Comparative sequence analysis as a tool for studying the secondary structure of mRNAs

Michelle F.Browner and Charles B.Lawrence

Department of Cell Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

Received 6 June 1985; Accepted 29 October 1985

ABSTRACT

Analysis of phylogenetically conserved secondary structure has been important in the development of models for the secondary structure of structural RNAs. In this paper, we apply this type of analysis to several families of informational RNAs to evaluate its usefulness in developing secondary structure models for mRNAs and mRNA precursors. We observed many conserved helices in all mRNA groups analyzed. Three criteria were used to identify potential helices which were not conserved solely because of coding sequence constraints, and may therefore be important for the structure and function of the RNA. These results suggest that this approach will be useful in deriving secondary structure models for informational RNAs when used in conjunction with other complementary techniques, and in designing experiments to determine the functional significance of conserved base pairing interactions.

INTRODUCTION

Several lines of evidence indicate that informational RNAs present transcript-specific signals to the cellular environment, which interact with specific proteins and particles that are involved in modifying and regulating the expression of the transcript. Sequence-dependent signals are thought to be important in the generation of mature 3' ends (1,2) and in the post-transcriptional methylation of specific adenosine residues (3). Also, the selection of appropriate splice-sites has been shown to be influenced by distant sequences (4), suggesting that the secondary structure of the molecule modulates the availability of the splice-site region (5). In addition to the transcript-specific signals required for modification of hnRNA, the packaging of the transcript into a ribonucleoprotein particle has also been shown to be non-random (6) and nucleotide-sequence specific (7). There are other examples of post-transcriptional regulation of RNAs which also indicate the importance of sequence-specific signals; the regulation of message accumulation (8), the rate of cytoplasmic turnover of specific RNAs (9), and the efficiency with which a specific transcript is translated (10).

These sequence-specific signals are likely to be formed, at least in part, by the conformation (secondary and tertiary structures) adopted by the RNAs in the nucleus and cytoplasm. To understand the nature of sequence-specific modification and regulation of informational RNAs it will be important to elucidate the secondary and tertiary structure the molecules form in their native particles. By analogy to the best studied RNP particles, the ribosomal subunits (reviewed in 11), it is likely that many of the techniques which have been useful for the study of ribosomal particles will be applicable to the study of informational RNPs. Successful models for the secondary structures of the small and large ribosomal RNAs have been developed using data from biochemical studies of particle structure (i.e. cross-linking and nuclease sensitivity) and from comparative sequence analysis of ribosomal RNAs from many diverse species (12). In fact, comparative sequence analysis has yielded the key data in the development of unique secondary structure models for the small and large rRNAs (11,13). This approach has also been used to develop secondary structure models for 5S (14) and 5.8S (15,16) RNAs.

Comparative sequence analysis is a powerful tool for developing RNA secondary structure models because it is used to identify regions of potential base pairing which are conserved in evolution. The rationale for examining families of informational RNAs with comparative sequence analysis is that sequence-specific interactions are important in the structure and function of RNA molecules. Base-pairing which has structural and functional significance would be expected to be conserved as the sequence evolves. Although there is some precedent for applying comparative analysis to eukaryotic informational RNAs, these studies have relied on looking at possible secondary structures in very small regions of the RNA molecule (17-19). However, it seemed to us that it was necessary to analyze the potential interactions in an entire mRNA molecule as there is no evidence to suggest which regions are more likely to be involved in functionally important secondary structures.

The present study was designed to determine if comparative sequence analysis, as has been applied to structural RNAs, could be useful in elucidating potentially important secondary structure interactions in informational RNAs. As models for this study, we have chosen gene families in which sequence information is available from a number of diverse species. In this preliminary study, we have also evaluated various criteria that assess the significance of conserved helices in informational RNAs.

ANALYSIS**mRNA Sequence Groups**

The computer analysis of conserved, potential secondary structures in mRNA sequences included the four gene groups shown in Table 1. These sequences were chosen for analysis primarily because complete genomic or cDNA sequences were available for at least four different species in each group. As indicated in the table, the sequences were either obtained from GenBank (the reference is the GenBank accession number) or the sequence was entered into our computer from a publication. In the cases where the genomic sequence was available, we constructed the mRNA using the sites indicated in the GenBank entry. For all the cellular sequences analyzed, the first base was the putative CAP site and the last base the polyadenylation signal (AAUAAA). In the case of the viral mRNAs, only the coding sequences were analyzed.

