A computer method for finding common base paired helices in aligned sequences: application to the analysis of random sequences

L.Chan¹, M.Zuker² and A.B.Jacobson*

Department of Microbiology, ¹Department of Computer Science, State University of New York at Stony Brook, Stony Brook, NY 11794, USA and ²The Institute for Biological Sciences, National Research Council of Canada, Ottawa K1A-0R6, Canada

Received September 11, 1990; Revised and Accepted December 11, 1990

ABSTRACT

We describe a new computer program that identifies conserved secondary structures in aligned nucleotide sequences of related single-stranded RNAs. The program employs a series of hash tables to identify and sort common base paired helices that are located in identical positions in more than one sequence. The program gives information on the total number of base paired helices that are conserved between related sequences and provides detailed information about common helices that have a minimum of one or more compensating base changes. The program is useful in the analysis of large biological sequences. We have used it to examine the number and type of complementary segments (potential base paired helices) that can be found in common among related random sequences similar in base composition to 16S rRNA from *Escherichia coli*. Two types of random sequences were analyzed. One set consisted of sequences that were independent but they had the same mononucleotide composition as the 16S rRNA. The second set contained sequences that were 80% similar to one another. Different results were obtained in the analysis of these two types of random sequences. When 5 sequences that were 80% similar to one another were analyzed, significant numbers of potential helices with two or more independent base changes were observed. When 5 independent sequences were analyzed, no potential helices were found in common. The results of the analyses with random sequences were compared with the number and type of helices found in the phylogenetic model of the secondary structure of 16S ribosomal RNA. Many more helices are conserved among the ribosomal sequences than are found in common among similar random sequences. In addition, conserved helices in the 16S rRNAs are, on the average, longer than the complementary segments that are found in comparable random sequences. The significance of these results and their application in the analysis of long non-ribosomal nucleotide sequences is discussed.

INTRODUCTION

The function of many large single-stranded RNAs is thought to be determined, at least in part, by their structure. A variety of methods have been developed to study the structure of large RNAs. The methods include theoretical approaches, such as computer modeling and comparative analysis of related sequences; and experimental approaches, such as chemical modification, chemical crosslinking, and electron microscopy. Most of these methods give information about secondary structure primarily, and each of the methods gives only partial information about the structure of the molecules being studied. For this reason, the data are most useful when several methods are used in conjunction with one another.

The phylogenetic method of studying RNA folding is based on the premise that structures with functional significance are conserved in evolution. Although, in some instances, hairpins with identical sequences can be conserved in distantly related RNAs, most often the nucleotide sequences of conserved helices will differ from one another but changes in base composition on one side of a helix will be compensated for by matching changes in base composition on the opposite side of the same helix. These changes are known as compensating base changes. It has been suggested that a given helix is proven when two or more independent compensating base changes can be demonstrated in related sequences (1). The method has proved to be useful in the analysis of secondary structure of ribosomal RNAs (1,2,3,4), in transfer RNAs (5), in snRNAs (6), in class I introns (7,8), and in the analysis of the conformation of the RNA moiety of the enzyme RNase P (9).

In our laboratory we are studying the structure and function of single-stranded RNAs from the RNA coliphages (10). As part of this work, we have been developing a variety of computer tools that facilitate analysis of the structure of the RNAs. Here, we describe a computer program that enables us to scan large, aligned sequences rapidly in order to identify conserved helices with compensating base changes. The program was developed to facilitate the analysis of large viral genomes where the number of sequences available for analysis is still rather limited. To test the program, we examined random sequences with properties that were similar to those of 16S ribosomal RNAs. The results of the analyses were surprising and are presented here both to

* To whom correspondence should be addressed

illustrate the use of the computer program and to show that the properties of the random sequences that are used in simulating a comparative sequence analysis can lead to markedly different results.

