

NCSU Libraries Reserve Room Cover Sheet

Reserve Module Filename: t1437

DRA Record Number: ASA 2239

**** BEST COPY AVAILABLE ****

WARNING CONCERNING COPYRIGHT RESTRICTIONS

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproduction of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship or research. If electronic transmission of reserve material is used for purposes in excess of what constitutes "fair use", that user may be liable for copyright infringement.

ATTN: REVISED COVER SHEET, 10 JAN 02

Course BCH 751

Instructor BROWN JAMES W

Source Title RNA STRUCTURE AND FUNCTION

Processed by M.K.

Date Completed 01/23/01

NORTH CAROLINA STATE UNIVERSITY LIBRARIES



S01939854 \$

Inferring RNA Structure by Phylogenetic and Genetic Analyses

François Michel and Maria Costa

Centre de Génétique Moléculaire du C.N.R.S.
91190 Gif-sur-Yvette, France

Comparative sequence analysis (CSA) is the process in which related macromolecular sequences are collected and compared in order to extract information about common underlying structures. The idea of inferring structure by comparing sequences is rooted in the well-verified fact that the architecture of macromolecules evolves much more slowly than their sequence. As a consequence, nature usually presents us with a significant fraction of all possible sequences from which particular structures or substructures may be generated and, in so doing, tempts us to guess what these structures might be. The sequences to be fed to CSA thus used to be the products of natural selection, but, as discussed in Baskerville et al. (this volume), they should increasingly come in the future from experiments in selection and evolution devised by humans. However, even though the source of variation is changing, the way CSA is carried out, and its limitations, should remain largely the same.

The classic example of a structure successfully predicted by CSA is the tRNA cloverleaf (Madison et al. 1966; RajBhandary et al. 1966; Zachau et al. 1966), but in fact, the Watson-Crick pairs themselves were inferred by applying the very principles of comparative analysis—invariance of higher-order structure—to possible pairings between the hydrogen-bond donor and acceptor groups of the bases (Watson and Crick 1953). CSA could, in principle, make use of any type of macromolecular sequences, and there are examples of CSA being successfully utilized to deduce interactions between proteins and nucleic acids (see, e.g., Nardelli et al. 1991). Nevertheless, the main use of CSA has been in inferring RNA structure. Part of the reason for that lies in the idiosyncrasies of the RNA molecule: The relationship between sequence and secondary structure (and, to some extent, tertiary structure) is far more explicit in RNA than in proteins. Historical circumstances have also played a part in the development of RNA comparative analysis. During the last 20 years or so, it has proved much easier to accumulate nucleic acid sequences

than to grow well-diffracting RNA crystals (there are, however, signs that this situation may be beginning to change [see Doudna et al. 1993; Pley et al. 1994a; Scott et al. 1995; Cate et al. 1996]). Nor has nuclear magnetic resonance (NMR) been able to provide us with structures of RNA molecules longer than 30–40 residues (see Puglisi and Puglisi, this volume). Thus, it is mainly CSA that made it possible to identify the secondary structure components of those complex RNAs with a structure conserved by evolution. A nonexhaustive list of secondary structure models established by CSA comprises, in addition to the tRNA cloverleaf, the models for 5S (Fox and Woese 1975), 16S (Woese et al. 1980; Stiegler et al. 1981; Zwieb et al. 1981), and 23S (Branlant et al. 1981; Glotz et al. 1981; Noller et al. 1981) ribosomal RNAs; for group I (Davies et al. 1982; Michel et al. 1982) and group II (Michel et al. 1982) self-splicing introns; for the RNase P RNA (James et al. 1988); for the 7S RNA of the signal recognition particle (for review, see Larsen and Zwieb 1991); and for the small nuclear RNAs at the heart of the spliceosome (for review, see Guthrie and Patterson 1988) and the telomerase RNA (Romero and Blackburn 1991). It is also primarily CSA that has provided us with a number of tertiary, noncanonical interactions in ribosomal RNAs (for review, see Gutell 1995) and with the current, working three-dimensional model of group I introns (Michel and Westhof 1990).

Many reviews (see Gutell 1993, 1995; Woese and Pace 1993; Gutell et al. 1994; Westhof and Michel 1994), most of them by Carl Woese and/or scientists who at one time or other were associated with him, have dealt with CSA. In this paper, we attempt to cover a diversity of types of molecules (rather than only ribosomal RNA and RNase P RNA), to emphasize some methodological problems that are not very often discussed, and to compare structures inferred by CSA with actual structures, when they are known. In addition to the phylogenetic prediction of RNA structure, we provide a few examples of the experimental verification of postulated structures by genetic analysis (understood here as the approach in which bases are substituted in a controlled way and the functional consequences are assessed).

OVERVIEW

The workings of CSA as applied to RNA may be summarized in a few sentences. Starting with a set of related molecules, blocks of partly conserved sequence are assumed to be homologous and aligned, subsets of positions (within or next to those blocks) whose base contents appear to

have changed repeatedly in a concerted way are sought, and finally, efforts are made to interpret observed patterns of covariation in terms of underlying molecular contacts. The interactions that may be inferred are essentially the ones between bases and consist mostly of classic (Watson-Crick and wobble) pairs; i.e., those pairings that form the secondary structure of the molecule, although careful analysis may reveal a number of elements of tertiary structure, in the form of noncanonical pairs and base triples. The process is iterative, in the sense that initial alignments can be refined by taking into account inferred structures, and improved alignments allow more subtle patterns of covariation to become apparent.

Despite the ease with which CSA may be described and the fact that, short of crystals, it constitutes by far the most efficient way to derive RNA secondary structure, there are nevertheless complex methodological issues associated with each individual step in this approach. These problems are discussed from a technical point of view in the section devoted to the automation of CSA, but some of them stand to reason. For instance, too little variation is uninformative, whereas too much variation will prevent alignment. In addition, deciding whether or not sections of two molecules are homologous may require far more effort and expertise than merely trying to compute the odds that a given extent of sequence similarity arose by chance. For example, an attempt (Banerjee et al. 1993) to prove the existence of a specific tertiary pairing in group I introns was flawed by the fact that segments of sequence labeled stem P2 correspond to a variety of nonhomologous structures, with distinct three-dimensional geometries relative to neighboring pairings, in different subgroups of group I introns (see Michel and Westhof 1990). Coming then to putative instances of covariation, it is clearly crucial to assess their significance. Unfortunately, estimates of statistical significance provided by such parameters as chi-square or mutual information (Chiu and Kolodziejczak 1991; see subsection on automation of covariation analysis) cannot suffice, for reckoning the absolute number of underlying events of concerted change is just as important as measuring the quality of correlations. Finally, although sites engaged in classic pairing (to the exclusion of additional interactions) readily betray themselves by the presence, in a sufficiently large and diverse sample, of all four Watson-Crick combinations (and few others, except for some G:U and U:Gs), there is no such simple rule for converting other patterns of covariation into structural information. Examples of geometric inference are provided and discussed while dealing with specific types of interactions. Before that is done, it is important to emphasize the specificities and limitations of CSA.

CSA Versus Other Approaches to RNA Structure

It is revealing to compare CSA with the other approaches used to investigate RNA structure. Experimental, physical-chemical methods such as chemical and enzymatic probing and crystallography typically yield static (in fact, average) pictures of macromolecules, more often in artificial than natural settings (although some chemical probing can be carried out *in vivo*; see, e.g. Nick and Gilbert 1985; Zaug and Cech 1995). Structural information is explicit or can be obtained with a limited amount of inference. At the other end of the methodological spectrum, *ab initio* calculations of secondary structure, using sets of thermodynamic values and heuristic (Nussinov and Jacobson 1980; Zuker and Stiegler 1981) or combinatorial (Ninio 1979) algorithms, can at best address those structures that are thermodynamically viable and kinetically accessible in the range of conditions under which the thermodynamic parameters that these methods utilize were derived. Not surprisingly, such calculations tend to work best (in the sense that they provide users with a large fraction of all biologically significant helical segments) when applied to those relatively few RNA molecules (most tRNAs, some group I and group II self-splicing introns, the RNA component of eubacterial RNase P) that are endowed with the ability to reach by themselves an active structure in salt solutions (Papanicolaou et al. 1984; Jaeger et al. 1989; Pace et al. 1989; Zuker et al. 1991; Konings and Gutell 1995).

