BIOINFORMATICS

Vol. 16 no. 6 2000
Pages 501–512

Phylogenetically enhanced statistical tools for RNA structure prediction

Viatcheslav R. Akmaev[1], Scott T. Kelley[2] and Gary D. Stormo[3]

[1]Dept. of Applied Mathematics, Box 526, University of Colorado, Boulder, CO 80309, USA, [2]Dept. of Molecular, Cellular and Developmental Biology, Box 347, University of Colorado, Boulder, CO 80309, USA and [3]Dept. of Genetics, Washington University School of Medicine, 660 S. Euclid, Box 8232, St. Louis, MO 63110, USA

Received on November 2, 1999; revised on January 20, 2000; accepted on January 28, 2000

Abstract

Motivation: *Methods that predict the structure of molecules by looking for statistical correlation have been quite effective. Unfortunately, these methods often disregard phylogenetic information in the sequences they analyze. Here, we present a number of statistics for RNA molecular-structure prediction. Besides common pair-wise comparisons, we consider a few reasonable statistics for base-triple predictions, and present an elaborate analysis of these methods. All these statistics incorporate phylogenetic relationships of the sequences in the analysis to varying degrees, and the different nature of these tests gives a wide choice of statistical tools for RNA structure prediction.*

Results: *Starting from statistics that incorporate phylogenetic information only as independent sequence evolution models for each position of a multiple alignment, and extending this idea to a joint evolution model of two positions, we enhance the usual purely statistical methods (e.g. methods based on the Mutual Information statistic) with the use of phylogenetic information available in the sequences. In particular, we present a joint model based on the HKY evolution model, and consequently a χ^2 test of independence for two positions. A significant part of this work is devoted to some mathematical analysis of these methods. We tested these statistics on regions of 16S and 23S rRNA, and tRNA.*

Availability: *The programs are available upon request.*

Contact: *slava@colorado.edu*

Introduction

Construction of an automated technique for RNA structure prediction is still an important problem in bioinformatics. During the past 20 years, a fairly large number of methods have been presented. These methods fall into two general categories: energy minimization methods (Jacobson and Zuker, 1993; Zuker and Sankoff, 1984; Tinoco *et al.*, 1971) and comparative sequence analysis methods (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). Comparative methods rather than energy methods, have been generally more successful and robust on large RNA molecules (Han and Kim, 1993; Le and Zuker, 1991). The potential usefulness of including phylogenetic information in comparative analysis has been recognized for some time, and several authors have attempted phylogenetic approaches to structure prediction (Muse, 1995; Gulko and Haussler, 1996; Akmaev *et al.*, 1999) . However, some of the techniques that have been developed before do not have automatic implementation of the phylogeny, and require manual intervention (James *et al.*, 1989; Woese *et al.*, 1983).

The usual approach to the statistical RNA structure prediction is to analyze columns of a multiple alignment, applying a standard log-likelihood ratio test. In essence, the log-likelihood ratio statistic compares if the sequence evolution of two positions (columns) is better described by the joint evolution model, or by the product of two independent evolution models applied to these two positions separately. This can be done with or without a phylogenetic tree describing the data. The simplest assumption is that the sequences have evolved from one common ancestor, and the time of evolution approaches infinity. In other words, this process assumes that the phylogeny of the sequences is a 'star' phylogeny with infinite branch lengths, and, essentially, disregards the phylogenetic relationships in the data. If the equilibrium probabilities of the nucleotides at each position are estimated by the sample frequencies, then the statistic, based on these assumptions, is essentially the Mutual Information (MI) statistic (Chiu and Kolodziejczak, 1991). Because MI ignores the phylogenetic information in the sequences, it tends to overestimate the amount of covariation between two positions, and accepts spurious correlations attributable to the phylogenetic relationships among the sequences (Lapedes *et al.*, 1999). A number of papers have illustrated how

inclusion of the phylogeny helps to reduce the effect of spurious correlations (Akmaev *et al.*, 1999; Gulko and Haussler, 1996).

The importance of the phylogenetic information has been well accepted among researchers. Unfortunately, the implementation of an evolution model, given a phylogenetic tree, is sometimes very complicated and requires a lot of computing power, especially for a joint evolution model of a pair of positions, and simplifying assumptions are usually made. For instance, a simplified joint distribution model was made by Muse (1995) and this method seems to work well in regions of RNA molecules with known base-pairs. Unfortunately, it is unclear how much noise the result would have for an unknown region, and whether it would be possible to distinguish the interacting and non-interacting pairs. Gulko and Haussler (1996) have implemented a true joint distribution model. However, the requirement of a training data set makes their approach difficult to use for molecules with unknown structure, and the dependence of the result on the training data might make the result itself inaccurate (e.g. the method might miss unusual interactions). We have proposed a method that combines a phylogenetic approach and a purely statistical procedure in one algorithm for RNA structure prediction (Akmaev *et al.*, 1999). The tests have shown that our method is superior to the MI methods and is much faster than any current application of a joint evolution model.

In this paper, we show some extensive analysis of the statistics, that appeared in a previous paper. Also, we consider an application of this approach to the prediction of triple interactions. Even though these methods have been shown to work well, we acknowledge that the implementation of the log-likelihood ratio statistic is a more conservative approach, and would allow us to get more accurate results. Thus, the final part of this paper is devoted to the joint evolution model formalism and implementation.

Statistical tools for analysis of RNA structure

Basic definitions

One of the tools we require is the ability to model sequence evolution at each position of a multiple alignment independently. A number of sequence evolution models have been developed in the past (reviewed by Swofford, 1998). The Jukes–Cantor one-parameter model is the simplest (Jukes and Cantor, 1969), while the GTR (Lanave *et al.*, 1984), general time-reversible model is the most general in the class of time-reversible models. Choosing between the complexity and accuracy, we settled on the Hasegawa–Kishino–Yano (HKY) model (Hasegawa *et al.*, 1985). The HKY model distinguishes between transitions and transversions, and has equilibrium probabilities for each nucleotide. This summarizes to a total of six parameters (two rate parameters and four equilibrium probabilities) for each column of a multiple alignment. The instantaneous rate matrix for this model is

$$Q = \begin{pmatrix} — & \alpha\pi_C & \beta\pi_G & \alpha\pi_U \\ \alpha\pi_A & — & \alpha\pi_G & \beta\pi_U \\ \beta\pi_A & \alpha\pi_C & — & \alpha\pi_U \\ \alpha\pi_A & \beta\pi_C & \alpha\pi_G & — \end{pmatrix} \qquad (1)$$

where α and β are the transversion and transition rates respectively, and π_as are the equilibrium probabilities of nucleotides a, where $a = A, C, G, U$.