MATERIALS AND METHODS

The algorithm

We describe how R previously aligned sequences, $S_1, S_2, \dots, S_R$, each with length N , are compared to identify common base paired regions with compensating base changes. The analysis is performed in three steps. First, the algorithm computes all the base paired regions that can be formed in each sequence. Second, the algorithm identifies all the base paired regions that are common among the sequences. Finally, the nucleotide patterns in each of the common helices are scanned to identify the conserved helices that have compensating base changes. Two options in the basic algorithm are available. The first option allows the user to find helices with G-U pairs. The second option allows the user to find conserved helices in subsets of the sequences.

To find the base paired regions, a complementary sequence S' is constructed for each sequence S , and a hashing scheme is applied to find the matching subsequences between S and S' . Subsequences of length l are examined, where l lies between a user defined helix length k and 6 ($k \leq l \leq 6$). For each l , a table is constructed in the form of a one dimensional array with 4^l entries. Each subsequence can be encoded into an integer between 1 and 4^l ; the integer identifies the subsequence pattern uniquely. The table is filled in a single pass through S by entering the starting position of each subsequence in S into the table at the location corresponding to the encoded value of the subsequence. The list of all base paired regions is found in a single pass through S' by pairing the appropriate values in the list at the table entry corresponding to the encoded value of each subsequence in S' with the starting position of the subsequence. This process is repeated until we have created a set of lists $\{L_k, L_{k+1}, \dots, L_6\}$ for each of the R sequences. Each list contains all the base paired regions for one value of l . A single base paired region is recorded in the list by a position pair (i, j) , where i and j are the starting positions of the two sides of a base paired region and $i < j$.

To find base paired regions that are common among the sequences, lists of helices with the same length are compared with one another. For each helix length l , a table in the form of a one dimensional array with N buckets is created. The data structure of each bucket is a list of position values; a counter value is associated with each position value. The table is filled in a single pass through the list L_l that was generated for the first sequence. For each item (i, j) in the list, j is added to bucket i and the counter value for j is set to 1. For each helix position pair (i', j') in the L_l lists that were generated for the remaining sequences, j' is checked for its presence in bucket i' . If j' is in bucket i' , the counter value for this j' is incremented by 1. For a position value j in bucket i , if the counter value for j equals R , (i, j) is the position pair of a common helix. Therefore, the helices that are common to all the sequences can be identified in a single pass through all the buckets. All the common helices identified by the above method are stored in a list. Helices in the list are examined and redundant shorter helices that are contained in longer helices are removed from the list. Helices of length greater than 6 (the maximum value of l) are also identified at this point by fusing helices that overlap with one another.

Table 1. Helices with compensating base changes.

Two sequences			
a)	* AAAGA AAACA	* UCUUU UGUUU	b) * * AUAGA AAACA
			* * UCUAU UGUUU
Three sequences			
c)	* * AAACA AAAGA AUAGA	* * UGUUU UCUUU UCUAU	d) * AAACA AAAGA AAAUA
			* UGUUU UUUUU UAUUU

Asterisks indicate positions where compensating base changes occur.

In the final step, the program generates lists of all helices that have a minimum of either one or two compensating base changes. Table 1 shows the types of base paired helices that are considered. Lists of helices with one or more compensating base changes will have all the helix types shown in the table. When the output is restricted to helices with a minimum of two compensating base changes per helix, the helices of types b, c, and d will be selected. When three or more sequences are compared, the program only recognizes compensating base changes that are located in a minimum of three different sequences (helices of type c and d).

The output from the program consists of two parts. The first part is shown in Figure 1. It is in the form of a diagram. The positions of individual helices are displayed horizontally above the nucleotide sequences. Asterisks are used to indicate the length of the helices that are common to all the sequences and the numbers between the asterisks are those assigned to individual helices as described below. When a given sequence has more than one complement, lower case letters are assigned to the alternate complementary regions. Overlapping helices are displayed on separate lines. The nucleotide differences between the first sequence and subsequent sequences are shown beneath the first sequence for each base paired region. Detailed information about individual base paired regions is provided in the second part of the output where the individual helices are listed in tabular form (Figure 2). Individual helices are sorted both according to length and distance from the 5' end of the sequence. Numbers are assigned to all helices so that the longest helices have the lowest numbers. The base sequence of each helix is shown along with its nucleotide position in the parent (non-aligned) sequence. In addition, the distance in nucleotides that intervenes between the complementary sides of the helix are shown. Each base paired region is tested individually to see whether it can be extended in either direction. Possible extensions are displayed in the output and the original size of the helix prior to extension is given with the helix label.

