribosomal RNA in protein synthesis (3). Underlying this recent appreciation of the functional aspects of RNA is a revitalization in the study of its structure; RNA structure is far more than a simple agglomeration of standard helical elements. Within the past few years we have witnessed the emergence of several new structural elements, including pseudoknots, non-canonical pairings, and tetra-loops. These recently discovered structural elements were elucidated by comparative sequence methods and substantiated by experimental methods (experimental work reviewed in (4)). The paths to understanding and deciphering additional structural elements can take us in different directions. While experimental paths have been effective, our current experimental methods preclude us from exploring in detail all of the diverse RNA structures that are theoretically possible. An alternative method, comparative sequence analysis, can infer structure possibilities from the sequence constraints imposed on a population of functionally and structurally homologous molecules.

Comparative sequence analysis is based on the biological paradigm that macromolecules are the product of their evolution. The process of mutation and selection explores the possible, and reveals the acceptable. We infer that functionally equivalent RNA molecules (e.g. tRNA) are structurally equivalent as well. Secondarily, we deduce that similar or homologous, higher-order structure can be derived from different primary structures. For our purposes here, the diversity in RNA primary structure is bounded; those RNA structural elements that are biologically meaningful are selected for and identified with this method. The experiments have been done for us; we are (simply) observing those products that have survived the evolutionary process. The comparative sequence method was first applied to tRNA (5–8). The resulting cloverleaf secondary structure was the only such structure in common with all of the known tRNA sequences. As the tRNA dataset grew larger, comparative methods were called upon again to infer a few tertiary interactions (9). All of the comparatively derived secondary structure pairings, and a few of the proposed tertiary interactions were subsequently verified when the yeast Phe-tRNA crystal structure was solved in high resolution (10, 11). Comparative methods have been used to infer secondary structure in other RNAs, including 5S (12), 16S (13–15), and 23S (16–18) ribosomal RNA. More recently, this method has been applied to other RNAs including Group I (19) and group II (20) introns, RNase P RNA (21), U RNAs (22), and 7S RNA (23).

The comparative sequence methods themselves are evolving, in parallel with, and in part due to the significant growth in size

and diversity in comparative sequence datasets, increases in our computing power, and refinements in the intrepretation of comparative results. Initially our approaches were very basic compared to current methods. Compensatory base changes were searched for by eye; those found within a potential helical element (A-U, G-C, and G-U pairings arranged in a contiguous, antiparallel arrangement) were scored positively. Those helices with two or more compensatory base changes at two or more helical positions were considered phylogenetically proven (24). Once the number of 16S sequences surpassed 30, which included representative samplings from the three domains [*Bacteria, Archaea*, and *Eucarya*] (25) and the two organelles [Chloroplasts and Mitochondria], we began a more systematic search for positional covariances (26). The nucleotide pattern at each column in the alignment was transformed into a simple number pattern reflecting the changes (or lack thereof) occuring at each position. These number patterns were then sorted or grouped for similarity. While this method itself was not exhausitive, it allowed us to identify more secondary structure basepairings, and adjust a few previously proposed pairings. Maybe the most important result from this method was the identification of the first set of tertiary base-base correlations in 16S rRNA (26, 27). The number of 16S and 23S rRNA sequences have continued to increase, by 1989 there existed several hundred 16S-like sequences, and nearly 100 23S-like sequences. Covariance analysis of this larger and phylogenetically diverse collection of 16S and 23S rRNA sequences yielded additional refinements in the proposed secondary structures while several new tertiary-like interactions were proposed (28–33). At this time we are confident in the majority of the proposed *Escherichia coli* 16S and 23S rRNA secondary structure basepairings. While we will continue to evaluate these secondary structures from a comparative perspective, our future efforts will focus primarily on the elucidation of other higher-order structure constraints.

Within the past few years, numerous correlations beyond the simple secondary structure base pairings have been deciphered in the 16S and 23S rRNAs. This list includes multiple examples of: non-canonical base pairings, lone canonical base pairings, pseudoknot-like structural elements, several base pairings that together form a parallel structure, and comparative evidence for helix–helix coaxial stacking (reviewed in (33)). While these 16S and 23S rRNA comparative structure results are profound and consistent with experimental studies, it is our belief that additional RNA structure remains to be identified. With an ever increasing rRNA sequence database and with the hope that more sophisticated correlation methods could reveal additional structural constraints, we have initiated an effort to develop improved methods and apply them to various RNA datasets. This communication describes quantitative comparative sequence methods and their application to tRNA and rRNA datasets. A subset of the initial results are presented by way of example to highlight the general potential of these newer protocols.

MATERIALS AND METHODS

Database and alignment

Input data for the correlation analysis is a set of aligned sequences. For the analysis described herein, we focus on three primary alignment sets, one each for tRNA, 16S rRNA, and 23S rRNA.

Comparative structure analysis requires an alignment of those sequences that make up the collection. The better the alignment, the more meaningful the information that can be discerned. Initially sequences are aligned for maximum primary structure homology. As secondary structure elements are identified and phylogenetically proven, these features, in addition to primary structure conservation, serve to constrain the juxtaposition of sequences. This process proceeds in an iterative fashion as more constraints define the character of an alignment, serving to resolve alignment uncertainities. Additional phylogenically distant sequences establish the limits of variability, discerning the allowable (and observable) states.

The tRNA aligned dataset is a modified version of the publically available tRNA sequence database prepared by Sprinzl *et al.* (34). [Modified in that we had to reorganize its format to be compatible with our alignment editor (AE2, developed by Tom Macke) and our correlation analysis tools. Small interpretative changes in the alignment of nucleotides were made as well]. This complete set contains 1710 aligned sequences, spanning the three primary lines of descent, the two organelles, and a few viral sequences. For the analysis presented here, we only include those tRNAs that are most regular, excluding the Mitochondrial and class-2 sequences. The Mitochondrial sequences contain an inordinant amount of structural variation; class-2 tRNAs contain an insertion of 10 or more nucleotides in the variable arm which appears to alter some of the known class-1 tertiary interactions (35). This smaller tRNA dataset contains 896 sequences.

Analyses on the 16S and 23S rRNAs were performed on aligned datasets collected from two sources. For the past 10 years one of us (RRG) has been collecting and maintaining 16S and 23S rRNA sequence alignments [R.R.G.-private collection, (36, 37)]. The phylogenetic diversity among these sequences is quite broad, with key representative sequences from each of the three primary lines of descent, and the two organelles. A larger collection of bacterial (Bacteria and Archaea) 16S rRNA aligned sequences (Ribosomal RNA Database project, (38, 39)) complemented this initial database. The total 16S rRNA collection contains 800 complete (or nearly so) sequences, while the 23S rRNA collection contains 150 sequences.

Calculation of mutual information and related measures

We base our identification of covariant positions on the mutual information observed between them. This measure has previously been described by Chiu and Kolodziejczak (40), but we repeat the definition here because we use somewhat different notation which is more convenient for the extensions described below.

Given an alignment of multiple sequences, we wish to determine the degree of covariation between two positions x and y. We begin by determining the frequency at which each base occurs at each position, f_{bx} and f_{by}. (Note that usually $b \in (A, C, G, T)$ but may also be a gap introduced for the alignment or an ambiguous base.) We also determine the frequencies of the pairs of bases occuring at positions x and y in the same sequence, f_{bxby}. If the two positions vary independently of one another then $f_{bxby} \approx f_{bx} \times f_{by}$. We are interested in measuring the divergence from independence. The Mutual Information in the positions x and y is defined as:

$$M(x,y) = \sum_{b_x b_y} fb_x b_y \ln \frac{fb_x b_y}{fb_x fb_y} \quad (1)$$

When M(x,y) is multiplied by the number of sequences in the alignment, a log-likelihood ratio of the form $\Sigma_i O_i \ln (O_i/E_i)$ *is obtained, where* O_i and E_i are some observed and expected values, respectively. Two times this log-likelihood ratio conforms

to a χ^2 distribution from which statistical significance can be easily calculated (41). It is worth noting that $M(x,y) \geq 0$, with the equality holding only in the case of the frequency of position pairs being exactly predicted by the frequencies of the independent positions. $M(x,y)$ is maximized when both positions are highly variable and also completely correlated. The correlations may be of any type, not limited to the known canonical pairings (i.e. A-U, G-C), although those are the most often observed and tend to be the strongest correlations. Two of the advantages of this method over some of the previous ones is that any types of correlations can be found and that correlations which are quantitatively low, but still significant, can also be found.