The diagonal elements of the matrix are such that the sum of the elements of each row is zero. The question is how do we determine the parameters of the model? Usually, the equilibrium probabilities are chosen to be the frequencies of the nucleotides at each particular position, but these estimates disregard phylogenetic information among the sequences. Instead of the frequencies, our estimates of the probabilities and the rates are the maximum likelihood estimates, given a phylogenetic tree. The likelihood function of the data at each position can be computed for a particular set of parameters such as the parameters in the HKY model (Schadt *et al.*, 1998). If Q is the instantaneous rate matrix, the transition probability matrix would be $e^{(Qt)}$, where t is a branch length between two nodes on the phylogenetic tree. The likelihood function is then maximized with respect to its parameters. The aspects of the maximization procedure have already been discussed elsewhere (Tillier, 1994). Equation (2) defines the maximum likelihood statistic at position i

$$\begin{aligned} &L_i^{\mathrm{HKY}}(T) = \max_{\alpha,\beta,\pi_A,\ldots,\pi_U} L_i^{\mathrm{HKY}}(\alpha,\ldots,\pi_U|T) \\ &\alpha > 0, \beta > 0, 0 \le \pi_a \le 1 \\ &\sum_{a\in\{A,\ldots,U\}} \pi_a = 1 \end{aligned} \qquad (2)$$

$L_i^{\mathrm{HKY}}(\alpha,\ldots,\pi_U|T)$ is the likelihood function of the data at position i as a function of the parameters of the model (equation (1)), and T is a phylogenetic tree. Throughout this paper we consider double and triple interactions between positions of a multiple alignment. Let us define a few other useful statistics for positions i, j and k, that describe such interactions from a statistical point of view:

$$\begin{aligned} L_i &= \prod_{l=1}^{N} f_{a_l} = \prod_{a=\{A,C,G,U\}} f_a^{N\cdot f_a} \\ L_{ij} &= \prod_{a,b=\{A,C,G,U\}} f_{ab}^{N\cdot f_{ab}} \\ L_{ijk} &= \prod_{a,b,c=\{A,C,G,U\}} f_{abc}^{N\cdot f_{abc}} \end{aligned} \qquad (3)$$

N is the number of sequences, f_a is the frequency of nucleotide a, where $a = A, C, G, U$ at site i, f_{ab} is the frequency of nucleotides a and b at sites i and j, f_{abc} is the frequency of nucleotides a, b and c at sites i, j and k. L_i is the maximum likelihood (ML) function at position i, given that all the sequences are independent. L_{ij} is the ML function at two positions under the same assumption, and L_{ijk} is the ML function at three positions of the multiple alignment.

To see what impact the presence of phylogenetic information has on the statistical analysis, we first compare the ML function, given the HKY model, and the ML function, assuming independence of the sequences

$$T_i = \log \frac{L_i^{\mathrm{HKY}}(T)}{L_i} \tag{4}$$

The T_i statistic depends on two parameters: (1) the phylogenetic tree; (2) the data at position (i). Any tree that reasonably describes the multiple alignment would fit the data better than a 'star' phylogeny with infinite branches. Therefore, we expect T_i to be positive, given an appropriate phylogenetic tree, in other words, a tree that would be reasonably close to the true phylogeny. On the other hand, if we fix the tree we can elaborate what happens if we vary the data at this position. There are two limiting cases: when the position is completely conserved, and where there is so much variation that the assumption of the independence becomes reasonable. The T_i statistic is close to zero in both cases. In these two cases the tree does not contain any information of what has happened at this position evolutionarily. If we consider the intermediate case when T_i is positive, it means that the tree phylogenetically describes the evolution at this position as opposed to the independence case where the evolutionary dependencies are not accounted for.

It is worth noting that the distribution of T_i depends on the tree T only in the numerator. Hence, it would be different for distinct phylogenies, and, at this point, it seems impossible to establish the distribution of T_i theoretically.

Analysis of the interacting pairs

The idea that mutational changes at interacting positions in homologous sequences are statistically correlated, because these positions interact in the structure, is the main underlying assumption of all comparative methods. All these methods make a comparison between independent and joint evolution models. The R_{ij} statistic we have developed (Akmaev *et al.*, 1999), which is presented in the formula

$$R_{ij} = -\log \frac{L_i^{\mathrm{HKY}}(T) L_j^{\mathrm{HKY}}(T)}{L_{i|j} L_{j|i}} \tag{5}$$

has been shown to be a better indicator of the correlation between two columns of a multiple alignment than the standard comparative methods that ignore phylogeny.

$L_{i|j}$ is the ML function of the data at position i, given the data at position j, under the 'star' phylogeny assumption.

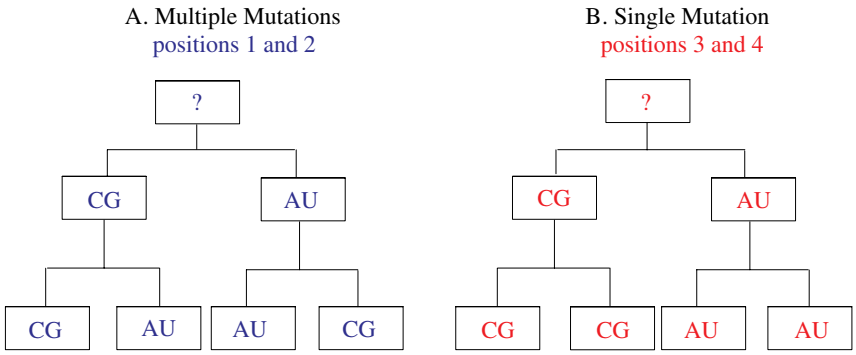
The R_{ij} statistic compares whether the data at positions i and j is better represented by the independent evolution model, or by the data at the other position. If two positions are correlated, the conditional MLs would be bigger than the independent, which would make R_{ij} positive. Why have we chosen to compare the product of the two independent MLs with the product of the two conditional ML functions, instead of using the more natural statistic shown here?

$$R'_{ij} = -\log \frac{L_i^{\mathrm{HKY}}(T) L_j^{\mathrm{HKY}}(T)}{L_{ij}} \tag{6}$$

To see the difference between these two statistics, we need to rewrite the expressions in terms of the basic blocks, defined in the previous section

$$\begin{aligned} R_{ij} &= 2 \cdot \log \frac{L_{ij}}{L_i L_j} - T_i - T_j \\ &= 2 \cdot N \cdot M_{ij} - T_i - T_j \\ R'_{ij} &= \log \frac{L_{ij}}{L_i L_j} - T_i - T_j \\ &= N \cdot M_{ij} - T_i - T_j \end{aligned} \tag{7}$$

