Birth of the snoRNPs: the evolution of the modification-guide snoRNAs

Denis L. J. Lafontaine and David Tollervey

Bacteria and eukaryotes adopt very different strategies to modify their rRNAs. Most sites of eukaryotic rRNA modification are selected by guide small nuclear RNAs (snoRNAs), while bacteria rely on numerous site-specific modification enzymes. This raises a ‘chicken and egg’ dilemma: how could a system of modification that requires a large number of snoRNA cofactors have developed? Did it arise in a de novo fashion, or evolve from a pre-existing protein-based system? The rRNA sequences are well conserved in evolution, but the pattern of modification is only moderately conserved, and many more sites are modified in eukaryotes than in bacteria; why is this so? We propose a model for the origins of the modification-guide snoRNAs that attempts to answer these questions.

IN ALL ORGANISMS, many nucleotides in the mature rRNAs undergo covalent, post-transcriptional modifications. These modifications occur on precursor RNAs during ribosome synthesis and are essential of three types: base methylation; methylation of the 2'-hydroxyl group of sugar residues (2'-O-methylation); and conversion of uridine residues to pseudouridine (Ψ) by base rotation. When the sites of these modifications were mapped in bacteria and eukaryotes, two striking features emerged for which there were no immediate explanations. (1) There are about ten times more sites of 2'-O-methylation and Ψ formation in eukaryotes than in bacteria; there are four methylated sugars and ten Ψ residues in the Escherichia coli rRNAs, compared with 50 of each in Saccharomyces cerevisiae and 100 of each in humans. By contrast, sites of base methylation are more frequent in bacteria. (2) Despite the fact that the sites of modification cluster in the active centres of the ribosome in all organisms, the actual nucleotides that are modified are not highly conserved. If, as might be assumed, the pre-rRNA modifications had been conserved along with the rRNAs, their number and locations should also have been conserved.

RNA modification in bacteria

The mechanisms for recognition of sites of Ψ formation in rRNAs and snoRNAs in E. coli appear to be very similar, and there is no evidence for the involvement of guide RNAs in site selection. The protein enzymes involved recognize the sequence and/or structure of their target site directly and usually modify several adjacent nucleotides. Moreover, most such enzymes can modify closely related sites present in different RNA species. For instance, E. coli TruB specifically modifies three adjacent nucleotides (positions 38, 39 and 40) in the anticodon arm of several different tRNAs, while TruB specifically modifies the conserved U55 to Ψ in the m5U34/ΨG loop in most tRNA species. Interestingly, the Ψ synthase RluA recognizes closely related sequences in the 23S rRNA at position 174, and in several tRNAs at position 32 (Ref. 4). Much less progress has been made in identifying the rRNA bacterial 2'-O-methylases, but the prediction is that the various positions are modified by individual protein enzymes.

RNA modification in eukaryotes

The mechanism of RNA modification in eukaryotes closely resembles that in bacteria, utilizing proteins that recognize specific sites in the rRNA, and the sites of RNA modification are well conserved in evolution. For example, the yeast rRNAΨ55 synthase, Pus4p, is the homologue of E. coli TruB (Ref. 5), and yeast Pus3p is closely related to E. coli TruA (Ref. 6). Pus1p modifies the anticodon loop and stem in several RNA species at sites that are not closely related in sequence. The anticodon-loop modifications require the presence of the pre-rRNA intron in cis, which suggests that it acts as an internal guide sequence. Like the E. coli enzymes, these yeast enzymes can recognize the authentic sites specifically in vitro in the absence of cofactors.

Recent results have, however, shown the modification of the eukaryotic rRNAs in a very different light. It turns out that the sites of Ψ formation and 2'-O-methylation are selected by site-specific base pairing with snoRNAs. At each of the ~91 Ψ sites and ~106 2'-O-methylated sites in the human rRNA (Ref. 1), a specific snoRNA is predicted to base-pair with the rRNA precursor (the pre-rRNA) at the site of modification.