In contrast with those approaches, CSA provides us merely with lists of statistical constraints, which then need to be converted into meaningful structural information—the subject of this review. On the positive side, because the constraints uncovered by CSA were generated by selection, they are guaranteed to reflect function and, at least in case evolution took place in natural settings, to be biologically significant. However, the drawback of *in vivo* evolution is that much of it may reflect the need to interact fruitfully with a number of other molecules, rather than merely make the appropriate intramolecular contacts and avoid unwanted ones. Even when the underlying interactions are purely confined to the RNA of interest, the various constraints observed may actually relate to successive, alternate states of that molecule: The picture generated by CSA is in no way bound to be static, and the structures inferred need not exist at the same time. In reality, molecules that, like the ribozyme component of many group I introns, often fold by themselves appear to have a largely static structure (for review, see Cech 1993), and a number of other molecules, e.g. ribosomal RNAs, have revealed only limited evidence of alternate secondary-structure states (but see Stebbins-Boaz and Gerbi 1991; Dammel and Noller 1993; Powers and Noller 1994; Lodmell et al.

1995): In all such cases, it is probably legitimate to group all identified classic base pairs together into a single so-called "secondary structure model." In contrast, complexes such as the spliceosome, in which proteins play an essential part, have been found to undergo major rearrangements of their RNA components during the processes in which they take part (for review, see Madhani and Guthrie 1994).

STRUCTURAL MOTIFS IDENTIFIED BY PHYLOGENETIC AND GENETIC ANALYSES

Classic Base Pairs Organized into Extended Pairings

The Watson-Crick base pair is CSA's favorite because it is the only base-base interaction to come in four strictly isosteric combinations and to be more often than not flanked by neighboring interactions with the same geometry. Thus, when potential helices are taken as statistical units and regarded as proven as soon as more than one compensatory change has been detected (Woese et al. 1980), very few sequences may be necessary to identify most pairings. In a somewhat extreme example, a tentative model of group II introns (Michel et al. 1982) that rested on only two (ideally divergent) sequences turned out to have been essentially correct. Refinement of such provisional models with additional sequence data seldom involves more than the addition of a few short and/or highly conserved pairings, for which no evidence could initially be obtained, plus the removal at some helix termini of a small number of pairs whose individual bases eventually turned out not to change in a concerted way. However, it should be recalled that because of the slower evolution of bases involved in multiple contacts, large numbers of sequences may prove necessary to establish the existence of alternate structural states of a molecule.

Among potential helices uncovered by CSA, the ones that involve sections of terminal or internal loops are often regarded as "tertiary" interactions (or, somewhat abusively, called "pseudoknots"; see below, "Coaxially Stacked Helices and Pseudoknots"). One reason for this is that such pairings cannot be accommodated in tree-like representations of secondary structure and, consequently, cannot be handled by the dynamic programming algorithms utilized to predict secondary structure *ab initio*. More pertinently, however, careful investigations have revealed that interactions of this type tend to give way at lower temperatures than most other pairings involving canonical base pairs (see e.g., Jaeger et al. 1993 and references therein), so that there is some justification in regarding them as components of tertiary, rather than secondary, structure.

Nonetheless, as far as CSA is concerned, the only meaningful criteria to evaluate potential extended base pairings are whether they involve antiparallel strands and are supported by compensatory base changes.

In addition to Watson-Crick base pairs, regular RNA helices typically include between 5% and 10% of G:U base pairs, which may therefore be regarded as "classic" as well (other base-base appositions occur much more infrequently within helices). That does not mean that all G:U pairs have equivalent roles in RNA structure and evolution. As surmised long ago, and confirmed by careful analysis of phylogenetic data (Rousset et al. 1991), many G:U pairs appear to constitute mere evolutionary intermediates between G:C and A:U, having frequencies and evolutionary life spans in the expected range for slightly deleterious alleles. In contrast, some G:U pairs are clearly under positive selective pressure (Gautheret et al. 1995b). Thus, a few invariant or nearly invariant G:U pairs in ribosomal RNA and self-splicing introns are suspected to be part of the active centers of these molecules. There are also sites where G:U pairs alternate with A:C "mismatches" (protonation of the N1 position of A would allow two hydrogen bonds to be formed with C, with the resulting pair having the same geometry as a G:U wobble pair). Several examples are known in ribosomal RNA (for review, see Gutell 1995), and another instance was recently uncovered in the preferred target motif for GAAA terminal loops (Costa and Michel 1995; Tanner and Cech 1995; see section below on tetraloops). Interestingly, all positions at which G:U and A:C interchange are located at the ends of helices, i.e., at sites where selection for an unusual geometry is nothing unexpected. The 5' cleavage site of group I ribozymes, at which almost all known group I introns have a U:G base pair, provides another example of positive selection for the wobble geometry: Of all other combinations of natural base pairs, C:A is the one that retains the highest fraction of the wild-type activity (Doudna et al. 1989; requirement for a wobble geometry, and also for the amino group of the G, was recently confirmed by the use of nonnatural bases, see Strobel and Cech 1995).

Tertiary Base Pairs

In keeping with common usage, we regard here as tertiary any interaction that is not part of an extended pairing between antiparallel strands. It usually becomes apparent, when large numbers of sequences are available, that the pattern of variation at some of the sites that do not belong to putative helices is constrained to some extent by the situation at one or

several other sites. Provided that the observed statistical biases cannot be ascribed to historical coincidence (because they recur in independently evolving lineages of sequences) and the sites involved are distant in secondary structure models, cases of mutually restricted variation can reasonably be assumed to reflect either direct contact or, at least, close spatial proximity. In contrast, correlated changes affecting nucleotides located close to one another in secondary structure (for instance, facing each other at the end of a helix) must be treated with caution, for they might just as well reflect the necessity for the bases involved *not* to interact; only when mutual variation is severely restricted, and in such a way that a geometrical interpretation can readily be proposed, does it become possible to assume base-pairing.

In this section, we do not attempt to draw up an exhaustive list of sites where tertiary base pairs have been shown, or are suspected, to exist, but by concentrating on the commonest patterns of covariation, we rather discuss the potential for, and difficulties involved in, structural inference.

Lone Watson-Crick Pairs

Some rare sites that are not part of putative helices covary nevertheless with one another according to Watson-Crick rules (for ribosomal RNA review, see Gutell 1995). Although the existence of a "lone" pair may safely be assumed in such cases, the geometry of that pair is likely to remain uncertain. Thus, the 15-48 pair of tRNA could be predicted by Levitt (1969) because it can be either G:C or A:U (and much more rarely, Y:R), but was misinterpreted as having the purine in *syn*, rather than being of the "reversed" type, with riboses arranged in *trans* (both geometries lead to the same, locally parallel orientation of strands; see Westhof 1992). The lesson here is that even though the various possible reversed Watson-Crick pairs are not strictly isosteric, they may still all be sufficiently compatible with the surrounding architecture to be replaced by one another during evolution. Following this realization, a lone pair in subgroup IA introns that comes in all Watson-Crick combinations, but with a huge excess of G:C, was interpreted as being a reversed, *trans* pair (Jaeger et al. 1993).

A lone Watson-Crick pair was recently demonstrated by genetic analysis to participate in recognition of the CCA extremity of tRNA by 23S ribosomal RNA (Samaha et al. 1995). Interestingly, recognition of tRNA by RNase P also involves base pairing with the CCA sequence, but at least two consecutive base pairs could be shown to exist in that case (Kirsebom and Svärd 1994).

Purine-purine Pairs

By far the commonest purine-purine interaction in the RNA structures that has been determined at atomic resolution is the so-called "sheared" G:A pair, in which the hydrogen-bond donor and acceptor groups at positions 6 and 7 of the A contact the acceptor and donor groups at positions 3 and 2 of the G, respectively. The first G and last A of GNRA terminal loops (Westhof et al. 1989; Heus and Pardi 1991) and also of the structurally related GUAAUA hexaloop (Huang et al. 1996) have been reported to form sheared G:A pairs. Loop E of 5S RNA (Wimberley et al. 1993) and the structurally related sarcin-binding loop of 28S rRNA (Szewczak et al. 1993) both include a sheared G:A pair. Juxtaposed sheared G:A pairs have been observed in synthetic oligoribonucleotides (SantaLucia and Turner 1993; Katahira et al. 1994) and in the hammerhead ribozyme, in which they extend stem II (Pley et al. 1994a; Scott et al. 1995). Not all G:A pairs in known RNA structures are of the sheared type, however. In tRNA Phe, the G:A pair at the proximal end of the anticodon stem has a Watson-Crick-like geometry, with one hydrogen bond between N6(A) and O6(G) and another between N1(A) and N1(G). NMR data collected by Walter et al. (1994) suggest that tandem GA:GA pairs in the middle of helices can adopt either sheared or Watson-Crick-like geometries, depending on the nature of neighboring base pairs.