M_{ij} is the MI at positions i and j. If two positions are independent, then R_{ij} is equal to R'_{ij} because MI, in this case, equals to zero. On the other hand, R_{ij} is bigger than R'_{ij} for a pair of correlated positions. This observation indicates that R_{ij} would probably better suit our purposes of sorting correlated and independent pairs. If we analyze the expression for the R_{ij} statistic, we would see that if the phylogenetic tree does not contain any information, i.e. it is close to a 'star' phylogeny with infinite branches, the last two terms disappear and R_{ij} is equivalent to the well-known χ^2 log-likelihood ratio statistic (Figure 1, positions 1 and 2). If the tree is critical (Figure 1, positions 3 and 4) and the positions are independent, then the first term vanishes and the R_{ij} becomes negative. If the two positions are correlated, the first term would be non-zero, as is supposed to happen with the log-likelihood ratio test. Figure 1 shows how the R_{ij} statistic uses the phylogenetic information in this analysis. In the case of multiple mutations (Figure 1A), the tree does not help to explain phylogenetic dependencies between these sequences. The T_1 and T_2 statistics are zero. On the other hand, it describes these dependencies well



No\Stat.	L_i	L_i^{HKY}	T_i	R_{ij}	R'_{ij}	M_{ij}
1	0.015625	0.015625	0	8.31776	4.15888	0.69315
2	0.015625	0.015625	0			
3	0.015625	0.066979	1.4555	5.40676	1.24788	0.69315
4	0.015625	0.066979	1.4555			

Fig. 1. Diagram showing the evolution of two pairs of positions down the same phylogenetic tree. All the branches are assumed to be equal. Positions 1 and 2 (A) show more evidence of correlated change than positions 3 and 4 (B). The table contains the values of six statistics for these positions.

for positions 3 and 4, and the T_i statistic is subsequently greater. Naturally, M_{ij} does not distinguish between these two cases, but R_{ij} has different values for the pairs (1,2) and (3,4). The correlation is more significant between sites 1 and 2. This separation between the values of R_{ij} for these two types of data becomes more noticeable with a bigger number of sequences.

If we compare the expressions in equations (7), we can easily derive the relationship between R'_{ij} and R_{ij}

$$R_{ij} = R'_{ij} - \log \frac{L_i L_j}{L_{ij}}$$
$$= -\log \frac{L_i^{\mathrm{HKY}}(T) L_j^{\mathrm{HKY}}(T)}{L_{ij}} - \log \frac{L_i Lj}{L_{ij}} \qquad (8)$$

It is interesting to examine why the addition of the log-likelihood ratio on the right-hand side of equation (8) makes the statistic more reliable. If we consider two cases, one in which the positions are independent, and another in which the positions are perfectly correlated, we can see that in the first case $R_{ij} = R'_{ij}$, while in the second case $R_{ij} > R'_{ij}$. This observation indicates that R_{ij} should give us a better separation between interacting and non-interacting positions than R'_{ij}. Ideally, we would like to have a joint evolution model in this log-likelihood ratio statistic

$$R_{ij}^{\mathrm{ideal}} = -\log \frac{L_i^{\mathrm{HKY}}(T) L_j^{\mathrm{HKY}}(T)}{L_{ij}^{\mathrm{joint}}(T)} \qquad (9)$$

The R'_{ij} statistic uses an approximation to the joint model likelihood function, which, in essence, is just the ML function of a pair of positions given a 'star' phylogeny with infinite branch lengths. Therefore, R'_{ij} would be less than R_{ij}^{ideal} for reasonable phylogenetic trees. Thus, the last term in equation (8) compensates this approximation for correlated positions, and does not change it for non-interacting pairs. This, in turn, gives a clearer separation between the distributions of the R_{ij} statistic for interacting and non-interacting positions. We have performed a number of tests that showed the R_{ij} statistic gives much better results than the R'_{ij} statistic without increasing the number of false positives (Akmaev *et al.*, 1999, and other unpublished results).

To summarize, the R_{ij} method has proven to be a good first-order approximation to the standard log-likelihood ratio statistic involving a joint evolution model. This approach allows us to better isolate correlated positions out of the total number of pair-wise combinations that we can then explore in more detail with a better model (e.g. joint evolution).

A phylogenetic approach to base-triple prediction

Base-triples are among the essential tertiary interactions in RNA structure. A number of base-triple predictions have been made in tRNA, group I introns, 16S and 23S RNA (Levitt, 1969; Gutell *et al.*, 1994; Michel and Westhof, 1990; Gautheret *et al.*, 1995). Although

methods have improved sufficiently to identify some of the tertiary interaction in these molecules, prediction of base-triples is still not a very reliable procedure. Most base-triple predictions have been made using pair-wise comparative analyses. The principle works very well in the detection of Watson–Crick pairs and other secondary structure interactions such as GU pairs or tetraloops. Unfortunately, it has been shown that base-triples often lack the strict pattern of covariation observed in RNA base-pairs. For example, the interaction of a Watson–Crick base-pair with a third base occurs through different types of non-canonical interactions, such as Hoogsteen pairing (Hayashi *et al.*, 1982), so that a mutation in one position does not necessarily imply mutations in the other positions (Gutell *et al.*, 1994). That is why these rare simultaneous mutations require more delicate handling, and incorporation of phylogenetic information in the analysis becomes even more important for base-triple prediction. A number of attempts to construct an algorithm for triple interaction predictions have been made using mathematical (Gautheret *et al.*, 1995) and biochemical (Conn *et al.*, 1998) models. Here we present a statistical formalism for the analysis of base-triple interactions in alignments of homologous RNA sequences that exploits the phylogenetic relationships among the sequences.

Once again, the ideal χ^2 test would be to consider the log-likelihood ratio statistic for triple interactions

$$S_{ijk}^{\text{ideal}} = -\log \frac{L_i^{\text{HKY}}(T) L_j^{\text{HKY}}(T) L_k^{\text{HKY}}(T)}{L_{ijk}^{\text{triple}}(T)} \tag{10}$$

Unfortunately, implementation of a joint evolution model for triples appears even less realistic than for doubles. Hence, we need to make an approximation for the ML function under the joint model L_{ijk}^{triple}. There are a number of possible ways to approximate this statistic. The three formulas of equation (11) represent some of them

$$\begin{aligned} S'_{ijk} &= -\log \frac{L_i^{\text{HKY}}(T) L_j^{\text{HKY}}(T) L_k^{\text{HKY}}(T)}{L_{ijk}} \\ S''_{ijk} &= -\log \frac{L_i^{\text{HKY}}(T) L_j^{\text{HKY}}(T) L_k^{\text{HKY}}(T)}{L_{i|jk} L_{j|ik} L_{k|ij}} \\ S_{ijk} &= -\log \frac{L_i^{\text{HKY}}(T) L_j^{\text{HKY}}(T) L_k^{\text{HKY}}(T)}{L_{jk|i} L_{ik|j} L_{ij|k}} \end{aligned} \tag{11}$$