Two options were introduced into the basic program. The first option enables the user to choose whether or not to allow the wobble pair G-U in addition to A-U and G-C base pairs. With this option, more than one subsequence may complement each subsequence in S . To find the additional complementary subsequences, the first step in the algorithm has been modified. The list of all base paired regions is still found in a single pass through S' . However, for each subsequence in S' , the algorithm must identify all possible matching subsequences rather than only one. The encoded value of each matching subsequence is a table entry. Since there are multiple matching subsequences for each subsequence in S' , there are also multiple table entries for each subsequence in S' . Base paired regions are now identified by

[illegible]

Figure 1. An example from the first section of output generated by the program. Two random sequences that were 4000 nucleotides long and 80% similar to one another were analyzed for the total number of complementary segments with 4 or more base pairs and two or more compensating base changes. The complementary segments contained G-U pairs. A total of 104 helices were found in the analysis. The segment from nucleotides 3500–3800 is shown here.

LIST OF HELICES WITH COMPENSATING BASE CHANGES

Sequence 1: (FROM 1 TO 4001)

Sequence 2: (FROM 1 TO 4001)

HELIX NUMBER 1 (length prior to extension is 11)							
	Start No	Sequence	End No	Intervening No	Start No	Sequence	End No
Random 1	3565	GAAUCUCGCGUGA	3576	182*	3759	UCGCGUAGGUUC	3770
Random 2	3566	AUUUUGCGGGA	3576	182*	3759	UCCCGUAGGGU	3769
HELIX NUMBER 2 (length prior to extension is 8)							
	Start No	Sequence	End No	Intervening No	Start No	Sequence	End No
Random 1	524	UACAUGUG	531	1263*	1795	UGCAUGUA	1802
Random 2	524	UACUGGUG	531	1263*	1795	UGCUGGUG	1802
HELIX NUMBER 3 (length prior to extension is 7)							
	Start No	Sequence	End No	Intervening No	Start No	Sequence	End No
Random 1	662	CAUCUGUGG	670	787*	1458	UCGCGGAGUG	1466
Random 2	662	CAACGGU	668	791*	1460	GCUGUUG	1466
HELIX NUMBER 4 (length prior to extension is 7)							
	Start No	Sequence	End No	Intervening No	Start No	Sequence	End No
Random 1	853	CCUUCUG	859	1398*	2258	CGGAGGG	2264
Random 2	853	CUUUCGG	859	1398*	2258	CCGAGAG	2264
HELIX NUMBER 5 (length prior to extension is 7)							
	Start No	Sequence	End No	Intervening No	Start No	Sequence	End No
Random 1	898	UCUGCCUC	905	2344*	3250	GAGGUGGG	3257
Random 2	898	UGUGCCG	904	2346*	3251	UGGCACG	3257
HELIX NUMBER 6 (length prior to extension is 7)							
	Start No	Sequence	End No	Intervening No	Start No	Sequence	End No
Random 1	954	CCCCCGUUC	962	1824*	2787	GAGUGGGGG	2795
Random 2	955	CCGCCUU	961	1826*	2788	AGGGUGG	2794

Figure 2. An example from the second section of output generated by the program. The sequences that were used to generate the output are described in Figure 1. The six longest base paired segments from the 104 that were found are shown.

pairing the appropriate values in the list at each of the multiple table entries with the starting position of the subsequence.