Can CSA be of any help in uncovering potential G:A pairs and assigning them to a particular structural type? Observations of G:A and A:G pairs replacing each other at the end of a helix, and alternating in evolution with canonical base pairs (Gutell 1995; see also Michel and Westhof 1990), point to a Watson-Crick-like geometry. On the other hand, a number of sites in ribosomal RNA, also located at the ends of helices, appear to accept either GA:GA or AA:GA (and occasionally AA:AA) but not the other combinations of As and Gs, and this was interpreted as reflecting selection for sheared G:A pairs (Gautheret et al. 1994). Part of the argument was that the geometry of sheared G:A pairs makes it difficult to accommodate a canonical base pair 5' of the A. In addition, the G of a sheared G:A pair may be replaced by an A (at the expense of one of the hydrogen bonds of the pair), but the A may not be replaced by a G (those GNRA terminal loops that are known to be under selection pressure in group I and group II introns are occasionally replaced by the ANRA sequence but never by GNRG [see Michel et al. 1989a; Michel and Westhof 1990]).

Additional types of purine-purine interactions in natural RNAs are suggested by phylogenetic analysis, which has uncovered several instances of exchanges between G:G and A:A in ribosomal RNAs (see

Gutell 1995), as well as a case of restricted variation in subgroup IC introns, where two sites, one next to helix P7 and the other separated from this helix by two nucleotides, appear to have undergone frequent coordinated changes from A:A to G:G and back, with one of the two G:A combinations missing (Michel and Westhof 1990). This interaction was found by Green and Szostak (1992) to have been replaced by a Watson-Crick pair in an in-vitro-selected molecule. There is also genetic evidence for G:G pairs being formed in the Rev-binding site of HIV RNA (Bartel et al. 1991), in both the antigenomic and genomic ribozymes of the hepatitis δ RNA (Been and Perrotta 1995) and between the first and last nucleotides of nuclear pre-messenger introns (Parker and Siliciano 1993). In all cases, A:A and A:C, but not a number of other base combinations, were shown to be acceptable substitutes for G:G (in the case of HIV, evidence was originally obtained by in vitro selection). Although two possible geometries for the G:G pair are compatible with these data (see Sanger 1984), NMR studies of the Rev-binding site point to a symmetric arrangement, with the Watson-Crick sides of the purines interacting and glycosyl bonds in *trans*. However, it remains to be settled whether such a base pair is accommodated next to a helix by having one of the bases in *syn* (Peterson et al. 1994) or else by a locally parallel orientation of the two strands (Battiste et al. 1995).

Pyrimidine-pyrimidine Pairs

Several Y:Y pairings that betray themselves by strict covariation of U:U and C:C must play important roles in ribosomal RNA (Gutell et al. 1994; in fact, one of these base pairs connects domains IV and V of the large ribosomal RNA). Isosteric U:U and C:C pairs can be generated either with a reversed Watson-Crick orientation, with riboses in *trans*, or, if one of the cytosines is protonated (SantaLucia and Turner 1991), with the classic arrangement of bases and riboses seen in extended helices.

Base Triples

CSA has a fair record in identifying the base triples that arise when a third base contacts a canonical base pair from either the deep or shallow groove sides (the deep and shallow grooves of the A-RNA helix are often, and somewhat abusively, called "major" and "minor" by analogy with B-DNA). Five years before the first three-dimensional structure of tRNA was established, at a time when only 14 tRNA sequences were available, Levitt (1969) successfully predicted one of the three deep-

groove base triples commonly found in type I tRNAs (a second triple interaction was predicted, but turned out not to exist). In group I introns, four base triples were proposed to be generally conserved and to play a crucial role in the assembly of the group I catalytic core (Michel and Westhof 1990; Michel et al. 1990). At least three of these triples have now been substantiated by experimental approaches (Michel et al. 1990; Chastain and Tinoco 1993; Green and Szostak 1994), even though their exact geometry remains uncertain. Still in group I introns, several additional cases of mutually restricted variation suggesting triple interactions were spotted by breaking down the set of 87 sequences then available into subgroups (Michel and Westhof 1990). All of these suspected triples have now been confirmed using larger datasets (Gautheret et al. 1995a; F. Michel, unpubl.), and the ones involving tetraloops are extensively discussed in the next section. Finally, several base triples were recently mentioned to have been identified in ribosomal RNA (see Gutell 1995).

Although such successes would seem to suggest that base triples generate clear-cut patterns of covariation, careful analyses (see, e.g., Gautheret et al. 1995a) point to serious difficulties. tRNA sequence databases as well as known tRNA structures indicate that even within type I tRNAs, base triples are not as strictly conserved as most classic base pairs are (Brennan and Sundaralingam 1976). Much worse, base triples, which often come in groups of contiguous interactions, readily rearrange during evolution: In the complex of tRNA Ser (GGA) with its synthetase, base pair 13-22 does not interact with position 46, as in other available type I structures, but rather with position 9 (which usually contacts the next base pair, 12-23).

Even when base triples can be assumed to be generally conserved, as in group I introns between helix P4 and the J6/7 connecting segment, it is often unclear why some base combinations should be acceptable and others not. Combinatorial effects, consisting in fine adjustment of one triple interaction in response to slight geometric alterations of the neighboring interaction, have repeatedly been invoked to account for covariation data (Michel et al. 1990; Green and Szostak 1994; Gautheret et al. 1995a), but no specific scheme could be proposed (the problem is compounded in many group I introns by the fact that one of the P4:J6/7 base-triples is a quadruple interaction [Flor et al. 1989; Michel and Westhof 1990; Cate et al. 1996]). Such networks of local correlations as generated by coevolution and rearrangement of neighboring triples have been argued to constitute the hallmark of triple interactions (Gautheret et al. 1995a), but in practice, they obscure identification and confuse geometric interpretation.

One might have hoped that the introduction of base substitutions in a well-defined structural context would help clarify rules for the formation of triples in the deep and shallow grooves. Unfortunately, disruption of tRNA triples in tRNA Phe had only modest effects on aminoacylation (Sampson et al. 1990), whereas the use of base substitutions (Michel et al. 1990) and *in vitro* selection (Green and Szostak 1994) to explore group I intron triples mostly confirmed that natural combinations are the ones that tend to work best in self-splicing and related reactions. Direct measurements (see, e.g., Jaeger et al. 1993) of the extent to which specific substitutions destabilize the overall tertiary structure of the molecules under scrutiny may prove more informative. Much can be expected also of the use of nonnatural bases. In group I introns, proof that the guanine base of the guanosine cofactor of self-splicing contacts the guanine of a conserved G:C pair in helix P7 could be obtained by replacing the former with 2-amino purine and the latter by A:U, G:U, and A:C (Michel et al. [1989a]; whether this G:G:C base triple is best described as approximately coplanar or, rather, somewhat coaxial, remains debatable: Compare Yarus et al. [1991] with von Ahsen and Noller [1993]).

Coaxially Stacked Helices and Pseudoknots

Such is the preference for bases to stack upon their neighbors that even base-paired segments that appear separate in secondary structure models may end up as components of a single continuous stack in space (see, e.g., Pley et al. 1994a). *A fortiori*, any two potential helices that remain contiguous throughout evolution may be suspected of coaxial stacking. However, a much stronger case can be made when two neighboring helices happen to vary in size among related molecules, but in such a way that their overall length remains constant. That is observed for helices P1 and P2 in several subgroups of group I introns (Michel and Westhof 1990) and two pairs of helices in ribosomal RNA (for review, see Gutell 1995).

In the so-called pseudoknots originally identified at the 3' end of plant RNA viruses by CSA (Pleij et al. 1985), the bases on the 3' side of a terminal loop engage in classic base-pairing with an external segment of sequence. Coaxial stacking of the resulting helix with the one at the base of the loop was proposed to be an essential feature of such structures long before this stacking was confirmed by the NMR studies of Puglisi et al. (1990; for review, see Pleij 1990; ten Dam et al. 1992; Westhof and Jaeger 1992). Base-pairing involving terminal loops is often found in large RNAs. Whenever the paired bases happen to abut on the

3' branch of the supporting helix, as is the case for instance for the two exon-binding sequences of subgroup IIA introns (Michel et al. 1989b), it is reasonable to contemplate the possibility of continuous stacking.