L_{ijk} is the joint ML of three positions under the assumption of a 'star' phylogeny with infinite branches. $L_{i|jk}$ is the ML of the data at position i, given the data at positions j and k. This statistic represents how well the data at one site is predicted by the data at two other sites. $L_{ij|k}$ is the ML at positions i and j, conditioned on the data at position k. This is a measure of the prediction of the data at positions i and j from the data at position k. The first statistic approximates the joint triple evolution model by a 'star' phylogeny with infinite branches and estimates the equilibrium probabilities by the frequencies of the nucleotides. The second statistic compares the independent ML under the evolution model to the ML conditioned on the other two positions, similarly to the R_{ij} statistic. The last one has the same idea as the second one but the ML function of the data at two positions is conditioned on the data at the third position. To see more clearly the difference between these three statistics, we express them in terms of the basic blocks, defined earlier

$$\begin{aligned} S'_{ijk} &= -\log \frac{L_i L_j L_k}{L_{ijk}} - T_i - T_j - T_k \\ S''_{ijk} &= -3 \cdot \log \frac{L_i L_j L_k}{L_{ijk}} - T_i - T_j - T_k \\ &\quad -N \cdot (M_{ij} + M_{ik} + M_{jk}) \\ S_{ijk} &= -3 \cdot \log \frac{L_i L_j L_k}{L_{ijk}} - T_i - T_j - T_k \\ &\quad + \log(L_i L_j L_k) \end{aligned} \tag{12}$$

To understand how each of these statistics behave, consider the two opposite cases: (1) these three positions are independent; (2) they are perfectly correlated. In the first case, the first term in each expansion disappears, since L_{ijk} is equal to the product of the three independent MLs ($L_i L_j L_k$). The last term in the S''_{ijk} statistic also vanishes because MI for two independent positions is equal to zero. Thus, the formulae simplify to

$$\begin{aligned} S'_{ijk} &= -T_i - T_j - T_k \\ S''_{ijk} &= -T_i - T_j - T_k \\ S_{ijk} &= -T_i - T_j - T_k + \log(L_i L_j L_k) \end{aligned} \tag{13}$$

It is clearly seen that $S_{ijk} \leq S'_{ijk} = S''_{ijk}$, when the positions are all independent. The difference is the entropy of the distributions at positions i, j, and k that disregard phylogeny.

In the second case, when the positions are perfectly correlated, the independent MLs are equal to each other ($L = L_i = L_j = L_k$). The joint ML functions are also equal to L, namely, $L_{ijk} = L$ and $N \cdot M_{ij} = -\log L$. Thus, formulas can be rewritten as

$$\begin{aligned} S'_{ijk} &= -2 \cdot \log L - T_i - T_j - T_k \\ S''_{ijk} &= -3 \cdot \log L - T_i - T_j - T_k \\ S_{ijk} &= -3 \cdot \log L - T_i - T_j - T_k \end{aligned} \tag{14}$$

Equations (14) show the following relationship between these statistics for perfectly correlated positions, $S_{ijk} = S''_{ijk} \geq S'_{ijk}$. Another interesting case worth considering is when two positions are correlated and the third base

is independent of the first two. After some algebra, the conclusion is that $S_{ijk} \leq S'_{ijk} \leq S''_{ijk}$, if we assume that positions i and j form a base-pair, and position k is independent of the other two. This shows the tendency of the S''_{ijk} statistic to overestimate the amount of covariation between three positions if two of these positions covary, even though there is the explicit subtraction of MI in the expression. These calculations demonstrate that the S_{ijk} statistic has bigger values for interacting positions and lower values for non-interacting positions out of these three statistics. Therefore, we assume it would allow us to get better predictions.

Figure 2 shows how these three statistics and the MI statistic for three positions behave in a region of 23S rRNA with a known base-triple (Conn *et al.*, 1998). The four histograms in Figure 2 present the output of these statistics for this region. Figure 2A is the S_{ijk} statistic, B is S'_{ijk}, C is S''_{ijk}, and D is the MI statistic, which does not incorporate phylogenetic information. The phylogeny of this data resembles a balanced tree with uniform branches, hence the MI statistic is expected to work well on this kind of data set, and, indeed, MI outperforms S'_{ijk} and S''_{ijk} (Figure 2B, C, and D). Although if we were to detect this triple, we would have to accept about 40 false triples, which is not a decent signal-to-noise ratio. The S_{ijk} statistic, on the other hand, has a range of values much wider than S'_{ijk} and S''_{ijk} (these three histograms are on the same scale). This, in turn, indicates that the S_{ijk} statistic has much better specificity than the other statistics. The test shows that the real triple is distinguishable with the S_{ijk} statistic.

Certainly we realize that this might have happened accidentally due to the nature of this particular data set. So we tested it on a number of other data sets that have known triples. Figure 3 shows the result for a tRNA base-triple prediction. Figure 3A shows the results of the S_{ijk} statistic applied to two tRNA data sets: aspartic acid tRNA and phenylalanine tRNA, each containing around 80 sequences taken from eukaryotes, eubacteria, and archae, and also including mitochondria and chloroplast sequences. The analysis disregarded positions that had less than 10% variation (i.e. at least 90% of the sequences had the same nucleotide). That is why only two triples were analyzed, though these two types of tRNA have more base-triples. The problem with tRNA data sets is that non-synonymous tRNAs have slightly different tertiary structures (Gautheret *et al.*, 1995). Therefore, the number of homologous molecules available to us is sometimes not enough to apply comparative analysis due to the lack of variation at positions of interest. However, the results show that the S_{ijk} statistic is able to distinguish triple interactions even in small data sets. The same type of statistic, disregarding phylogenetic information, has also been applied to these tRNA sets

$$M'_{ijk} = -\log \frac{L_i L_j L_k}{L_{ij|k} L_{ik|j} L_{jk|i}} \qquad (15)$$

The effect of phylogenetic information, used in this test, is not dramatic but is still apparent. If we were to catch all true positives, we would get about 15 false positives in the case of the S_{ijk} statistic (Figure 3A), and about 50 in the case of the M'_{ijk} statistic (Figure 3B), even though the reliability of the trees, based on 80 positions, is doubtful. The S'_{ijk} and S''_{ijk} statistics performed much poorer than either of them.