We actually calculate $M(x,y)$ using the following formula which is more efficient and has other useful interpretations:

$$M(x,y) = H(x) + H(y) - H(x,y) \quad (2)$$

where H is an entropy term (i.e., $H = -\Sigma_b f_b \ln f_b$). We calculate $H(x)$ for all positions and store it in an array. We then calculate $H(x,y)$ for all pairs of positions (subject to $x < y$, since it is a symmetric measure). $M(x,y)$ is constrained by the relationship:

$$M(x,y) \leq \min[H(x),H(y)] \quad (3)$$

which means that $M(x,y)$ is bounded above by the variability of the least variable position. If either position x or y is non-random, $M(x,y)$ is less than its maximum possible value even for completely correlated positions. In the extreme of x or y being invariant, $M(x,y) = 0$ regardless of their interactions and we can learn nothing about them from comparative methods alone. We would like to be able to identify positions that may not be highly correlated as measured by $M(x,y)$, but are as correlated as they can be given the limited variability of the individual positions. That is, there may be positions that are constrained for reasons other than their interaction with another position, but we would still like to find that interaction if it exists. For this reason we calculate two other numbers:

$$R_1(x,y) = \frac{M(x,y)}{H(x)} \quad (4)$$

and

$$R_2(x,y) = \frac{M(x,y)}{H(y)} \quad (5)$$

Both R values are in the range 0 to 1 and, in general, $R_1(x,y) \neq R_2(x,y)$. (Note that $R_1(x,y) = R_2(y,x)$, which allows us to determine all values of R_1 and R_2 while maintaining the constraint $x < y$ during the calculation). Thus we have transformed the symmetric $M(x,y)$ into asymmetric values for the purpose of finding subtle correlations that may otherwise be missed. This will let us see correlations between positions that are nearly invariant, but whenever they do change they do so in a coordinated way. In addition, by splitting the effects of position x and y, we can see things that are only partially correlated and subtle. As a simple example, imagine that position x had the frequencies $f_b = 0.25$ for all b, whereas at postion y the frequencies were $f_A = 0.75$, $f_G = 0.25$. Then $H(x) = 1.39$ and $H(y) = 0.56$, which also means $M(x,y) \leq 0.56$. Now suppose that when position y is a G, position x is also a G, and whenever y is an A position x is not a G. Then $R_1(x,y) = 0.4$, but $R_2(x,y) = 1.0$ because $M(x,y)$ is as large as it can be, given $H(y)$. Another way to think about it is that knowing the base at position x provides absolute knowledge about the base at position y, but the reverse is not true. Whenever $M(x,y)$ is large, both R values will be also, and in general $M(x,y)$ values should be examined first because they are more reliable. However, some correlated positions will be missed if only $M(x,y)$ values are used and some of those can be identified using the R values. But we do not have a good way to estimate the significance of the R values, and they can produce false-positives. That is, in addition to increasing the 'signal' of some true interactions, they also increase the 'noise' associated with coincidental correlations. Therefore, R values are best used to identify potentially interesting correlations that are missed by $M(x,y)$ values and are worthy of further examination.

As stated above, a probability can be calculated based on $M(x,y)$, the sample size and the degrees of freedom (the product of [the number of observed characters − 1] at each position). In the current calculation we are not accounting for the number of mutational events required for the data given a phylogenetic tree, but rather we consider each observed state to be independent. These probability values should, therefore, be taken as lower bounds. That is, the true probability, given the phylogenetic tree, of observing the $M(x,y)$ values by chance is usually higher, and therefore less significant, than our calculations indicate. While the probability values are clearly important because they provide rough measures of the significance of any correlation found, they are computationally much more intensive than the other numbers and we only calculate them for position pairs that appear interesting by other criteria. We have written two related programs that perform the calculations. MIXY (for Mutual Information of X and Y) goes through all combinations of positions (subject to $x < y$) and returns on each line of the output file:

$$x\ y\ M(x,y)\ R_1(x,y)\ R_2(x,y)$$

Of course, most of the combinations are uninteresting, so rather than save the entire output to a file (which would have 2850 lines even for tRNA sequences of 76 bases in length) we usually run it through a filter that only saves values above some user-specified thresholds for the M and/or R values. Another filter we often used is called *Nbest* which saves only the N (user-specified, typically about 4 or 5) highest correlations for each position. These may be based on either the M or R values. Note that sorting by R_1 can give very different results from sorting by R_2. For some position x, sorting of $R_1(x,y)$ will rank the positions y according to how well they predict the base at x. Sorting of $R_2(x,y)$ will rank the positions y by how well they are predicted given the base at position x.

The related program, *IMIXY* (for *I*nteractive *MIXY*) displays the $M(x,y)$, $R_1(x,y)$, $R_2(x,y)$ entropy and probability values, in addition to the actual frequencies of the pairs for any two positions in an alignment. Once a pair of potentially interesting positions have been identified with *MIXY*, *IMIXY* can be used interactively, with the user requesting the two positions of interest, or the filtered output from *MIXY* can be used as input.

The graphics displayed herein were generated with one of two programs. The RNA structure diagrams (Figures 4−8) were composed with *XRNA*, an interactive 2-D and 3-D RNA structure drawing and viewing program developed by Bryn Weiser [University of California, Santa Cruz, under development and currently unpublished]. The contour and 3D surface plots (Figure 2) were drawn with the IDL graphics package [IDL: Interactive Data Language, version 2.0. Research Systems, Inc., Boulder, CO 80303.]. The alignments were created and manipulated with

A

```
nucs per seq: 76
num of seqs: 896
    x     y     M(xy)    R1(xy)   R2(xy)
-----------------------------------------
    53    61    0.085    0.953    0.908
```

B

```
number of sequences: 896
53,61
           53     A      C      G      U      -      N   H(x)
              0.018  0.000  0.982  0.000  0.000  0.000  0.090
   61        ----------------------------------------------
    A  0.000 | 0.000  0.000  0.000  0.000  0.000  0.000
    C  0.981 | 0.000  0.000  0.981  0.000  0.000  0.000
    G  0.000 | 0.000  0.000  0.000  0.000  0.000  0.000
    U  0.019 | 0.018  0.000  0.001  0.000  0.000  0.000
    -  0.000 | 0.000  0.000  0.000  0.000  0.000  0.000
    N  0.000 | 0.000  0.000  0.000  0.000  0.000  0.000
 H(y)  0.094

   G*C(98.1), A*U(1.8), G*U(0.1)

   H(xy)   M(xy)   R1(xy)  R2(xy)   -lnP
   0.098   0.085   0.953   0.908   79.206
```

Figure 1. Mutual Information analysis of positions 53 and 61 of tRNA (*S.cerevisiae* phenyalanine numbering). The tRNA dataset contains 896 sequences. **A**: The output from *MIXY*. **B**: The output from *IMIXY*. The table in the center shows the frequency of each base at each position and their entropies (H), as well as the frequency of each combination of bases. The line below the table lists, in order of frequency, all of the base combinations that exist. The final line contains the $H(x,y)$, $M((x,y)$, $R_1(x,y)$, and $R_2(x,y)$ values and $-ln\ P$, the negative logarithm of the probability of observing the $M(x,y)$ value by chance, given the assumptions discussed in the text. This probability calculation includes the sample size and the degrees of freedom, but does not take into account the phylogenetic relationship between the sequences and, therefore, overestimates the significance to some extent.

the program *AE2* (Alignment editor, version 2), developed by Tom Macke (Scripps Clinic, CA) [under development and currently unpublished].