Can reengineering of molecules offer arguments in favor of coaxial stacking? As demonstrated by Murphy et al. (1994), whose experiments come closest to what may be regarded as proof in these matters, when the P4-P5-P6 domain of group I introns is redesigned using circular permutation in such a way that helical segments P4 and P6 become a single continuous helix, its overall structure remains unaltered (and that despite the fact that the helical twist at the P4-P6 junction must have been modified to some extent).

Tetraloops and Tetraloop-interacting Motifs

Probably the most unexpected feature that has emerged from CSA of large structured RNA is the magnitude of the constraints nature exerts on the length and sequence of terminal loops. Despite the fact that calculations based on statistical mechanics had suggested that terminal loops of 7–8 nucleotides should be thermodynamically favored (Turner and Sugimoto 1988 and references therein), the first secondary-structure models of ribosomal RNAs revealed a predominance of four-nucleotide loops (tetraloops). This bias is largely attributable to a huge excess of loop sequences belonging to just two families, UNCG and GNRA (for review, see Woese et al. 1990; N is any nucleotide and R stands for a purine). GNRA loops abound as well in RNase P RNAs, and also in the two groups of self-splicing introns (in which UNCG loops are not so common). In addition, both UNCG and GNRA loops are found as part of prokaryotic transcription terminators (d'Aubenton-Carafa et al. 1990). Following these observations, those two families of tetraloops were targeted for physical-chemical studies, which eventually resulted in NMR-derived three-dimensional models (Cheong et al. 1990; Heus and Pardi 1991). Both UNCG and GNRA loops have well-defined, compact structures, whose common features include base-pairing between the first and last nucleotides and stacking of at least three of the four bases on their neighbors (crystal-derived structures [Pley et al. 1994a,b; Scott et al. 1995; Cate et al. 1996] now exist for GNRA loops, the diagnostic features of which were originally predicted by modeling based on chemical accessibilities [Westhof et al. 1989]).

Despite these structural similarities, different rationales for the abundance of UNCG and GNRA loops in natural RNAs had by then been proposed. On the one hand, UUCG loops had been found to bring

exceptional stability to the hairpin structure of which they are a part (Tuerk et al. 1988). On the other, it was suggested, based on CSA and subsequent modeling of group I introns (Michel and Westhof 1990), that GNRA loops participate in RNA tertiary structure by interacting with specific base pairs in the shallow groove of helices. With regard to the possibility that GNRA loops also stabilize hairpin structures, something which is all too often taken for granted in the literature, the following facts apply: (1) Inclusion of bonuses for GNRA tetraloops does improve the predictive power of *ab initio* folding programs based on minimization of calculated energies (Jaeger et al. 1989; Konings and Gutell 1995); unfortunately, this only means that it is a reasonable strategy to favor structures that include commonly observed components, and has nothing to say about the possible basis for nature's preferences. (2) Melting studies of oligonucleotides have failed to provide truly conclusive evidence that GNRA loops markedly stabilize RNA hairpin structures: Contrast discussions in Antao et al. (1991) and SantaLucia et al. (1992).

In the two instances of covariation uncovered by Michel and Westhof (1990), a specific base pair within a helix tends to be G:C when the third base of a distant GYRA loop is an A, whereas A:U is overwhelmingly preferred when the same third base is a G. Moreover, the next base pair in the 3' direction along the helix is usually found to be G:C. On the basis of these observations and the proposed architecture for group I introns, a model was suggested in which the last two bases of GNRA loops contact two consecutive purine-pyrimidine pairs in the shallow groove of the target helix. Since then, two instances of interactions between GNAA loops and helices have been found as packing contacts in two distinct crystals of the hammerhead ribozyme (Pley et al. 1994b; Scott et al. 1995). Comparison of the crystal-derived structures with the model of Michel and Westhof (1990) (see Fig. 1) illustrates both the power and limitations of CSA. The loop interacts with two consecutive G:C pairs in the shallow groove of the helix, and purine strands are in parallel orientation, just as predicted, but the model and structure have only one hydrogen bond in common, and that is the one responsible for covariation of the third base of the loop and its purine target (in the crystal structure, all but one of the seven other hydrogen bonds implicated in binding of the loop to the helix involve at least one 2' OH group and therefore could not, in principle, have been deduced by CSA).

Following the initial observations of Michel and Westhof (1990), additional instances of the same type of covariation were uncovered at several sites within group I and group II introns, and in most cases, the proposed contacts have been vindicated by experiments (Jaeger et al.

its preferred 11-nucleotide partner than other GNRA loops have for their specific targets (Costa and Michel 1995; Chanfreau and Jacquier 1996). The structural basis for such an efficient and specific recognition has now become apparent, because the 160-nucleotide group I intron domain, the structure of which was recently solved at atomic resolution by Cate et al. (1996), includes one instance of a GAAA loop interacting with its 11-nucleotide receptor.

The universality of tetraloops and tetraloop-binding motifs as building blocks of RNA architecture is becoming more and more obvious as similar types of interactions are being revealed by CSA of structured RNAs other than the self-splicing introns. Analysis of RNase P RNA sequences from natural microbial populations suggests that GNRA loops at the tips of helices P14 and P18 dock from opposite directions into the same helix (P8), where they could contact two adjacent sets of two base pairs each (Brown et al. 1996). In addition, most RNase P RNAs from gram-positive bacteria have a GAAA loop at the end of stem P2, and that loop probably interacts with stem P10.1a via the same 11-nucleotide motif that was identified in self-splicing introns (Tanner and Cech 1995). Finally, in 16S ribosomal RNA, the third nucleotide of the 1266-1269 GNRA loop (*Escherichia coli* numbering) has been reported to covary with the 1311-1326 base pair in the manner originally described for GYNA loops in group I introns (Gutell 1995).

Still, most of the GNRA loops in ribosomal RNA (and some of those in self-splicing introns as well) have not yet been implicated in RNA tertiary contacts, even though many of them are highly conserved by evolution. One likely possibility is that at least some of the rRNA GNRA loops are contacted by proteins. Ricin, a cytotoxic protein that cleaves 28S ribosomal RNA, specifically recognizes GAGA loops (Glück et al. 1992). It is also a safe bet that additional RNA partners of GNRA loops remain to be identified. Indeed, the 16S 863-866 tetraloop, when GAAA, would contact positions 571 and 570 in a way unrelated to the interactions described heretofore (see Gutell 1995). At the same time, it is worth stressing that all the other currently known tertiary motifs involving GNRA loops belong to the same structural family, in the sense that they appear to share a common contact between a G:C base pair and the last A of the loop (see Fig. 2 in Costa and Michel 1995). Just as is seen, for instance, with interactions between zinc fingers and DNA, evolution is bound to favor those structural motifs that are organized into closely knit families, since this makes it possible for the specificity of recognition of molecular partners to change and evolve without the molecules ever taking major leaps in sequence space. Ironically, the very feature that al-

lowed GNRA loops to invade so many sites in a collection of independently evolving RNA molecules is also the feature that has made CSA so successful in tracing their partners.

AUTOMATION OF CSA

Even when the appropriate sequence databases already exist, and computers are called upon for the more repetitive tasks, CSA can remain a time-consuming process, and especially so when evidence for tertiary interactions, rather than only contiguous classic base pairs, is sought. There have been attempts to optimize not only CSA *per se*, i.e., the search for covariations with possible implications for structure, but also each stage of the overall deciphering process it is part of and which may stretch all the way from the collection of appropriate sequences to the building of three-dimensional models. These endeavors have often met with some significant measure of success, but serious difficulties were also encountered, and these mishaps make it easier to grasp the intellectual challenges ahead in a field where human expertise still remains the benchmark.

Automated Sequence Collection

More often than not, the sequences that constitute the raw materials for CSA analyses were determined for other purposes and the resulting datasets may be too small or too biased—many sequences have not diverged to appropriate extents—for sophisticated analyses. In such cases, the missing data may be obtained by sequencing additional representatives of the molecular family of interest from organisms with the right extent of phylogenetic divergence. This method should work when the phylogeny of molecules and that of the organisms that encode them are congruent. Unfortunately, this cannot be assumed to be generally the case for multi-gene families or transposable elements, such as group I or group II self-splicing introns. Another more indiscriminate, but highly efficient, approach was recently pioneered by Brown *et al.* (1996). They used consensus primers to amplify by polymerase chain reaction (PCR) part of the gene for RNase P RNA from natural populations of microorganisms, and then cloned and sequenced a number of molecules from the resulting mixtures. The 52 novel sequences thus obtained, which have more than doubled the size of the RNase P RNA database, are highly diverse and, when compared with one another, revealed two novel tertiary contacts in the molecule.