Around 150 different species of snoRNA are predicted to be present in human nuclei. This surprisingly large number of snoRNAs can be divided into two major classes, designated the box C+D snoRNAs and the box H+ACA snoRNAs, on the basis of conserved sequence elements and conserved predicted secondary structures4,5 (Figs 1 and 2). The only known exception is the RNA component of the ribonucleoprotein endonuclease Rnase MRP (Ref. 10). Each box H+ACA snoRNA can bind to the pre-rRNA at either one or two sites of Ψ formation6,7 and, similarly, each box C+D snoRNA can bind at either one or two sites of 2'-O-methylation8,9. The structures of the snoRNA-pre-rRNA hybrids formed by the members of each class are predicted to be rather similar (Figs 1 and 2). This suggests that, at each site of Ψ formation or 2'-O-methylation, the snoRNA-pre-rRNA interaction creates a common structure that might, in principle, be recognized by a single Ψ synthase and a single 2'-O-methylase. This situation is potentially analogous to the recognition of related sites in multiple RNAs by the E. coli modification enzymes.

For both classes of snoRNA, the in vitro function is sensitive to perturbation of the sequence of the snoRNA in the pre-rRNA interaction region. This indicates that recognition of the snoRNA-rRNA interaction by a specific modifying enzyme(s) is structure sensitive. Most box C+D snoRNAs form at least
The snoRNAs are associated with proteins in small nuclear ribonucleoprotein particles (snRNPs). Four proteins are common to all the box H snoRNAs (snRNPs). Four proteins that base pair with the rRNA (Ref. 12). Inhibition of modification by subtle mutations in the snoRNA is structure sensitive and is mediated by mismatches within this region (14,17). Similarly, the function of the box H-ACA snoRNA is sensitive and is inhibited by subtle mutations in the snoRNA, even beyond the sequences in box H-ACA snoRNPs: Cbf5p (NAP57 in vertebrates) and Gar1p (Y. Henry and M. Caizergues-Ferrer, pers. comm.) and Gar1p are required for the stability of the snoRNA (Ref. 25). These proteins are common to all the box H-ACA snoRNPs: Cbf5p (NAP57 in vertebrates) and Gar1p (Y. Henry and M. Caizergues-Ferrer, pers. comm.) and Gar1p are required for the stability of the snoRNA (Ref. 25). Like Cbf5p, Gar1p is required for the stability of the snoRNA but is not required for the stability of the snoRNAs but is required for the function of the snoRNPs (21–22). Recombinant Gar1p binds to box H-ACA snoRNAs in vitro through a conserved central domain (25). The terminal regions of Gar1p contain glycine/arginine repeat structures (GAR domains) that stabilize the protein–snoRNA interaction. The GAR domain of another nucleolar protein, nucleolin, binds RNA and destabilizes RNA secondary structure in vitro (24,26). On depletion of Gar1p, the box H-ACA snoRNA snR30 dissociated from the pre-rRNA (Ref. 22). Together these data suggest that Gar1p binds to the snoRNAs through its central domain, while the GAR domains might stabilize the snoRNA–pre-rRNA interaction, possibly also helping to open the complex structure of the pre-ribosomal particle. This process might be facilitated by Rok1p, a putative ATP-dependent helicase that interacts genetically with Gar1p (Ref. 25).