RESULTS AND DISCUSSION

Several points are implicit in our application and interpretation of these correlation analysis methods:

- The nucleotide is the basic unit. Each of the 4 nucleotides are treated as separate entities.
- Correlating positions are identified regardless of the flanking positions' patterns of change, or known structural context.
- All 16 possible base-base covariations between two positions are noted. There is no inherent bias towards A-U and G-C pairings.
- This analysis reveals positional correlations; we subsequently infer base-base interactions in the context of higher-order structure.
- Comparative analysis *per se* does not address, nor does it imply that all proposed interactions occur simultaneously.
- Positional covariations involving a larger number of phylogenetic compensatory base substitutions (i.e. events) are considered more significant than those with a smaller number of such coordinated changes. In this communication, we do not calculate the actual number of coordinated changes that have occurred during evolution.

To establish the validity of this approach, we initially apply these algorithms to tRNA, a molecule for which a detailed three-dimensional structure is known, and for which a large and encompassing collection of aligned and biologically diverse sequences currently exists. Our initial goal is to determine how many of the tRNA secondary and tertiary interactions can be identified with these methods. Next we address larger RNA molecules for which a large comparative collection of sequences is known. Equally important we want to analyze molecules that have been studied extensively in the past with comparative and experimental methods, so we can compare and contrast the results from our newer methods. For this, we choose the 16S and 23S rRNA and provide several examples that demonstrate the utility of the method.

Correlation analysis of tRNA

Three values, $M(x,y)$, $R_1(x,y)$, and $R_2(x,y)$ are calculated for every pair of positions in the alignment set. For a molecule the length of tRNA (76 nucleotides: Yeast-Phe numbering) there are 2850 pairs. This number increases as the square of the sequence length, so that for 16S rRNA, which contains 1542 nucleotides, there are 1,188,111 pairs. Examples of the output from the *MIXY* and *IMIXY* programs are displayed in Figure 1. Each x and y comparison is output on a single line as shown in Figure 1A for the closing base pair of the tRNA TψC stem (positions 53 and 61). The frequencies of all pairing types for positions 53 and 61 are shown in Figure 1B, as generated by the program *IMIXY*. In this example, the base pair positions 53 and 61 are nearly invariant, but the few base substitutions that do occur at these positions are coordinated, maintaining canonical pairings. The $M(x,y)$ value is low, since the degree of variation is small. However, the R_1 and R_2 values are close to their maximum

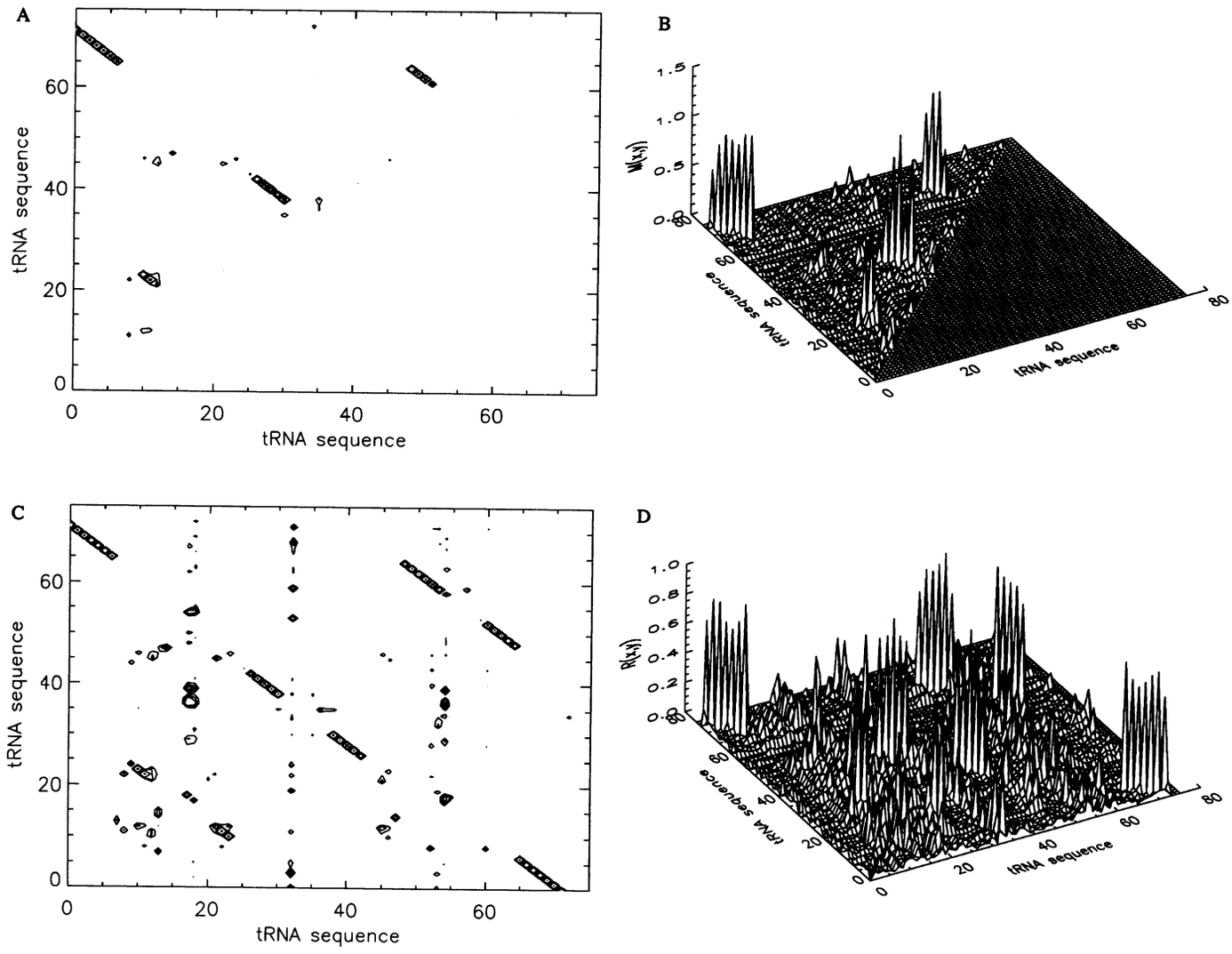

Figure 2. Graphical display of $M(x,y)$ and R values. Only values above 0.2 are displayed on the Contour plots. **A**: Contour plot of $M(x,y)$ values. **B**: Surface plot of $M(x,y)$ values. **C**: Contour plot of R values. The values are determined by taking the values from $M(x,y)$ (shown in part A) and replicating them symmetrically into the other half of the matrix, and then dividing each row by the entropy of the position on the vertical axis. As described in the text, $R_1(x,y) = R_2(y,x)$, so this plot shows both values. If the vertical axis is considered to be position x, then the plot is of $R_1(x,y)$; if the vertical axis is considered to be position y then the plot is of $R_2(x,y)$. Sorting by $R_1(x,y)$ is equivalent to sorting within rows and sorting by $R_2(x,y)$ is equivalent to sorting within columns. **D**: Surface plot of R values. The values are the same as in part C, but displayed as a 3-D plot.

value of 1.000 because nearly of all changes at one position are compensated for by a change at the other position.

Within the large number of calculated $M(x,y)$, $R_1(x,y)$, and $R_2(x,y)$ values from the *MIXY* program, a small percentage are expected to be biologically as well as statistically significant. Various methods are being explored to determine which values are most significant. In this communication, we address two such methods: 1) contour and surface plots, and 2) a simple sorting of the highest correlations for each position in an alignment. While high correlation values are significant, lower values, when viewed in the proper context can be significant as well.

The $M(x,y)$ and R values from the tRNA dataset are plotted on contour and surface plots in Figure 2. This graphical perspective emphasizes the relationships of the high values which are the peaks over the X-Y plane. tRNA $M(x,y)$ values are displayed in Figures 2A and 2B, while the R values are displayed in Figures 2C and 2D. The most prominent feature are the four clusters of highly correlating positions, which represent the four tRNA helices. Many of the lower peaks, in Figures 2A and 2B, that are not associated with one of the four tRNA stems do associate with known tertiary interactions. However, a few of these lower correlating values are not associated with any of the known base−base interactions. The R plots (2C, 2D) contain many highly correlating positions not found in the $M(x,y)$ plot, including several tertiary interactions missed in the $M(x,y)$ plot. For example the tertiary pairs 18/55 and 19/56 appear in the R plots but are missed in the $M(x,y)$ plots due to the high conservation found at these positions.