One obvious drawback of this approach is that sections of the molecule distal to the primers are lost. In addition, the possibility exists of generating recombinant molecules during the amplification process (one instance of a chimeric sequence was indeed observed). Finally, sequences that match the primers poorly or come from organisms with low population densities are unlikely to be recovered. On the positive side, the source organisms need not be cultivated and particular sequence subgroups within a family of molecules can be targeted by using specific primers.

Automated Covariation Analysis

Assuming the sequences of interest have already been aligned (see next section), the problem then is to sort out "meaningful" covariations from "spurious" ones by using some estimate of statistical significance. The difficulties involved in choosing the right estimate are best grasped by realizing that superficially similar datasets can result from completely different histories. Just consider two aligned positions with roughly equal frequencies of two possible bases each, and assume that the contents of one site can be fully predicted from those of the other site; in other words, a case of perfect apparent covariation (Fig. 2). One possible explanation for this situation is that it arose out of necessity; i.e., the two combinations of bases observed are the only ones that support function (this is the underlying assumption when using CSA to infer structure). However, the same pattern could just as well have been generated through historical coincidence, with the base at each site having changed just once, but the two substitutions having occurred by chance in the same deep branch of the evolutionary tree.

Clearly, the way to distinguish between these two possibilities, i.e., chance or necessity, is to find out whether multiple changes were involved in generating the observed patterns (the problem does not exist, of course, if the sequences being compared have been generated by *in vitro* selection, without ungoing mutagenesis). Numbers of substitutions at a site can be estimated readily, although with some uncertainty, when a preestablished phylogenetic tree is available, but even when molecules cannot be assumed to have had the same history as the genomes that encode them, it remains possible to build a phylogenetic tree from the set of aligned sequences. Tree-building methods using principles of parsimony or maximum likelihood (for review, see Swofford et al. 1996) in effect sort out sites according to probable numbers of events: So-called "homoplasic" sites, which have undergone too many substitutions to par-

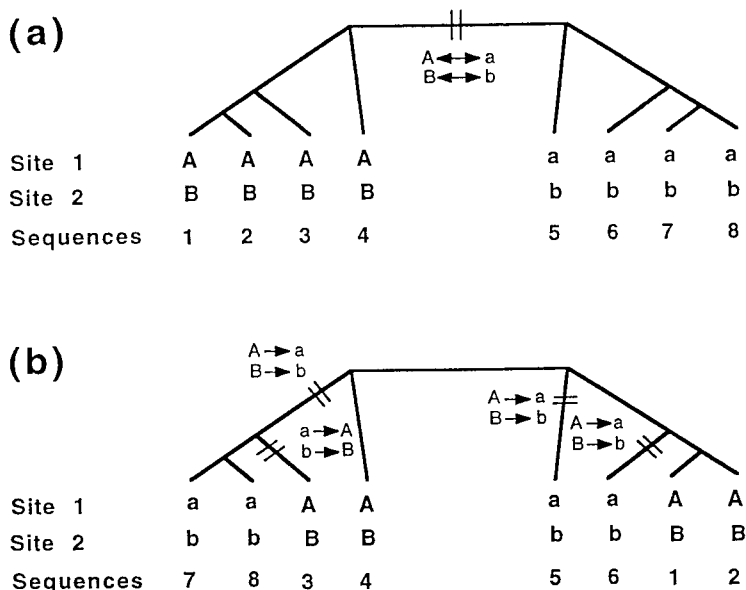


Figure 2 Different possible interpretations of a case of perfect covariation depending on the underlying evolutionary tree. (a) An unrooted tree topology that requires no more than two substitutions to be postulated in order to account for the situation at sites 1 and 2. (b) An alternate tree topology that requires the assumption that no less than eight substitutions occurred. These could consist of four pairs of coordinated changes, which strongly suggests that the nature of the residue at one site is constrained by the situation at the other site. In contrast, the situation in *a* does not allow a conclusion in favor or against such a constraint.

ticipate in phylogenetic reconstruction, are precisely the ones with the greatest potential for inferring structure.

One might conclude from this discussion that searching for apparent coordination of events should be privileged. However, interacting bases need not change simultaneously provided some intermediate states can be tolerated. For instance, as already mentioned, G:U pairs provide a pathway by which G:C and A:U pairs can be exchanged without base-pairing ever being lost. Only if such intermediates are sufficiently short-lived (sufficiently deleterious) to be infrequent in sequence samples will events appear to be concerted (for quantitative analyses and discussions, see Rousset et al. 1991).

In conclusion, any estimate of the significance of a covariation should take into account both the extent of statistical bias in the distribution of bases and the numbers of underlying substitutions. Unfortunately, no single parameter has yet been proposed that would unify these two

aspects, and currently available approaches either privilege one or the other side of the coin, or else are empirical compromises. Thus, the computer program of Winker et al. (1990) assumes a phylogenetic tree is already available and uses it to record and compare lists of changes; only thereafter does the program eliminate pairs of sites that, despite suggestive evidence of synchronous change, happen to tolerate multiple, conflicting combinations of bases. In contrast, Gutell and coworkers have consistently looked upon statistical biases in the distribution of bases at two or more sites as primary indicators of covariation. Mutual information, first introduced to the field by Chiu and Kolodziejczak (1991), was advocated as a more sensitive measure of statistical bias (Gutell et al. 1992), but a recent analysis of triple interactions by Gautheret et al. (1995a) uses traditional chi-square analysis (rare combinations are a source of problems whatever the estimate being used). In the latter work, phylogenetic screening of covariations was also introduced as a secondary sieve, and group I intron data were reanalyzed using essentially the same division into phylogenetic groups and subgroups as in the analysis of Michel and Westhof (1990; these authors had sorted out covariation data according to subgroups in order to provide lower bounds for the numbers of substitution events). Note, however, that the estimate—ratio of concerted changes over total number of changes—devised by Gautheret et al. (1995a) is highest for the inconclusive situation in which a single, possibly concerted change is observed.

Automated Alignment

Alignment procedures based exclusively on comparison of sequences are inadequate except for very closely related RNA molecules. The main reason for this is the high frequency with which deletions and insertions occur compared to substitutions during evolution of noncoding RNAs. It follows that reaching a reliable alignment requires that at least secondary structure, and possibly also some elements of tertiary structure, be taken into account. However, since identification of secondary structure components by CSA implies that sequences have been aligned beforehand, the problem is clearly a circular one. Methods exist for adding new sequences to a preexisting alignment by taking both sequence and secondary structure into account (see, e.g., Corpet and Michot 1994). A practical, but not fully rigorous solution to the problem of simultaneously building a multiple alignment and deriving a common secondary structure was proposed by Eddy and Durbin (1994). At the root of this approach is the elaboration of what the authors call a "covariance model,"

in effect a probabilistic model that precisely describes both sequence and secondary structure constraints on a family of molecules, takes into account the likelihood of deletions and insertions, and can also be used to identify additional members of the family in sequence databases. Covariance models are built in the same iterative way CSA is carried out by humans. Starting from a provisional alignment, a structure is sought that maximizes the total extent of covariation, using previously mentioned mutual information as an estimate of statistical bias; parameters of the model are optimized and, finally, sequences are realigned.

Covariance models have been reported to be highly successful in aligning tRNA sequences *ab initio*, finding a consensus secondary structure, and identifying tRNAs in primary sequence data. Unfortunately, their use is currently limited to sequences not much longer (150–200 nucleotides, according to the authors) than those of tRNAs because of computational demands. A more irremediable limitation, which stems from the dynamic programming algorithms used both to build covariance models and to align sequences to a model, is the impossibility to deal with other than tree-like structures: Interactions such as pseudoknots that, because they straddle elements of secondary structure, are commonly viewed as belonging to tertiary structure, cannot be incorporated into that framework (note, however, that when an alignment is available, methods other than dynamic programming exist for finding the combination of sites that will result in maximizing total mutual information; see Cary and Stormo 1995).

Automated Model Building

The major limitation of current computer-based methods in carrying out automatic alignment remains their inability to summon all relevant information (both on the molecule of interest and RNA structure in general) when having to decide whether two putative substructures should be regarded as homologous. Another area in which human expertise has not yet been seriously challenged is structural reconstruction from covariation patterns. Although guidelines are being elaborated to help infer various types of base-base pairing and triple interactions (see, e.g., Gautheret et al. 1994, 1995a,b), interpretation remains subject to context, and it must be kept in mind that covariation does not necessarily imply base-base contacts. Nor should it be assumed that all contacts inferred exist at the same time: Part of the power of CSA is its ability to address alternative molecular states, whose possible existence must eventually be experimentally tackled.