Figure 1

(a) The predicted structures of the hybrids between each box H-ACA snoRNA and the pre-rRNA at the sites of modification look strikingly similar. The sequences flanking the site of modification are base-paired to the snoRNA in a complex pseudoknot structure, while the base that is to be modified is located by mismatch within this region (14,17). Similarly, the function of the box H-ACA snoRNA is sensitive and is inhibited by subtle mutations in the snoRNA, even beyond the sequences in box H-ACA snoRNPs: Cbf5p (NAP57 in vertebrates) and Gar1p (Y. Henry and M. Caizergues-Ferrer, pers. comm.) and Gar1p are required for the stability of the snoRNA (Ref. 25). These proteins are common to all the box H-ACA snoRNPs: Cbf5p (NAP57 in vertebrates) and Gar1p (Y. Henry and M. Caizergues-Ferrer, pers. comm.) and Gar1p are required for the stability of the snoRNA (Ref. 25). Like Cbf5p, Gar1p is required for the stability of the snoRNA but is not required for the stability of the snoRNAs but is required for the function of the snoRNPs (21–22). Recombinant Gar1p binds to box H-ACA snoRNAs in vitro through a conserved central domain (25). The terminal regions of Gar1p contain glycine/arginine repeat structures (GAR domains) that stabilize the protein–snoRNA interaction. The GAR domain of another nucleolar protein, nucleolin, binds RNA and destabilizes RNA secondary structure in vitro (24,26). On depletion of Gar1p, the box H-ACA snoRNA snR30 dissociated from the pre-rRNA (Ref. 22). Together these data suggest that Gar1p binds to the snoRNAs through its central domain, while the GAR domains might stabilize the snoRNA–pre-rRNA interaction, possibly also helping to open the complex structure of the pre-ribosomal particle. This process might be facilitated by Rok1p, a putative ATP-dependent helicase that interacts genetically with Gar1p (Ref. 25).
Over time, duplication of snoRNA genes and random mutations would have given rise to snoRNAs that had new sequence complementarities to the rRNA and that were capable of directing the modification of new sites in the pre-rRNA. A major advantage of this system over the bacterial, protein-only modification system is that eukaryotes could test the effects of modification at many more positions and maintain those that, either individually or collectively, were beneficial to ribosome synthesis and/or function. Only short snoRNAs and the rRNA (two stretches of 3–10 nucleotides; see Fig. 1) are required to direct $\Psi$ formation\textsuperscript{11,12}, and these would presumably have arisen fairly readily – much more so than in a protein-based system in which a new enzyme must evolve for each new site. Formation of $\Psi$ from uridine has no energy requirement, so this trial-and-error strategy would not constitute any major metabolic burden to the cell. Moreover, each site of rRNA modification must be recognized twice. It must be identified by base-pairing to the snoRNA, and the snoRNA-pre-rRNA hybrid must then have the correct structure to be recognized as an appropriate substrate by the snoRNA-associated $\Psi$ synthase. This double selection potentially allowed greater accuracy in the selection of sites of modification. Put simply, the system must make two errors for mis-modification to occur. Together, these features are probably responsible for the observation that eukaryotes have many more modified nucleotides in their rRNAs than do bacteria. 

**Origins of the 2'-O-methylation-guide snoRNAs**

The methylase(s) responsible for eukaryotic 2'-O-methylation have not yet been identified; neither Nop1p, Nog56p nor Nog58p possesses the sequence motifs that are characteristic of methyltransferases. However, the box C+D snoRNAs form a highly conserved predicted structure that brings the two protein-binding sites, box C and box D or box D' into well-defined positions with respect to the site of modification (Fig. 2). This positional information might be used by a common methylase to select the sites of modification.

We anticipate that the methylation-guide snoRNAs, like the $\Psi$-guide snoRNAs, are required to utilize a single ancestral box C+D snoRNA. A candidate for the primordial box C+D snoRNA is U3, which is not predicted to direct 2'-O-methylation but is required for processing of the pre-rRNA\textsuperscript{13,20}. In vertebrates, U3 is substantially more abundant than other snoRNA species. In yeast, U3 base-pairs to the 5'-external transcribed spacer region of the pre-rRNA and to the 5' region of the 18S rRNA, and these interactions are necessary for several early pre-rRNA-processing steps (Refs 37–39 and K. Sharma and D. Tollervey, unpublished). Like snR10, another box C+D snoRNA, U14, is required for pre-rRNA processing and modification\textsuperscript{15-42}.

**Evolutionary implications**

This model has clear implications for the relationship between the sites of modification in bacteria and eukaryotes. Because the sites of modification in the eukaryotic rRNA were generated by mutations in the snoRNAs, they were selected independently of the pre-existing sites. This can account for the poor correlation between the sites of RNA modification in bacteria and eukaryotes. The clustering of sites of modification in similar regions of the RNA is, therefore, predicted to be the result of convergent rather than divergent evolution.

Three sites of $\Psi$ and three sites of 2'-O-methylation are common to bacteria and eukaryotes. Guide snoRNAs have been identified in eukaryotes for most of these conserved positions. The human box H-A CA snoRNA U19 is predicted to direct $\Psi$ formation at sites equivalent to E. coli $\Psi$ 1915 and $\Psi$ 1917, while U65 (snK34 in yeast) is predicted to select a position equivalent to E. coli $\Psi$ 2457 (Refs 11, 63). The human box C+D snoRNA U31 (snR67 in yeast) is predicted to direct the 2'-O-methylation equivalent to E. coli Gm2251, and yeast snR70 is predicted to select a position equivalent to E. coli Cm1402 (Ref. 44 and T. Lowe and S. Eddy, pers. commun.). Therefore, these sites appear to have been independently selected twice in evolution, in bacteria and in eukaryotes, which presumably indicates that they are of particular importance.
functional importance. Yeast snR34 and snR67 are not essential for growth, however, which indicates that the corresponding modifications are dispensable for ribosome synthesis and function (Ref. 43 and T. Lowe and S. Eddy, pers. commun.).

