A proposal for the secondary structure of a variable area of eukaryotic small ribosomal subunit RNA involving the existence of a pseudoknot

Jean-Marc Neefs and Rupert De Wachter*

Department Biochimie, Universiteit Antwerpen (UIA), Universiteitsplein 1, B-2610 Antwerp, Belgium

Received July 9, 1990; Revised and Accepted September 3, 1990

ABSTRACT

Eukaryotic small ribosomal subunit RNAs contain an area of variable structure, V4, which comprises about 250 nucleotides in most species, whereas the corresponding area in bacterial small ribosomal subunit RNAs consists of about 64 nucleotides folded into a single hairpin. There is no consensus on the secondary structure of area V4 in eukaryotes, about 10 different models having been proposed. The prediction of a model on a comparative basis poses special problems because, due to the variability of the area in length as well as sequence, a dependable alignment is very difficult to achieve. A new model was derived by systematic examination of all combinations of helices that have been hitherto proposed, plus some new ones. The following properties of the helices were examined: transposability to all presently known sequences, presence of compensating substitutions, and thermodynamic stability. A model was selected by ranking all possible combinations of transposable helices according to the number of compensating substitutions scored. The optimal model comprises a pseudoknot and four hairpin structures. Certain species contain additional hairpins inserted between these structural elements, while in others the structure is partially or entirely deleted.

INTRODUCTION

The complete or nearly complete sequence of the small ribosomal subunit RNA (further abbreviated srRNA) has been published for 57 eukaryotes, 16 archaebacteria, 138 eubacteria, 12 plastids and 31 mitochondria. Literature references and sequence library accession numbers can be found in (1). The extensive effort put into the comparative sequence analysis of this ribosome constituent is warranted by the fact that it is gradually revealing more details of its secondary structure (1−3) and even elements of its tertiary (4,5) structure. On the other hand, this analysis has been put to use in the investigation of bacterial evolution (6), eukaryotic evolution (7), and indeed in the study of the evolutionary relationships among all life forms, including the origins of plastids and mitochondria (8−10). The structural and evolutionary studies are intertwined in the sense that the derivation of evolutionary trees requires a dependable sequence alignment as a starting point, and the establishment of such an alignment is enhanced by the knowledge of secondary structure landmarks, such as the boundaries of helices and loops and the existence of compensating substitutions in complementary strands.

The outline of the secondary structure model for eukaryotic srRNAs, shown in Fig. 1, allows to distinguish 48 'universal' helices, so termed because they are common to eukaryotic, archaebacterial and eubacterial srRNAs. Several authors have pointed to the alternation of conserved and variable areas. The variability applies to the local sequence, but also to the length of the helices and loops forming the local secondary structure. Eight such variable areas are distinguished in Fig. 1. One of these areas, V4, is situated between the 21st and the 22nd universal helix counting from the 5′ terminus. This part of the molecule is folded into a single hairpin in bacteria and most organelles. In eukaryotic srRNAs it has an average length of about 250 nucleotides, and its variability is such that no consensus has been hitherto reached on a local secondary structure model. Some authors (11−15) prefer not to define any structure for this area in their models, whereas others (7, 16−27) have made a variety of proposals. This situation is unsatisfactory for students of srRNA structure as well as for those using it as a molecular clock. One would like to know whether the folding of variable areas can be radically different in srRNAs from different species or taxa, or whether the basic pattern is uniform, variability being confined to the length of single- and double-stranded areas forming the local structure. The evolutionists, in addition, could take advantage of variable, i. e. rapidly evolving areas of ribosomal RNAs in order to elucidate relationships among recently diverged species, provided that they can construct dependable alignments based on a credible secondary structure model.

We have reexamined area V4 on the basis of all presently available sequence data and attempted to derive an optimal model. To this end, we have systematically examined all helices proposed in previous models and tested whether they can be transposed to all presently known eukaryotic sequences. All combinations

* To whom correspondence should be addressed

of helices satisfying this condition into sterically possible models were then examined and ranked according to the number of compensating substitutions characterizing them.

In order to avoid confusion, the intended meaning of a number of terms, used frequently in the following paragraphs, is defined below.

Standard base pair: a Watson-Crick pair or the wobble pair G·U.

Non-standard base pair: one of the 7 remaining combinations (U-U, U-C, C-C, C-A, A-A, A-G, G-G).

Helix segment: a part of a helix uninterrupted by interior loops or bulges, but possibly containing non-standard base pairs intercalated between two standard base pairs. The question whether such pairs actually form a stack with the surrounding pairs is discussed in (28).

Helix: a double-stranded area consisting of one helix segment or of several helix segments separated by bulges or internal loops. Helices are considered different and named differently if separated from each other by a section of a multibranched loop, a pseudoknot loop, or a single stranded area that does not form a loop.

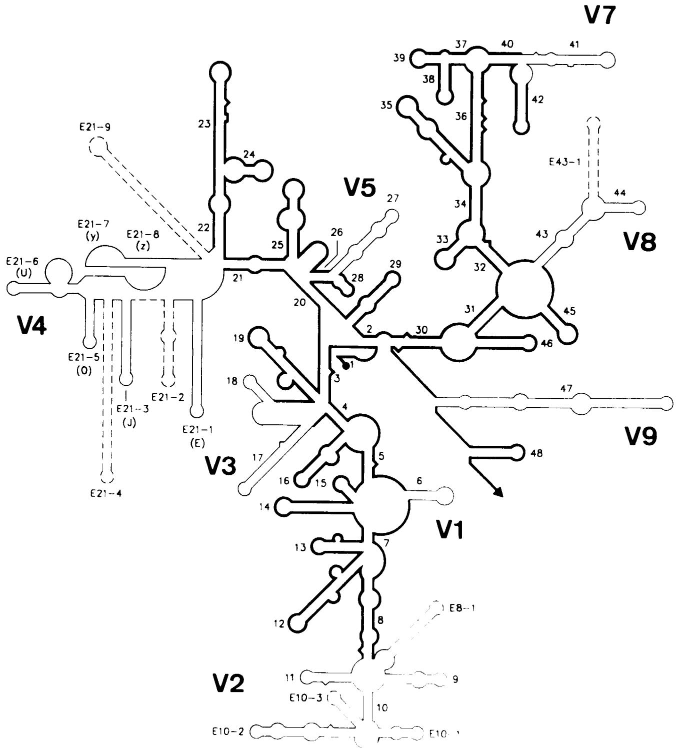

Fig. 1. Outline of the eukaryotic srRNA secondary structure. The helix numbering system is according to (1), i.e. universal helices (those common to eukaryotic and bacterial srRNAs) are numbered from 5′ to 3′ terminus. Eukaryote specific helices bear a number of the form Ea-b, where a is the number of the preceding universal helix and b is a serial number. Areas of relatively conserved sequence and secondary structure are drawn in bold lines. Variable areas are labelled V1 to V9 and drawn in thin lines (V6 exists only in prokaryotes). Helices found only in a limited number of species are drawn in broken lines. The structure chosen for area V4 is model EJOUyz derived in this study, and shown in Fig. 2h for human srRNA.

Existing models for area V4 and inventory of potential helices

The first model for the secondary structure of area V4 of eukaryotic srRNA was proposed in 1981 by Zwieb et al. (17) for the *Xenopus laevis* and *Saccharomyces cerevisiae* srRNA sequences and comprised 6 helices. In a survey of 10 eukaryotic srRNA sequences published in 1984, Nelles et al. (18) proposed the presence in this area of 3 helices, only one of which was taken from the model of Zwieb et al.. A large number of different models for area V4 were proposed in the period 1985–1988, some of them (19, 21, 24, 26) taking elements of the model of Zwieb et al. (17), others (23–25, 27) showing more resemblance to the partial model of Nelles et al. (18), all of them proposing variant or additional helices in some area. In a sequence compilation published in 1988 (16), the number of helices, supported by comparison of the 40 eukaryotic srRNA sequences available, was raised to 5. Most of the models that contain an original proposal for at least one helix are collected in Fig. 2, transposed to the human srRNA sequence. Fig. 3 gives an inventory of all the helices from the different models, delimited by line segments drawn under the complementary sequences in human srRNA. This allows to see which helices can exist simultaneously and which ones exclude each other because they use overlapping sequences. To the 25 helices used in various combinations in previously proposed models, we have added 3 more (labelled U, y, and z) in the course of the present investigation. All models containing a proposal for at least one original helix are listed in Table 1, followed by a description of their complete helix content. This table allows a more systematic comparison of the successively proposed models.

