

## A Phylogenetic Approach to RNA Structure Prediction

Viatcheslav R. Akmaev, Scott T. Kelley, and Gary D. Stormo

Dept. of Molecular, Cellular and Developmental Biology

University of Colorado, Boulder, CO 80309-0347

Phone: 303-492-1474, FAX: 303-492-7744

slava@ural.colorado.edu

kelleys@ural.colorado.edu

stormo@ural.colorado.edu

### Abstract

Methods based on the Mutual Information statistic (MI methods) predict structure by looking for statistical correlations between sequence positions in a set of aligned sequences. Although MI methods are often quite effective, these methods ignore the underlying phylogenetic relationships of the sequences they analyze. Thus, they cannot distinguish between correlations due to structural interactions, and spurious correlations resulting from phylogenetic history. In this paper, we introduce a method analogous to MI that incorporates phylogenetic information. We show that this method accurately recovers the structures of well-known RNA molecules. We also demonstrate, with both real and simulated data, that this phylogenetically-based method outperforms standard MI methods, and improves the ability to distinguish interacting from non-interacting positions in RNA. This method is flexible, and may be applied to the prediction of protein structure given the appropriate evolutionary model. Because this method incorporates phylogenetic data, it also has the potential to be improved with the addition of more accurate phylogenetic information, although we show that even approximate phylogenies are helpful.

### Introduction

Comparative methods have proven highly successful in predicting secondary structure in multiple alignments of RNA (Woese *et al.* 1983; Michel and Westhof 1990; Winker *et al.* 1990; Gutell *et al.* 1992; Gutell 1994; Cary and Stormo 1995; Gulko and Haussler 1996). The most commonly used comparative method, known as Mutual Information (MI), was introduced as a quantitative measure of correlation to help automate the process of secondary structure prediction (Chiu and Kolodziejczak 1991). MI methods predict structure by examining patterns of variation across sets of aligned sequences (Gutell *et al.* 1992). In essence, these procedures look for correlated (or compensating) mutations between pairs of positions in a sequence by calculating the amount of covariation between sequence positions. If two positions tend to have significant amounts of “mutual information” (i.e., as one position changes the other tends to change as well) this is an evidence that these positions are

interacting in the molecule and are likely to form Watson-Crick base-pairs or other interactions. With enough variation from numerous aligned sequences, it is possible to make structural predictions for very large RNA molecules (Gutell 1994).

Comparative methods are not only capable of determining basic helical structure, but are also very adept at predicting less intuitive structural motifs often missed by thermodynamic methods. These motifs include interactions such as pseudoknots, tertiary canonical and non-canonical base pairings, base-stacking, and tetra-loops. Many of these predictions have later been verified experimentally, demonstrating the predictive value of these methods (Chastain and Tinoco 1991; Lodmell *et al.* 1995).

Comparative methods can, and sometimes do, take into account the phylogenetic relationships of the sequences. This is usually done “by hand” in the sense that some minimum number of changes along the tree are required before a correlation is trusted (Woese *et al.* 1983). However, standard MI methods ignore the tree, treating all the sequences as if they had arisen from the same common ancestor at the same point in time. Thus, they implicitly assume that all of the variation in each sequence has evolved independently from all of the other sequences. The result of this is that MI methods will typically overestimate the true amount of correlation between sequence positions, and will accept “spurious correlations” (i.e., correlations attributable to a shared phylogenetic history) treating them as significant evidence of interaction and structure (Lapedes *et al.* 1997). Figure 1 illustrates how knowledge of the phylogenetic relationships among a group of sequences can be useful in identifying such false positives.

Using simulated data sets, Lapedes *et al.* (1997) demonstrated how ignoring phylogenetic signal can lead to false conclusions of non-independence. By simulating the evolution of sequences down two types of trees, one a “star” phylogeny and one a more realistic bifurcating phylogeny, the authors demonstrated that the more realistic phylogenetic tree “created” mutual information between sites even when such interactions were absent. Therefore, it appears that the inclusion of the phylogenetic tree into these comparative methods should improve the prediction of secondary structure by reducing effects of spurious correlations. In this paper, we describe an improved method of RNA struc-

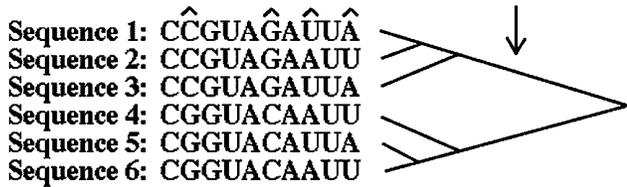


Figure 1: A hypothetical example showing covariation between positions in an RNA sequence. The figure shows a set of aligned homologous sequences. Using standard mutual information methods, in the absence of a phylogeny, positions 2 and 6 would appear to covary as would positions 8 and 10. In other words, each time there is a change in one position, there is a corresponding change in the other. However, the phylogeny on the right indicates that sequences 1-3 form a closely related group, as do sequences 4-6. If the common ancestor of all these sequences had a G at position 2 and a C at position 6, there would only be evidence for one independent origin of compensating mutations between these positions (indicated by the arrow). On the other hand, there are multiple instances of correlated mutations between positions 8 and 10 because related taxa are all dissimilar at these two positions indicating multiple independent origins of this correlation.

ture prediction analogous to mutual information techniques that incorporates phylogenetic information. By taking the tree into account, we show that the method reduces the number of false positives detected by non-phylogenetic methods, and increases the ability to detect truly correlated positions.