An alternative method of identifying significant correlating pairs ranks the N highest information values (either M or R values) for each individual position, using the program *Nbest*. Figure 3 displays a slightly abbreviated *Nbest* output from our tRNA

X	Y	M(xy)	R1(xy)	X	Y	M(xy)	R1(xy)	X	Y	M(xy)	R1(xy)	X	Y	M(xy)	R1(xy)	X	Y	M(xy)	R1(xy)
1	72	0.65	0.75 S	16	71	0.06	0.06	31	39	1.10	0.84 S	46	13	0.31	0.39	61	53	0.09	0.91 S
1	35	0.15	0.18	16	2	0.05	0.06	31	36	0.30	0.23	46	22	0.28	0.35 T	61	9	0.03	0.32
1	39	0.15	0.17	16	35	0.05	0.06	31	72	0.17	0.13	46	47	0.23	0.28	61	44	0.02	0.21
2	71	0.91	0.90 S	17	13	0.13	0.12	32	38	0.13	0.19	47	24	0.25	0.27	62	52	0.47	0.85 S
2	35	0.10	0.10	17	47	0.12	0.11	32	36	0.11	0.16	47	13	0.24	0.26	62	36	0.05	0.09
2	1	0.06	0.06	17	24	0.12	0.11	32	35	0.10	0.15	47	11	0.23	0.25	62	31	0.04	0.07
3	70	1.02	0.89 S	18	37	0.02	0.69	33	54	0.08	0.63	48	15	0.32	0.58 T	63	51	1.04	0.82 S
3	36	0.10	0.08	18	19	0.01	0.61	33	60	0.07	0.57	48	35	0.10	0.18	63	36	0.20	0.16
3	35	0.08	0.07	18	55	0.01	0.61 T	33	4	0.06	0.52	48	38	0.05	0.10	63	73	0.15	0.12
4	69	0.98	0.75 S	19	37	0.02	0.49	34	35	0.16	0.11	49	65	0.81	0.67 S	64	50	1.02	0.80 S
4	5	0.14	0.11	19	55	0.02	0.48	34	36	0.16	0.11	49	13	0.08	0.06	64	49	0.07	0.06
4	68	0.12	0.09	19	18	0.01	0.41	34	29	0.09	0.07	49	64	0.07	0.06	64	65	0.06	0.05
5	68	0.94	0.71 S	20	36	0.13	0.14	35	73	0.26	0.19	50	64	1.02	0.83 S	65	49	0.81	0.69 S
5	4	0.14	0.10	20	35	0.12	0.13	35	47	0.18	0.13	50	49	0.06	0.05	65	13	0.08	0.07
5	69	0.13	0.10	20	60	0.12	0.13	35	51	0.17	0.13	50	34	0.05	0.04	65	17	0.07	0.06
6	67	1.06	0.77 S	21	22	0.02	0.24	36	39	0.30	0.23	51	63	1.04	0.82 S	66	7	1.05	0.88 S
6	5	0.07	0.05	21	23	0.01	0.17	36	31	0.30	0.22	51	35	0.17	0.13	66	36	0.07	0.06
6	3	0.07	0.05	21	48	0.01	0.17	36	37	0.21	0.16	51	36	0.15	0.12	66	34	0.06	0.05
7	66	1.05	0.89 S	22	13	0.34	0.42 S	37	36	0.21	0.35	52	62	0.47	0.87 S	67	6	1.06	0.77 S
7	36	0.08	0.07	22	46	0.28	0.35 T	37	72	0.06	0.11	52	36	0.05	0.09	67	68	0.09	0.07
7	63	0.06	0.06	22	23	0.17	0.22	37	73	0.04	0.07	52	51	0.04	0.08	67	5	0.08	0.06
8	14	0.04	0.33 T	23	12	0.99	0.89 S	38	36	0.21	0.24	53	61	0.09	0.95 S	68	5	0.94	0.73 S
8	16	0.02	0.19	23	13	0.28	0.26	38	32	0.13	0.15	53	9	0.03	0.35	68	4	0.12	0.09
8	13	0.02	0.18	23	9	0.27	0.24 T	38	24	0.09	0.10	53	29	0.02	0.25	68	69	0.11	0.09
9	23	0.27	0.33 T	24	11	0.78	0.91 S	39	31	1.10	0.89 S	54	60	0.15	0.68	69	4	0.98	0.76 S
9	12	0.26	0.32	24	13	0.28	0.32	39	36	0.30	0.25	54	33	0.08	0.36	69	5	0.13	0.10
9	13	0.12	0.15	24	47	0.24	0.29	39	72	0.16	0.13	54	1	0.06	0.27	69	68	0.11	0.09
10	25	0.08	0.35 S	25	10	0.08	0.14 S	40	30	0.63	0.83 S	55	37	0.02	0.49	70	3	1.02	0.82 S
10	45	0.06	0.29 T	25	17	0.07	0.13	40	39	0.10	0.14	55	19	0.02	0.48	70	36	0.14	0.11
10	64	0.04	0.17	25	24	0.06	0.11	40	36	0.09	0.13	55	40	0.01	0.44	70	35	0.11	0.09
11	24	0.78	0.90 S	26	44	0.23	0.20 T	41	29	1.33	0.98 S	56	19	0.01	0.28 T	71	2	0.91	0.85 S
11	13	0.29	0.33	26	47	0.13	0.12	41	34	0.09	0.07	56	64	0.01	0.18	71	35	0.10	0.10
11	47	0.23	0.27	26	11	0.12	0.10	41	39	0.09	0.06	56	50	0.01	0.15	71	29	0.06	0.06
12	23	0.99	0.88 S	27	43	0.74	0.61 S	42	28	1.09	0.87 S	57	60	0.07	0.11	72	1	0.65	0.68 S
12	13	0.30	0.26	27	36	0.09	0.07	42	36	0.07	0.05	57	29	0.03	0.05	72	31	0.17	0.18
12	9	0.26	0.24	27	31	0.08	0.07	42	29	0.07	0.05	57	41	0.03	0.05	72	39	0.16	0.16
13	22	0.34	0.36 S	28	42	1.09	0.86 S	43	27	0.74	0.63 S	58	60	0.01	0.30	73	35	0.26	0.24
13	46	0.31	0.34	28	29	0.08	0.06	43	36	0.09	0.08	58	44	0.01	0.18	73	63	0.15	0.14
13	12	0.30	0.32	28	41	0.08	0.06	43	31	0.06	0.05	58	23	0.01	0.17	73	51	0.14	0.13
14	15	0.04	0.38	29	41	1.33	0.98 S	44	26	0.23	0.19 T	59	36	0.10	0.08	74	1	0.00	0.00
14	8	0.04	0.36 T	29	34	0.09	0.07	44	47	0.13	0.11	59	60	0.10	0.08	74	2	0.00	0.00
14	16	0.04	0.35	29	39	0.09	0.07	44	35	0.11	0.09	59	29	0.08	0.07	74	3	0.00	0.00
15	48	0.32	0.53 T	30	40	0.63	0.87 S	45	46	0.12	0.13	60	54	0.15	0.21	75	1	0.00	0.00
15	35	0.07	0.12	30	39	0.11	0.15	45	13	0.11	0.12	60	20	0.12	0.17	75	2	0.00	0.00
15	63	0.05	0.08	30	36	0.10	0.14	45	22	0.08	0.09	60	59	0.10	0.14	75	3	0.00	0.00

Figure 3. *Nbest* sorting of the $R_1(x,y)$ values for the tRNA dataset The three highest correlating $R_1(x,y)$ values for each position are displayed. The $M(x,y)$ values are shown for comparison. (Positions 74–76 are invariant and thus the $M(x,y)$ values are all zero. Position 76 is omitted to save space).