Detection of phylogenetically conserved base pairing

The determination of conserved potential helices in the groups of related sequences involved three steps: 1) the sequences in each gene family were aligned to determine homologous bases between sequences; 2) all potential helices with a stem of at least five base pairs (allowing G-U pairing) were calculated for each sequence in a group; 3) the position of each base pair in a helix was compared to the sequence alignment for each mRNA sequence. A helix was saved if it contained at least five base pairs that are conserved in homologous positions in all members of the sequence group.

The most important part of this method is the choice of a sequence alignment algorithm to find homologous nucleotides between the sequences of interest. Various methods for finding homologous regions between gene and protein sequences have been thoroughly reviewed by Waterman (28). It is clear from this survey that the definition of sequence homology is operational and depends upon the method and the choice of parameters (such as values for gap penalties) used for the analysis. On the other hand, when closely related sequences are compared, most methods will give a similar sequence alignment. For this reason, we aligned sequences using a modification of a method suggested by Staden (29). This method has the advantage over dynamic programming methods of aligning long sequences (1500 nucleotides in our case) in a single run. We also wished to identify helices conserved in all members of sequence groups composed of up to 12 sequences. This requires obtaining an alignment of several sequences taking each into account at the same time, a problem for which there is presently no practical solution (30). For the purposes of this study, we chose one member of the group as a "master" and aligned all other

TABLE 1
RNA FAMILIES

Group	Sequence ^a	Reference ^b
Globin mRNA Sequences:		
Alpha	Human	J00154
	Chimpanzee	(20)
	Goat	J00043
	Mouse	J00410
	Rabbit	J00658
	Chicken	J00855
Beta	Duck	J00923
	Human	J00164
	Goat	(21)
	Mouse	J00413
	Rabbit	J00659
	Chicken	J00860
Human Beta-like	Duck	J00926
	Beta	J00164
	Delta	J00167
	Epsilon	J00172
	Gamma	J00175
Actin mRNA Sequences:		
Alpha	Human Skeletal	J00068
	Human Cardiac	J00070-700073
	Rat	J00692
	Chicken	J00805
Beta	Human	(22)
	Chicken	(23)
	Rat	J00691
	Soybean	J01297
	Yeast	J01310
	Sea Urchin	J01166
	Amoeba	J01016
	Maize	J01238
Histone mRNA Sequences:		
H4	Human	(24)
	Mouse	J00422
	Sea Urchin (early)	J01172
	Sea Urchin (late)	J01175
	Yeast	(25)
	Wheat	(26)
	Tetrahymena	(27)
Virus Sequences:		
papovavirus family	SV40	J02400-J02410
	BKV	J02038
	Mouse Polyoma	J02288
	Hamster Polyoma	J02289

^a Refers to the species from which the mRNA sequence was obtained.

^b Specifies a GenBank accession number or reference number (in parentheses) from which the mRNA sequence was obtained.

TABLE 2
CONSERVATION OF COMMON HELICES IN mRNAs

Group	Stem ^a Length	Probability ^b	# of Common Helices Expected ^c	Observed ^d
α -globin Mammalian	5	0.05	19.	31
	6	0.03	3.	12
	7	0.02	0.7	1
	8	0.009	0.1	1
	9	0.005	0.03	1
α -globin All	5	0.0007	0.2	9
	6	0.0002	0.02	3
	7	0.00004	0.002	0
	8	0.000009	0.0001	0
	9	0.000002	0.00001	1
Actins Vertebrate Alphas & Betas	5	0.009	20.	249
	6	0.004	3.	39
	7	0.001	0.4	16
	8	0.006	0.05	0
	9	0.0002	0.008	1
Papovavirus T-antigen	5	0.0001	1.0	125
	6	0.00002	0.07	20
	7	0.000003	0.004	6
	8	0.0000006	0.003	1

^a The number of consecutive base pairs required to score helix.

^b The probability that a helix of a given stem length will be conserved in all sequences, calculated as described.

^c The number of helices expected from the probability calculations as described.

^d The number of conserved helices found in all sequences.

members to this sequence. Only nucleotides aligned in all members of the group were considered in the analysis. This effectively limited the analysis to nucleotides in well conserved regions of the sequence families.