### Phylogenetically Based Structure Prediction

In our approach to the correlation analysis of RNA sequences, we first estimate the independent rates of evolution for each position in the sequence along a phylogenetic tree. In particular, we apply the Hasegawa-Kishino-Yano (HKY85) (Hasegawa *et al.* 1985) model of DNA/RNA sequence evolution down a given tree for each site of the sequences. A number of DNA/RNA sequence evolution models have been developed during the last twenty years (reviewed by Swofford 1998). The well-known Jukes-Cantor model (Jukes and Cantor 1969) is the simplest of these, while the GTR, general time-reversible 11-parameter model is probably the most complex generalization that has been considered (Lanave *et al.* 1984). Surprisingly, the assumption of the same substitution rates over the sites of the sequences was a very common one in the past, presumably because this was less computationally demanding. However, with constantly increasing computational power, the opportunity to work with more general (i.e., more realistic) models becomes very attractive. We settled on the 6 parameter HKY85 model as a compromise between model complexity and computational efficiency.

The HKY85 model has six unknown parameters: the four nucleotides equilibrium frequencies, the mean substitution rate and the transition-transversion ratio. However, only five of these six parameters are independent because the nucleotide frequencies add up to one. The question is how do

we find the parameters of the model given a data set and a phylogeny estimate? Basically, the estimates of the parameters are the set of numbers that maximizes the likelihood function of the data given a set of aligned sequences and a phylogenetic tree. The numerical aspects of the maximization problem have been addressed in previous works (Schadt *et al.* 1998). The output is the optimal parameters of the HKY85 model and the maximum of the likelihood function for each position of the sequences. Because we optimize the parameters of the model separately for each position we are probably over-fitting to the data, especially when we have a small number of sequences. But because our use of these parameters is to identify pairs of positions that do not appear to evolve independently, the result is that our identification of correlated positions is conservative.

Given the rate parameters at each site, assuming that each position evolved independently of all others, the next step is to determine if two positions evolved independently or if they show some bias towards simultaneous mutations on the phylogenetic tree. Recently, efforts have been made to simplify the joint distribution approach to some kind of quasi-joint evolution models with some positive results, though the method clearly needs refining (Muse 1995). Some methods even try to work out the real joint distribution model for a pair of positions (Gulko and Haussler 1996). Gulko and Haussler’s results were very impressive. However, in our approach we do not make many of the simplifications that they used (e.g., the discrete branch lengths, the same rates of evolution at all positions of a multiple alignment as well as the high dependence of the result on the training dataset, and the requirement of a training dataset itself).

Certainly, the ideal approach would be to determine the optimal evolution parameters for each potential pair of interacting positions, and then choose the “good ones” based on how much better they fit the data. However, this is impractical on many datasets for two reasons. First, the number of parameters for joint evolution is larger; how large depends on the model used to describe evolution of pairs. Second, and more important, is the great number of potential pairs one has to examine to identify the set of true pairs. For example, in a sequence of length 1000, there are at most 500 pairs of interacting positions. (This assumes each base interacts with at most one other position. While this ignores base-triple interactions, it is a reasonable limit to the expected number of interactions in RNA.) But the total number of pair-wise combinations is about 500,000. Determining the maximum likelihood parameters for all of those potential pairs, in order to find out which pairs are not well described by independent evolution, would be prohibitive. Our approach is much faster and can identify the likely interacting positions, including information from the phylogenetic tree, which can then be examined in more detail.

In this paper, we describe a statistical estimation method of structure prediction that combines a phylogenetic approach, utilizing the HKY85 model, with a purely statistical approach that does not take into account the phylogeny. The likelihood ratio of these two approaches essentially asks whether the evolution of two positions is better described by a model that takes the tree into account, but assumes inde-

pendence, or a model that ignores the tree but assumes interactions. We show that this method is a more reliable predictor of interacting positions, on both real and simulated datasets, than standard MI-based methods. We demonstrate its use on three example datasets.