In the second option, conserved base paired regions are identified if they are common to at least M out of R sequences where M is a user specified value less than R . To accomplish

this, the second step in the algorithm has been modified in two ways. In the original version described above, a common helix is identified when the counter value is exactly R . In this version, a conserved helix is identified when the counter value is at least M . In addition, since we accept helices that need not be conserved

in the first sequence, we must compare additional sequences with one another. These additional comparisons are performed by repeating the second step of the algorithm. Instead of being performed only once, this step is now repeated $R-M+1$ times. For each helix of length l and each time p that the second step is performed, a table is constructed from the L_l list for the p^{th} sequence. Items in the L_l lists for the sequences S_{p+1} to S_R are checked with the values stored in the table and the counter values are incremented as described in the second step of the algorithm. Finally, a list of all base paired regions that are common to at least M sequences is obtained. The sequences in M are identified in the tabular output (Figure 2) by an asterisk adjacent to the intervening number. This option is useful in identifying conserved helices where some members may have G-A or C-A pairs.

The program is written in Pascal and the current version is implemented on a Sun 3/60 Workstation with 12M of RAM. The input to the program consists of a set of aligned sequences in separate files. An insertion into the sequence to align the sequences can be in the form of a ., an N, or an X. In the current implementation the program can analyze 20 sequences that are each 500 nucleotides long for helices with three or more base pairs in approximately 2 minutes. A similar analysis of 20 sequences that are each 1500 nucleotides long takes approximately 20 minutes. The program can be obtained from the corresponding author upon request.

Random Sequences

Two types of random sequences were used for the analyses described in this report. In the first type, the bases were chosen independently using a random number generator with the only constraint that the expected base composition be equal to that of 16S ribosomal RNA from *Escherichia coli*. In second type, the base composition was also that of 16S rRNA, but in addition, the sequences were 70% or 80% similar to one another. The similar sequences were generated by exchanging pairs of nucleotides at random in a parental sequence that had the same base composition as 16S rRNA. The parental sequence was obtained by a random shuffling of the nucleotide sequence of 16S rRNA. Each individual random sequence was obtained by exchanging random pairs of nucleotides in the parental sequence until the required level of difference between the new sequence and the parental sequence was reached. Pairs of nucleotides that were identical to one another were not exchanged and pairs of nucleotides were exchanged only once in each sequence.

RESULTS

The analyses that are presented here were performed on sets of random sequences that were of two different lengths. The first set consisted of sequences that were 1542 nucleotides long, the exact size of 16S rRNA from *Escherichia coli*. The second set consisted of sequences that were 4000 nucleotides long. In addition, as indicated in the Methods section, two different methods were used to generate the random sequences. One method generated sequences in an independent manner using a random number generator. The second method yielded random sequences that were similar to one another. Most of the analyses reported here were done with sequences that were 80% similar to one another, because biological sequences with this level of similarity are used to start a comparative analysis (11).

The random sequences were analyzed both for the total number of complementary segments that could be found in common

among them, and for the fraction that had two or more independent base changes. The different properties of the sequences affect the number of complementary segments that are found in common among them as well as the fraction that have two or more independent base changes. The total number of complementary segments with three or more base pairs that were found in common among the random sequences are shown in Table 2. Results are shown for comparisons of 2, 3 and 5 sequences. Twenty random sequences were used in each analysis and each value in the table represents the mean of 5 separate computer runs. The results show that large numbers of complementary regions are found in common among random sequences that are 80% similar to one another, and fewer helices are found in common among sequences that are independent. The difference was most striking for the analyses of 5 sequences; several thousand complementary segments were found in common among the random sequences that were similar to one another, and no complementary segments were found in common among the independent random sequences.

All of the complementary segments listed in Table 2 were examined further to see whether they had two or more independent base changes. Table 3 shows the number of complementary segments with two or more independent base changes that were found in each analysis. The table also shows the range in size of the longest complementary segments that were found in the individual computer runs. The results that were obtained differed depending on the type and number of sequences that were analyzed. Most of the common complementary segments found between two sequences (Table 2), had two or more independent base changes when the sequences were independent. In contrast, when the random sequences were similar to one another, less than one percent of the common complementary segments had two or more independent base changes. In addition, the analysis showed that some of the common complementary regions that were found were surprisingly long. For example, the longest common segments with 2 or more independent base changes that were found for pairs of similar random sequences that were 4000 nucleotides long, ranged in size from 9 to 11 base pairs. An example of a helix with 11 base pairs that was found in an analysis of two similar random sequences can be seen in Figure 1.