Many conserved helices were observed for the globin, actin and papovavirus early mRNA groups. This data is summarized in Table 2. The number of conserved helices for these groups was so large that they could not all possibly be incorporated into a unique secondary structure model. We therefore considered several methods for evaluating their possible involvement in forming the native secondary structure of the RNA molecule.

Expected frequency of conserved helices

The members within a sequence group have significant homology with each

other, thus a certain number of helices are expected to be conserved simply because the potential base pairing involves nucleotides in highly conserved regions of the sequence and has no functional significance in the formation of RNA secondary structure. We therefore determined the expected frequency of conserved helices in each group due to primary sequence conservation.

The probability of finding a region of potential secondary structure to be conserved in all members of a group of sequences was calculated first. The sequences in the group to be analyzed are aligned as described above. Then the master sequence in the group is compared individually to each of the other sequences to compile the nucleotide changes which occur between the aligned positions in the two sequences. This information is then used to calculate the probability that a potential base pair in the master sequence will also be a base pair in the homologous positions of the compared sequence. The product of these probabilities for each of the comparisons in the sequence group is the probability that a potential base pair in one of the sequences will be a base pair in the homologous positions of all sequences in the group (P_{bp}). The probability that a stem length of n will be conserved in all sequences (P_n) then is:

$$P_n = (P_{bp})^n$$

The maximum number (M_n) of possible conserved helices of length n will be equal to the sum of the actual number of potential helices of length $\geq n$ in the master sequence, or:

$$M_n = (\sum_{n}^{\infty} N_n)$$

where N_n is the actual number of possible helices of length n in the master sequence alone. The expected number of conserved helices of length n (E_n) in a sequence group is then:

$$E_n = P_n * M_n$$

Table 2 shows the expected frequencies of conserved helices calculated for the globin, actin and papovavirus early mRNA groups. The table shows only representative examples as we found similar results in all the mRNA groups analyzed. A comparison of the number of conserved helices actually found versus the frequency of helices predicted, indicated the conserved helices

occur in informational mRNAs at frequencies considerably greater than predicted. For example, in the α -globin group when all sequences were analyzed, there are 0.02 helices with a stem of 6 predicted to be conserved based on the calculated homologies. However, we observed 3 conserved helices in this group.

We have calculated the expected number of conserved helices using a straightforward calculation for the probability that a base pair will be conserved in all sequences. Fitch has shown that the expected number of potential helices in a single sequence is influenced by taking into account asymmetries in nearest neighbor frequencies (31) which are characteristic of gene sequences (32,33). The calculation of the expected frequencies of conserved helices above is not affected by this statistical property as we use the actual number of potential helices of a given length (N_n) that occur in the master sequence as a starting point for the calculation rather than a predicted value for the number of potential helices in the sequence. Thus, the excess of observed conserved helices over expected must be due to some other property of the sequence; such as a constraint resulting from encoding a polypeptide in the sequence, or some of the helices may actually be involved in forming functionally important regions of secondary structure in the molecule.

Occurrence of compensating base changes in conserved helices

In the case of structural RNAs (rRNA), a conserved helix is given higher significance if there is at least one compensating base change found in the rRNA sequences examined (11,12,13). As shown in Table 3, we have identified compensating base changes (two base transitions or transversions) in some of the groups of cellular mRNAs and in all of the viral mRNAs examined. There are more compensating base changes in the conserved interactions for both of the viral mRNAs examined compared to the number found in the cellular mRNA groups. The difference between the number of compensating base changes in cellular and viral mRNAs is most probably due to the fact that the average sequence homology for the viral sequences is less than the average sequence homology calculated for the cellular mRNA groups. For informational RNAs it is not obvious what value the occurrence of a compensating base change will have in proving a helix, as informational RNAs have an additional constraint on sequence variability because they also encode a protein sequence.

The influence of amino acid sequence on the detection of conserved helices

The nucleotide sequence in a conserved base paired region of an informational RNA specifies an amino acid sequence, therefore, the apparent

TABLE 3
TRANSITIONS AND TRANSVERSIONS IN COMMON HELICES

Group	Stem Length ^a	# of Conserved Helices ^b	# of Transitions ^c	# of Transversions
α-Globin:				
Mammalian	5	46	4	2
All	5	3	0	0
β-Globin:				
Mammalian	5	164	9	2
All	5	28	2	0
β-Like:				
Human	5	77	1	3
Actins:				
Alphas & Betas	5	56	1	0
Betas	5	64	1	0
Vertebrate	5	289	15	1
Papovavirus:				
T-antigen (early)	5	152	24	24
VPI (late)	5	78	10	17

^a The minimum number of consecutive base pairs required.