It remains possible that some sites of \( \Psi \) or 2'-O-methylation in the euakaryotic rRNAs are selected directly by the original protein enzymes. There can, however, be few such positions because predicted snoRNAs exist for the large majority of sites of modification. For example, guide snoRNAs have currently been identified for 51 of the 55 RNA 2'-O-methyl groups in yeast (T. Lowe and S. Eddy, pers. commun.).

Base methylation of rRNA could be very similar in eukaryotes and bacteria. The only known eukaryotic rRNA base-modifying enzyme is Dim1p, which generates the conserved m\( ^{6} \)Am\( ^{6} \)A doublet at the 3' end of the yeast 18S rRNA (Ref. 45). Dim1p is highly homologous to the \( E. \) coli dimethylase KsgA and is, indeed, able to function in rRNA dimethylation in \( E. \) coli\( ^{46} \). There is no evidence that Dim1p requires a guide snoRNA for its function in yeast, and we predict that this will be the case for most base-modifying enzymes.

Pre-rRNA processing at the 5' and 3' ends of the mature rRNAs in bacteria and Archaea is coordinated by the formation of extended stems between the flanking sequences, which generate the cleavage sites for RNase III (Ref. 10). The corresponding pre-rRNA cleavages in eukaryotes are also highly coordinated but no equivalent stems can be formed. Interactions with the snoRNAs that are required for processing (e.g. U3 and snR30) could bring the ends of the mature rRNAs together, mimicking in \( E. \) coli the interactions generated in cis by the formation of the long duplexes in \( E. \) coli and Archaea. We propose that the modification-guide snoRNAs are derived from such ancestral pre-rRNA-processing snoRNAs. This would explain why the snoRNAs involved in pre-rRNA processing and the modification-guide snoRNAs share common elements of RNA primary and secondary structure and common proteins.

A clear analogy can be drawn between the proposed origins of the modification-guide snoRNAs and previous proposals for the origins of RNase MRP (Ref. 10). Eukaryotic RNase MRP cleaves the pre-rRNA in internal transcribed spacer I and is closely related to the RNP endonuclease RNase P that, like TruII and Pus4p, processes pre-tRNAs in bacteria and in eukaryotes. It is believed that duplication of an ancestral RNase P RNA gene early in euakaryotic evolution gave rise to the RNA components of both enzymes\( ^{43} \), which retain common structural features and are associated with eight common proteins\( ^{47} \). In most bacteria and Archaea, RNase P can process the pre-rRNA because of the conserved presence of a RNA in the internal transcribed-spacer region. Following gene duplication, co-evolution of RNase MRP and the pre-rRNA might have led to the loss of the spacer RNA but retention of the RNase MRP cleavage site in the pre-rRNA.

When did the snoRNAs arise?

The gene duplication that gave rise to ChlP and Pus4p could, in principle, have occurred in an Archaron rather than a eukaryote. However, inspection of the complete genomic sequences of \( M. \) thermoautotrophicum, \( A. \) fulgidus and \( M. \) jannaschii reveals only single TruII-like open-reading frames. Moreover, no clear homologue of the other common box H\( ^{1} \)ACA snoRNP protein, Gar1p, is present in these genomes. By contrast, homologues of Nop1p and Nop58p are present in the genomes of \( M. \) thermoautotrophicum, \( A. \) fulgidus and \( M. \) jannaschii (Nop1p from \( M. \) jannaschii...
is 41% identical, 53% similar to yeast Nop1p over its full length; *M. jannaschii* Nop58p is 33% identical, 45% similar to yeast). These genes are likely to be co-expressed as an operon in *M. thermoautotrophicum* and *A. fulgidus*, and their co-regulation would strongly support their functional conservation. These observations suggest that Archaea possess homologues of the box C1D snoRNAs but not the box H1ACA snoRNAs (Table I).