The analysis of 5 sequences that were 80% similar to one another also yielded results that were unexpected. Eighteen complementary segments with 2 or more independent base changes were found in common among sequences that were 1542 nucleotides long; one hundred segments were found in common when the sequences were 4000 nucleotides long. These results are shown in more detail in Table 4 where the complementary segments are sorted by size. Approximately 80% of the common complementary segments that were found for sequences that were 1542 nucleotides long have only 3 base pairs. Less than 1% of the complementary segments were 5 base pairs long. When the sequences were 4000 nucleotides long, however, 8% of the common complementary segments that exhibited two or more independent base changes were at least 5 base pairs long.

It is interesting to compare the results of the simulations with random sequences with the distribution of conserved helices that are present in the phylogenetic model of 16S ribosomal RNA from *Escherichia coli* (Table 4). The individual helices in the 16S rRNA model were tabulated according to the same rules that were used for the random sequences. The helices contain no single-stranded loops and G-A pairs are treated as interior loops. By these criteria, 79 helices with 3 or more base pairs were

Table 2. Number of common complementary segments in random sequences.

Sequence	Segment Length	Number of Sequences		
		2	3	5
1542-I	≥ 3	2774 ± 88	149 ± 23	0
1542-S80	≥ 3	18,278 ± 719	10,990 ± 813	4409 ± 203
4000-I	≥ 3	18,741 ± 412	1052 ± 84	4 ± 3
4000-S80	≥ 3	116,661 ± 1403	70,543 ± 1717	27,285 ± 734

Each random sequence has the same base composition as 16S rRNA from *Escherichia coli*. The numerical designation of each sequence indicates its length in nucleotides. Sequences with the suffix I were generated independently. Sequences with the suffix S80 were 80% similar to one another (see text for details). Each value in the table represents the average of 5 computer runs. The mean and sample standard deviation are shown. The complementary segments contain potential G-U pairs.

Table 3. Common complementary segments with 2 or more independent base changes that are found in random sequences.

Sequence	Segment Size	Number of Sequences Analyzed					
		2		3		5	
		Number	Range	Number	Range	Number	Range
1542-I	≥ 3	1901	6–8	130	4–5	0	–
1542-S80	≥ 3	38	5–8	16	5–7	18	4–5
4000-I	≥ 3	12,614	7–8	896	5	3	3–4
4000-S80	≥ 3	267	9–11	96	6–9	100	5–7

The sequences used in this analysis were the same as those used for the analysis shown in Table 2. For each set of sequences the mean number of common complementary segments with 2 or more independent base changes is shown. Each value represents the mean of 5 computer runs. Values in the range column give the length in nucleotides of the longest common complementary segments that were found in the individual computer runs.

identified in the secondary structure model of 16S rRNA from *Escherichia coli* that has been established by the phylogenetic approach (4). Approximately 50% of these helices have 5 or more base pairs. Clearly, many more helices are conserved in common in 16S rRNAs than are found in common among our random sequences; most of the conserved helices are also larger than the common complementary segments that were found in the random sequences.

DISCUSSION

The analysis of random sequences presented here shows that our new program is fast and can be used to analyze large nucleotide sequences relatively easily. The results also illustrate the importance of choosing appropriate random sequences to simulate the biological situation that is being studied. In our studies, the number of conserved complementary segments with two or more independent base changes that are obtained when two independent random sequences are analyzed is very high, and much lower values are obtained when similar random sequences are analyzed (table 3). The lower values are closer to those obtained with related biological RNAs, and the similar random sequences thus appear to provide a more realistic model for the biological situation than independent random sequences. In contrast to the results that are obtained in the analysis of two sequences, the number of conserved complementary segments that are found in common among multiple sequences is far too low when independent random sequences are analyzed. Here again, the analysis of similar random sequences gives a more realistic estimate of the number of helices that might be found in common by chance along in related biological sequences.