^b Total number of conserved helices with minimum stem length.

^c Two base transitions.

conservation of base pairing may simply be a result of constraints in the nucleotide sequence to encode conserved amino acids in the resulting protein. To examine this possibility for the conserved hairpins in our sequence groups, we developed a method to evaluate the conserved stem of each hairpin individually with respect to the flexibility of the nucleotide sequence in the stem to change without altering the resulting amino acid sequence. This procedure gives a value for each stem, a codon score, which suggests whether the sequence is conserved because of amino acid coding requirements or for some other reason (i.e. RNA secondary structure).

A table was prepared which breaks down every possible codon into individual positions and the possible nucleotide bases which can pair with the nucleotide in each position. A value is assigned to each possible base pair, equivalent to the number of ways the codon nucleotide can mutate if paired with the specified nucleotide and remain a codon for the same amino acid, but

TABLE 4
CODON SCORE TABLE FOR CUG (LEUCINE)

Codon base	Position ^a	Paired base ^b /value ^c	Paired base ^b /value ^c
C	1	G/0	-----
U	2	A/0	G/0
G	3	C/3	U/2

^a Position in codon.

^b Base paired with codon base in helix.

^c Value is the number of ways the codon base can mutate preserving the leucine codon, but disrupting the base pairing.

disrupt the potential base pair. These values are a measure of the flexibility the nucleotides in a stem have to change and preserve the amino acid sequence, without regard to secondary structure. The table for one codon, CUG (leucine), is illustrated in Table 4. For position one, there is one mutation which will result in a codon for the same amino acid (C to U); however, this also preserves the potential base pairing with G resulting in a value of 0 (zero) for this position. All leucine codons have a U in a position 2; thus, the position cannot mutate and the score for this position is 0 whether paired with an A or G. In position 3, the G can mutate to an A, G or C (3 ways) and result in a leucine codon. If it was originally paired with a C, all 3 changes disrupt the base pair (score=3), if paired with a U only 2 of the changes disrupt the base pair (score=2).

To evaluate each conserved hairpin, the reading frame of both halves of the conserved stem is predicted so that for each nucleotide in the stem its position in a specific codon can be determined. A score for the stem is determined by summing values in the codon score table corresponding to each nucleotide in the stem. Stems with low scores indicate that the sequence in the stem is likely to have been conserved to preserve the amino acid sequence. Stems with high scores have substantial flexibility to mutate while preserving the amino acid sequence. Thus, the stems with high codon scores may be conserved to preserve the potential base pairing.

Codon scores were calculated for conserved helices found in the various mRNA sequence groups. Figure 1 shows a histogram of scores for the globin sequences. Results were qualitatively the same for the actins and papovavirus mRNA groups. Most of the conserved helices have scores with values between 2 and 7, indicating that the nucleotides involved could mutate several different

FIGURE 1: Histogram of codon scores for the globin mRNAs. The number of conserved helices observed in each group to have the specified codon score. A. Alpha-globin, B. Beta-globin, C. Beta-like globin; mammalian sequences, striped bars; all sequences, closed bars.

ways to preserve the amino acid sequence but disrupt the potential base pairing. Several helices have relatively high scores (9-11) indicating considerable flexibility to change. These helices would be the most likely to have been conserved to preserve base pairing.

This method was developed to assist us in evaluating the likelihood that a conserved helix within a protein coding region has a functional role in forming the native secondary structure of the RNA. The value of this approach is not known at present because there are no well developed models of informational RNA secondary structure that we can use for comparison. Should the use of codon scores prove to be valuable, the relatively simple calculation we have described could be refined to take into account the asymmetric usage of possible codons for a given amino acid (34) and the relative interchangeability of closely related amino acids (35).