## Materials and Methods

### Data Sets

Sequence alignments of tRNA's and 16S small subunit rRNA's were obtained by downloading from the world wide web (tRNA: (Sprinzl *et al.* 1998) 16S rRNA: (Van de Peer *et al.* 1998)). For the analysis of tRNA sequences, we used 300 sequences representing a diverse array of taxa. For the analysis of 16S rRNA, we selected approximately 150 bacterial sequences from 12 bacterial families. The 16S sequences were selected such that the phylogeny of these sequences would resemble the tree of Figure 5B. In other words, we selected groups of closely related sequences from among several distantly related bacterial families.

For the 16S rRNA sequences, we used the alignments as they were presented in the web sites (as we did for the tRNA sequences) except that we removed positions in the 16S bacterial sequence alignments at which 20% or more of the sequences contained gaps. This was necessary because individual sequences often contained > 70% gap positions due to the fact that the database contained the alignments of thousands of 16S rRNA sequences, including many distantly related sequences. Although alignments between distantly related sequences in the database (e.g., Bacteria to Eucarya) were probably not completely reliable, the alignments within the bacterial sequences appeared much more reliable and contained few gaps once the common gap positions were removed. For the 16S rRNA structure predictions, we focused on a small and well known part of the molecule that corresponded to the region between nucleotides 404 and 547 of the *E. coli* sequence (Gutell 1994).

We also tested the methods in this paper on randomly generated sequence data. This data was generated by the program SEQ-GEN (Rambaut and Grassly 1997) which allowed us to evolve sequences down a phylogenetic tree using the HKY model of evolution. (In this case, the parameters of the model were predetermined and were the same for all positions.) We used a phylogenetic tree based on 66 16S rRNA bacterial sequences, including branch lengths, as input to the program, and the program allowed us to evolve multiple data sets of varying length using this tree.

### Phylogenetic Analyses

Phylogenetic analyses were performed using PAUP\* (reviewed by Swofford 1998). We used the neighbor joining (NJ) criterion for all phylogenetic analyses because we needed to generate a large number of phylogenetic estimations with an enormous number of taxa (> 150). Other methods, such as parsimony or maximum likelihood, would have been too slow. The NJ procedure of PAUP\* also generated branch length estimates which were necessary for our methods. Phylogenetic analyses were based on the same data set that we used to predict RNA structure. For the tRNA

sequences, this meant that the analysis was based on only 99 base positions, while we used the entire 16S rRNA sequence, approximately 1,500 bases, to generate trees for the prediction of 16S rRNA structure. Although we doubt the overall reliability of the trees used in the analyses, particularly in the case of the tRNA structure predictions which were based on little data, our methods were robust for all the trees generated with a number of different models of evolution (see Results). Furthermore, one of the main conclusions of this work is to demonstrate that using information from the phylogenetic tree, even if it is only approximately correct, leads to better predictions of structure than methods that ignore the tree altogether.

### Test statistic

In the first step of our method, the input data are a set of aligned sequences and a tree generated by PAUP\*. The output is the optimal parameters of the HKY85 model and the maximum of the likelihood function for each position of the sequence. It is worth noting that, since the positions are assumed to be evolving at different rates, this method requires more data than if we assumed the same rates for all the sites of the sequences. However, the increased demand in the amount of data necessary is not as drastic as might have been anticipated.

The likelihood function computed in the first step is then compared with the likelihood of the data at position  $i$  conditioned on the data at position  $j$ , for all pairs  $(i, j)$ . In essence, we are comparing whether the prediction of the data at position  $i$ , given the data at position  $j$ , is better or worse than the prediction of the data at position  $i$  based on its independent evolutionary model. Mathematically, the following statistic represents the comparison:

$$R_{i|j} = -\log \frac{L_i^{HKY}(\mu^*, k^*, \pi_A^*, \dots, \pi_U^*)}{L_{i|j}} \quad (1)$$

where:

$$L_i^{HKY}(\mu^*, \dots, \pi_U^*) = \max_{\mu, k, \pi_A, \dots, \pi_U} L_i^{HKY}(\mu, \dots, \pi_U) \quad (2)$$

$\mu$  is the mean substitution rate,  $k$  is the transition-transversion ratio and  $\pi_a$  is the frequency of nucleotide  $a$ , where  $a = \{A, C, G, U\}$ .

$$L_{i|j} = \prod_{\substack{a = \{A, \dots, U\} \\ b = \{A, \dots, U\}}} f_{a|b}^{N \cdot f_{ab}} \quad (3)$$

$f_{a|b}$  is the frequency of nucleotide  $a$  at position  $i$  given nucleotide  $b$  at position  $j$ .  $f_{ab}$  is the frequency of the pair of nucleotides  $ab$  and  $N$  is the number of sequences.