For the sake of completeness it should be noted that in a number of species, viz. *Drosophila melanogaster*, *Acanthamoeba castellanii*, *Euglena gracilis*, *Trypanosoma brucei*, *Leishmania donovani*, *Crithidia fasciculata*, and *Naegleria gruberi*, there are relatively long insertions in area V4. There are 3 points where such insertions are observed, indicated in Fig. 3 on the human srRNA sequence. Each of them results in the presence of one extra hairpin structure. Since the insertions fall outside most of the 28 helices listed, and near the boundary of the remaining ones, their presence does not add extra branching points to the proposed models.

Attempting a systematic search for the optimal secondary structure model

For the derivation of a secondary structure model for area V4 of all presently known eukaryotic srRNA sequences, we started from an alignment similar to the one published in (16), but supplemented with all the new eukaryotic sequences known, as listed in the most recent compilation (1). However, redundant sequences for the same species, such as the human and rat sequences determined by different authors, were eliminated. Of the 13 sequences available for the genus *Tetrahymena*, which contain no notable differences in area V4, only that of *T. thermophila* was retained. The resulting alignment for area V4 contained sequences from 53 species.

We originally attempted to make *tabula rasa* of all models hitherto proposed and therefore proceeded as follows. A computer program was written that compares all possible pairs of columns of the alignment of area V4 and tests the aptitude of the nucleotides present to form a standard base pair. The number of pairs of columns to be compared in a sequence alignment of length N is $(N-3).(N-4)/2$, taking into account that the minimum size of a hairpin loop is 3 nucleotides. If base pairing

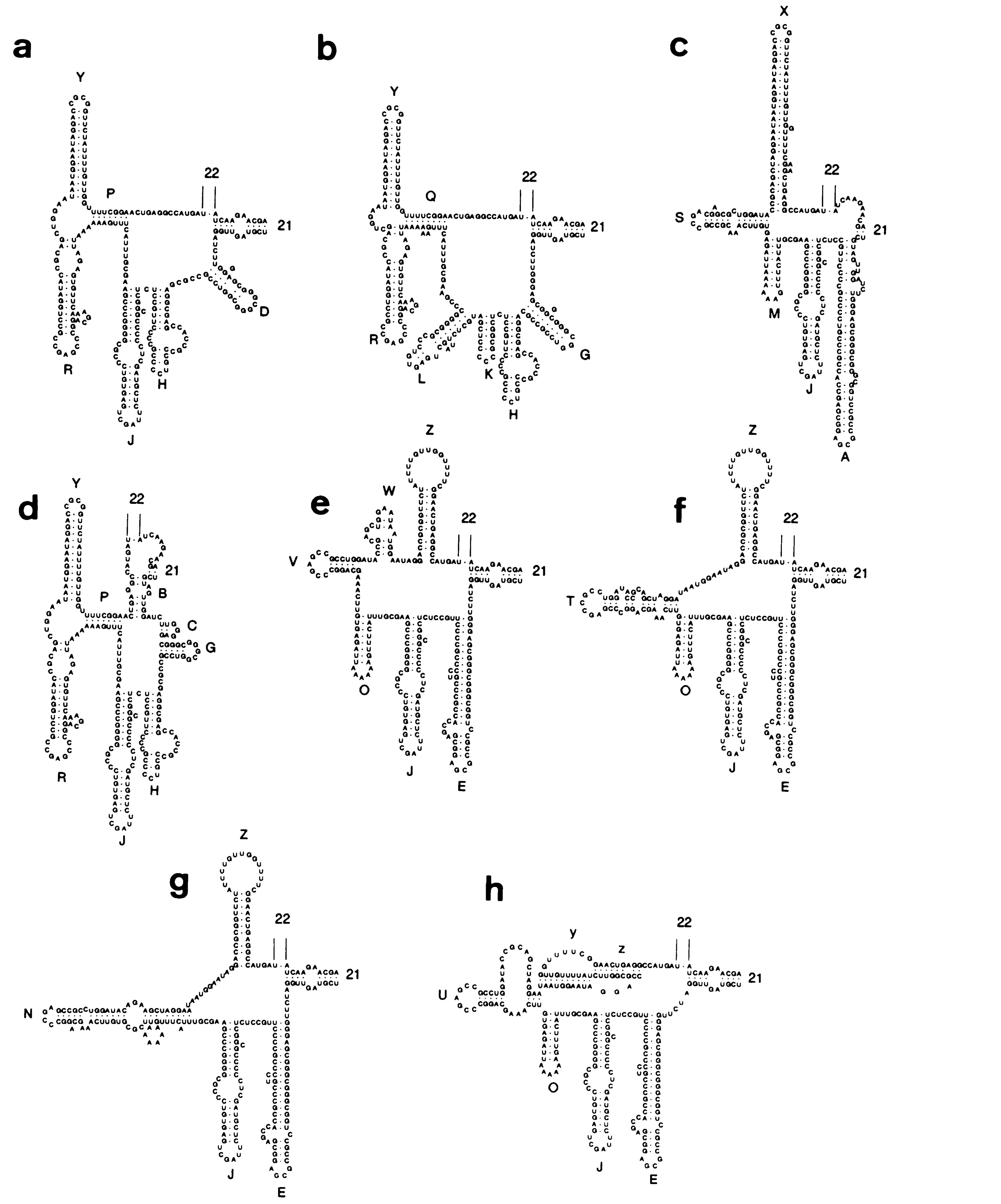

Fig. 2. Secondary structure models proposed for area V4 of eukaryotic srRNA. In order to facilitate comparison, all models are transposed to the human srRNA regardless of the species for which they were proposed by their authors. Helices are named A to z in the order of the position of their 5'-terminal nucleotide. Universal helix 21 precedes area V4 but is truncated in models (c) and (d). Universal helix 22 follows area V4. The structures shown are those listed in Table 1 except for the partial model of Nelles et al (18), and the model of Choi (19) which is equal to model (d) minus helices C and G; (a) Zwieb et al. (17); (b) Herzog and Maroteaux (21); (c) Gonzalez and Schmickel (22); (d) Rairkar et al. (26); (e) Ellis et al. (23); (f) Hendriks et al. (25); (g) Johansen et al. (27); (h) this paper.

is possible in all sequences for a pair of columns a and b, these alignment positions are considered as the location of a potential base pair. The search is then extended to columns a+1 and b−1 of the alignment to see if the complementarity exceeds a single base pair to form a potential helix segment. If it does, the search is continued until the complementarity stops at positions a+n and b−n, and positions a and b are recorded as the starting points of a potential helix segment of length n. If it does not, positions