Specific examples of phylogenetically conserved helices

The conserved helix observed in the α -globin mRNA with a relatively high codon score, 11, is shown diagrammatically in Figure 2. The conserved base paired region of this helix has a stem length of 9 base pairs, the longest

FIGURE 2: Conserved secondary structures in alpha globin mRNAs. The potential base paired region and the mRNAs with this structure are shown, along with the transitions observed in two of the mRNAs. The bar indicates the initiating AUG; nt, nucleotide; TAA, stop codon.

conserved stem in the α -globin groups. There are 2 single base transitions in 2 of the 7 sequences examined. The conserved helix involves a base pairing between the initiating AUG and a region 17 base pairs upstream (5') from the stop codon. Further evidence that this helix may be structurally important is provided by the nuclease sensitivity studies of Pavlakis, et al. (36). These investigators have reported that the initiating AUG in mouse and rabbit α -globin mRNA is inaccessible to cleavage by exogenous nucleases. In comparison, the initiating AUG of the mouse and rabbit β -globin mRNAs is quite susceptible to cleavage by nucleases. In our computer analysis, we found a relatively poor, conserved interaction involving the β -globin initiating AUG; a stem length of 6 in the β -globin mRNAs versus a stem length of 9 in the α -globin sequences. We have recently examined the Xenopus laevis α -globin sequence (GenBank accession number J00972) for the occurrence of a homologous helix. Although there was base pairing observed between the two regions in the Xenopus sequence, the potential interaction was much less extensive (stem length of four) than the interaction observed in the other 7 sequences examined. Figure 3 shows one of the conserved interactions found in the early mRNAs in the papovavirus family. This base-paired region is of interest for two reasons: 1) the codon score for this interaction is 10, a relatively high codon score, and 2) there are two transversions in addition to two transitions which preserve the interaction as shown in the figure.

We have analyzed 7 histone H4 mRNA sequences (Table 1). For the analysis of the H4 sequences it was not possible to align the tetrahymena mRNA sequence with the other 6 mRNAs. To get around this problem, we calculated the conserved helices in the other 6 sequences. From this analysis only one

FIGURE 3: Conserved secondary structures in papovavirus family early mRNA. The potential base paired region and the compensating base changes for each sequence are shown. The position of the secondary structure is as indicated for SV40 T-antigen mRNA.

potential helix was identified and the codon score was calculated to be 0. We also calculated all possible helices in the tetrahymena H4 mRNA. As the single helix identified in the group of H4 mRNAs involved the initiating AUG

Conserved Secondary Structure in Histone H4 mRNAs

FIGURE 4: Conserved secondary structure in histone H4 mRNAs. The base paired region observed in each H4 sequence are shown. The position of the secondary structure is as indicated for the human mRNA. The homologous nucleotides are aligned for each sequence, with the amino acid of the codon indicated.

codon, base-paired to a region approximately 40 base pairs upstream from the stop codon, we were able to identify a similar potential helix in the tetrahymena mRNA. The structure identified is shown in Figure 4. The conserved helix involves a two-base transition, the base pair U-A is changed to G-U, in the tetrahymena sequence. This transition preserved the helix, but alters the codon involved in the interaction. Thus, there is a compensating base change which effects the coding sequence, but preserves the base pairing. The occurrence of non-Watson-Crick base pairing (G-U) has been allowed in the comparative sequence analysis of rRNA (11,37). Indeed, there are predicted helices in which G-U pairing involves over half of the base pairs in a stem (11). There is speculation that other non-Watson-Crick base pairing may be involved in RNA secondary structure. As of now there is no experimental evidence to suggest a functional role for the three potential structures identified above. However, in light of the criteria that have been used to prove a helix in rRNA, these potential interactions deserve attention.

Stability of conserved helices

We have chosen not to use pairing and stacking energies in evaluating possible conserved helices. Any predicted helix will be just one component of a complex secondary structure whose stability will be a function of all of the secondary and tertiary interactions in the molecule. However, the use of folding algorithms which predict low free-energy structures may be valuable when used with comparative analysis to develop secondary structure models (see Discussion).