Once a list has been constituted of putatively coexisting interactions, one is left with the daunting prospect of having to incorporate all relevant information into a working three-dimensional model. Although this aspect is beyond the scope of this review (see instead Gautheret and Cedergren 1993; Westhof and Michel 1994), it may be mentioned that by far the best prospects for an alternative to human expertise reside in constraint satisfaction algorithms (Major et al. 1991) because of their built-in ability to incorporate all types of potentially relevant information. At present, however, human intervention remains essential both in formulating problems that are simple enough for the available programs to cope with and in sorting out outputs according to such criteria as the likelihood of proposed solutions being compatible with physically reasonable folding processes.

PERSPECTIVES

Are current attempts to codify and automate CSA worth the effort? As far as secondary structure is concerned, this is not an issue: CSA is by far the most efficient way to infer secondary structure, and programs that seek potential Watson-Crick pairs, with or without prior alignment of the sequences to be searched, are neither very difficult to conceive nor absolutely indispensable (although they do save much time). On the other hand, such large numbers of sequences are necessary to uncover tertiary interactions that exhaustive searches cannot reasonably be carried out without at least some automation. Automation is also a prerequisite for any sophisticated analysis of the patterns and rates of evolutionary change: It is on our very ability to model evolution, both nucleotide by nucleotide and for combinations of sites, that progress in uncovering structurally significant constraints (and also in fact, in phylogenetic reconstruction; see Swofford et al. 1996) will rest. Better descriptions of evolutionary change that acknowledge the existence of evolutionary intermediates and make it possible to compare the contributions of different base combinations to molecular fitness (see, e.g., Rousset et al. 1991) should help as well in the interpretation of statistical constraints in a stereochemical framework.

On the practical side, however, the future of CSA in inferring RNA tertiary structure should depend primarily on whether or not the sequences of molecules of interest will accumulate faster than physical-chemical methods can solve their structures. Recent progress made in crystallography and NMR might have tipped the balance in favor of experimental approaches, but the current development of RNA-based in-

vitro-selection experiments is likely to generate enough sequences to reverse this trend, at least for the moment. In the long run, however, as the issue increasingly becomes that of our ability to predict the conformation of an RNA molecule directly from its sequence, we can expect that not only structural databases, but also the comparative databases and the experience gained in exploiting them (if only to spot biologically significant structural motifs) should prove invaluable.

ACKNOWLEDGMENTS

We are grateful to all our colleagues who supplied preprints and reprints, and to Luc Jaeger, Bruno Sargueil, and Eric Westhof for critical reading of the manuscript.

REFERENCES

- Antao, V.P., S.Y. Lai, and I. Tinoco, Jr. 1991. A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Res.* **19**: 5901–5905.
- Banerjee, A.R., J.A. Jaeger, and D.H. Turner. 1993. Thermal unfolding of a group I ribozyme: The low-temperature transition is primarily disruption of tertiary structure. *Biochemistry* **32**: 153–163.
- Bartel, D.P., M.L. Zapp, M.R. Green, and J.W. Szostak. 1991. HIV-1 Rev regulation involves recognition of non-Watson-Crick base pairs in viral RNA. *Cell* **67**: 529–536.
- Battiste, J.L., R. Tan, A.D. Frankel, and J.R. Williamson. 1995. Assignment and modeling of the Rev response element RNA bound to a Rev peptide using ¹³C-heteronuclear NMR. *J. Biomol. NMR* **6**: 375–389.
- Been, M.D. and A.T. Perrotta. 1995. Optimal self-cleavage activity of the hepatitis delta virus RNA is dependent on a homopurine base pair in the ribozyme core. *RNA* **1**: 1061–1070.
- Branlant, C., A. Krol, M.A. Machatt, J. Pouyet, J.P. Ebel, K. Edwards, and H. Kössel. 1981. Primary and secondary structures of *Escherichia coli* mre 600 23S ribosomal RNA. Comparison with models of secondary structure for maize chloroplast 23S rRNA and for large portions of mouse and human 16S mitochondrial rRNAs. *Nucleic Acids Res.* **9**: 4303–4324.
- Brennan, T. and M. Sundaralingam. 1976. Structure of transfer RNA molecules containing the long variable loop. *Nucleic Acids Res.* **3**: 3235–3251.
- Brown, J.W., J.M. Nolan, E.S. Haas, M.A.T. Rubio, F. Major, and N.R. Pace. 1996. Comparative analysis of ribonuclease P RNA using gene sequences from natural microbial populations reveals tertiary structural elements. *Proc. Natl. Acad. Sci.* **93**: 3001–3006.
- Cary, R.B. and G.D. Stormo. 1995. Graph-theoretic approach to RNA modeling using comparative data. In *Proceedings of the 3rd International Conference on Intelligent Systems for Molecular Biology* (ed. C. Rawlings et al.), pp. 75–80. AAAI Press, Menlo Park, California.
- Cate, J.H., A.R. Gooding, E. Podell, K.H. Zhou, B.L. Golden, C.E. Kundrot, T.R. Cech,

- and J.A. Doudna. 1996. Crystal structure of a group I ribozyme domain—Principles of RNA packing. *Science* **273**: 1678–1685.
- Cech, T.R. 1993. Structure and mechanism of the large catalytic RNAs: Group I and group II introns and ribonuclease P. In *The RNA world* (ed. R.F. Gesteland and J.F. Atkins), pp. 239–269. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Chanfreau, G. and A. Jacquier. 1996. An RNA conformational change between the two chemical steps of group II self-splicing. *EMBO J.* **15**: 3466–3476.
- Chastain, M. and I. Tinoco, Jr. 1993. Nucleoside triples from the group I intron. *Biochemistry* **32**: 14220–14228.
- Cheong, C., G. Varani, and I. Tinoco, Jr. 1990. Solution structure of an unusually stable RNA hairpin, 5' GGAC(UUCG)GUCC. *Nature* **346**: 680–682.
- Chiu, D.K.Y. and T. Kolodziejczak. 1991. Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci.* **7**: 347–352.
- Corpet, F. and B. Michot. 1994. RNAlign program: Alignment of RNA sequences using both primary and secondary structures. *Comput. Appl. Biosci.* **10**: 389–399.
- Costa, M. and F. Michel. 1995. Frequent use of the same tertiary motif by self-folding RNAs. *EMBO J.* **14**: 1276–1285.
- Dammel, C.S. and H.F. Noller. 1993. A cold-sensitive mutation in 16S rRNA provides evidence for helical switching in ribosome assembly. *Genes Dev.* **7**: 660–670.
- d'Aubenton-Carafa, Y., E. Brody, and C. Thermes. 1990. Prediction of rho-independent *Escherichia coli* transcription terminators: A statistical analysis of their RNA stem-loop structures. *J. Mol. Biol.* **216**: 835–858.
- Davies, R.W., R.B. Waring, J.A. Ray, T.A. Brown, and C. Scazzocchio. 1982. Making ends meet: A model for RNA splicing in fungal mitochondria. *Nature* **300**: 719–724.
- Doudna, J.A., B.P. Cormack, and J.W. Szostak. 1989. RNA structure, not sequence determines the 5' splice-site specificity of a group I intron. *Proc. Natl. Acad. Sci.* **86**: 7402–7406.
- Doudna, J.A., C. Grosshans, A. Gooding, and C.E. Kundrot. 1993. Crystallization of ribozymes and small RNA motifs by a sparse matrix approach. *Proc. Natl. Acad. Sci.* **90**: 7829–7833.
- Eddy, S.R. and R. Durbin. 1994. RNA sequence analysis using covariance models. *Nucleic Acids Res.* **22**: 2079–2088.
- Flor, P.J., J.B. Flanagan, and T.R. Cech. 1989. A conserved base pair within helix P4 of the *Tetrahymena* ribozyme helps to form the tertiary structure required for self-splicing. *EMBO J.* **8**: 3391–3399.
- Fox, G.E. and C.R. Woese. 1975. 5S RNA secondary structure. *Nature* **256**: 505–507.
- Gautheret, D. and R. Cedergren. 1993. Modeling the three-dimensional structure of RNA. *FASEB J.* **7**: 97–105.
- Gautheret, D., S.H. Damberger, and R.R. Gutell. 1995a. Identification of base-triples in RNA using comparative sequence analysis. *J. Mol. Biol.* **248**: 27–43.
- Gautheret, D., D. Konings, and R.R. Gutell. 1994. A major family of motifs involving G•A mismatches in ribosomal RNA. *J. Mol. Biol.* **242**: 1–8.
- . 1995b. G•U base pairing motifs in ribosomal RNA. *RNA* **1**: 807–814.
- Glutz, C., C. Zwieb, and R. Brimacombe. 1981. Secondary structure of the large subunit ribosomal RNA from *Escherichia coli*, *Zea mays* chloroplast, and human and mouse mitochondrial ribosomes. *Nucleic Acids Res.* **9**: 3287–3306.
- Glück, A., Y. Endo, and I.G. Wool. 1992. Ribosomal RNA identity elements for ricin A-