dataset. Only the three best $R_1(x,y)$ values for each tRNA position are shown. [The $M(x,y)$ values are shown as a comparison with the ranked $R_1(x,y)$ values. The $R_2(x,y)$ values along with the values for the conserved position 76 were deleted to save space]. All secondary structure base pairings are denoted with an *S*, while tertiary base–base interactions are denoted with a *T*. There are several interesting points to make from this analysis. First, all positions involved in secondary structure base pairings correlate best with their cognate pairing partner. This is true for both $M(x,y)$ and $R(x,y)$ values. Second, the highest $M(x,y)$ and $R(x,y)$ values for all base paired positions in the acceptor, anticodon, and TψC stems are significantly larger than the second highest value. However, for two of the base pairs in the D stem (10/25 and 13/22) the $R(x,y)$ values are much less than the values for other positions involved in base pairing. Furthermore, the second and third highest information values for positions in the D stem (10–13/22–25) are generally much greater than the second highest values in the other helical positions. Additional correlations involving the D stem are discussed again below. Third, the majority of the tertiary interactions are identified in Figure 3. All but one of the known tertiary interactions appear in the list of the 3-best correlations in at least one orientation. The tertiary pair 54/58 is missed due to its high conservation, a limitation of any comparative method. These positions are nearly invariant and the few changes that occur at one position or the other are not necessarily compensated for by changes at the tertiary pairing partner. We note that there are several other pairs that appear interesting by the data in Figure 3, but it is also possible that they represent false-positives, as described earlier. We have only pointed out the examples where we know interactions occur and they can be noticed in the *R* values, even if the *M* values are low.

The results from the tRNA correlation analysis presented in Figure 3 are summarized on the secondary structure diagram in Figure 4. Blackened arrows represent interactions with the highest $R_1(x,y)$ ranking. Open arrows indicate interactions ranked within

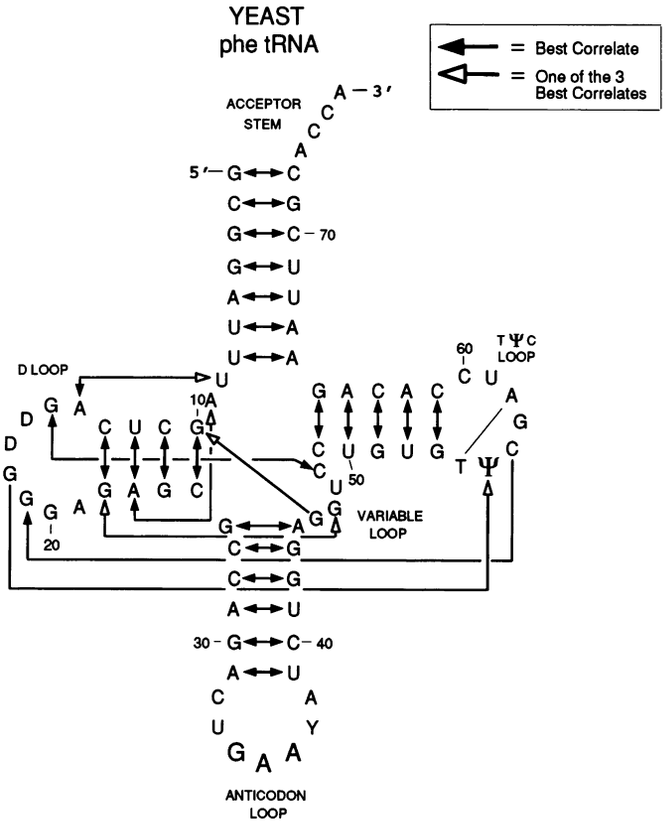

Figure 4. Summary of *Nbest* results from Figure 3 displayed on the tRNA cloverleaf (see text for explanation).

the 3-best. In addition to all of the secondary structure base pairings noted above, both positions involved in the two tertiary interactions, 15/48 and 26/44, correlate best with their base pairing partner. Except for 54/58, the remaining tertiary interactions are also identified by such analysis.

Given the success in identifying known tRNA base−base interactions, we also ask what significant correlations exist that do not associate with known tRNA structural constraints. Analysis of the $M(x,y)$ and R values in Figure 2 reveals several peaks that are above background and that do not map to existing base pairings. Analysis of the *Nbest* results in Figure 3 reveals some interesting high values in the anticodon loop region and the D stem. By way of example, we address these latter two regions, utilizing *Nbest* methods to establish additional correlations.

In the case of the anticodon loop, we focus on position 36, the 3′end of the anticodon. The ten positions with $M(x,y)$ values within 50% of the top value were ranked in Figure 5A. These positions are displayed on the secondary structure diagram in Figure 5B, with position 36, the focal position denoted with a triangle. The positions with the two highest $M(x,y)$ values are denoted with large closed circles, while the next 8 highest ranking positions are denoted with smaller closed circles. A three-dimensional perspective of this tRNA is shown in Figure 5C. Some of these correlations involving this focal position and positions in the anticodon loop and helix have been noted previously (42, 43). The other base pair identified (51−63) occurs on the same face as positions 36 and the 11/24 base pair in the 3D structure. In addition, it is intriguing to note that mutations at position 24 affect codon-anticodon specificity (44, 45).

A

X	Y	M(xy)	
36	39	0.305	-\ BP: 39/31
36	31	0.297	-/ BP: 31/39
36	37	0.214	
36	38	0.207	
36	63	0.203	---\ BP: 63/51
36	11	0.177	-\ | BP: 11/24
36	24	0.165	-/ | BP: 24/11
36	34	0.157	|
36	35	0.157	|
36	51	0.153	---/ BP: 51/63

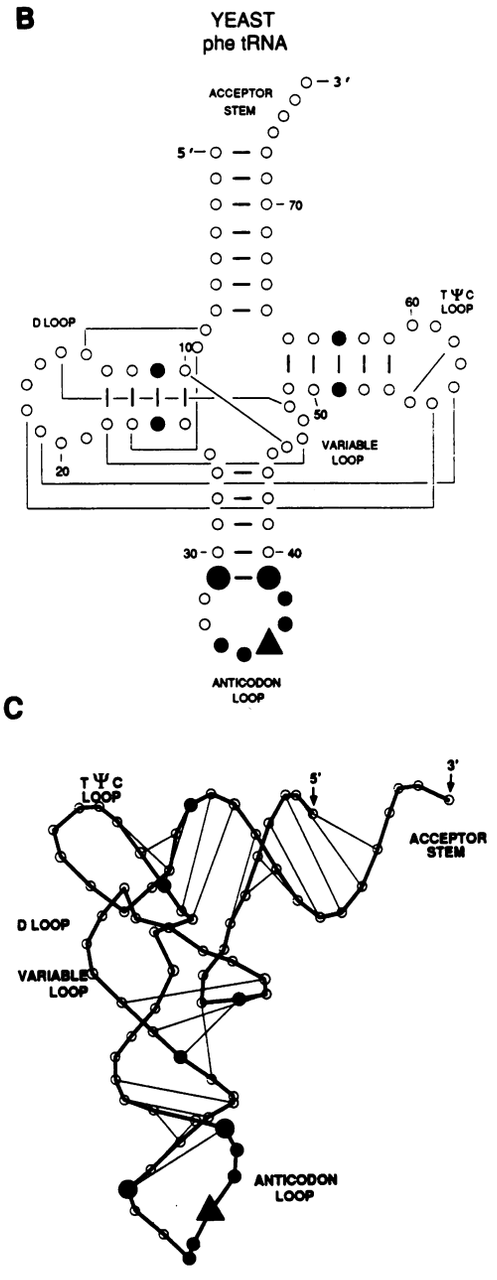

Figure 5. A: *Nbest* analysis of position 36 in tRNA. The ten highest $M(x,y)$ values are displayed. Positions which base-pair with each other are indicated at the right. **B**: Secondary Structure Diagram for tRNA. Position 36 is diagrammed with a triangle, positions 31 and 39 with large closed circles, and positions 37, 38, 63, 11, 24, 34, 35, and 51 are shown with smaller closed circles. **C**: Three-dimensional diagram of tRNA. The symbol mapping is the same as in part B.