Discussion

Base-triple prediction is a much more delicate procedure than prediction of pair interactions. As has been shown in this section, not every statistic would be very useful or effective for this analysis, and, obviously, more extensive incorporation of phylogenetic information in the analysis would greatly improve triple prediction. Even the way the phylogenetic information is used in the S_{ijk} statistic, improves the prediction accuracy over a standard frequency-based approach. This method appears to have better specificity and sensitivity than methods that disregard phylogenetic information. Certainly, implementation of a joint triple evolution model would be a much more rigorous approach, but the current state of computing facilities does not make such an approach reasonable at the moment. However, there are a number of different auxiliary methods that might aid to confirm these predictions. Gautheret *et al.* (1995) considered some of these methods, and proposed the neighbor-effect model which could provide more evidence for triple prediction. Though the neighbor-effect statistic is not a perfect indicator of a base-triple presence, it has been shown to be effective in some of the triple predictions of tRNA and regions of group I introns.

It is remarkable that triples with S_{ijk} statistic values close to the real triples usually come from two or more base-pairs. If two base-pairs show correlation due to different factors not attributable to the physical interaction, this would give four possible triple combinations with values well above the values for the pool of the independent positions. Fortunately, this type of correlation can be easily detected and discarded. It is worth mentioning that these statistics can not be successfully applied to positions with a significant number of gaps. If one needs to analyze this type of position, then the sequences that do not have gaps at these sites have to be separated from the data set, and analyzed independently.

Statistic based on joint evolution model

Although the methods proposed in the previous section are robust, it seems unlikely that the distribution of any

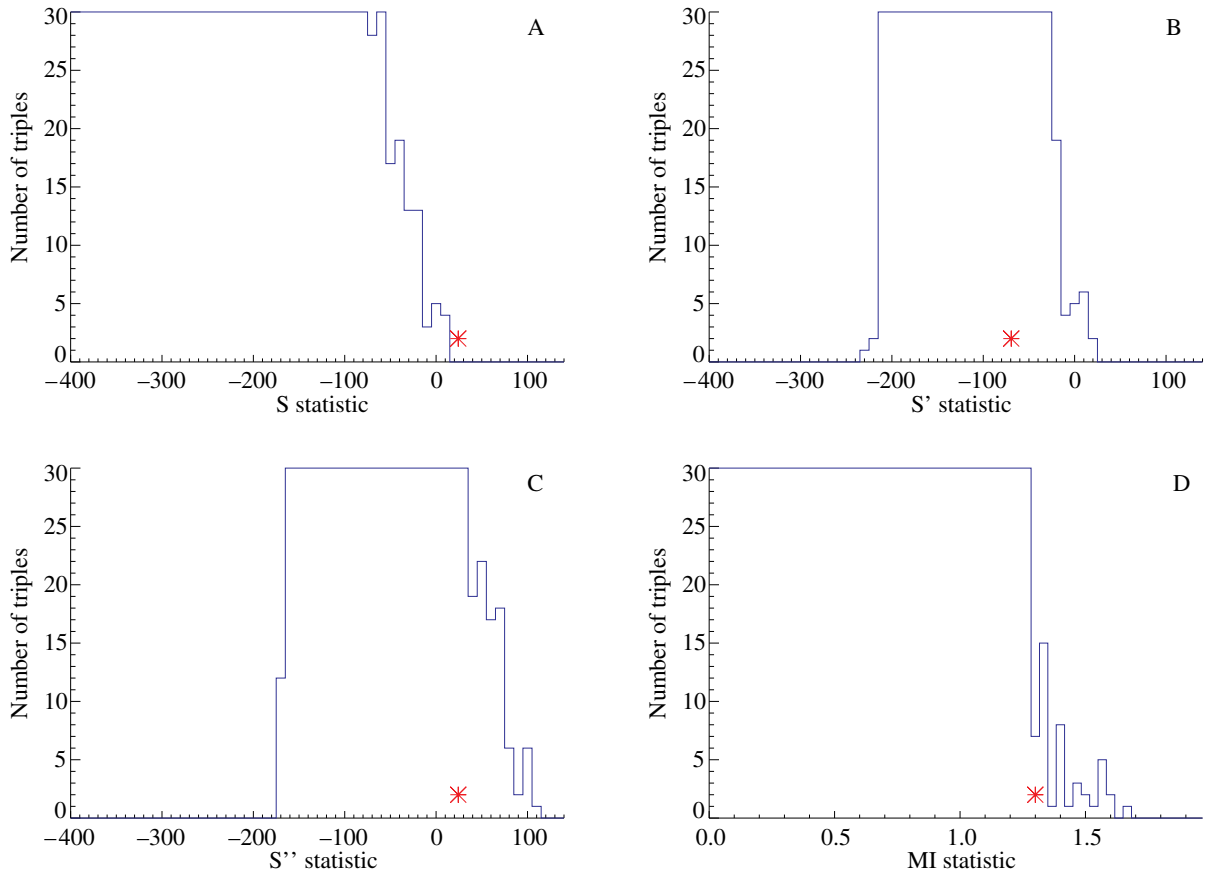


Fig. 2. Histograms showing the behavior of four different statistics in a region of 23S rRNA. All possible triples were used to make the histograms, the plots were cut at 30. The values of each statistic for a known triple are marked by asterisks.

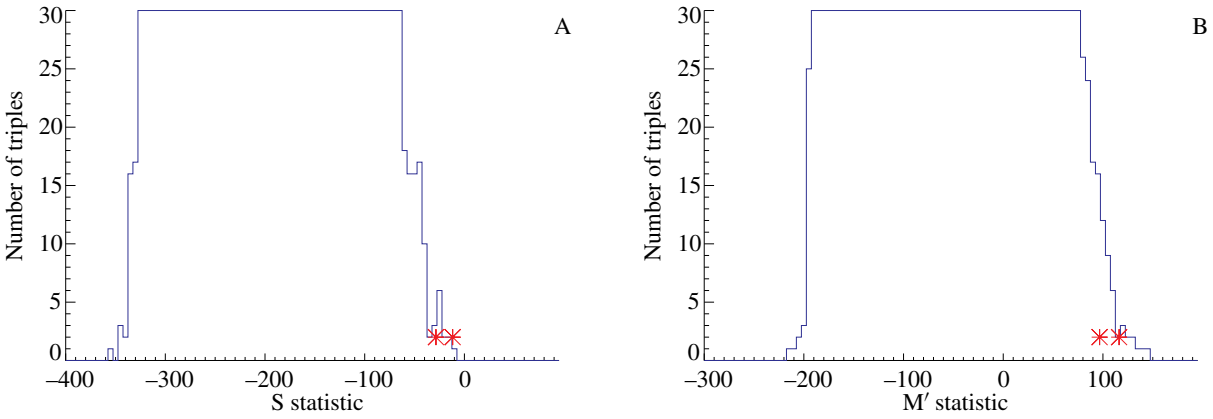


Fig. 3. Histograms showing the output of the S_{ijk} statistic and the same type MI statistic for tRNA data sets. All possible triples were used to make the histograms, the plots were cut at 30. The values of each statistic at known triples are marked by asterisks.