Table 4. Distribution of common complementary segments in random sequences and in the phylogenetic model of 16S rRNA from *Escherichia coli*.

Sequence	3	Complementary Segment Size			Total
		4	5	≥ 6	
1542-S80	14	3	1	0	18
4000-S80	72	20	6	2	100
16S rRNA	18	21	8	32	79

Five sequences were used for each analysis. The results shown for random sequences represent the mean of 5 computer runs. The compilation of the data for the 16S rRNA is described in the text.

The presence of two independent base changes in a set of conserved helices is commonly considered proof that these secondary structures have biological significance and are likely to exist within the cell. The simulations described here with random sequences suggest that these criteria may not be stringent enough especially for the analysis of long sequences. It should be noted, however, that the similar random sequences that were used in our analysis were all prepared from one parental sequence in a single step. Biological sequences that are compared with one another in a phylogenetic analysis of RNA secondary structure have not diverged in parallel from one common ancestor and a true simulation would need to replicate the exact manner in which the sequences have diverged from one another. Further analysis of this problem is needed.

Despite these limitations, the analysis of similar random sequences that is reported here provides useful insight into some of the problems that are likely to be encountered in the early steps of a comparative secondary structure analysis with large RNA sequences. When only two sequences are compared with

one another, both the number of common complementary segments that have two or more complementary base changes and the size of the segments that are obtained are relatively large and there is no way to distinguish between real and fortuitous helices unless other information, such as the results from chemical modification, can be used in conjunction with the comparative sequence analysis. As the number of sequences that are available for analysis increases, these problems diminish. Nevertheless, our simulations show that when as many as five similar random sequences are analyzed, a significant number of large complementary segments with two or more independent base changes can be found in common among them.

The results of our simulations with random sequences were also compared with the number and distribution of helices found in the phylogenetic model of 16S rRNA (Table 4). Our analysis showed that the conserved helices with 2 or more compensating base changes that were found in 16S rRNA sequences were both larger on average and more frequent than those found in the simulations with random sequences. It should be noted, however, that the information shown in the table is based on the analysis of more than 400 sequences, and one might not expect to find the full set of the helices in any subset of 5 sequences if each helix must have 2 or more compensating base changes. There are several reasons for this. An alignment based on 5 sequences might not coincide exactly with an alignment based on 400 sequences. In addition, many of the helices that are conserved among the sequences might have fewer than 2 compensating base changes. When we examine subsets of 5 aligned 16S rRNA sequences for helices with two or more compensating base changes (data not shown), we do not find the full set of helices that are present in the phylogenetic model but we always find many more helices in common among the biological sequences than are found in common among similar random sequences. We also find some extra helices that are not in the phylogenetic model. The presence of these extra helices is consistent with the analysis of similar random sequences that is discussed in this manuscript.

All of the methods that identify common helices in related sequences depend heavily on the manner in which the sequences are aligned with one another. In sequences that are similar to one another, the alignment is straightforward, and can be accomplished by eye, or with some of the newer computer programs for multiple sequence alignment (12,13,14,15). To align the coding regions of viral genomes, we align the viral proteins on the basis of their amino acid composition using the Dayhoff PAM 250 matrix (16). Files containing the corresponding nucleic acid sequences with appropriate insertions are generated automatically. Regions of the sequences that align poorly with one another are more problematic and must be aligned by eye. Several investigators have based this alignment on the conservation of secondary structure rather than sequence similarity (reviewed in 11). In the case of messenger RNAs this approach is difficult since the conservation of protein structure must be considered in parallel with nucleic acid structure.

Two computer programs other than the one described here have been developed for the identification of conserved helices in related RNAs. Gutell, Noller, Woese and collaborators have developed a program that identifies covariant base changes in related sequences for their studies with 16S and 23S ribosomal RNAs (4,17). This method is particularly useful when many sequences are known and has been used by these authors to confirm individual base pairs in conserved helices as well as to identify potential tertiary interactions. A new program that makes use of the graphics capabilities of modern workstations has also

been described recently by Waterman (18). The program is interactive and allows the user to set parameters that will enable him to identify helices of different types. The output of the analysis is displayed graphically on the computer screen. The program can identify helices that have interior and bulge loops and it appears to be well suited for the identification of helices in regions of the sequences that are not easy to align with one another. The program may be easiest to use with relatively short sequences.