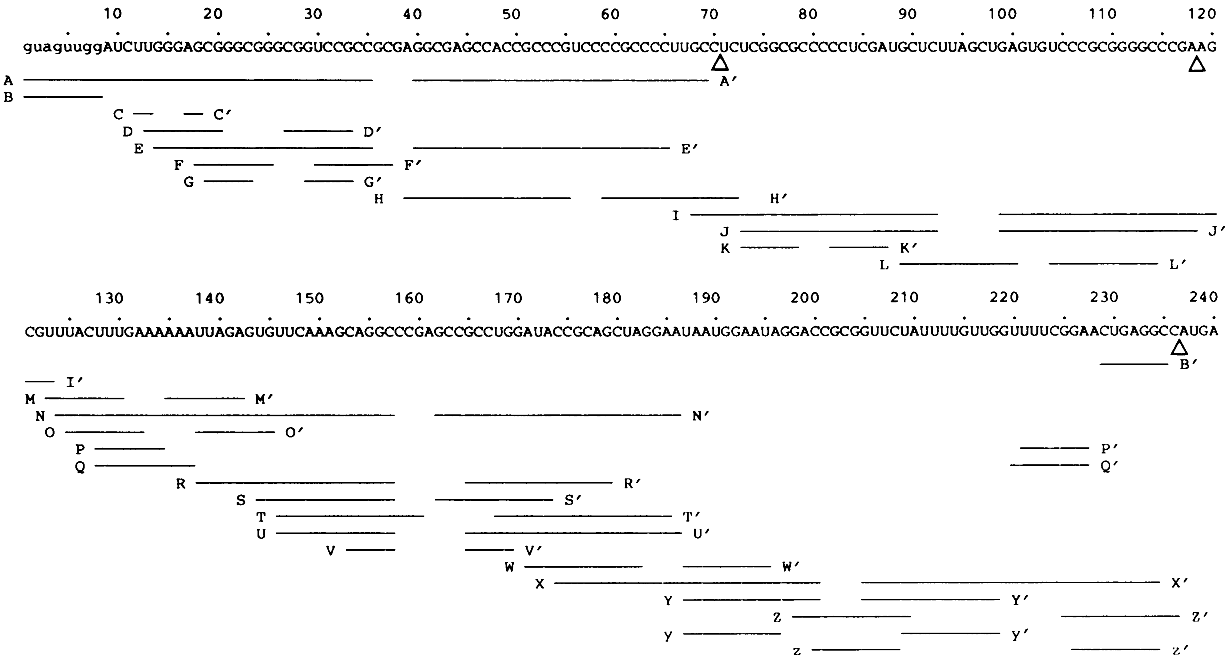

Fig. 3. Compatibility of helices in different models for area V4. The extent of each helix is defined by the line segments drawn under area V4 of the human srRNA sequence. The first 8 nucleotides printed in lower case belong to universal helix 21 in all models except (c) and (d) of Fig. 2. Segments A and A′ define complementary strands of helix A, etc. The lines are not interrupted for internal loops or bulges, which can be seen in Fig. 2. The point of insertion of extra helices, which occur in a few species only, are indicated by triangles below the *H. sapiens* sequence.

a and b are abandoned and the search is resumed at the next pair of columns, a and b+1, until all columns have been compared. For each potential base pair recorded, the presence or absence of compensating substitutions is noted.

The idea was to list all possible helix segments and then to select as the optimal secondary structure model the combination of segments that scores the largest number of compensating substitutions. However, it soon became evident that this approach runs into two practical difficulties.

First, the proof of a base pair by the observation of compensating substitutions is not absolute, but relative. This is illustrated by the comparison of pairs of columns in the following three imaginary alignments:

```
..A.....U..     ..A.....U..     ..A.....U..
..A.....U..     ..A.....U..     ..A.....C..
..C.....G..     ..A.....G..     ..A.....G..
..U.....A..     ..U.....A..     ..C.....A..
..U.....A..     ..U.....A..     ..U.....A..
```

Comparison of the two columns in the leftmost alignment shows unequivocal evidence for the existence of a base pair, since the bases in the two columns are always complementary and there are compensating substitutions. In the middle alignment, the presence of the non-standard base pair A·G in one sequence weakens the evidence for base pairing somewhat. However, it can still be concluded that the two positions form a base pair, and that the pair A·G can replace Watson-Crick or wobble pairs to a certain extent in helices. This was actually the reasoning followed by Noller and Woese (2) when they inferred the existence of A·G pairs in 16S rRNA. In the rightmost alignment, the imagination must be stretched a little further if one is to conclude to the existence of base pairing, with 60% of the sequences having a non-standard base pair. The example illustrates the fact that when trying to prove a helix on the basis of compensating substitutions, one encounters a continuum of cases ranging from solid proof to obvious disproof. Hence it is necessary to establish a criterion, such as a minimum fraction of sequences that must show standard base pairing in order for the pair to be accepted. Rejection or acceptance is always arbitrary to some extent, all the more so because the fraction of sequences showing standard base pairing at a given site depends on the composition of the sequence set available for comparison at the time the test is performed.

The second difficulty arises from the fact that an area such as V4 is variable in length as well as in sequence, which requires the introduction of many gaps in the local alignment. Indeed,

Table 1. Inventory of helices proposed in various models

Model (a)	Helix (b)																											
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	y	Z	z
Zwieb et al. (17)				D				H		J						P		R							Y			
Nelles et al. (18)										*					O												Z	
Choi (19)		B		*				*		*						*		*							*			
Herzog & Maroteaux (21)						F		*			K	L					Q	*							*			
Gonzalez & Schmickel (22)	A									*			M						S					X				
Ellis et al. (23)					E				I						*							V	W				*	
Hendriks et al. (25)					*				*						*					T							*	
Rairkar et al. (26)		*	C				G	*		*						*		*							*			
Johansen et al. (27)					*				*					N													*	
This paper					*					*					*						U					y		z

(a) The models are listed in chronological order of their proposal. Models (20,24) consisting merely of helices proposed by other authors are not listed here.
(b) For each model, helices originally proposed by its author(s) are indicated by a character, helices taken from anterior models by an asterisk. The structure of the helices can be found in Figg. 2 and 3. Certain helices consist of other smaller helices plus additional base pairs or helix segments: thus I includes J, U includes V, X includes Y which itself includes y, and Z includes z.

the average length of area V4 in mammalian srRNAs is 245 nucleotides, but in our present alignment (1) which contains 62 eukaryotic sequences, the area occupies 460 positions, not counting the long insertions present in a small set of species. Normally, the boundaries of secondary structure elements are used as landmarks to facilitate alignment, but then one has to postulate a secondary structure model before starting to align, and the structures deduced by running a computer program on the resulting alignment will be biased in favour of that model.

As a result of these difficulties, the program did not lead to a plausible model when it was applied to the alignment for area V4. The number of potential helix segments detected depended, as expected, on the criteria set for acceptance of a base pair, i.e. the fraction of sequences where it must be a standard pair. If the criteria were severe, the sets of complementary sequences resulting from a search formed loose but knotted networks of short segments, incompatible with the usual appearance of an RNA secondary structure. When the criteria were gradually relaxed, the number of possible combinations rose very fast. By the time that the criteria were lenient enough to allow an amount of base pairing expected for a stable secondary structure, a huge number of possible combinations of segments resulted, with such marginal differences in the total number of compensating substitutions as to make a choice of an optimal model meaningless.

A different approach to the derivation of a local secondary structure model for a variable area in large ribosomal subunit RNA was followed by Bachellerie and coworkers (29). They made partial alignments of the area, each alignment comprising a group of species, viz. the archaebacteria, eubacteria, plastids, and a number of eukaryotic taxa. The idea is that alignment of a set of relatively related sequences is more dependable and requires less gaps, hence the observation of compensating substititions is facilitated. After taxon-specific models have been constructed, it is possible to deduce a general structure from which taxon-specific structures can be derived by insertion or deletion of helices at certain sites. This approach, which we applied succesfully to other areas of srRNA (unpublished), gave no satisfactory results in the case of area V4. The reason, we assume, is that the evolution of area V4 has witnessed a very high ratio of insertion and deletion events to substitutions. As a consequence, partial alignments optimized for primary structure similarity do not show enough compensating substitutions, a fraction of these being obliterated by numerous insertions and deletions that are placed erroneously in the alignment process. An additional difficulty peculiar to area V4 is that there is no equivalent structure in prokaryotes to rely upon.