Another possibility for these observed correlations, other than structural constraints, is some external factor causing these bases to vary coordinately. For example, proper interactions with aminoacyl tRNA synthetases requires that those enzymes be able to distinguish their correct set of tRNAs from others and this is accomplished, at least partially, through the use of specific sequences in various parts of the 0tRNA structure (46, 47). By dividing the tRNA database into iso-accepting classes and running MIXY on those separately we may be able to distinguish correlations that are for internal structural reasons from those that are for interactions with external factors.

Positions 13 and 22 form the closing base pair in the D stem; position 22 interacts with position 46, forming a base−triple. Establishing nucleotide 13 as our focal position, we observe 7 positions with $M(x,y)$ values within 50% of the top value. These are diagrammed in Figure 6. The top two correlating positions are denoted with large closed circles while the remaining 5 positions are shown as small closed circles. It is interesting to note that the top two correlating positions, 22 and 46, together with position 13, form the base pair triple interaction. The remaining five positions are all packed into a tight 3-D cluster, as diagrammed in Figure 6C.

Correlation analysis of rRNA

Application of these new methods on 5S, 16S, and 23S rRNA datasets are at an early stage. The vast majority of the interactions previously proposed for 16S and 23S rRNA (33) have also emerged from *Nbest* analysis. A few of those correlations that substantiate and complement previous findings are discussed here, to further illustrate some of the structural constraints that can be identified with these methods.

Previously, an intriguing structural constraint was identified in 16S rRNA involving hairpin loops comprised of 4 nucleotides, commonly known as the *tetra loop*. It was noticed that the majority of hairpin loops of size four (in 16S rRNA) usually fall into one of three classes, GNRA, UNCG, and CUUG (48). One loop in particular, at positions 83−86, varies quite extensively throughout the eubacterial domain, although in over 93% of these 343 sequences the loop is constrained to a GCAA, UUCG, or CUUG sequence. The closing base pair of the underlying helix is constrained as well: UUCG is always closed by a C-G pair, CUUG by a G-C pair, and GCAA is closed predominantly (but not always) with a A-U pair. *Nbest* analysis of the top five $M(x,y)$ values for positions 82−87 from the current eubacterial 16S dataset is shown in Figure 7A; a graphical representative of this loop is diagrammed in Figure 7B. In this latter figure, the position that correlates highest for each position from 82−87, is shown with a blackened arrow, while those positions with rankings from 2 to 5 are connected with open arrows. Several interesting points are worth noting. First, all of the top five correlating values for these six positions (30 in total) point to one of the other positions under consideration except three, and these all point to the underlying base pair 81−88. Second, for the four positions in the tetra loop, all of the information values for the five highest ranking positions are within 50% of the top value. Third, the two highest values associate positions 82 and 87, which are base paired with one another. Fourth, in this example as well as the previous one for tRNA, high correlation values do not necessarily imply a base pair, but suggest that certain structures are constrained by the context of their surrounding nucleotides and constitute a structural domain that varies coordinately.

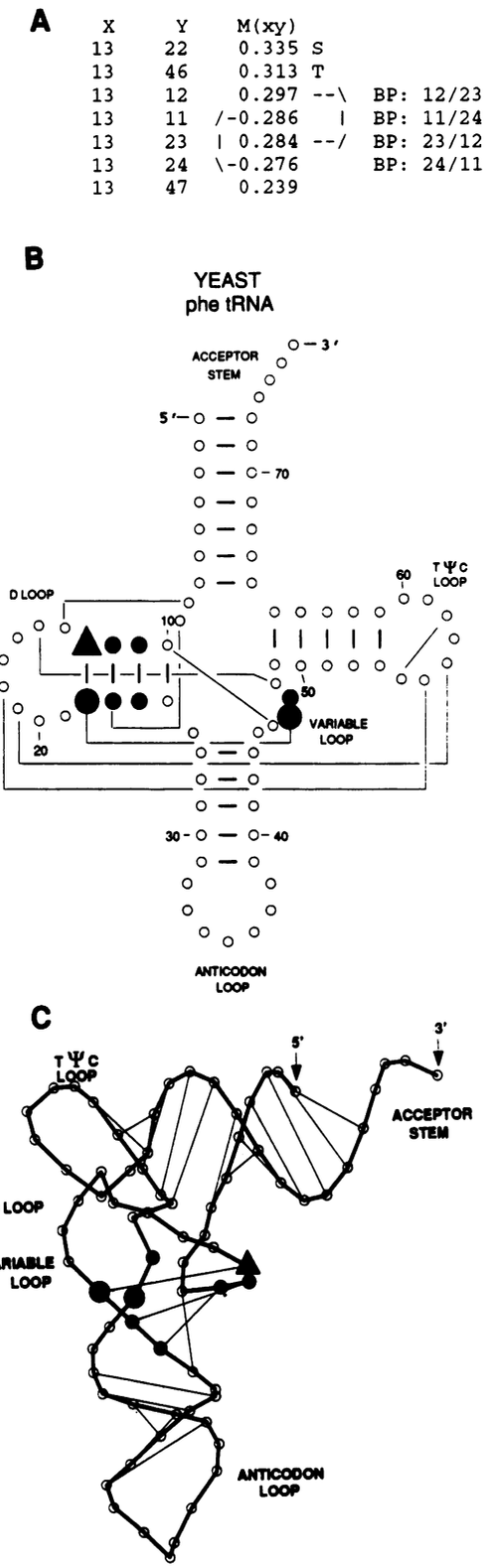

Figure 6. **A**: *Nbest* analysis of position 13 in tRNA. The seven highest $M(x,y)$ values are displayed. Positions that base-pair with each other are indicated at the right. **B**: Secondary Structure Diagram for tRNA. Position 13 is diagrammed with a triangle, positions 22 and 46 with large closed circles, positions 12, 11, 23, 24, and 47 are shown as smaller closed circles. **C**: Three-dimensional diagram of tRNA. The symbol mapping is the same as in part B.

A

X	Y	M(xy)	X	Y	M(xy)
82	87	1.062 S	85	83	0.777
82	83	0.728	85	87	0.663 -\ BP:87/82
82	85	0.642	85	82	0.642 -/ BP:82/87
82	81	0.361-\ BP:81/88	85	84	0.423
82	88	0.353-/ BP:88/81	85	86	0.354
83	85	0.777	86	83	0.396
83	82	0.728-\ BP:82/87	86	85	0.354
83	87	0.714-/ BP:87/82	86	87	0.344-\ BP:82/87
83	84	0.409	86	84	0.340 |
83	86	0.396	86	82	0.282-/ BP:87/82
84	85	0.423	87	82	1.062 S
84	83	0.409	87	83	0.714
84	86	0.340	87	85	0.663
84	82	0.286-\ BP:82/87	87	88	0.402 BP:88/81
84	87	0.267-/ BP:87/82	87	86	0.344

B

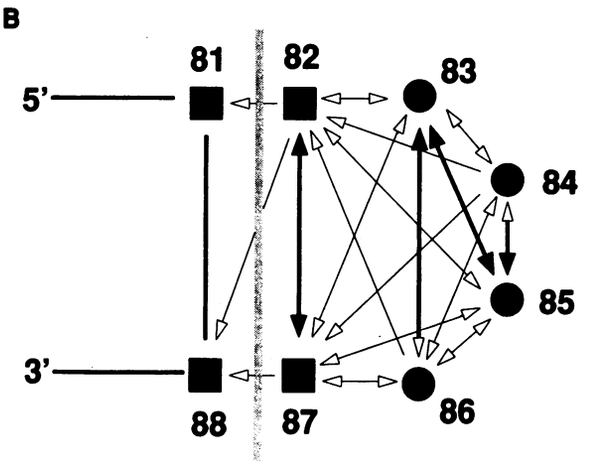

Figure 7. A: *Nbest* analysis of positions 82–87 in 16S rRNA (*Escherichia coli* numbering). The five highest *M(x,y)* values for each position are displayed. **B**: Secondary Structure diagram, with arrows connecting the highest correlating pairs. Blackened arrows denote the highest correlating value for each position. Open arrows denote other correlating pairs. Closed boxes and circles distinguish nucleotides involved in secondary structure pairing (boxes) from those in the tetra-loop (circles).