- chain recognition and catalysis. Analysis with tetraloop mutants. *J. Mol. Biol.* **226**: 411–424.
- Green, R. and J.W. Szostak. 1992. Selection of a ribozyme that functions as a superior template in a self-copying reaction. *Science* **258**: 1910–1915.
- . 1994. *In vitro* genetic analysis of the hinge region between helical elements P5-P4-P6 and P7-P3-P8 in the *sunY* group I self-splicing intron. *J. Mol. Biol.* **235**: 140–155.
- Gutell, R.R. 1993. Comparative studies of RNA: Inferring higher-order structure from patterns of sequence variation. *Curr. Biol.* **3**: 313–322.
- . 1995. Comparative sequence analysis and the structure of 16S and 23S rRNA. In *Ribosomal RNA: Structure, evolution, processing, and function in protein biosynthesis* (ed. R.A. Zimmermann and A.E. Dahlberg), pp. 109–126. CRC Press, Boca Raton, Florida.
- Gutell, R.R., N. Larsen, and C.R. Woese. 1994. Lessons from an evolving ribosomal RNA: 16S and 23S rRNA structure from a comparative perspective. *Microbiol. Rev.* **58**: 10–26.
- Gutell, R.R., A. Power, G.Z. Hertz, E.J. Putz, and G.D. Stormo. 1992. Identifying constraints on the higher-order structure of RNA: Continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.* **20**: 5785–5795.
- Guthrie, C. and B. Patterson. 1988. Spliceosomal snRNAs. *Annu. Rev. Genet.* **22**: 387–419.
- Heus, H.A. and A. Pardi. 1991. Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science* **253**: 191–194.
- Huang, S., Y.-X. Wang, and D.E. Draper. 1996. Structure of a hexanucleotide RNA hairpin loop conserved in ribosomal RNAs. *J. Mol. Biol.* **258**: 308–321.
- Jaeger, J.A., D.H. Turner, and M. Zuker. 1989. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci.* **86**: 7706–7710.
- Jaeger, L., F. Michel, and E. Westhof. 1994. Involvement of a GNRA loop in long-range RNA tertiary interactions. *J. Mol. Biol.* **236**: 1271–1276.
- Jaeger, L., E. Westhof, and F. Michel. 1993. Monitoring of the cooperative unfolding of the *sunY* group I intron of bacteriophage T4. The active form of the *sunY* ribozyme core is stabilized by multiple interactions with 3' terminal intron components. *J. Mol. Biol.* **234**: 331–346.
- James, B.D., G.J. Olsen, J. Liu, and N.R. Pace. 1988. The secondary structure of ribonuclease P RNA, the catalytic element of a ribonucleoprotein enzyme. *Cell* **52**: 19–26.
- Katahira, M., M. Kanagawa, H. Sato, S. Uesugi, S. Fujii, T. Kohno, and T. Maeda. 1994. Formation of sheared G:A base pairs in an RNA duplex modelled after ribozymes, as revealed by NMR. *Nucleic Acids Res.* **22**: 2752–2759.
- Kirsebom, L.A. and S.G. Svård. 1994. Base pairing between *Escherichia coli* RNase P RNA and its substrate. *EMBO J.* **13**: 4870–4876.
- Konings, D. and R.R. Gutell. 1995. A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. *RNA* **1**: 559–574.
- Larsen, N. and C. Zwieb. 1991. SRP-RNA sequence alignment and secondary structure. *Nucleic Acids Res.* **19**: 209–215.
- Levitt, M. 1969. Detailed molecular model for transfer ribonucleic acid. *Nature* **224**: 759–763.
- Lodmell, J.S., R.B. Gutell, and A.E. Dahlberg. 1995. Genetic and comparative analyses

- reveal an alternative secondary structure in the region of nt 912 of *Escherichia coli* 16S rRNA. *Proc. Natl. Acad. Sci.* **92**: 10555–10559.
- Madhani, H.D. and C. Guthrie. 1994. Dynamic RNA-RNA interactions in the spliceosome. *Annu. Rev. Genet.* **28**: 1–26.
- Madison, J.T., G.A. Everett, and H.K. Kung. 1966. On the nucleotide sequence of yeast tyrosine transfer RNA. *Cold Spring Harbor Symp. Quant. Biol.* **31**: 409–416.
- Major, F., M. Turcotte, D. Gautheret, G. Lapalme, E. Fillion, and R. Cedergren. 1991. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science* **253**: 1255–1260.
- Michel, F. and E. Westhof. 1990. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.* **216**: 585–610.
- Michel, F., A. Jacquier, and B. Dujon. 1982. Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. *Biochimie* **64**: 867–881.
- Michel, F., K. Umesono, and H. Ozeki. 1989a. Comparative and functional anatomy of group II catalytic introns—A review. *Gene* **82**: 5–30.
- Michel, F., A.D. Ellington, S. Couture, and J.W. Szostak. 1990. Phylogenetic and genetic evidence for base triple formation in the catalytic domain of group I introns. *Nature* **347**: 578–580.
- Michel, F., M. Hanna, R. Green, D.P. Bartel, and J.W. Szostak. 1989b. The guanosine binding site of the *Tetrahymena* ribozyme. *Nature* **342**: 391–395.
- Murphy, F.L. and T.R. Cech. 1994. GAAA tetraloop and conserved bulge stabilize tertiary structure of a group I intron domain. *J. Mol. Biol.* **236**: 49–63.
- Murphy, F.L., Y.-H. Wang, J.D. Griffith, and T.R. Cech. 1994. Coaxially stacked RNA helices in the catalytic center of the *Tetrahymena* ribozyme. *Science* **265**: 1709–1712.
- Nardelli, J., T.J. Gibson, C. Vesque, and P. Charnay. 1991. Base sequence discrimination by zinc-finger DNA-binding domains. *Nature* **349**: 175–178.
- Nick, H. and W. Gilbert. 1985. Detection in vivo of protein-DNA interactions within the *lac* operon of *Escherichia coli*. *Nature* **313**: 795–798.
- Ninio, J. 1979. Prediction of pairing schemes in RNA molecules—Loop contributions and energy of wobble and non-wobble pairs. *Biochimie* **61**: 1133–1150.
- Noller, H.F., J. Kop, V. Wheaton, J. Brosius, R.R. Gutell, A.M. Kopylov, F. Dohme, W. Herr, D.A. Stahl, R. Gupta, and C.R. Woese. 1981. Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Res.* **9**: 6167–6189.
- Nussinov, R. and A.B. Jacobson. 1980. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci.* **77**: 6309–6313.
- Pace, N.R., D.K. Smith, G.J. Olsen, and B.D. James. 1989. Phylogenetic comparative analysis and the secondary structure of ribonuclease P RNA—A review. *Gene* **82**: 65–75.
- Papanicolaou, C., M. Gouy, and J. Ninio. 1984. An energy model that predicts the correct folding of both the tRNA and the 5S RNA molecules. *Nucleic Acids Res.* **12**: 31–44.
- Parker, R. and P.G. Siliciano. 1993. Evidence for an essential non-Watson-Crick interaction between the first and last nucleotides of a nuclear pre-mRNA intron. *Nature* **361**: 660–662.
- Peterson, R.D., D.P. Bartel, J.W. Szostak, S.J. Horvath, and J. Feigon. 1994. ¹H NMR studies of the high-affinity Rev binding site of the Rev responsive element of HIV-1 mRNA: Base pairing in the core binding element. *Biochemistry* **33**: 5357–5366.
- Pleij, C.W.A. 1990. Pseudoknots: A new motif in the RNA game. *Trends Biochem. Sci.*

15: 143–147.