Combining previously proposed helices into an optimal model

Since a purely systematic search for a model was thwarted by the fact that no objective sequence alignment can be achieved *a priori* for area V4, we took a more pragmatic approach. All the models previously proposed can be applied, with more or less success, to each of the known srRNA sequences. This is illustrated in Fig. 2 with the transposition of models to the human sequence. A partial alignment can then be made for each of the 28 helices found in the available models (Table 1), the presence of compensating substitutions examined, and the helices combined into new models which can be ranked according to their content of compensating substitutions. Although this procedure is less rigorous because it does not examine all possibilities, it can be assumed that the best base pairing opportunities probably are represented among the helices proposed by ten or so independent model builders, each examining area V4 in one or more different sequences.

Transposability of helices

In order to find the optimal combination, we first examined for each of the 28 helices listed in Table 1 and defined in Fig. 3, whether it is transposable to all presently known srRNA structures. To this end, area V4 was drawn for 31 of the 53 sequences considered (a single representative being chosen for taxa such as vertebrates and angiosperms), according to 6 different models, viz. those of Zwieb et al. (17), Herzog and Maroteaux (21), Gonzalez and Schmickel (22), Ellis et al. (23), Hendriks et al. (25), and Johansen et al. (27). The work, though extensive, was facilitated by the facts that sequences within a taxon are often very similar and that models partly overlap in helix content. The set of models chosen covers all 28 helices of Table 1 except B, C, and G. The structure of the latter helices could be examined on partial models since C and G are alternative structures for helix D, and complementarity B is easy to localize since it is at the boundary of the variable area V4.

Certain helices look quite plausible in the srRNA of the species for which a given model was proposed, but it is evident that they cannot be fitted to srRNAs from species of certain other taxa, sometimes even to species belonging to the same taxon. When one tries to do so, one is left with an unstable structure, retaining only a few base pairs interspersed with several, sometimes adjacent, non-standard pairs. In other cases the helix must be bent into quite different shapes, regarding the size and/or position of bulges and interior loops present, in order to be fitted to different sequences. On these grounds, helices C, F, G, H, I, K and L were discarded from the list of potential helices because they are transposable to less than 3/4 of the examined sequences. In the case of helix I, which is an elongated version of helix J, only the part present in excess over helix J cannot be transposed. For the remaining set of 21 transposable helices, the number of compensating substitutions was counted after their structures were aligned for all the sequences available.

Setting criteria for the proof of a base pair by compensating substitutions

As illustrated with an example above, the proof of a base pair by compensating substitution can be weakened to a certain extent by the fact that in a fraction of the examined sequences the Watson-Crick or wobble pair is replaced by a non-standard pair. At which point should the proof be invalidated because the fraction of sequences where the pair is non-standard is too large? In order to find a solution to this problem, we examined each base pair of each of the 21 helices that survived the transposability test. For each pair showing at least one compensating substitution, the fraction of sequences where the pair is standard was noted. The same statistic was made for helix 47, which is an established helix of the srRNA secondary structure (Fig. 1), but also constitutes a variable area, V9. The results are plotted in the form of histograms in Fig. 4, showing the distribution of compensated base pairs as a function of the fraction of the sequences where they are complementary. It can be seen in Fig. 4a that most of the base pairs of helix 47 are standard pairs in more than 80% of the sequences examined. Hence it seems reasonable to instore cutoff values in the vicinity of 80% for acceptance of a base pair supported by compensating substitutions.

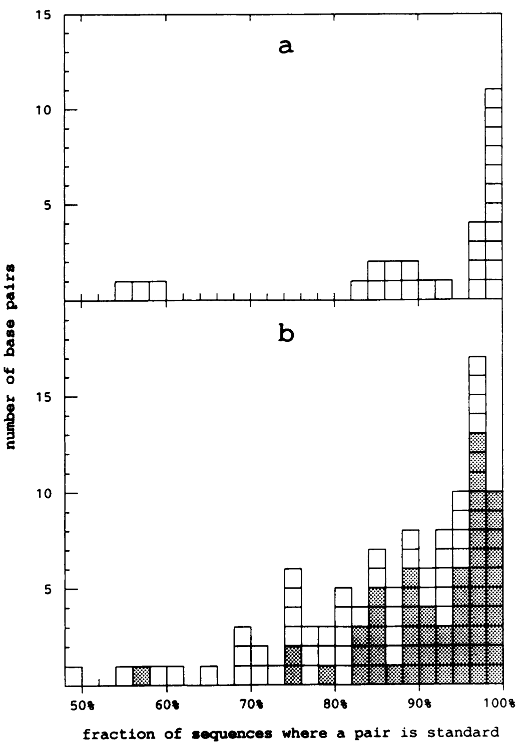

Fig. 4. Distribution of base pairs as a function of the fraction of sequences where they exist as standard pairs. Each square represents one base pair of a helix. A square placed in the 80–82% interval, means that the base pair is a standard pair (G·C, A·U or G·U) in 43 of the 53 eukaryotic srRNAs considered (81%). Histogram (a) gives the distribution for helix 47, which is universal but variable in structure (area V9 in Fig. 1). Histogram (b) comprises all base pairs belonging to the 21 transposable helices (see text) proposed for area V4. The shaded area covers the base pairs of model EJOUyz.

	A	B	D	E	J	M	N	O	P	Q	R	S	T	U	V	W	X	Y	y	Z	z
A	$\boxed{1}$	0	0	0	$\boxed{1}$	$\boxed{1}$	1	1	1	1	1	$\boxed{1}$	1	1	1	$\boxed{1}$	1	1	1	$\boxed{1}$	1
B		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
D			1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
E				1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
J					$\boxed{1}$	$\boxed{1}$	1	1	1	1	1	$\boxed{1}$	1	1	1	$\boxed{1}$	1	1	1	$\boxed{1}$	1
M						$\boxed{1}$	0	0	0	0	0	$\boxed{1}$	1	1	1	$\boxed{1}$	1	1	1	$\boxed{1}$	1
N							1	0	0	0	0	0	0	0	0	0	0	1	1	1	1
O								1	0	0	0	0	1	1	1	1	1	1	1	1	1
P									1	0	1	1	1	1	1	1	0	1	1	0	0
Q										1	1	1	1	1	1	1	0	1	1	0	0
R											1	0	0	0	0	0	0	1	1	1	1
S												$\boxed{1}$	0	0	0	$\boxed{1}$	1	1	1	$\boxed{1}$	1
T													1	0	0	0	0	1	1	1	1
U														1	0	0	0	1	1	1	1
V															1	1	1	1	1	1	1
W																$\boxed{1}$	0	0	0	$\boxed{1}$	1
X																	1	0	0	0	0
Y																		1	0	0	0
y																			1	0	1
Z																				$\boxed{1}$	0
z																					1

Fig. 5. Compatibility matrix of the 21 transposable helices. The value 1 is assigned to matrix elements corresponding to two compatible helices, i.e. helices that can occur simultaneously in a model (see Fig. 3), the value 0 to all other elements. The path followed by the computer program to derive the first model, AJMSWZ, is indicated by the boxed elements.