The program that we have described here is probably most useful when used in conjunction with other methods for the analysis of RNA secondary structure. We have a small program that checks the output of programs that predict RNA secondary structure (19) to see whether structures that are predicted are consistent with data from comparative sequence analysis. Another simple program under development would eliminate all conserved helices that are inconsistent with chemical modification data. The basic program could also be modified relatively easily to score for H type pseudoknots (20). This modification would allow us to perform simulations with random sequences to determine how many pseudoknots could be found by chance in related RNAs.

ACKNOWLEDGEMENTS

We thank Cher-Win Lin and Hui-Yu Wang for their contributions to early versions of the program. We thank L.G. Caro for reading an early version of the manuscript. The research described here was supported in part by NIH grant GM3482509 to A.B.J. M. Zuker is a fellow of the Canadian Institute for Advanced Research.

REFERENCES

1. C. R. Woese, R. Gutell, R. Gupta, and H. F. (1983) *Noller*, Microbiol. Rev. 47, 621-669.
2. Fox, G. E. and Woese, C. R. (1975) *Nature* (London) 256, 505-507.
3. Gutell, R. R. and Fox, G. E. (1988) *Nucleic Acids Res.* 16, Suppl. r175-269.
4. Gutell, R. R., Weiser, B., Woese, C. R., and Noller, H. F. (1985) *Prog. Nucl. Acid Res. and Mol. Biol.* 32, 155-216.
5. Sprinzl, Hartmann, T., Meissner, F., Moll, J., and Vorderwulbecke, T., (1987) *Nucleic Acids Res.* 15, Suppl. r53-188.
6. Guthrie, C., and Patterson, B. (1988) *Ann Rev. Genet.* 22,387-419.
7. Cech, T. R. (1988) *Gene* 73, 259-271.
8. Waring, R. B., and Davies, R. W. (1984) *Gene*, 277-291.
9. James, B. D., Olsen, G. J., Liu, J., and Pace, N. R. (1988) *Cell* 52, 19-26.
10. Jacobson, A. B., Zuker, M., and Hirashima, A. (1988) In B. Dudock and M. Inouye, eds., *Molecular Biology of RNA: New Perspectives*, Academic Press, pp. 331-354.
11. James, B. D., Olsen, G. J., and Pace, N. R. (1989) In Dahlberg, J and Abelson, J., eds., *Methods in Enzymology, RNA Processing (Part A) General Methods*. Academic Press, Vol.180 pp. 227-239.
12. Barton, G. J. (1990) In Doolittle, R. F. (ed), *Methods in Enzymology, Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, Academic Press, Vol. 183, pp. 403-428.
13. Corpet, F. (1988) *Nucleic Acids Res.* 16, 10881-10890.
14. Taylor, W. R. (1990) In Doolittle, R. F. (ed), *Methods in Enzymology, Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, Academic Press, Vol. 183, pp. 456-474.
15. Feng, D. and Doolittle, R. F. (1990) In Doolittle, R. F. (ed), *Methods in Enzymology, Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, Academic Press, Vol. 183, pp. 375-387.
16. Schwartz, R. M. and Dayhoff, M. O. (1978) In Dayhoff, M. O. ed., *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Washington, D. C. Vol. 5, Supp. 3., pp. 353-358.
17. Gutell, R. R. and Woese, C. R. (1990) *Proc. Nat. Acad. USA* 87, 663-667.
18. Waterman, M. S. (1988) In Waterman, M. S. ed., *Mathematical Methods for DNA Sequences*, CRC Press, Boca Raton, Fla. pp.185-223.
19. Zuker, M. (1989) *Science* 244, 48-52.
20. Pleij, C. W. A. (1990) *Trends in Biochem. Sci.* 15, 143-147.