A

X	Y	M(xy)
317	334	0.765
317	307	0.608
318	333	0.435
318	314	0.257
319	323	0.402
319	367	0.266
323	319	0.402
323	153	0.285
333	318	0.435
333	314	0.282
334	317	0.765
334	307	0.573

B

```
317,334
G*C(47.6), A*U(40.2), C*G(6.1), G*U(2.4), -*-(1.2), G*-(1.2),
A*C(1.2)

H(xy)   M(xy)   R1(xy)  R2(xy)   -lnP
1.142   0.765   0.820   0.784   50.610


318,333
C*G(84.1), U*A(9.8), G*C(2.4), -*-(1.2), U*G(1.2), C*C(1.2)

H(xy)   M(xy)   R1(xy)  R2(xy)   -lnP
0.624   0.435   0.833   0.810   25.490


319,323
C*G(82.9), G*C(9.8), -*-(1.2), G*-(1.2), U*U(1.2), U*G(1.2),
G*G(1.2), C*A(1.2)

H(xy)   M(xy)   R1(xy)  R2(xy)   -lnP
0.705   0.402   0.735   0.717   20.088
```

C

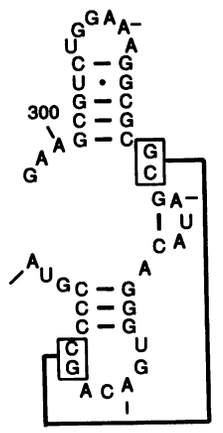

Figure 8. A: *Nbest* example for a complex structure in the 317–334 region of 23S rRNA (*Escherichia coli* sequence and numbering). **B**: Abbreviated *IMIXY* output for the pertinent positions in part A. **C**: Secondary Structure Diagram of this region of the 23S rRNA, positions 298–340. Position 300 is indicated, tic marks identify every tenth base.

It has been found experimentally that certain tetra loop sequences, closed by a short underlying helix, confer exceptional stability to small oligonucleotides (49–51). Recently, the three-dimensional structures for three sequences from two classes of tetra loops, UUCG, and GCAA and GAAA (GNRA) have been solved by NMR (52, 53), revealing in part, how such a structure could stabilize RNA. These experimental results suggest, at least partially, why the stability and structure of these local RNA motifs are important and thus selected.

Recently additional structure has been elucidated with comparative methods (33) for the L4 cross-linking region of *Escherichia coli* 23S rRNA (54). This unique structure is presented to demonstrate the unusual base pairing constraints that can be elucidated by such methods.

Positions 298–340 of the 23S rRNA are at the end of an extended secondary structure helix. Two helices were known previously within this region, namely 301–305/312–316 and 325–327/335–337. As shown in the *Nbest* and *IMIXY* analysis in Figure 8, the following pairs were all found to covary with one another, in a strict canonical fashion: 317/334, 318/333, and 319/323. These pairings, as diagramed in Figure 8C, make up an unusual structure. The two helices 317–318/333–334 and 325–327/335–337 form a pseudoknot structure, with the lone pair interaction, 319/323, nested in the middle. If all of these interactions occur simulataneously, several helix–helix coaxial stackings are possible, making for a very complex structure. This example is one of many comparatively derived structural elements that can be studied by site-directed mutational analysis. Another

unusual pseudoknot structure in the 16S rRNA (30) was proposed from comparative methods, and was subsequently shown to be correct (55).

CONCLUSIONS AND PROSPECTUS FOR THE FUTURE

The elucidation of higher-order structural constraints by comparative methods has evolved over the years, in part as a function of the significant growth in the various homologous RNA alignment datasets, large increases in available computing power, enhancements in the basic covariance algorithms, and refinements in our implementation and interpretation of the analysis. Various comparative methods have been developed and implemented on tRNA, 16S rRNA and 23S rRNA datasets (24, 26, 40, 42, 54). These comparative methods, in combination, have identified many structural elements within those RNAs (see (33) for rRNA discussion). The types of structural elements identified are:

- Standard Watson–Crick base pairing.
- Arrangement of canonical pairings into contiguous, antiparallel structural elements (i.e. secondary structure helices).
- *Tetra-loop* hairpin loops.
- Tertiary canonical and non-canonical pairings.
- Non-canonical pairings and their replacements, including: G-U ⟷ A-C, U-U ⟷ C-C, A-G ⟷ G-A, G-G ⟷ A-A, etc.
- Multiple examples of lone canonical base pairs (i.e. pairings that are not immediately contiguous with another structural pairing).
- Pseudoknot structures, usually involving canonical pairings.
- Canonical and non-canonical pairings arranged in a parallel (vs. the usual antiparallel) orientation.
- Suggestive evidence for helical coaxial stacking.

These classes of structural elements, as inferred from positional covariances, can be appreciated by our current experimental understanding of RNA structure. Beyond this list, we have presented several examples of a more intiguing set of correlations; weaker correlations that do not infer structural base pairings, but instead identify more than two positions that are evolutionarily constrained [Figures 5 and 6, tRNA anticodon loop and D helix constraints]. It is interesting to note that experimental studies have suggested some linkage between the positions identified here, however a structural explanation remains to be found.

The methods presented here, which have grown out of previous comparative protocols, can decipher additional RNA structural information for those RNAs that have been studied previously by these methods, as well as addressing other homologous RNA datasets that have not been analyzed by such methods. Although detailed three-dimensional structural information is known for some tRNAs, comparative analysis could well reveal additional tRNA structural refinements, such as identifying subtle structural features that distinguish each amino-acid accepting tRNA class, the so called tRNA identity problem. The secondary structures for the 16S and 23S rRNA are largely resolved. At this time, our focus is on other higher-order structural details, such as tertiary interactions. As eluded to earlier, other types of contraints could be found as well in the rRNAs.

These methods also reveal more than the mere existence of a helix; they reveal constraints on the actual pairing possibilities. For example, each base pair in the 16S rRNA has a unique pattern of allowable pairing types (i.e. G-C, A-U, G-U, etc.). Some contain all possible canonical pairings, others allow for G-U pairings in addition to canonical pairings. Some are constrained to a subset of the canonical pairings, implying additional structural constraints exist for those base pairs. We anticipate that different RNA datasets (for example, Group I introns and RNase P RNA) will reveal different types of RNA structural constraints. Thus not only will solutions to different RNA molecules emerge from such analyses (e.g. 16S rRNA), a broader sampling of structural elements will emerge as well.

Elucidating these structural refinements will require large increases in our RNA datasets. In the midst of the sequencing revolution, the amount of comparative sequence data will continue to increase at faster rates for a variety of homologous RNA types, presenting us with this opportunity. Our computational methods themselves will also need to evolve. For example, the methods discussed here do not integrate the number of mutational events that underly the correlation values calculated. By knowing a minimal phylogenetic tree for the sequences being analyzed, we can enhance our correlation values for those pairs for which multiple phylogenetic changes have occurred throughout evolution (57). The work presented here suggests that such analysis will continue to reveal more constraints on RNA structure.

ACKNOWLEDGMENTS

We gratefully acknowledge George Hartzell, Steve Carley, and Tony Direnzo, for assisting in various stages of the development of the correlation analysis computer programs discussed here. In addition, we wish to acknowledge Bryn Weiser (for XRNA) and Tom Macke (for AE2) for their programming expertise. These latter two programs greatly facillitate the generation of RNA structure display and primary structure alignments. We also appreciate the insightful comments from the two referees.

This research has been supported by NIH grants HG00249 and GM28755 (awarded to G.D.S.), GM48207 (awarded to R.R.G.), and the Keck Foundation, whose funds have provided additional support for this project.

R.R.G. is an Associate in the Program in Evolutionary Biology of the Canadian Institute for Advanced Research (CIAR). We also thank the W.M.Keck Foundation for their generous support of RNA science on the Boulder campus.