Table 2. Transposable helices: length and compensating substitutions

Helix	Length (a)	Number of base pairs											
		70% standard (b)				80% standard (b)				90% standard (b)			
		S_2	S_1	S_0	Total (c)	S_2	S_1	S_0	Total (c)	S_2	S_1	S_0	Total (c)
A	27	14	2	1	17	13	1	0	14	4	0	0	4
B	6	1	4	0	5	1	3	0	4	1	3	0	4
D	7	3	0	0	3	2	0	0	2	1	0	0	1
E	20	19	0	0	19	17	0	0	17	13	0	0	13
J	16	16	0	0	16	16	0	0	16	12	0	0	12
M	8	1	0	4	5	1	0	3	4	1	0	2	3
N	21	4	6	5	15	3	5	5	13	3	4	3	10
O	8	1	5	1	7	1	5	1	7	0	5	1	6
P	7	5	1	1	7	4	1	1	6	1	0	0	1
Q	8	3	1	1	5	2	1	1	4	0	0	1	1
R	12	4	7	0	11	2	7	0	9	2	4	0	6
S	11	1	3	3	7	0	3	3	6	0	1	2	3
T	8	2	4	2	8	2	4	2	8	1	2	1	4
U	8	4	3	1	8	3	3	1	7	3	3	1	7
V	5	2	2	1	5	2	2	1	5	2	2	1	5
W	8	0	3	0	3	0	1	0	1	0	1	0	1
X	27	12	4	8	20	9	3	7	19	4	2	2	8
Y	14	9	2	4	15	6	2	3	11	2	1	1	4
y	10	7	0	3	10	6	0	3	9	2	0	1	3
Z	12	8	1	0	9	8	1	0	9	6	1	0	7
z	9	8	1	0	9	8	1	0	9	6	1	0	7

(a) This is the number of base pairs (standard and non-standard) that the helix possesses in human srRNA. (cf. Fig. 2).
(b) The number of base pairs in a helix depends on the criterion for acceptance of a base pair. "70% standard" means that two bases in opposite strands are considered as a pair if they belong to the set G·C, A·U, G·U in at least 70% of the sequences. A sequence where the base pair is deleted is not included in the calculation of this ratio.
(c) For each acceptance criterion (70%, 80%, 90%), the table lists 3 figures:
S_2: number of base pairs where at least one case of full compensation is observed: both bases are substituted but the complementarity is preserved (e.g. A·U to G·C, or A·U to U·G).
S_1: number of base pairs where one observes substitution of one of the bases with preservation of the complementarity (e.g. A·U to G·U).
S_0: number of base pairs where no substitution is observed that preserves complementarity.
Total=S_2+S_1+S_0. In general, "total" differs from "length" because the number of base pairs in human srRNA is different from the number of base pairs that are standard in > 70% (80%, 90%) of the set of eukaryotic sequences.

Table 3. Models ranked according to number of compensating substitutions.

Model (a)	Number of base pairs											
	70% standard (b)				80% standard (b)				90% standard (b)			
	S_2	S_1	S_0	Total (c)	S_2	S_1	S_0	Total (c)	S_2	S_1	S_0	Total (c)
AJMUyz	50	6	8	64	47	5	6	58	28	4	4	36
AJOUyz	50	11	5	66	47	10	4	61	27	9	3	39
AJN yz	49	9	8	66	46	7	7	60	27	5	4	36
AJR yz	49	10	3	62	45	9	2	56	26	5	1	32
AJMTyz	48	7	9	64	46	6	7	59	26	3	4	33
AJOTyz	48	12	6	66	46	11	5	62	25	8	3	36
BEJPRY	54	14	4	72	46	13	3	62	31	8	1	40
BEJPUY	54	10	5	69	47	9	4	60	32	7	2	41
BEJPTY	52	11	6	69	46	10	5	61	30	6	2	38
BEJQRY	52	14	4	70	44	13	3	60	30	8	2	40
BEJQUY	52	10	5	67	45	9	4	58	31	7	3	41
BEJPSY	51	10	7	68	44	9	6	59	29	5	3	37
BEJMUY	50	9	8	67	44	8	6	58	32	7	4	43
BEJOUY	50	14	5	69	44	13	4	61	31	12	3	46
BEJQTY	50	11	6	67	44	10	5	59	29	6	3	38
BEJN Y	49	12	8	69	43	10	7	60	31	8	4	43
BEJQSY	49	10	7	66	42	9	6	57	28	5	4	37
EJMVX	50	6	12	68	45	5	10	60	32	4	5	41
EJOVX	50	11	9	70	45	10	8	63	31	9	4	44
EJMSX	49	7	14	70	43	6	12	61	30	3	6	39
EJMUyz	55	4	7	66	51	4	6	61	37	4	4	45
EJOUyz	55	9	4	68	51	9	4	64	36	9	3	48
EJN yz	54	7	7	68	50	6	7	63	36	5	4	45
EJR yz	54	8	2	64	49	8	2	59	35	5	1	41
EJMTyz	53	5	8	66	50	5	7	62	35	3	4	42
EJOTyz	53	10	5	68	50	10	5	65	34	8	3	45
EJMSyz	52	4	9	65	48	4	8	60	34	2	5	41
EJMU Z	48	4	5	57	45	4	4	53	35	4	3	42
EJOU Z	48	9	2	59	45	9	2	56	34	9	2	45
EJN Z	47	7	5	59	44	6	5	55	34	5	3	42
EJR Z	47	8	0	55	43	8	0	51	33	5	0	38
EJMT Z	46	5	6	57	44	5	5	54	33	3	3	39
EJMVWZ	46	6	5	57	44	4	4	52	34	4	3	41
EJOT Z	46	10	3	59	44	10	3	57	32	8	2	42
EJOVWZ	46	11	2	59	44	9	2	55	33	9	2	44
EJMSWZ	45	7	7	59	42	5	6	53	32	3	4	39
BDJPSW	26	11	4	41	23	8	4	35	15	5	2	22
BDJQVW	25	10	2	37	23	7	2	32	16	6	2	24
BDJQSW	24	11	4	39	21	8	4	33	14	5	3	22
BDJMVW	23	9	5	37	22	6	4	32	17	6	3	26
BDJOVW	23	14	2	39	22	11	2	35	16	11	2	29
BDJMSW	22	10	7	39	20	7	6	33	15	5	4	24

(a) The upper part of the table lists the 36 models (on a total of 162) supported by the largest number of compensating substitutions. For comparison, the lower part lists the 6 worst models.
(b) Criteria as in Table 2.
(c) Degree of compensation in the observed substitutions, as defined in Table 2. Total=S_2+S_1+S_0.

Combining helices and evaluating the resultant models

Fig. 5 shows a compatibility matrix for the 21 helices that can be transposed to at least 40 of the 53 eukaryotic srRNAs (>

Table 4. Stability of 11 helices found in the optimal models for area V4 and in other areas of srRNA (a).

Species	Helices of area V4 (b)										Universal helices (c)			
	E	J	M	N	O	R	S	T	U	yz	27	47	41	38
Homo sapiens	-24.5	-19.1	5.2	3.4	0.3	2.5	-5.9	-2.8	4.6	-5.4	-10.2	-58.0	-6.6	5.9
Artemia salina	-10.8	-16.0	5.2	8.2	0.3	7.7	2.9	9.4	0.7	-4.1	-14.6	-23.6	-7.5	5.9
Drosophila melanogaster	-14.8	-7.2	5.7	1.3	0.4	2.2	---	6.0	0.4	-3.3	-8.1	-21.7	-6.5	6.6
Glycine max	-20.8	-10.8	5.7	5.3	0.3	1.5	-3.8	-0.2	4.2	-11.5	-12.4	-28.4	-18.0	5.9
Chlorella vulgaris	-24.4	-19.8	5.7	-1.6	0.9	-3.9	-7.2	2.0	6.1	-10.4	-13.7	-30.9	-14.1	3.2
Saccharomyces cerevisiae	-15.4	2.5	5.2	6.7	0.3	-0.6	0.3	-1.2	5.9	-4.7	-5.3	-27.0	-10.0	5.9
Prorocentrum micans	-12.4	-8.0	5.2	7.3	0.9	1.4	1.7	0.4	6.1	-9.3	-2.2	-31.6	-14.9	5.9
Tetrahymena thermophila	-8.6	-6.7	7.6	-2.6	2.1	-2.0	-3.3	-1.4	1.8	-5.5	-11.6	-17.9	-9.3	5.9
Physarum polycephalum	-11.8	-11.7	---	-2.8	0.1	6.4	-2.6	-0.6	5.9	-2.0	-3.0	-26.9	-10.7	7.1
Plasmodium bergei II	-2.6	-1.9	5.9	2.5	0.3	3.8	-1.7	2.1	6.2	-4.2	-7.0	-21.9	-23.5	7.3
Crithidia fasciculata	-9.2	1.4	7.9	8.5	1.7	16.1	7.3	8.2	8.8	-3.0	7.8	-17.4	-8.9	6.7
Euglena gracilis	-16.4	3.3	---	2.3	1.4	1.7	4.6	4.1	2.5	-6.9	4.9	-31.6	-28.1	5.9