If position  $i$  is independent of position  $j$ , then  $L_{i|j} = L_i$ , and  $R_{i|j}$  would be always negative for trees reasonably describing the data. As an example of unreasonable tree, we can consider a tree that has zero branch length between taxa with different nucleotides.  $L_i^{HKY}$  given this tree is zero, no matter what the parameters are, which makes  $R_{i|j}$  to be  $+\infty$ . Referring to Figure 1, assuming the tree fits the data,

the likelihood  $L_2^{HKY}$  at position 2 would be much bigger than the likelihood  $L_{10}^{HKY}$  at position 10. Thus, considering these two sites independently of the other positions, i.e.,  $L_{i|j} = L_i$  and  $R_{i|j} = R_i$ , where  $i=2,10$ ,  $R_{10}$  would be bigger than  $R_2$  even though  $L_2 = L_{10}$ . In this case,  $R_i$  represents the amount of variation at position  $i$ . The more variation at site  $i$  is present in the data, the bigger  $R_i$  is. This indicates that this approach would allow us to get rid of spurious correlations associated with the phylogeny, as illustrated in Figure 1.

Since two sites are involved, the statistic below is the output for each pair of positions:

$$R_{ij} \equiv R_{i|j} \cdot R_{j|i} = -\log \frac{L_i^{HKY} L_j^{HKY}}{L_{i|j} L_{j|i}} \quad (4)$$

$R_{ij}$  will be 0 if the two methods give equal values for the likelihood of the data. Notice that this will also be true for any completely conserved positions. The value will be negative if the tree-based independent model fits the data better, and will be positive in the opposite case. Therefore 0 is an appropriate threshold to use for predicting interacting pairs. A number of tests and simulations of this method have been completed which we present in the following sections.

## Results

To verify the accuracy of our methods, we performed tests on three different data sets: a tRNA data set, a 16S rRNA data set, and a simulated data set. We first asked whether our method was able to accurately predict the tRNA and 16S rRNA secondary structure. Then we asked how well our method performed compared with standard mutual information (MI) methods for all three data sets. In all the tests, we excluded positions that were invariant in the sequences we used because these positions contained no information about correlation. In fact, no comparative methods, by themselves, are able to predict interactions at positions where there is no variation (Cary and Stormo 1995).

### tRNA data set

$R_{ij}$  statistic for each pair of positions was calculated by our CgHKY (correlations given the HKY model) program. For a tRNA molecule 99 nucleotides in length there are 4851 pairs of positions to examine. The data set that we have tested the program on contained 300 tRNA molecules from numerous different, and distantly related, organisms. Although the number of calculations was fairly big, the program required only about 3 minutes on a Sun Ultra 30 station to complete the computations.

CgHKY was able to predict the majority of known interactions in the tRNA secondary structure (Figure 2). Due to a lack of variation, we could not predict several of the known base pairs (8:14, 18:55, 19:56) and base triples (9 with 12:23, 45 with 10:25). In these cases, at least one of the positions was completely conserved (no variation) and provided no data for the analysis. Thus, CgHKY was able to predict all but one (54:58) of the interactions in tRNA that were possible to predict. Figure 3 shows a distribution of

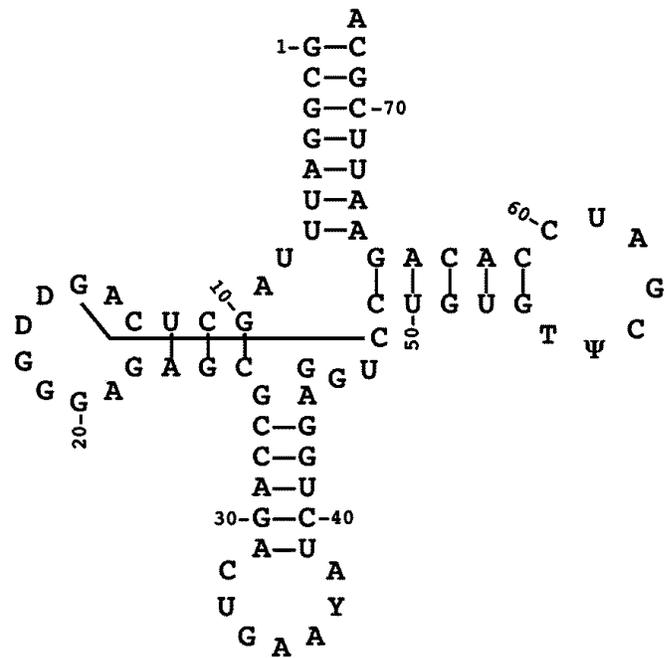


Figure 2: Secondary structure diagram for tRNA. The solid lines represent the interactions predicted by CgHKY with the variable region excluded from the analysis. All these interactions have been discovered in the crystal structure.

the  $R_{ij}$  statistic output for the tRNA data from the CgHKY program.

The figure reveals an almost perfect separation between the independent and the interacting pairs, and indicates a threshold on the border of the two sets. Interestingly, the known interacting pair CgHKY did not predict that had variation (54:58) ends up right in the middle of a supposedly non-interacting group of positions on the interval [-30,0]. This result suggests that the threshold should not be treated as an absolute, and that efforts should be made to further investigate sites just below this threshold as potentially interacting positions.