- Pleij, C.W.A., K. Rietveld, and L. Bosch. 1985. A new principle of RNA folding based on pseudo-knotting. *Nucleic Acids Res.* **13**: 1717–1731.
- Pley, H.W., K.M. Flaherty, and D.B. McKay. 1994a. Three-dimensional structure of a hammerhead ribozyme. *Nature* **372**: 68–74.
- . 1994b. Model for an RNA tertiary interaction from the structure of an intermolecular complex between a GAAA tetraloop and an RNA helix. *Nature* **372**: 111–113.
- Powers, T. and H.F. Noller. 1994. Selective perturbation of G530 of 16S rRNA by translational miscoding agents and a streptomycin-dependence mutation in protein S12. *J. Mol. Biol.* **235**: 156–172.
- Puglisi, J.D., J.R. Wyatt, and I. Tinoco, Jr. 1990. Conformation of an RNA pseudoknot. *J. Mol. Biol.* **214**: 437–453.
- RajBhandary, U.L., A. Stuart, R.D. Faulkner, S.H. Chang, and H.G. Khorana. 1966. Nucleotide sequence studies on yeast phenylalanine sRNA. *Cold Spring Harbor Symp. Quant. Biol.* **31**: 425–434.
- Romero, D.P. and E.H. Blackburn. 1991. A conserved secondary structure for telomerase RNA. *Cell* **67**: 343–353.
- Rousset, F., M. Pélandakis, and M. Solignac. 1991. Evolution of compensatory substitutions through G•U intermediate state in *Drosophila* rRNA. *Proc. Natl. Acad. Sci.* **88**: 10032–10036.
- Samaha, R.R., R. Green, and H.F. Noller. 1995. A base pair between tRNA and 23S rRNA in the peptidyl transferase center of the ribosome. *Nature* **377**: 309–314.
- Sampson, J.R., A.B. DiRenzo, L.S. Behlen, and O.C. Uhlenbeck. 1990. Role of the tertiary nucleotides in the interaction of yeast phenylalanine tRNA with its cognate synthetase. *Biochemistry* **29**: 2523–2532.
- Sänger, W. 1984. *Principles of nucleic acid structure*. Springer-Verlag, New York.
- SantaLucia, J., Jr. and D.H. Turner. 1991. Stabilities of consecutive A•C, C•C, G•G, U•C, and U•U mismatches in RNA internal loops: Evidence for stable hydrogen-bonded U•U and C•C+ pairs. *Biochemistry* **30**: 8242–8251.
- . 1993. Structure of (rGGCGAGCC)₂ in solution from NMR and restrained molecular dynamics. *Biochemistry* **32**: 12612–12623.
- SantaLucia, J., Jr., R. Kierzek, and D.H. Turner. 1992. Context dependence of hydrogen bond free energy revealed by substitutions in an RNA hairpin. *Science* **256**: 217–219.
- Scott, W.G., J.T. Finch, and A. Klug. 1995. The crystal structure of an all-RNA hammerhead ribozyme: A proposed mechanism for RNA catalytic cleavage. *Cell* **81**: 991–1002.
- Stebbins-Boaz, B. and S.A. Gerbi. 1991. Structural analysis of the peptidyl transferase region in ribosomal RNA of the eukaryote *Xenopus laevis*. *J. Mol. Biol.* **217**: 93–112.
- Stiegler, P., P. Carbon, J.P. Ebel, and C. Ehresmann. 1981. A general secondary structure model for procaryotic and eucaryotic RNAs of the small ribosomal subunits. *Eur. J. Biochem.* **120**: 487–495.
- Strobel, S.A. and T.R. Cech. 1995. Minor groove recognition of the conserved G•U pair at the *Tetrahymena* ribozyme reaction site. *Science* **267**: 675–679.
- Swofford, D.L., G.J. Olsen, P.J. Waddell, and D.M. Hillis. 1996. Phylogenetic inference. In *Molecular systematics*, 2nd edition (ed. D.M. Hillis et al.), pp. 407–514. Sinauer Associates, Sunderland, Massachusetts.
- Szewczak, A.A., P.B. Moore, Y.-L. Chan, and I.G. Wool. 1993. The conformation of the

- sarcin/ricin loop from 28S ribosomal RNA. *Proc. Natl. Acad. Sci.* **90**: 9581–9585.
- Tanner, M.A. and T.R. Cech. 1995. An important RNA tertiary interaction of group I and group II introns is implicated in gram-positive RNase P RNAs. *RNA* **1**: 349–350.
- ten Dam, E., C.W.A. Pleij, and D. Draper. 1992. Structural and functional aspects of RNA pseudoknots. *Biochemistry* **31**: 11665–11676.
- Tuerk, C., P. Gauss, C. Thermes, R. Groebe, M. Gayle, N. Guild, G. Stormo, Y. d'Aubenton-Carafa, O.C. Uhlenbeck, I. Tinoco, Jr., E.N. Brody, and L. Gold. 1988. CUUCGG hairpins: Extraordinary stable RNA secondary structures associated with various biochemical processes. *Proc. Natl. Acad. Sci.* **85**: 1364–1368.
- Turner, D.H. and N. Sugimoto. 1988. RNA structure prediction. *Annu. Rev. Biophys. Biophys. Chem.* **17**: 167–192.
- von Ahsen, U. and H.F. Noller. 1993. Methylation interference experiments identify bases that are essential for distinct catalytic functions of a group I ribozyme. *EMBO J.* **12**: 4747–4754.
- Walter, A.E., M. Wu, and D.H. Turner. 1994. The stability and structure of tandem GA mismatches in RNA depend on closing base pairs. *Biochemistry* **33**: 11349–11354.
- Watson, J.D. and F.H.C. Crick. 1953. A structure for deoxyribonucleic acid. *Nature* **171**: 737–738.
- Westhof, E. 1992. Westhof's rule. *Nature* **358**: 459–460.
- Westhof, E., and L. Jaeger. 1992. RNA pseudoknots. *Curr. Opin. Struct. Biol.* **2**: 327–333.
- Westhof, E. and F. Michel. 1994. Prediction and experimental investigation of RNA secondary and tertiary foldings. In *RNA-protein interactions* (ed. K. Nagai and I.W. Mattaj), pp. 25–51. Oxford University Press, Oxford, United Kingdom.
- Westhof, E., P. Romby, P.J. Romaniuk, J.-P. Ebel, C. Ehresmann, and B. Ehresmann. 1989. Computer modeling from solution data of spinach chloroplast and of *Xenopus laevis* somatic and oocyte 5S rRNAs. *J. Mol. Biol.* **207**: 417–431.
- Wimberley, B., G. Varani, and I. Tinoco, Jr. 1993. The conformation of loop E of eukaryotic 5S ribosomal RNA. *Biochemistry* **32**: 1078–1087.
- Winker, S., R. Overbeek, C.R. Woese, G.J. Olsen, and N. Pfluger. 1990. Structure detection through automated covariance search. *Comput. Appl. Biosci.* **6**: 365–371.
- Woese, C.R. and N.R. Pace. 1993. Probing RNA structure, function and history by comparative analysis. In *The RNA world* (ed. R.F. Gesteland and J.F. Atkins), pp. 91–117. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Woese, C.R., S. Winker, and R.R. Gutell. 1990. Architecture of ribosomal RNA: Constraints on the sequence of tetraloops. *Proc. Natl. Acad. Sci.* **87**: 8467–8471.
- Woese, C.R., L.J. Magrum, R. Gupta, R.B. Siegel, D.A. Stahl, J. Kop, N. Crawford, J. Brosius, R.R. Gutell, J.J. Hogan, and H.F. Noller. 1980. Secondary structure model for bacterial 16S ribosomal RNA: Phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res.* **8**: 2275–2293.
- Yarus, M., M. Illangsekare, and E. Christian. 1991. An axial binding site in the *Tetrahymena* precursor RNA. *J. Mol. Biol.* **222**: 995–1012.
- Zachau, H.G., D. Dütting, H. Feldmann, F. Melchers, and W. Karau. 1966. Serine specific transfer ribonucleic acids. XIV. Comparison of nucleotide sequences and secondary structure models. *Cold Spring Harbor Symp. Quant. Biol.* **31**: 417–424.
- Zaug, A.J. and T.R. Cech. 1995. Analysis of the structure of *Tetrahymena* nuclear RNAs in vivo: Telomerase RNA, the self-splicing rRNA intron, and U2 snRNA. *RNA* **1**: 363–374.

- Zuker, M. and P. Stiegler. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **14**: 299–315.
- Zuker, M., J.A. Jaeger, and D.H. Turner. 1991. A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nucleic Acids Res.* **19**: 2707–2714.
- Zwieb, C., C. Glotz, and R. Brimacombe. 1981. Secondary structure comparisons between small subunit ribosomal RNA molecules from six different species. *Nucleic Acids Res.* **9**: 3621–3640.