(a) The change in free energy associated with helix formation, in kcal/mol at 37°C, was calculated according to (32). As these authors provide no method for calculating the free energy of a pseudoknot, we computed this as follows: only one initiation of base pairing is assumed and the energy of the loops is calculated as for hairpin loops. The nucleotide following helix z is considered as an unpaired terminal nucleotide, since it probably cannot form a terminal mismatch with the opposite strand, which usually forms a short loop comprising only 3 nucleotides. If the free energy of the loops in pseudoknot formation is computed according to Abrahams *et al.* (33), he stabilities improve, ΔG values being 2.48 kcal/mol lower on average. (b) The helices listed are those occurring in the family of models EJ..yz (see text and table 3). No ΔG value is listed if the helix cannot be satisfactorily transposed to the srRNA of a species. (c) These helices, termed 'universal' because they are present in eukaryotic as well as bacterial srRNAs, are situated in the generally accepted part of the srRNA secondary structure model (Fig. 1). The ΔG values are listed for comparison with the helices proposed for area V4.

3/4). Two helices i and j are considered compatible if they are formed from non-overlapping sequences (see Fig. 3), hence can coexist in a model. In this case, the value 1 is assigned to element (i,j) of the matrix, in the opposite case, the value 0. Next, a computer program scans the matrix and finds all possible combinations of compatible helices, i.e. potential secondary structure models. The path followed to find the first combination is indicated in the matrix. In row A it can be seen that the first helix compatible with A is J. Scanning row J, one finds helix M as the next helix, compatible with A as well as J, etc.. The first model thus assembled is AJMSWZ. The second model is AJMSWz, where the last helix, Z, added to the combination AJMSW is replaced by the next compatible one, z. Subsets of more complete sets are ignored. In this way, the 21 helices can be combined into 162 different models.

Table 2 lists, for each helix, the number of base pairs with and without compensating substitutions. These data are listed according to three criteria for acceptance of a base pair, corresponding with a cutoff value of 70%, 80% and 90% in the histogram of Fig. 4b. The program now uses these data in order to rank all the potential models according to the total number of compensating substitutions in the constituent base pairs. Table 3 lists the models that obtain the best scores. The list was limited to the 36 best models on a total of 162 as follows: the model showing the largest number of compensating substitutions is helix combination EJMUyz. This scores 55 compensating substitutions if one accepts all base pairs that are standard in at least 70% of the sequences, 51 and 37 compensating substitutions if the acceptance limit is shifted to 80% and 90% respectively. The list of models was extended to all those numbering up to 6 compensating substitutions less than the optimal model under at least one of the 3 acceptance criteria (70%, 80%, 90% standard base pairing). The worst model, also listed in Table 3, is BDJMSW. It scores 22, 20, or 15 compensating substitutions depending on the base pair acceptance criterion. Most of these substitutions are due to helix J, which is an obligatory constituent of all models, since the alternative helices I, K and L (Fig. 3) have been rejected due to their poor transposability.

Selecting the most probable model

The models in Table 3 are listed according to structural families. As an example, the first group of 6 models all start with helices A and J at the 5'-end and have helices y and z at their 3'-end. The second family has the structure BEJ...Y, and so on. Within each family, the models are ranked in descending order of compensating substitutions, taking the leftmost column of values as a guide.

The following considerations allow to further narrow the number of plausible models. The models of the first family share the presence of helix A at the 5'-terminus of area V4. All models of second family of models share helix B, which forms a long range interaction joining both ends of area V4. Both helices A and B use nucleotides that are also involved in the formation of helix 21 (see Fig. 1), which does not belong to area V4 but is a universal helix common to eukaryotic and prokaryotic srRNAs. Helix 21 possesses only one base pair proven by compensating substitutions among eukaryotic sequences. However, a helix of very similar structure, i. e. two segments of 4 base pairs separated by a symmetrical internal loop, also exists in prokaryotes, where its existence is proven by compensating substitutions in all 8 base pairs. It therefore seems extremely likely that helix 21 is a universal constituent of the srRNA secondary structure, and unwise to sacrifice it in favour of helices A or B, which in addition do not have a convincing record of compensating substitutions

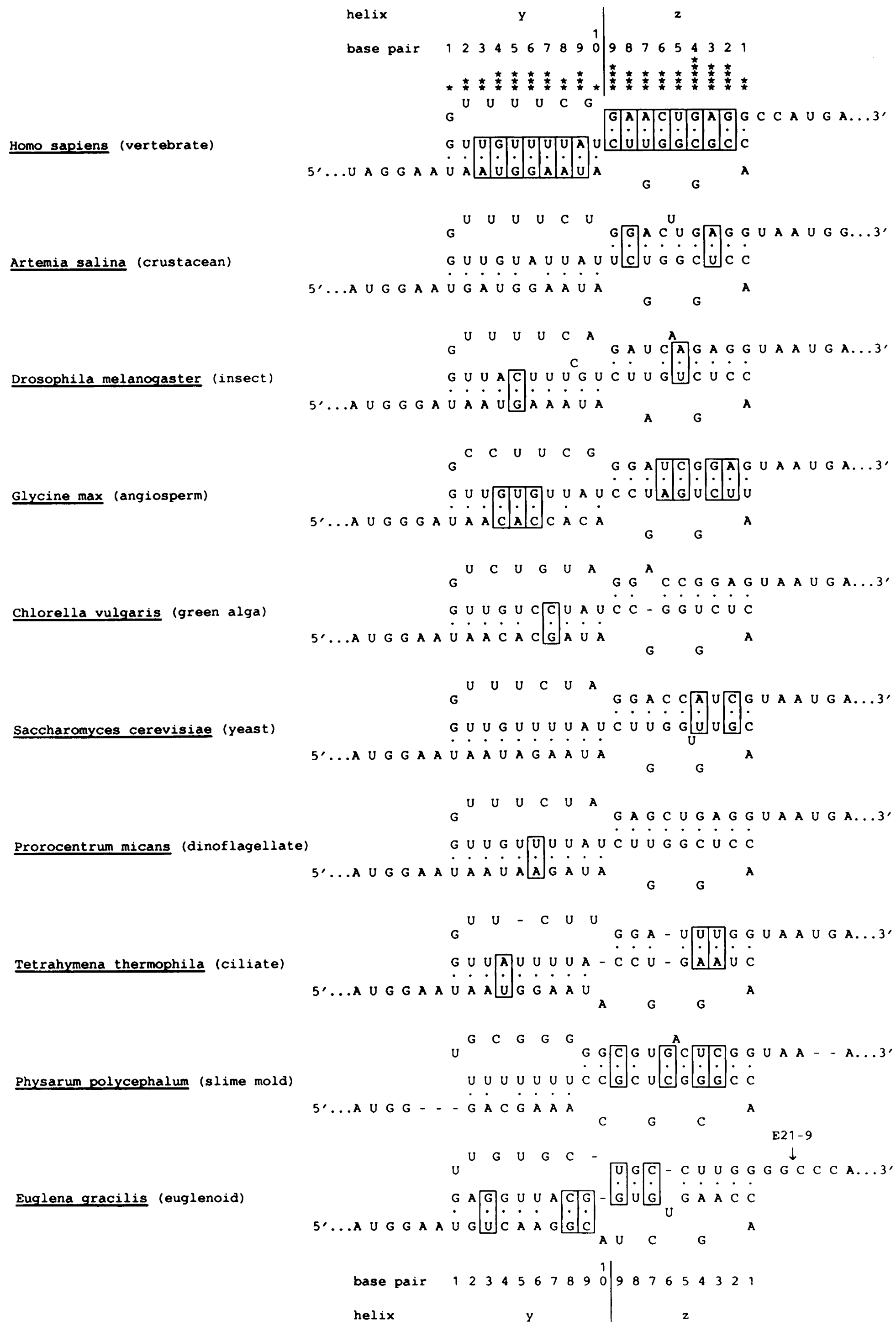

Fig. 6. Structure of pseudoknot yz in 10 phylogenetically distant species. The pseudoknot structures are aligned in order to show the sequence homology. Base pairs supported by compensating substitutions are boxed in the human sequence and in one of the species where the substitution occurs. The number of asterisks above each base pair denotes the number of different standard pairs observed at this position. The occurrence of two different pairs involves compensation at some sites (e.g. U·A and C·G in pair 8 of helix y) but not at others (e.g. G·C and G·U in pair 1 of helix z). The point of insertion of extra helix E21–9 is indicated on the *E. gracilis* structure.