REFERENCES

1. Cech,T.R. (1990) (Nobel Lecture). *Angew. Chem. Int. Ed.Engl.* **29**, 759–768.
2. Pace,N.R. and Smith,D.K. (1990) *J.Biol.Chem.* **265**, 3587–3590.
3. Noller,H.F., Hoffarth,V. and Zimniak,L. (1992) *Science* **256**, 1416–1419.
4. Chastain,M. and Tinoco,I. (1991) *Prog. Nucleic Acid Res. and Mol. Biol.* **41**, 131–177.
5. Holley,R.W., Apgar,J., Everett,G.A., Madison,J.T., Marquisee,M., Merrill,S.H., Penswick,J.R. and Zamir,A. (1965) *Science* **147**, 1462–1465.
6. Madison,J.T., Everett,G.A. and Kung,H.K. (1966) *Cold Spring Harbor Symposia on Quantitative Biology*. Vol. XXXI, pp. 409–416.
7. Zachau,H.G., Dutting,D., Feldmann,H., Melchers,F. and Karau,W. (1966) *Cold Spring Harbor Symposia on Quantitative Biology*. Vol. XXXI, pp. 417–424.
8. RajBhandary,U.L., Stuart,A., Faulkner,R.D., Chang,S.H. and Khorana H.G. (1966) *Cold Spring Harbor Symposia on Quantitative Biology*. Vol. XXXI, pp. 425–434.
9. Levitt,M. (1969) *Nature* **224**, 759–763.
10. Quigley,G.J. and Rich,A. (1976) *Science* **194**, 796–806.
11. Kim,S.-H. (1976) *Prog. Nucleic Acid Res. and Mol. Biol.* **17**, 181–216.
12. Fox,G.E. and Woese,C.R. (1975) *Nature* **256**, 505–507.
13. Woese,C.R., Magrum,L.J., Gupta,R., Siegel,R.B., Stahl,D.A., Kop,J., Crawford,N., Brosius,J., Gutell,R., Hogan,J.J. and Noller,H.F. (1980) *Nucleic Acids Res.* **8**, 2275–2293.
14. Stiegler,R., Carbon,P., Zuker,M., Ebel,J.P. and Ehresmann,C. (1980) *C.R. Acad. Sci. (Paris) Ser.D.* **291**, 937–940.

15. Zwieb,C., Glotz,C. and Brimacombe,R. (1981) *Nucleic Acids Res.* **9**, 3621–40.
16. Noller,H.F., Kop,J., Wheaton,V., Brosius,J., Gutell,R.R., Kopylov,A.M., Dohme,F., Herr,W., Stahl,D.A., Gupta,R., and Woese,C.R. (1981) *Nucleic Acids Res.* **9**, 6167–6189.
17. Glotz,C., Zwieb,C. and Brimacombe,R. (1981) *Nucleic Acids Res.* **9**, 3287–3306.
18. Branlant,C., Krol,A., Machatt,M.A., Pouyet,J., Ebel,J.P., Edwards,K. and Kossel H. (1981) *Nucleic Acids Res.* **9**, 4303–4324.
19. Michel,F. and Westhof,E. (1990) *J.Mol.Biol.* **216**, 585–610.
20. Michel,F., Umesono,K. and Ozeki,H. (1989) *Gene* **82**, 5–30.
21. Brown,J.W., Haas,E.S., James,B.D., Hunt,D.A. and Pace,N.R. (1991) *J.Bacteriol.* **173**, 3855–3863.
22. Guthrie,C. and Patterson,B. (1988) *Annu. Rev. Genet.* **22**, 387–419.
23. Zwieb,C. (1989) *Prog. Nucleic Acid Res. and Mol. Biol.* **37**, 207–234.
24. Noller,H.F. and Woese,C.R. (1981) *Science* **212**, 403–411.
25. Woese,C.R, Kandler,O. and Wheelis,M.L. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 4576–4579.
26. Gutell,R.R., Weiser,B., Woese,C.R. and Noller,H.F. (1985) *Prog. Nucleic Acid Res. and Mol. Biol.* **32**, 155–216.
27. Gutell,R.R., Noller,H.F. and Woese,C.R. (1986) *EMBO J.* **5**, 1111–1113.
28. Leffers,H., Kjems,J., Ostergaard,L., Larsen,N. and Garrett,R.A. (1987) *J. Mol. Biol.* **195**, 43–61.
29. Haselman,T., Camp,D.G. and Fox,G.E. (1988) *Nucleic Acids Res.* **17**, 2215–2221.
30. Woese,C.R. and Gutell,R.R. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 3119–3122.
31. Haselman,T., Gutell,R.R., Jurka,J. and Fox,G.E. (1989) *J. Biomol. Struct. Dynamics* **7**, 181–186.
32. Gutell,R.R. and Woese,C.R. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 663–667.
33. Gutell,R.R., Larsen,N. and Woese,C.R. (1993) In Zimmermann,R.A. and Dahlberg,A.E (eds), *Ribosomal RNA: Structure, Evolution, Gene Expression and Function in Protein Synthesis*. CRC Press, Boca Raton, FL. in press.
34. Sprinzl,M., Dank,N., Nock,S. and Schon,A. (1991) *Nucleic Acids Res.* **19 (Suppl.)**, 2127–2171.
35. Dock-Bregeon,A.C., Westhof,E., Giege,R. and Moras,D. (1989) *J. Mol. Biol.* **206**, 707–722.
36. Gutell,R.R., Schnare,M. and Gray,M. (1990) *Nucleic Acids Res.* **18 (Suppl.)**, 2319–2330.
37. Gutell,R.R., Schnare,M. and Gray,M. (1992) *Nucleic Acids Res.* **20 (Suppl.)**, 2095–2109.
38. Olsen,G.J., Larsen,N. and Woese,C.R. (1991) *Nucleic Acids Res.* **19**, 2017–2021.
39. Olsen,G.J., Overbeek,R., Larsen,N., Marsh,T.L., McCaughey,M.J., Maciukenas,M.A., Kuan,W.-M., Macke,T.J., Xing,Y. and Woese,C.R. (1992) *Nucleic Acids Res.* **20**, 2199–2200.
40. Chiu,D.K.Y. and Kolodziejczak,T. (1991) *CABIOS* **7**, 347–352.
41. Press,W.H., Flannery,B.P., Teukolsky,S.A. and Vetterling,W.T. (1988) *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge University Press.
42. Olsen,G.J. (1983) Ph.D. thesis. University of Colorado.
43. Yarus,M. (1982) *Science* **218**, 646–652.
44. Hirsh,D. and Gold,L. (1971) *J. Mol. Biol.* **58**, 459–468.
45. Smith,D. and Yarus,M. (1989) *J. Mol. Biol.* **206**, 503–511.
46. McClain,W.H. and Foss,K. (1988) *Science* **240**, 793–796.
47. Schulman,L.H. (1991) *Prog. Nucleic Acid Res. and Mol. Biol.* **41**, 23–87.
48. Woese,C.R, Winker,S. and Gutell,R.R. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 8467–8471.
49. Tuerk,C., Gauss,P., Thermes,C., Groebe,D.R., Gayle,M., Guild,N., Stormo,G., D'Aubenton-Carafa,Y., Uhlenbeck,O.C., Tinoco,I., Brody,E.N. and Gold,L. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 1364–1368.
50. Antao,V.P., Lai,S.Y., Tinoco,I. Jr (1991) *Nucleic Acids Res.* **19**, 5901–5905.
51. Antao,V.P. and Tinoco,I.Jr (1992) *Nucleic Acids Res.* **20**, 819–824.
52. Cheong,C. Varani,G. and Tinoco,I.Jr (1990) *Nature* **346**, 680–682.
53. Heus,H.A. and Pardi,A. (1991) *Science* **253**, 191–194.
54. Brimacombe,R., Greuer,B., Mitchell,P., Osswald,M., Rinke-Appel,J., Schüler,D. and Stade,K. (1990) In Hill *et al.* (eds), *The RIBOSOME: Structure, Function & Evolution*. American Society for Microbiology. pp. 93–106.
55. Powers,T. and Noller,H.F. (1991) *EMBO J.* **10**, 2203–2214.
56. Haselman,T., Chappelear,J.E. and Fox,G.E. (1988) *Nucleic Acids Res.* **16**, 5673–5684.
57. Winker,S., Overbeek,R., Woese,C.R., Olsen,G.J. and Pfluger,N. (1990) *CABIOS* **6**, 365–371.