### 16S rRNA dataset

For the 16S rRNA we present the results of the analysis of 144 sites in the 16S rRNA, corresponding to position 404 through position 547 of *E. coli* (Figure 4).

Again, the CgHKY program was able to predict the majority of the secondary structure in this data set (Figure 4). With the 16S data set, we also compared the output of the CgHKY program with that of the MIXY program. The MIXY program ("Mutual Information between X and Y") calculates the Mutual Information (MI) for each pair of position, which is a commonly used technique for prediction of RNA secondary structure (Gutell *et al.* 1992). However, the Mutual Information method does not account for phylogenetic features of the data. This weakness of the MI approach has been pointed out in recent papers (Lapedes *et al.* 1997). The MI statistic works very well if the phylogenetic tree of

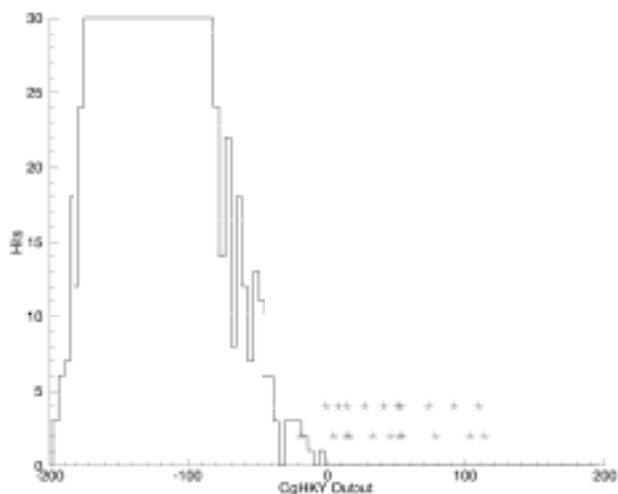


Figure 3: The histogram of the CgHKY output cut at the level 30. The solid line is the histogram itself for the positions without interactions. Each asterisk represents the score for a single pair of interacting sites, and their height is irrelevant.

the data set approximately resembles a star phylogeny.

Figure 5 shows two realistic phylogenetic trees, one with uniform branches, and one with clusters of closely related sequences. In order to compare CgHKY and MIXY on non-uniform data, we collected groups of closely related sequences from different bacterial “families” whose phylogenetic relationships would approximate the kind of tree shown in Figure 5.B in which CgHKY should outperform MIXY. The test dataset we used contained 155 16S rRNA sequences. Both approaches (MIXY and CgHKY) were applied to this data, and the results are plotted in Figure 6.

Figure 6.A shows the histogram of the  $R_{ij}$  statistic from the CgHKY program, while Figure 6.B shows the histogram of the MI statistic for the same set of positions. Almost perfect separation is observed in the CgHKY output (Figure 6.A). The two hits around mark 10, that are seen on the interacting pairs side, are pairs 486-504 and 486-541, Figure 4. It is worth noting that positions 504 and 541 are interacting and, in fact, they are right in the stem region. If we look at these three sites for each sequence in the data set we would see that they have UCG nucleotides at positions 486-504-541 in 92% of the sequences and it has AGC combination in the rest of the sequences. At first we interpreted this as evidence of a base-triple interaction. However, examination of a larger dataset showed this not to be true. Therefore those two points are false positives. But note that they are the only false positives when the threshold is set low enough to capture all of the true positives (i.e., with no false negatives).

The MIXY histogram does not show as clear a separation between the non-interacting and interacting sets of positions as does the CgHKY output (Figure 6.B). There are a couple of outliers near to the 300 mark that are not distinguishable

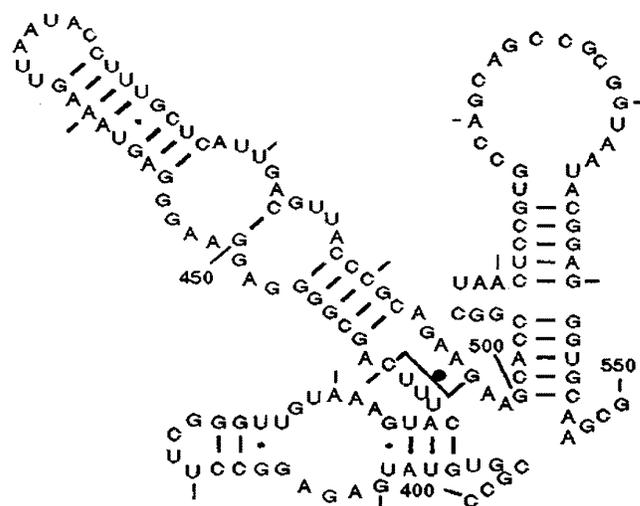


Figure 4: The 400-550 region of 16S ribosomal RNA. The variable stem and loop region, position 455 through position 477, were excluded from the analysis, as were the invariant positions. Figure modified from Gutell (1994).