(see Table 2). On these grounds we eliminate the families of models comprising helices A and B.

Helices E and J occur in all models of the remaining 3 families, account for an impressive number of compensating substitutions in proportion to their length, and are satisfactorily transposable to all known sequences possessing area V4. Their existence therefore seems difficult to refute. The family of models with the structure EJ…yz has an appreciably better score of compensating substitutions than the families EJ…X and EJ…Z. Helices y and z can form a pseudoknot structure, discussed in detail below, which is satisfactorily transposable to all eukaryotic srRNA sequences that possess area V4 and shows many compensating substitutions.

Within the EJ…yz model family, the choice of the remaining

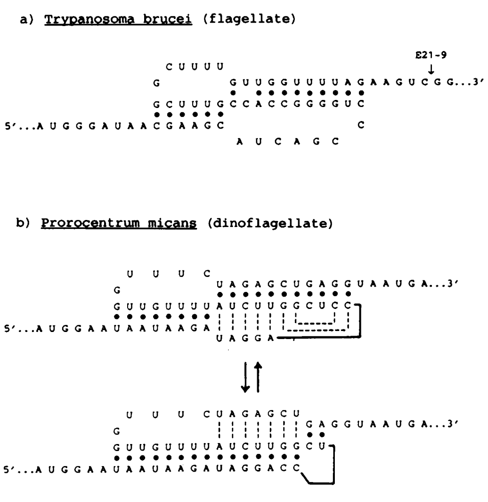

Fig. 7. Peculiarities of the pseudoknot structure. a) Example of a pseudoknot structure with deviant lengths of helices y and z and of the connecting loops, with respect to the structures shown in Fig. 6. In this species, the extra helix E21−9 (cf. Fig. 1) is inserted at the point indicated. b) Possibility of the exchange of base pairs between the two helices of a pseudoknot. In each structure, one of the helices is extended to its maximum length at the expense of the other. The dissolved base pairs existing in the alternative structure are indicated by broken lines. By turning around the sugar phosphate backbone one by one, the 7 bases of the sequence AUCUUGG could gradually lengthen one helix at the expense of the other. In the lowermost structure, the 5′-proximal loop would be reduced to a single phosphodiester bond. This type of base pair exchange is a possibility in most species, but usually over a shorter range.

part of the structure is less straightforward. Assuming that the set of 21 potential helices (Table 2) considered is exhaustive and that no additional possibilities have been overlooked, the following 7 combinations remain possible:

<table>
<tr><td rowspan="3">E J</td><td>M</td><td>S
T
U</td><td rowspan="3">y z</td></tr>
<tr><td>O</td><td>T
U</td></tr>
<tr><td colspan="2">N
R</td></tr>
</table>

The combination EJMUyz is the one that achieves the best ranking in Table 3, but the other models of the family follow quite closely, the worst one having only three compensating substitutions less than the best one. In order to compare the merits of each helix, we not only consider the number of compensating substitutions (Table 2) but also its transposability, and its thermodynamic stability, listed in Table 4 for 12 species. Although helices M, N, O, R, S, T, and U belong to the set of 21 helices transposable to at least 3/4 of the eukaryotic sequences, the transposition often necessitates some distortion of the helix structure, such as a change in size or location of a bulge or internal loop, in the case of helices S and T. Helix M cannot be transposed to certain species such as *Euglena gracilis* and *Physarum polycephalum* because the resulting structures would possess rows of adjacent non-standard base pairs. As can be seen in Table 4, the thermodynamic stability is very variable for most helices, but this phenomenon also applies to helices 27, 41 and 47, which are established helices in other variable areas of the srRNA secondary structure model (Fig. 1).

Helices M and O, which both can be combined with T or with U, are built for a large part from the same nucleotides (see Fig. 3). Both contain few compensating base pairs (see Table 2), which is due to the sequence conservation in this part of area V4. Helix O then seems preferable due to its better thermodynamic stability and transposability. Even helix O does not have a convincing stability, but similar cases are found in established helices of the srRNA secondary structure model, e.g. helix 38 which is listed in Table 4 for comparison.

If helix O is adopted, the rest of the structure must be occupied either by T or by U. The latter solution seems preferable in view of the fact that T can be transposed only at the expense of structural variation and only 2 in 8 base pairs are supported by compensating substitutions. There then remains the possibility of replacing the set OU by a single helix, either N or R. According to Table 3, models EJNyz and EJRyz are slightly less advantageous in terms of compensating substitutions than model EJOUyz. It should nevertheless be stressed that the choice among the 7 combinations listed above is less obvious than the selection of helices E, J and y-z, and it seems well possible that additional sequence data gathered in the future change the preferences expressed here, or even point to new possibilities not detected at present.

The pseudoknot

Of the 21 helices that are satisfactorily transposable to all srRNAs (Table 2), Y and Z have a good record of compensating substitutions, and either one or the other occurs in most of the models hitherto proposed (Fig. 2). Y and Z cannot occur simultaneously since they use overlapping sequences (see Fig. 3) and hence are incompatible. However, by stripping Y and Z of 4 and 2 base pairs respectively, they can be shortened to the non-overlapping versions y and z, while the combined number of compensating substitutions is reduced by just 2 units at most (see Table 2, least stringent criteria). They can then be combined to form a pseudoknot. This type of higher-order RNA structure (reviewed in 30) has been found at the 3′-terminus of certain plant viral RNA's and at the intron-exon boundary of rRNA precursors and mitochondrial mRNAs. It is also formed by helices 1 and 2 of srRNA (Fig. 1), where it is proven by 6 compensating substitutions in 9 base pairs.

The structure of the yz pseudoknot is displayed in Fig. 6 for srRNAs from 10 organisms chosen from phylogenetically distant taxa, showing that the structure is satisfactorily transposable and strongly supported by compensating substitutions. The 5′-proximal helix y is generally ten base pairs long, the 3′-proximal helix z nine base pairs. Both helices are shortened by deletions in certain species. Insertions also occur and in several cases give rise to the presence of a bulge about half way helix z. Conversely, the bulges in *Chlorella vulgaris* and *Euglena gracilis* are the result of a deletion in one of the strands of helix z. The 5′-proximal loop connecting y to z is usually three nucleotides long. This should be long enough to bridge the distance of approximately 15 Å between the A in base pair 10 of helix y and the C in base pair 1 of helix z (ref. 30 and Pleij, personal communication). The 3′-proximal loop most frequently has a length of 7 nucleotides.

In species that diverge early in eukaryotic evolution, according to phylogenetic studies based on srRNA sequences (8, 31), the pseudoknot is also present. However, it sometimes has a deviant structure with regard to the length of the helices or the connecting loops. In the slime mold *Physarum polycephalum*, helix z seems to be extended at the expense of helix y but it is still possible to align the pseudoknot structure satisfactorily with that in other species (Fig. 6). Pseudoknots that are more difficult to align with those in Fig. 6 are found among e.g. the flagellates, for which an example of a structure is shown in Fig. 7a. On the other hand, in the majority of known sequences an exchange of base pairs among helices y and z is conceivable. This is illustrated with the *Prorocentrum micans* sequence in Fig. 7b and it could mean that the pseudoknot is actually a dynamic structure.