DISCUSSION

The use of comparative sequence analysis has been shown to be a very powerful tool in predicting secondary structure models for structural RNAs. Comparative sequence analysis involves a comparison of the possible duplex interactions in one rRNA with the duplex structures predicted for a rRNA sequence from a different species. If the same helix is predicted in all the rRNA sequences examined, the existence of the duplex region is given increased significance if there are compensating base changes that preserve the duplex structure at the same position (11,13). The usefulness of the comparative sequence analysis in predicting functionally important duplex structures in informational RNAs has not previously been examined. To determine if comparative sequence analysis could be used in predicting secondary structures in informational RNAs, we have completed a preliminary analysis of four mRNA groups. We analyzed only mRNAs, as it is necessary to align the sequences

that are compared, and the divergence of intron sequences in natural gene sequences makes the sequence alignment in these regions impossible. This preliminary analysis identified conserved, base-paired interactions in all the mRNA groups examined. The analysis of the globin, actin and papovavirus mRNAs identified many conserved interactions (Table 2); many of the observed helices are mutually exclusive, making it necessary to evaluate the conserved interactions to determine which may be more significant in a functional and/or structural role.

We have used three criteria to assess the significance of the conserved helices: 1) Helices of a given stem length occur at a frequency greater than expected given the homology between the members of the sequence group (Table 2). This may be of limited value in identifying important base-paired regions because of the statistical properties of the protein coding sequence and the conservation of the primary sequence in the protein coding regions. 2) Compensatory base changes were observed in conserved helices from some of the mRNA groups examined (Table 3). We found more compensating base changes in the mRNA groups having less homology between their members (the papovavirus mRNAs). This result suggests that it will be valuable to use relatively divergent sequences to evaluate conserved helices in informational RNAs by the criteria of compensating base changes. 3) The conservation of base-paired regions solely due to conservation of amino acid coding sequence was assessed by scoring the codons involved in the base-pairs on their flexibility to mutate, preserving the amino acid but disrupting the potential base-pair (Figure 1). The calculation of codon scores allowed us to identify conserved stems which were probably not conserved only because of coding sequence constraints. The value of this criterion in screening potential helices has yet to be determined.

Although there are limitations to each of the criteria we have used in evaluating the conserved helices, they have been useful in identifying interesting potential secondary structure interactions (Figure 2-4). The identification of such potential secondary structure features will be very useful in designing broader approaches for predicting secondary structure models of informational RNAs. Indeed, comparative sequence analysis must be coupled with other techniques, as has been done for rRNA structure studies, to develop secondary structure models. Biochemical studies on native RNP particles, such as cross-linking and nuclease sensitivity studies, can be used to find base-paired and single-stranded regions. Data from comparative sequence analysis may also be utilized in designing experiments to test the

functional significance of particular base pairing interactions. For example, potentially base-paired regions can be mutated in an expression vector containing the gene of interest and the effect of disrupting the base-paired region examined using transient expression assays. Using various expression vectors, it would also be possible to examine the structure of nuclear and cytoplasmic particles involved in packaging the mutated RNA sequences. Other useful tools for predicting secondary structure models for informational RNAs are RNA folding algorithms. These algorithms can be used to predict minimum free energy folding of a region of an RNA molecule, given constraints provided by biochemical data (reviewed in 38). While none of these tools are sufficient by themselves to elucidate the native secondary structure of an informational RNA, combining these approaches will be productive in defining many of the important functional regions of informational RNA structure.

ACKNOWLEDGEMENTS

We acknowledge Steve Delaune for typing the manuscript and Jeff Rosen for helpful criticism.

This work was supported by Grant AI19578 from the National Institutes of Health.

REFERENCES

1. McDevitt, M.A., M.J. Imperiale, H. Ali and J.R. Nevins. (1984) *Cell*, 37, 993-999.
2. Sadofsky, M. and J.C. Alwine. (1984) *Mol. Cell Biol.*, 4, 1460-1468.
3. Horowitz, S., A. Horowitz, T.W. Nilsen, T.W. Munns and F.M. Rottman. (1984) *Proc. Natl. Acad. Sci.*, 81, 5667-5671.
4. Khoury, G., P. Gruss, R. Dahr and C. Lai. (1979) *Cell*, 18, 85-92.
5. Wieringa, B., F. Meyer, J. Reiser and C. Weissman. (1982) *Gene Regulation*, B.W. O'Malley (ed.), pp 65-85, Academic Press, New York.
6. Beyer, A.L., O.L. Miller and S.L. McKnight. (1980) *Cell*, 20, 75-84.
7. Augenlicht, L.H. (1979) *Biochem.*, 18, 3780-3786.
8. Groudine, M. and C. Casimir. (1984) *Nuc. Acids Res.*, 12, 1427-1445.
9. Heintz, N., H.L. Sive and R.G. Roeder. (1983) *Mol. Cell. Biol.*, 3, 539-550.
10. Ray, B.K., T.G. Brendler, S. Adya, S. Daniels-McQueen, J.K. Miller, J.W.B. Hershey, J.A. Grifo, W.C. Merrick and R.E. Thack. (1983) *Proc. Natl. Acad. Sci.*, 80, 663-667.
11. Noller, H.F. (1984) *Ann. Rev. Biochem.*, 53, 119-162.
12. Brimacombe, R., P. Maly and C. Zwieb. (1983) *Prog. Nuc. Acid Res.*, 28, 1-48.
13. Noller, H.F. and C.R. Woese. (1981) *Sci.*, 212, 403-411.
14. Fox, G.E. and C.R. Woese. (1975) *Nature*, 256, 505-507.
15. Ursi, D., A. Vandenberghe and R. DeWachter. (1982) *Nuc. Acids Res.*, 21, 2335-2343.
16. Vaughn, J.C., S.J. Sperbeck, W.J. Ramsey and C.B. Lawrence. (1984) *Nuc. Acids Res.*, 27, 7479-7490.