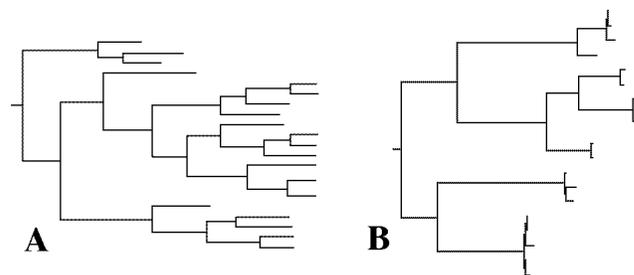


Figure 5: Two examples of different phylogeny. Tree A has almost uniform branches, while tree B contains distantly related groups of species.

among the independent pairs, and there is also some noise on the interacting set side. If we set the threshold so as to have no false negatives, there are over 40 false positives. A better threshold might end up with 3 false positives and false negatives each. Therefore the sensitivity and specificity of CgHKY is improved over the standard MI approach.

### Simulation

A tree of sequences similar to that shown in Figure 5.B was chosen for the simulation. The tree contained 66 taxa (leaves), and each sequence was 300 positions long which makes a total of 44850 pairs of positions. The sites were evolved independently of each other down the phylogenetic tree using the program SEQ-GEN (see Methods).

Given the 300 site long randomly generated data set, we then created correlated positions within this data set. The odd numbered sites remained unchanged, but the even numbered sites were changed so that each

even numbered position would be correlated to the previous position as either a Watson-Crick or a GT pair. A portion of this data set is shown below:

```
1:GCTGATCGGTTGATGTGCGCGTCGG...
2:GCTGATCGGTTGATGTGCGCGTCGG...
3:GCTGATATGTTGATGTGCGCGTCGG...
4:GCTGATCGGTTGATGTGCGCGTCGG...
5:GTTGATCGGTTGATGTGCGCGTCGG...
6:GTTAATATGCTGATGTGCGCGTCGG...
7:ATCGATCGTAATGCTGGCGCTGTAG...
8:ATCGATCGTAATGCTGGCGTTGGCG...
9:ATCGATCGTAATGCTACGGTTGCGG...
```

Figure 7 shows the results from the CgHKY and the MIXY program on the simulated data set. As can be seen, there is very little overlap in the  $R_{ij}$  statistic scores for interacting and non-interacting positions (Figure 7.A), while the MI statistic has a wide overlap where it is not possible to distinguish which pairs are correlated and which are not (Figure 7.B). Therefore on the simulated data CgHKY also has improved sensitivity and specificity compared to MI.

## Discussion

The phylogenetically based method of structure prediction we describe in this paper proved to be quite effective in predicting RNA molecular structure. We demonstrated that the method accurately predicted 90% of the tRNA structure (Figure 2) and 92% the partial 16S rRNA structure we analyzed (Figure 4). The interactions the method failed to predict were either the result of a lack of variation in the data set at those positions (e.g., positions 18 and 54 of tRNA, positions 507 and 524 of 16S rRNA) or because the structure was not present in all of the sequences analyzed. For example, we excluded the stem-loop region (bases 455 to 477, Figure 4) because it is missing in large fraction of the sequences in the dataset, and if we included it we obtained numerous spurious correlations. However, if we examined only those sequences that contained that stem-loop region, we did obtain the correct base-pairing structure.

Occasionally, there were some true correlations that we could not distinguish from non-interacting positions. For instance, positions 54 and 58 of the tRNA molecule (Figure 2) are known to interact, but they fell within a group of apparently non-interacting positions. However, a closer look at this group of positions showed that many of them are correlated or potentially correlated in the molecule. For instance, this group contains pairs of positions from the anticodon region. Because of the nature of the data (the sequences were taken from different anticodon families), changes at the anticodon positions are probably correlated, since a change from one anti-codon sequence to another often requires simultaneous change at these three codon positions. Thus, there may be some true correlations between these other positions that have little to do with structure.

In a comparison of our method with the results of a traditional MI program (MIXY), we found the tree based method to be generally more robust. The inclusion of phylogenetic information improved the ability to distinguish between correlated and independently evolving sites in both real (Figure 6) and simulated (Figure 7) data sets. This suggests that the tree based method is better able to differentiate between true correlations and those attributable to phylogenetic noise. We note, however, that the MI methods we used performed quite well in most circumstances, though it appears that our method is more robust over a wider range of conditions and provided a clearer threshold by which to evaluate correlated and non-correlated positions.