CONCLUSIONS

The procedure described above results in the selection of an optimal model for area V4 of small ribosomal subunit RNA, reached by systematic scrutiny of all conceivable combinations of potential helices that have been proposed in this area. The most important criterion in the selection process is to maximize the number of compensating substitutions, but other criteria, such as transposability of helices to all available sequences, and thermodynamic stability, are taken into account. A more systematic approach to the problem of model selection was applied by Studnicka et al. (34) to 5S rRNA. This was possible because alignment of the available 5S rRNA sequences is relatively straightforward. It fails in the case of area V4 of srRNA because the large variability in length, combined with variability in sequence, results in too many alignment possibilities.

In principle it cannot be excluded that the actual structure contains helices that have gone undetected by all the investigators who have tried to devise models for the area. However, in view of the large number of models proposed (Table 1, Fig. 2), and the fact that authors have tried to apply these to sequences from very diverse organisms, it seems unlikely that any complementarity of sizable length and stability has been overlooked. Nevertheless, the model that we propose for area V4 (Fig. 2h) cannot be considered as definitive in the sense that the evidence for the constituent helices is of uneven quality. Helices E, J, and the pseudoknot yz are supported by an impressive number of compensating substitutions. In contrast, the structure of the area extending between helix J and pseudoknot yz seems less well established. Although the combination EJOUyz seems most probable at the moment, it will be necessary to test this in the light of future sequence evidence as it becomes available.

Area V4 occurs in all hitherto examined eukaryotic srRNAs except that of the microsporidian *Vairimorpha necatrix*, where it is entirely deleted. On the other hand, the srRNAs of certain species, such as *Drosophila melanogaster*, *Acanthamoeba castellanii*, *Euglena gracilis*, and the flagellates, contain relatively long insertions in area V4. These do not interfere with the model that we propose, since they result in additional hairpins intercalated between helices E and J, J and O, and z and 22 (see Fig. 1). In the diplomonad *Giardia lamblia*, area V4 comprises only 98 nucleotides, which is less than half the average length of the area in other eukaryotes. The latter species is the only one for which no complete model with the structure EJOUyz can be formed. A pseudoknot structure is possible, but the remaining sequence is too short to accommodate 4 additional hairpin structures as in other species. Several more simple structures can be imagined, but it is not possible to find an obvious correspondence with the helices transposable to other srRNAs, especially since the sequence is rich in G and C and bears little resemblance in primary structure to area V4 in other species.

The method for stepwise derivation of an optimal model described above is a general one which could be applied to other areas of structural variability in the srRNA molecule (Fig. 1), to similar areas existing in large ribosomal subunit RNA (35), and, for that matter, to investigation of secondary structure of any type of RNA for which a sufficient number of primary structures is available.

ACKNOWLEDGEMENTS

Our research was supported by the Incentive Program for Fundamental Research in the Life Sciences of the Belgian Office for Science Policy Programming (grant BIO/03), and by the Fund for Medical Scientific Research. J. Neefs holds an IWONL scholarship.

REFERENCES

1. Neefs,J., Van de Peer,Y., Hendriks,L. and De Wachter,R. (1990) Nucleic Acids Res., **18**, 2237–2317.
2. Noller,H.F. and Woese,C.R. (1981) Science, **212**, 403–411.
3. Gutell,R.R., Weiser,B., Woese,C.R. and Noller,H.F. (1985) Prog. Nucl. Acid Res. Mol. Biol., **32**, 155–216.
4. Gutell,R.R., Noller,H.F. and Woese,C.R. (1986) The EMBO J., **5**, 1111–1113.
5. Woese,C.R. and Gutell,R.R. (1989) Proc. Natl. Acad. Sci. USA, **86**, 3119–3122.
6. Woese,C.R. (1987) Microbiol. Rev., **51**, 221–271.
7. Hendriks,L., Van Broeckhoven,C., Vandenberghe,A., Van de Peer,Y. and De Wachter,R. (1988) Eur. J. Biochem., **177**, 15–20.
8. Cedergren,R., Gray,M.W., Abel,Y. and Sankoff,D. (1988) J. Mol. Evol., **28**, 89–112.
9. Gray,M.W. (1988) Biochem. Cell Biol., **66**, 325–348.
10. Van de Peer,Y., Neefs,J. and De Wachter,R. (1990) J. Mol. Evol., **30**, 463–476.
11. Gunderson,J.H. and Sogin,M.L. (1986) Gene, **44**, 63–70.
12. Schnare,M.N., Collings,J.C. and Gray,M.W. (1986) Curr. Genet., **10**,405–410.
13. Gunderson,J.H., Sogin,M.L., Wollett,G., Hollingdale,M., De la Cruz,V.F., Waters,A.P. and McCutchan,T.F. (1987) Science, **238**, 933–937.
14. Chan,Y.-L., Gutell,R.R., Noller,H.F. and Wool,I.G. (1984) J. Biol. Chem., **259**, 224–230.
15. Spangler,E.A. and Blackburn,E.H. (1985) J. Biol. Chem., **260**, 6334–6340.
16. Dams,E., Hendriks,L., Van de Peer,Y., Neefs,J., Smits,G., Vandenbempt,I. and De Wachter,R. (1988) Nucleic Acids Res., **16**, r87- r173.
17. Zwieb,C., Glotz,C. and Brimacombe,R. (1981) Nucleic Acids Res., **9**, 3621–3640.
18. Nelles,L., Fang,B.-L., Volckaert,G., Vandenberghe,A. and De Wachter,R. (1984) Nucleic Acids Res., **12**, 8749–8768.
19. Choi,Y.C. (1985) J. Biol. Chem., **260**, 12769–12772.
20. Eckenrode,V.K., Arnold,J. and Meagher,R.B. (1985) J. Mol. Evol., **21**, 259–269.
21. Herzog,M. and Maroteaux,L. (1986) Proc. Natl. Acad. Sci. USA, **83**, 8644–8648.
22. Gonzalez,I.L. and Schmickel,R.D. (1986) Am. J. Hum. Genet., **38**, 419–427.
23. Ellis,R.E., Sulston,J.E. and Coulson,A.R. (1986) Nucleic Acids Res., **14**, 2345–2364.
24. Hancock,J.M., Tautz,D. and Dover,G.A. (1988) Mol. Biol. Evol., **5**, 393–414.
25. Hendriks,L., De Baere,R., Van Broeckhoven,C. and De Wachter,R. (1988) FEBS Lett., **232**, 115–120.
26. Rairkar,A., Rubino,H.M. and Lockard,R.E. (1988) Biochemistry, **27**, 582–592.
27. Johansen,T., Johansen,S. and Haugli,F.B. (1988) Curr. Genet., **14**, 265–273.
28. Ninio,J. (1979) Biochimie, **61**, 1133–1150.
29. Michot, B., Qu, L.-H. and Bachellerie, J.-P. (1990) Eur. J. Biochem. **188**, 219–229
30. Pleij,C.W.A., Rietveld,K. and Bosch,L. (1985) Nucleic Acids Res., **13**, 1717–1731.
31. Hendriks,L., Goris,A., Neefs,J., Van de Peer,Y., Hennebert,G. and De Wachter,R. (1989) System. Appl. Microbiol., **12**, 223–229.
32. Freier,S.M., Kierzek,R., Jaeger,J.A., Sugimoto,N., Caruthers,M.H., Neilson,T. and Turner,D.H. (1986) Proc. Natl. Acad. Sci. USA, **83**, 9373–9377.
33. Abrahams,J.P., van den Berg,M., van Batenburg,E. and Pleij,C. (1990) Nucleic Acids Res., **18**, 3035–3044.
34. Studnicka,G.M., Eiserling,F.A. and Lake,J.A. (1981) Nucleic Acids Res., **9**, 1885–1904.
35. Gutell,R.R. and Fox,G.E. (1988) Nucleic Acids Res., **16**, r175–269.