17. Trapnell, B.C., P. Tolstoshev and R.G. Crystal. (1980) *Nuc. Acids Res.*, 8, 3659-3672.
18. Naroa, H., N.J. Deacon and D. Buckle. (1980) *J. Theor. Biol.*, 85, 83-97.
19. Davies, R.W., R.W. Waring, J.A. Ray, T.A. Brown and C. Scazzochio. (1982) *Nature*, 300, 719-724.
20. Liebhaber, S. and K. Begley. (1983) *Nuc. Acids Res.*, 11, 8915-8925.
21. Schon, E., M. Cleary, J. Haynes and J. Lingrel. (1981) *Cell*, 27, 359-369.
22. Ponte, P., S. Ng, J. Engel, P. Gunning and L. Kedes. (1984) *Nuc. Acids Res.*, 12, 1687-1696.
23. Kost, T.A., N. Theodorakis and S.H. Hughes. (1983) *Nuc. Acids Res.*, 11, 8287-8301.
24. Zhong, R., R.G. Roeder and N. Heintz. (1983) *Nuc. Acids Res.*, 11, 7409-7425.
25. Smith, M.M. and O.S. Andrésson. (1983) *J. Mol. Biol.*, 169, 663-690.
26. Tabata, T., K. Sasaki and M. Iwabuchi. (1983) *Nuc. Acids Res.*, 11, 5865-5875.
27. Bannon, G.A., J.K. Bowen, M-Yao and M. Gorovsky. (1984) *Nuc. Acids Res.*, 12, 1961-1975.
28. Waterman, M.S. (1984) *Bull. Math. Biol.*, 46, 467-472.
29. Staden, R. (1982) *Nuc. Acids Res.*, 10, 2951-2961.
30. Waterman, M.S. (1984) *Bull. Math. Biol.*, 46, 567-577.
31. Fitch, W.M. (1983) *Nuc. Acids Res.*, 11, 4655-4663.
32. Smith, T.F., M.S. Waterman and J.R. Sadler. (1984) *Nuc. Acids Res.*, 11, 2205-2220.
33. Nussinov, R. (1984) *Nuc. Acids Res.*, 12, 1749-1763.
34. Grantham, R., C. Gautier and M. Gouy. (1980) *Nuc. Acids Res.*, 8, 1893-1912.
35. Dayhoff, M.O., R.M. Schwartz and B. Orcutt. (1979) *Atlas of Protein Sequence and Structure*, Dayoff, M.O., Eds., Vol. 5, Suppl. 3, pp 345-352, National Biomedical Research Foundation, Washington, D.C.
36. Pavlakis, G.N., R.E. Lockard, N. Vamvakopoulos, L. Rieser, U.L. Raj-Bhandary and J.N. Vournakis. (1980) *Cell*, 19, 91-102.
37. Woese, C.R., L.J. Magrum, R. Gupta, R.B. Siegel, D.A. Stahl, J. Kop, W. Crawford, J. Brosius, R. Gutell, J.J. Hogan and H.F. Noller (1980) *Nuc. Acids Res.*, 8, 2275-2293.
38. Zuker, M. and D. Sankoff. (1984) *Bull. Math. Biol.* 46, 591-621.