of those statistics will be established. As we have already mentioned, all these statistics depend on the particular phylogenetic tree only in the numerator. Therefore, the distribution of each statistic also depends on the tree, which makes it impossible to find out statistical properties of these distributions. Thus, if we would like to have an

ability to base our analysis on a well-known distribution, it is necessary to consider a standard log-likelihood ratio test. These tests are usually used in cases where there is a needs to decide whether two events are independent or not. Equation (16) shows this statistic

$$\chi^2 \approx -2 \cdot \log \frac{L_i^{\text{ind}} L_j^{\text{ind}}}{L_{ij}^{\text{joint}}} \tag{16}$$

The number of degrees of freedom for this statistic depends on the number of estimated parameters in the numerator and denominator. In general, it is the difference between the number of independent parameters in the denominator and the number of independent parameters in the numerator. This idea applies in the case of the MI statistic for RNA data sets, which is approximately a multiple of a χ^2 distribution with 9 degrees of freedom.

Unfortunately, the MI methods disregard phylogenetic information. In this section, we present a test which is similar to the MI statistic, but incorporates the phylogeny of the data in the independent likelihood functions as well as in the joint likelihood function.

Joint evolution model for a pair of positions

In the case of independent likelihood functions (equation (16)), we use the ML functions given the HKY model, namely L_i^{HKY} and L_j^{HKY}. The hard question is the joint evolution model. By the joint evolution model, we refer to a model that describes the evolution of two positions down a phylogenetic tree. If in the case of independent models we had four states of the system (A, C, G, U), here the number of states increases quadratically and raises from 4 to 16 $(AA, AC, \ldots, UU)$. But this is not the biggest problem: what is more important is the fact that in order to get an approximate χ^2 statistic, the joint model has to be consistent with the independent evolution model. This means that if we use the HKY evolution model for each position separately, we are also supposed to use the same kind of model for a pair of positions. This brings us to the following instantaneous substitution rate matrix

$$Q^{\text{joint}} = \begin{pmatrix} — & \alpha\pi_{AC} & \beta\pi_{AG} & \alpha\pi_{AU} & \alpha\pi_{CA} & \cdots \\ \alpha\pi_{AA} & — & \alpha\pi_{AG} & \beta\pi_{AU} & 0 & \cdots \\ \beta\pi_{AA} & \alpha\pi_{AC} & — & \alpha\pi_{AU} & 0 & \cdots \\ \alpha\pi_{AA} & \beta\pi_{AC} & \alpha\pi_{AG} & — & 0 & \cdots \\ \alpha\pi_{AA} & 0 & 0 & 0 & — & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ 0 & 0 & 0 & \alpha\pi_{AU} & 0 & \cdots \end{pmatrix} \tag{17}$$

As in the independent model, α is the transversion rate, β is the transition rate. π_{ab}s are the equilibrium probabilities of nucleotide pairs. These parameters sum to 18 unknown parameters, although only 17 of them are independent, since the equilibrium probabilities sum to 1. The diagonal elements of the matrix are such that the sum of all elements in each row is zero. Each element Q_{ij}^{joint} is the instantaneous rate of mutation from state i to state j. The states are numbered in the natural order: $AA, AC, AG, AU, CA, CC, \ldots, UU$. In this model, we assume that it is impossible to have a double mutation in one instant (e.g. the instantaneous substitution rate from state AA to state UU equals zero). It is also worth mentioning that this model assumes the same transition and transversion rates for both positions. This assumption is reasonable because we use this model for base-pair identification, and as we saw in numerous tests, positions that form a base-pair tend to have approximately equal rates. Moreover, this assumption helps us to get rid of two extra parameters.

The transition probability matrix for this instantaneous rate matrix would be

$$P^{\text{joint}}(t) = e^{Q^{\text{joint}} \cdot t} \tag{18}$$

Unfortunately, this matrix (each element of which is a function of time) can not be found explicitly as a function of this model's parameters. This implies that for each set of parameters it is necessary to perform the complete eigensystem decomposition of the matrix Q^{joint}. Fortunately, there is a way to reformulate this problem in terms of a symmetric matrix (we would like to thank Alan Lapedes for this idea). Eigensystem decomposition for a symmetric matrix is a much easier and faster procedure, and there are numerous methods that deal with this problem (Atkinson, 1988).

Once we have calculated the transition probability matrix, exactly the same procedure, as in the case of one position, may be applied to find the likelihood function of the data at two positions given this joint sequence evolution model.

Log-likelihood ratio statistic

Before we use the joint likelihood model described in the previous section, there is still the question of the unknown parameters of this model. In the case of the independent model, we estimated the parameters by the maximum likelihood estimates. To be consistent with this approach, we have to estimate the parameters of the joint model by the maximum likelihood estimates, namely

$$L_{ij}^{\text{HKY}}(T) = \max_{\alpha,\beta,\pi_{AA},\ldots,\pi_{UU}} L_{ij}^{\text{HKY}}(\alpha, \ldots, \pi_{UU} | T) \tag{19}$$

This maximization is, obviously, the most time-consuming part of the algorithm. Keeping in mind that neither gradient nor Hessian matrix are available for this function, there are a very limited number of optimization methods

that may be applied here. Moreover, there is one more complication in this problem, namely that the parameters have very strict constraints

$$\begin{aligned} &\alpha > 0, \beta > 0 \\ &0 \le \pi_{ab} \le 1, ab \in \{AA, \ldots, UU\} \\ &\sum_{ab \in \{AA,\ldots,UU\}} \pi_{ab} = 1 \end{aligned} \tag{20}$$

Since this number of parameters is usually overfitting the data at any particular position, the set of numbers that maximizes the likelihood function is, in most cases, on the boundary of the feasible region. Thus, application of Newton's method with the approximated gradient vector and the Hessian matrix is not possible in this case. Instead of this, we have implemented Powell's direction set method (Brent, 1973). Unfortunately, this method alone usually gets stuck, not reaching the maximum, because of the constraints of the problem (equations (20)). However, the combination of this method with the simplex method works very well (Nelder and Mead, 1965), and requires just about 5 min on a Sun Ultra 30 station to maximize the likelihood function for a pair of positions in a data set of about 150–200 sequences.