Although the method we describe here is generally superior to the purely statistical method, there are some disadvantages. First, it is hard to establish what distribution the  $R_{ij}$  statistic has, while the MI statistic is known to be approximately a multiple of a  $\chi^2$ -statistic. This issue obscures a possible claim of statistical significance in the identification of the interacting positions. Second, to get comparable values for the  $R_{ij}$  statistic for different pairs of positions, the data set has to have about the same number of gaps at the positions that are considered. This is because the likelihood functions, given the phylogenetic tree, for a position with 10% gaps and a position with 50% gaps have a different order of numbers and are virtually incomparable without a proper normalization. One possible way to deal with this problem is to introduce a gap as a fifth character. Unfortunately this approach would make positions with a significant number of gaps correlated with each other. The sites in the variable stem region of 16S rRNA (Figure 4) would be correlated with each other, in all possible combinations, and they tend to have bigger values for the  $R_{ij}$  statistic with the other positions. This is also the case for the MI statistic.

On the other hand, this method can be applied to the prediction of higher order interactions, such as base triples, and may be able to discover interactions not found by MI methods. To make the statistic more realistic for these types of interactions, the comparison between the independent likelihood and the likelihood given the probabilities of a nucleotide at some position conditioned on the data at the other two or three sites might be considered. Furthermore, the phylogenetically based method has the potential to be enhanced with improved phylogenetic information. Indeed, the information on structure and correlation provided by this prediction method may be used to make better phylogenies of the data.

Finally, it is also worth mentioning that the method can be extended to the prediction of protein higher order structure. There have been some attempts to construct a reasonable substitution rate matrix for amino acid mutations (Dayhoff *et al.* 1978), but all the practical ones require a lot of assumptions which sometimes lead to undesirable simplifications. Hopefully, with the constantly increasing computing facilities we shall be able to implement a reasonable protein sequence evolution model.

## Acknowledgements

We would like to thank Alan Lapedes and three anonymous referees for helpful discussions and insights. This work was supported by a grant from NIH, HG00249.

## References

- Cary, R. B. and Stormo, G. D. 1995. Graph-theoretic approach to RNA modeling using comparative data. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, 75–80.
- Chastain, M. and Tinoco, I., Jr. 1991. Structural elements in RNA. *Prog. Nucl. Acids Res. and Mol. Bio.* 41:131–177.
- Chiu, D. K., Kolodziejczak, T. 1991. Inferring consensus structure from nucleic acid sequences. *Comp. Appl. Biosci.* 7:347–352.
- Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. 1978. A model of evolutionary change in proteins. In Dayhoff, M. O., ed., *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation. 345–352.
- Gulko B., Haussler D. 1996. Using multiple alignments and phylogenetic trees to detect RNA secondary structure. *Pac. Symp. Biocomput.* 350–367.
- Gutell, R. R., Power, A., Hertz, G. Z., Putz, E. J., Stormo, G. D. 1992. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.* 20(21):5785–5795.
- Gutell, R. R. 1994. Collection of small subunit (16S and 16S-like) ribosomal RNA sequences. *Nucl. Acids Res.* 22:3502–3507.
- Hasegawa, M., Kishino, H., Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 21:160–174.
- Jukes, T. H. and Cantor, C. R. 1969. Evolution of protein molecules. In Munro, H. N., ed., *Mammalian Protein Metabolism*. Academic Press. 21–132.
- Lanave, C., Preparata, G., Saccone, C. and Serio, G. 1984. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* 20:86–93.
- Lapedes, A.S., Giraud, B.G., Liu, L.C., Stormo, G.D. 1997. Correlated mutations in protein sequences: phylogenetic and structural effects. In *Proceedings of the IMS/AMS 1997 International Conference on Statistics in Computational Molecular Biology*. 33
- Lodmell, J. S., Gutell, R. R. and Dahlberg, A. E. 1995. Genetic and comparative analyses reveal an alternative secondary structure in the region of nt 912 of *Escherichia coli* 16S rRNA. *Proc. Natl. Acad. Sci.* 92:10555–10559.
- Michel F. and Westhof E. 1990. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative analysis. *J. Mol. Biol* 216(3):585–610.
- Muse, S. V. 1995. Evolutionary analyses of DNA sequences subject to constraints on secondary structure. *Genetics.* 139:1429–1439.
- Rambaut, A. and Grassly, N. C. 1997. Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Applic. Biosci.* 13:235–238.
- Schadt, E. E., Sinsheimer, J. S., Lange, K. 1998. Computational advances in maximum likelihood methods for molecular phylogeny. *Genome Research* 8:222–233.
- Sprinzi, M., Horn C., Ioudovitch A. and Steinberg S. 1998. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* 26:148–153.
- Swofford, D. 1998. *PAUP\*: Phylogenetic analysis using parsimony (\*and other methods)*. Sunderland, Massachusetts: Sinauer Associates, version 4 edition.
- Van de Peer, Y., Caers, A., De Rijk, P., De Wachter, R. 1998. Database on the structure of small ribosomal subunit RNA. *Nucleic Acids Res.* 26:179–182.
- Winker S., Overbeek R., Woese C.R., Olsen G.J., Pfluger N. 1990. Structure detection through automated covariance search. *Comput. Applic. Biosci.* 6:365–371.
- Woese, C. R., Gutell, R. R., Gupta, R. and Noller, H. F. 1983. Detailed analysis of the higher-order structure of the 16S-like ribosomal ribonucleic acids. *Microbiol. Rev.* 47:621–669.