In agreement with this model, there is a dramatic discrepancy in the numbers of cytosine residues and 2'-O-methyl groups in *S. solfataricus* and *S. acidocaldarius*. The number of cytosine groups in *S. solfataricus* and *S. acidocaldarius* is low, and close to that of bacteria such as *E. coli* (Refs 48 and 49). By contrast, the number of 2'-O-methyl groups is much higher than in bacteria and close to the number in eukaryotes (Refs 48 and 49). *S. solfataricus* has a total of 67 2'-O-methyl modifications in the rRNA, in comparison with ~55 modifications in yeast and four in *E. coli*. It has been suggested that the 2'-O-methyl groups act to stabilize RNA stems by constraining the sugar residues into the more rigid C3'-endo conformation (Ref. 50). In extreme thermophiles there might, therefore, have been strong selection for the development of a system that would permit the 2'-O-methylation of many sites in the rRNA. Recent evolutionary models propose that the eukaryotes arose from the thermophilic Archaea (Ref. 51) and they would, therefore, have inherited this system of RNA modification.

Interestingly, while Cbf5p-depleted strains are not impaired in the formation of wobble U snRNAs (C. Branlant, pers. commun.), a potential 2'-O-methyl-ation-guide snoRNA has been identified for the U6 spliceosomal snRNA (Ref. 16). Whether this is an ancient feature remains to be determined.

**Back to the RNA world**

The RNA-world hypothesis envisages that, at a relatively early stage in the evolution of life, organisms existed in which RNA molecules served both as genome and as enzymes. Following the discovery of catalytic RNAs, this model became widely accepted. In the light of this model, it has generally been assumed that, throughout evolution, the functions once performed by RNAs have been progressively taken over by proteins. Therefore, surviving RNAs are often regarded as being ancient relics – survivors from the RNA world. In the model that we have presented, the snoRNAs have a very different evolutionary history, arising in a de novo fashion in Archaea or eukaryotes and progressively taking over the functions of protein enzymes.

**Conclusions**

We propose that, early in eukaryotic evolution, a tRNA w synthase that previously recognized a target sequence in *cis* within pre-tRNAs acquired the ability to recognize an interaction in *trans* between an snoRNA base-paired to the rRNA. The bulk of the snoRNA species then arose by duplication of an ancestral snoRNA gene. A similar process probably gave rise to the class of methylation-guide snoRNAs, although it might have occurred in Archaea. The high flexibility and greater accuracy predicted for the snoRNPs-based system of modification allowed many more sites of modification to be selected in eukaryotes; these sites are, however, related to the sites of modification in bacteria by convergent, rather than divergent, evolution.

**Acknowledgements**

We thank C. Branlant, M. Caizergues-Ferrer, S. Eddy, S. Fournier, Y. Henry, T. Kiss, T. Lowe, S. Massenet and J. Ni for allowing us to quote unpublished results, and members of the Tollervey lab for critical reading of the manuscript.
References
3 Nurse, K., et al. (1995) RNA 1, 102–112
7 Simos, G. et al. (1996) EMBO J. 15, 2270–2284
11 Girard, J. P. et al. (1992) EMBO J. 11, 673–682
12 Schimmang, T. et al. (1989) EMBO J. 8, 2955–2965
20 Cathomen, T. et al. (1998) Nature Genet. 18, 537–537
21 Girard, J. P. et al. (1992) EMBO J. 11, 673–682
22 Bouzouan-Antenni, C. et al. (1997) EMBO J. 16, 4770–4776

Table I. Birth of the snoRNPs

<table>
<thead>
<tr>
<th></th>
<th>Bacteria</th>
<th>Archaea</th>
<th>Eukaryotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>P sites</td>
<td>10 (E. coli)</td>
<td>9 (S. aureus)</td>
<td>43 (yeast)</td>
</tr>
<tr>
<td>H1-ACA snoRNPs</td>
<td>Single TruI homologue; no clear homologue of Gar5p</td>
<td>Single TruI homologue; no clear homologue of Gar5p</td>
<td>Cto5p and TruI homologues associating with box H1-ACA snoRNPs</td>
</tr>
<tr>
<td>2′-O-Methyl sites</td>
<td>4 (E. coli)</td>
<td>67 (S. aureus)</td>
<td>55 (yeast)</td>
</tr>
<tr>
<td>C/D snoRNPs</td>
<td>No clear homologues of hop5 and hop18p</td>
<td>Homologues of hop5 and hop18p – in some species</td>
<td>No clear homologues of hop5 and hop18p associated with box C/D snoRNPs</td>
</tr>
</tbody>
</table>

We speculate that the large rise in numbers of 2′-O-methyl groups in the Archaea followed the appearance of the box C/D class of guide snoRNPs in this kingdom. By contrast, the box H1-ACA snoRNPs might have arisen in eukaryotes.