Given the maximum likelihood estimates, we can consider the following log-likelihood ratio test

$$H_{ij} = -2 \cdot \log \frac{L_i^{\mathrm{HKY}}(T) L_j^{\mathrm{HKY}}(T)}{L_{ij}^{\mathrm{HKY}}(T)} \tag{21}$$

This statistic would be an asymptotic χ^2 test if the samples were independent. However, if the number of sequences is fairly big, it seems reasonable to suggest that at least some groups of these sequences are independent. So this requirement of independence does not appear to be prohibitive in this case. One also needs to say that, in case of two independent positions, this statistic may have negative values because of the equal-rates assumption in the joint model. Therefore, this statistic is approximately a χ^2 statistic for positions with equal-rate parameters (though this still needs a rigorous proof). To find out the number of degrees of freedom, one needs to calculate the number of independent parameters in the numerator and the denominator (equation (21)). The numerator has eight equilibrium probabilities and four rate parameters. However, only six of these probabilities and two rate parameters are independent, so this sums up to eight independent parameters. The denominator has 18 parameters, 17 of which are independent. Hence, the assertion is that H_{ij} is approximately a χ^2 statistic with 9 degrees of freedom in the case of two positions with equal transition and transversion rates.

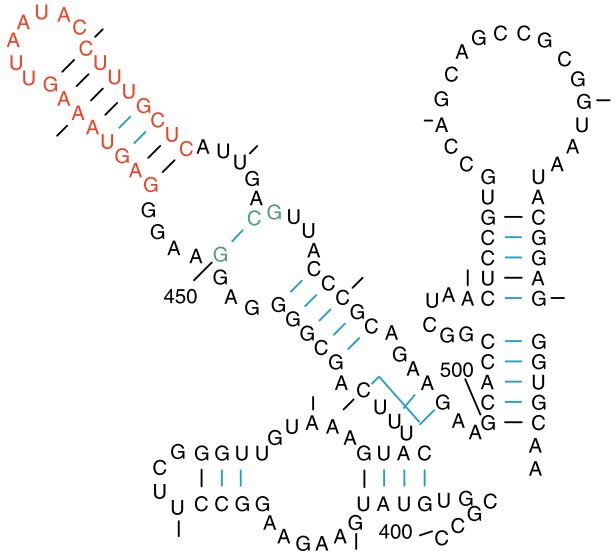


Fig. 4. A small portion of the *E.coli* 16S rRNA, starting from position 400 to position 547. The region starting at position 455 and up to position 477 is the variable stem–loop region. The base-pairs that have been analyzed by the R_{ij} and H_{ij} methods are indicated by light connecting lines. The dark lined base-pairs (except the variable stem–loop region) did not show up in our analysis because one or both positions were conserved in the data set we used.

H_{ij} *statistic test*

To test the H_{ij} statistic, we examined the same set of 16S ribosomal RNA as in our previous paper (Akmaev *et al.*, 1999). Sequence alignment for this region of 16S rRNA was downloaded from the world wide web (Van de Peer *et al.*, 1998). This alignment has approximately 150 bacterial sequences from 12 bacterial families. It is worth mentioning that the phylogenetic tree we used was generated using all positions of the 16S rRNA multiple alignment for better accuracy. In previous work, we showed by using this data set that the R_{ij} statistic outperforms the MI statistic (Akmaev *et al.*, 1999). Here, we compare the R_{ij} and the H_{ij} methods to see if the H_{ij} statistic, besides having a statistical significance measure, makes more accurate predictions and is worth the trouble of its calculation.

Figure 4 shows the particular region of the 16S rRNA we were interested in. Unfortunately, we could not consider all existent base-pairs because some of these positions were more than 95% conserved, and these were disregarded in the analysis. The positions 455 through 477 form a variable stem–loop region that was omitted in our previous work. Predictions inside this variable region are complicated by the fact that about half of the sequences have gaps at these positions.

As mentioned previously, we consider the R_{ij} statistic as

i	j	R_{ij}	H_{ij}	i	j	R_{ij}	H_{ij}
406	436	109.3	105	502	543	107.3	103
407	435	23.8	40	503	542	58.3	65
408	434	-5.6	42	504	511	-25	5
416	421	-25.9	-4	504	534	-30.2	-1
416	482	-29.1	-4	504	540	-19.1	4
417	426	35.3	35	504	541	12	11
418	425	60.9	57	511	540	28	29
421	526	-43.2	-4	511	541	-25	6
422	482	-33.4	3	513	538	72.2	74
427	446	-41.4	11	514	532	-35.9	2
427	494	-37.7	8	514	537	5.7	15
438	496	29.6	39	534	541	-30.2	-2
440	497	-28	46	540	541	-19.1	5
442	492	74	104	457	473	-102.8	10
443	491	16.9	84	457	474	-127.7	6
444	490	94	110	457	475	-56.5	29
445	489	64	60	457	476	-115.3	18
446	488	-5.7	18	457	489	-189.6	2
450	483	17.6	33	457	490	-245.6	0.1
483	484	-29.6	29	457	491	-241.3	14
486	504	12	10	458	473	-77.3	4
486	511	-25	6	458	474	-50	24
486	534	-30.2	-1	458	475	-94	13
486	540	-19.1	7	458	476	-141.6	0.1
486	541	12	9	458	489	-189.8	0.3
501	544	42.8	37	458	490	-218	2

Fig. 5. This table shows the values of R_{ij} and H_{ij} statistics for pairs within a region of 16S rRNA (Figure 4) that have R_{ij} above -50. The positions colored blue form base-pairs. The green values of the statistics are correct predictions, the red numbers are false predictions. In the bottom-right corner of the table, we also consider pairs from the variable stem–loop region, which is missing in many of the sequences, and compare the results for H_{ij} and R_{ij}.

the first-order approximation to the joint model approach. To see what difference the implementation of a joint evolution model makes (the H_{ij} statistic), we extended the acceptance threshold to make sure we included all known base-pairs. The table of results for R_{ij} and H_{ij} methods is shown in Figure 5. If we assume 95% significance for the H_{ij} statistic, this would set the threshold at about 17 (assuming that H_{ij} has a χ^2 distribution with 9 degrees of freedom). This means that if the value of H_{ij} is >17 we reject the null hypothesis (that the positions are independent), and if it is <17, then we accept it. In this case, we can see that among these pairs (ignoring the variable region for the moment) H_{ij} rejects two true positives (pairs 504–541, and 514–537) and accepts one false positive (pair 483–484), although it is seen that the values for 504–541 and 514–537 are the highest among the values for independent pairs. On the other hand, if we set the threshold for R_{ij} at zero, then the R_{ij} method rejects three true positives (408–434, 440–497, 446–488) and accepts two false positives (486–504, 486–541). Thus, the H_{ij} statistic is about 5% more accurate than R_{ij}.

In Figure 5 the lower-right corner of the table represents analysis of the variable stem–loop region. We took two positions (457 and 458) and calculated H_{ij} and R_{ij} for pairs of these two positions with positions inside the variable region and with positions outside this region.