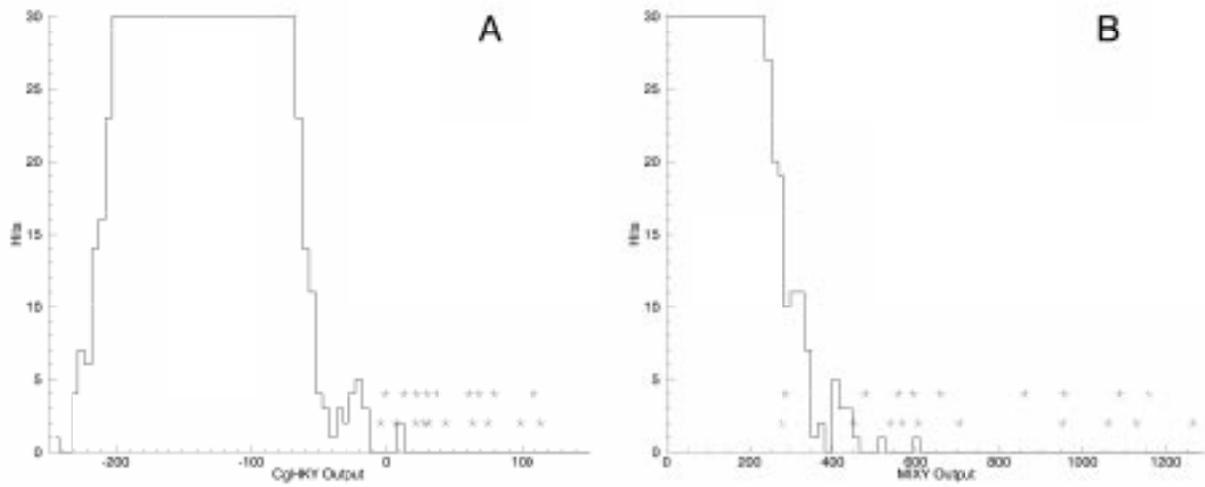


Figure 6: The histograms of the CgHKY and MIXY outputs for the 16S rRNA data set. The solid lines are the histograms for the positions without interactions. Each asterisk represents the score for a single pair of interacting sites, and their height is irrelevant.

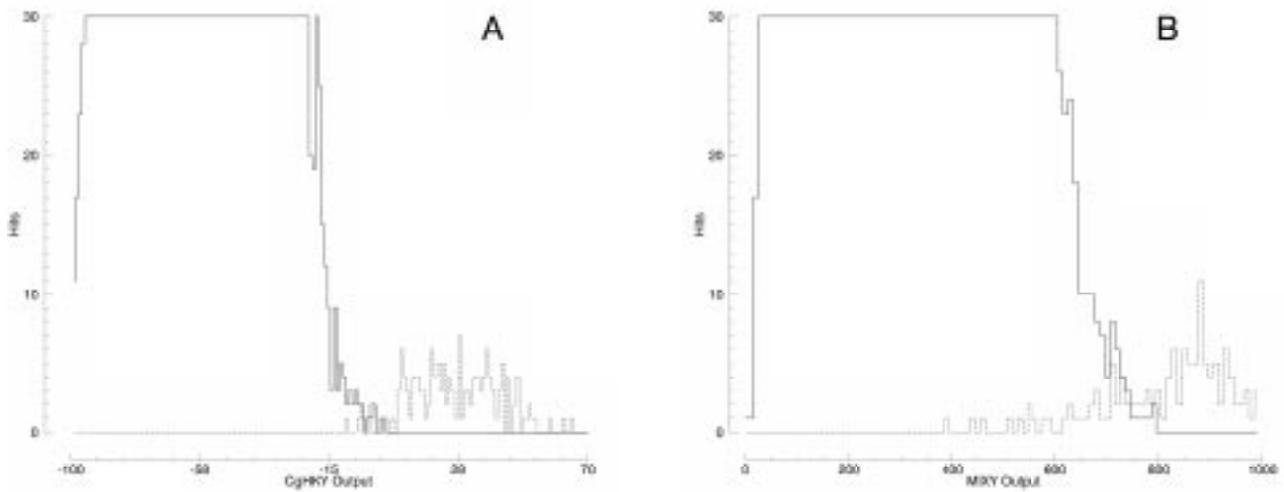


Figure 7: The histograms of the CgHKY and MIXY outputs for the simulated data sets. The solid lines are the histograms for the data set without interactions. The dashed lines represent the output for the data set with interacting sites only.