One of the major weaknesses of the R_{ij} statistic is that, although positions with lots of gaps still have the largest R_{ij} values with the correct pair (e.g. 457–475, 458–474; Figure 5), these values are still indistinguishable from the independent positions without gaps. There are a number of ways to treat gaps in these methods. For instance, a gap may be considered as a fifth character, or it may be removed from the analysis completely. Unfortunately, there is no way to make a consistent implementation of gaps for the denominator and the numerator of the R_{ij} statistic. The values of the R_{ij} statistic, which are shown in Figure 5, were obtained by implementing gaps as an extra character in the conditional likelihoods, and disregarding them in the independent likelihood functions (equation (5)). This is a more conservative approach that does not result in many false positives at those positions. On the other hand, the H_{ij} statistic is implemented in such a way as to treat this problem consistently, and is able to predict these two base-pairs on the same scale. From this reasoning, it is clear that H_{ij} should outperform R_{ij} in the very important case of positions with gaps, and our tests show this to be the case (Figure 5). Certainly, the power of the test drops with the decreased amount of information (half of the sequences we used have gaps at positions 457 and 458), and there is a false prediction even in this small example, but the improvement of H_{ij} over R_{ij} in this case of numerous gaps is remarkable.

Discussion

Our analyses indicate that the H_{ij} statistic does improve the accuracy of predictions over R_{ij} especially in the case of gaps. One of the main advantages of the H_{ij} statistic is that there is a way to establish a significance of this test (even though some assumptions are required). In the case of R_{ij}, it is really impossible to say if a value of -50 in one data set is less significant than a value of 50 in another data set. This is why the question of a threshold for the R_{ij} statistic arises for each particular set of sequences. This problem is even more noticeable for the base-triple technique that we presented in the previous section. Moreover, as we said before, the H_{ij} test treats positions with gaps properly. To illustrate how important this problem is one might say that just 5% of gaps can make the result of the R_{ij} statistic so biased that it would be impossible to make any predictions at this position.

Although H_{ij} improves predictions accuracy somewhat, it is not clear that this improvement is enough to compensate for the time necessary to evaluate all pairs of positions. Instead we recommend a first pass with R_{ij}, followed by the more robust H_{ij} test on the higher R_{ij} correlations.

Acknowledgements

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

References

Akmaev,V.R., Kelley,S.T. and Stormo,G.D. (1999) A phylogenetic approach to RNA structure prediction. In *Proceedings of the Seventh International Conference on Intelligent Systems forMolecular Biology,* pp. 10–17.

Atkinson,K.E. (1988) *An Introduction to Numerical Analysis*. 2nd edn, John Wiley, New York, NY.

Brent,R.P. (1973) *Algorithms for Minimization Without Derivatives*. Prentice-Hall, Englewood Cliffs, NJ.

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,* pp. 75–80.

Chiu,D.K. and Kolodziejczak,T. (1991) Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci.*, **7**, 347–352.

Conn,G.L., Gutell,R.R. and Draper,D.E. (1998) A functional ribosomal RNA tertiary structure involves a base-triple interaction. *Biochemistry*, **37**, 11980–11988.

Gautheret,D., Damberger,S.H. and Gutell,R.R. (1995) Identification of base-triples in RNA using comparative sequence analysis. *J. Mol. Biol.*, **248**, 27–43.

Gulko,B. and Haussler,D. (1996) Using multiple alignments and phylogenetic trees to detect RNA secondary structure. *Pac. Symp. Biocomput.*, 350–367.

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

Gutell,R.R., Larsen,N. and Woese,C.R. (1994) Lessons from an evolving RNA: 16S and 23S structures from a comparative perspective. *Microbiol. Rev.*, **58**, 10–26.

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

Han,K. and Kim,H. (1993) Prediction of common folding structure of homologous RNAs. *Nucl. Acids Res.*, **21**, 1251–1257.

Hasegawa,M., Kishino,H. and Yano,T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **21**, 160–174.

Hayashi,Y., Nakanishi,M., Tsuboi,M., Fukui,T., Ikehara,M., Tazawa,I. and Inoue,Y. (1982) Hydrogen exchange kinetics of nucleic acids. Double and triple helices with Hoogsteen-type basepairs. *Biochim. Biophys. Acta*, **698**, 93–99.

Jacobson,A.B. and Zuker,M. (1993) Structural analysis by energy dot plot of a large mRNA. *J. Mol. Biol.*, **233**, 261–269.

James,B.D., Olsen,G.J. and Pace,N.R. (1989) Phylogenetic comparative analysis of RNA secondary structure. *Methods Enzymol.*, **180**, 227–239.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In Munro,H.N. (ed.), *Mammalian Protein Metabolism* Academic Press, pp. 21–32.

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. and Stormo,G.D. (1999) Correlated mutations in protein sequences: phylogenetic and structural effects. In *Proceedings of the IMS/AMS 1997 International Conference on Statistics in Computational Molecular Biology.* Vol. 33, Monograph Series of the Institute for Mathematical Statistics, Hayward, CA, pp. 236–256.

Le,S.Y. and Zuker,M. (1991) Predicting common foldings of homologous RNAs. *J. Biomol. Struct. Dyn.*, **8**, 1027–1044.

Levitt,M. (1969) Detailed model for transfer ribonucleic acid. *Nature (Lond.)*, **224**, 759–763.

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**, 585–610.

Muse,S.V. (1995) Evolutionary analyses of DNA sequences subject to constraints on secondary structure. *Genetics*, **139**, 1429–1439.

Nelder,J.A. and Mead,R. (1965) *Computer J.*, **7**, 308.

Schadt,E.E., Sinsheimer,J.S. and Lange,K. (1998) Computational advances in maximum likelihood methods for molecular phylogeny. *Genome Res.*, **8**, 222–233.

Reviewed by Swofford,D. (1998) *PAUP*: Phylogenetic analysis using parsimony (*and other methods)*. Massachussetts, Sunderland, Sinauer Associates, version 4.

Tillier,E.R.M. (1994) Maximum likelihood with multiparameter models of substitution. *J. Mol. Evol.*, **39**, 409–417.

Tinoco,I.Jr., Uhlenbeck,O.C. and Levine,M.D. (1971) Estimation of secondary structure in ribonucleic acids. *Nature*, **230**, 363–367.

Van de Peer,Y., Caers,A., De Rijk,P. and De Wachter,R. (1998) Database on the structure of small ribosomal subunit RNA. *Nucl. Acids Res.*, **26**, 179–182.

Winker,S., Overbeek,R., Woese,C.R., Olsen,G.J. and Pfluger,N. (1990) Structure detection through automated covariance search. *Comput. Appl. 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.

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