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 tRNAs. Most sites of eukaryotic tRNA 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 tRNA 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 RNAs undergo covalent, post-transcriptional modifications. These modifications occur on precursor RNAs during ribosome synthesis and are essentially 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 were 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 RNA, compared with 50 of each in Saccharomyces cerevisiae and 100 of each in human RNA. (2) Despite the fact that the sites of modification cluster in the active centre of the ribosome in all organisms, the actual nucleotides that are modified are not highly conserved. If, as might be assumed, the pre-tRNA modifications had been conserved along with the tRNAs, their number and locations should also have been conserved.

RNA modification in bacteria

The mechanisms for recognition of sites of Ψ formation in tRNAs and tRNAs 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 m5S1PCG loop in most tRNAs. Interestingly, the Ψ synthase RluA recognizes closely related sequences in the 23S rRNA at position 746, and in several tRNA at position 32 (Ref. 4). Much less progress has been made in identifying the RNA bacterial 2'-O-methylases, but the prediction is that 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 tRNA, and the sites of tRNA modification are well conserved in evolution. For example, the yeast tRNAΨ55 synthase, Pus4p, is the homologue of E. coli TruB (Ref. 5), and yeast Pus3p is closely related to E. coli TrnUa (Ref. 6). Pus1p modifies the anticodon loop and stem in several tRNA species at sites that are not closely related in sequence. The anticodon-loop modifications require the presence of the pre-tRNA 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 tRNA 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 ~100 2'-O-methylated sites in the human tRNA (Ref. 1), a specific snoRNA is predicted to base-pair with the tRNA precursor (the pre-tRNA) at the site of modification. Around 150 different species of snoRNA are predicted to be present in human nucleoli. 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 structures (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-tRNA at either one or two sites of Ψ formation and and, similarly, each box C+D snoRNA can bind at either one or two sites of 2'-O-methylation. The structures of the snoRNA-pre-tRNA 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-tRNA 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 vivo function is sensitive to perturbation of the sequence of the snoRNA in the pre-tRNA interaction region. This indicates that recognition of the snoRNA-tRNA interaction by the modifying enzyme(s) is structure sensitive. Most box C+D snoRNAs form at least
The predicted structures of the hybrids between each box H-ACA snoRNA and the pre-rRNA at the sites of Ψ formation 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 by rotation about the N1-C6 axis is free to interact with the Ψ synthase. The base pairing involves two short stretches of 3-10 nucleotides on both sides of the base to be modified, the distance between this position and the conserved elements on the snoRNA (box H or ACA) is usually of 14 nucleotides. (b) Ψ formation by base rotation.

The importance of the GAR domain and the function of Gar1p

Gar1p in that it contains a GAR domain for 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). Three proteins that are common to the box C-ACA snoRNAs have also been identified: Nop1p (fibulorin in vertebrates) and two homologous proteins, Nop56p and Nop58p (Refs 26–28 and D. L. J. Lafontaine and D. Tollervey, unpublished). Like Chl5p, Nop58p is required for the stability of the snoRNAs but is required for the function of the snoRNPs. Nop1p also resembles Gar1p in that it contains a GAR domain that might help to mediate snoRNA–pre-rRNA interactions.

The striking differences between the rRNA-modification systems in bacteria and eukaryotes pose a dilemma: how could the eukaryotic modification system, which requires a huge number of guide snoRNPs, have arisen from a pre-existing protein-based modification system? Analysis of yeast Chl5p offers a possible escape from this apparent paradox. Chl5p and its mammalian homologue, NAP57, are localized in the nucleolus and eukaryotes, which suggests that they are the rRNA Ψ synthases. Genetic depletion of Chl5p inhibits pre-rRNA processing and blocks formation of Ψ in the pre-rRNA, which indicates that this is indeed the case.

From the protein homology, it seems clear that both Psp4p and Chl5p are derived from a TruB-like Ψ synthase. It also seems inescapable that eukaryotes once possessed a single box H-ACA snoRNA from which the many snoRNAs species of this class were derived. We propose that duplication of the gene encoding a TruB-like RNA Ψ synthase gave rise to the ancestors of Chl5p and Psp4p. While Psp4p retained its function in rRNA modification, Chl5p acquired the ability to recognize a hybrid between an ancestral box H-ACA snoRNA and the pre-rRNA. This primordial snoRNA would, presumably, have previously fulfilled some other role in ribosome synthesis (e.g. in ribosomal assembly or pre-rRNA processing). A candidate for this ancestral snoRNA is small, the only known essential box H-ACA snoRNA, which is required for pre-rRNA processing in yeast but is not predicted to direct Ψ formation in the pre-rRNA (Ref. 32, 33 and T. Kiss, pers. commun.). Another box H-ACA snoRNA, snR110, is required for normal processing of the 18S rRNA and for the formation of a Ψ residue in the 18S rRNA.
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 in which a new enzyme must evolve for each new site. Formation of Ψ from uracil 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 Ψ 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, Nop56p nor Nop58p 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 C and 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 Ψ-guide snoRNAs, arose by duplication of 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\(^\text{35,36}\). In vertebrates, U3 is substantially more abundant than other snoRNA species. In yeast, U3 basepairs 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\(^\text{33,42}\).

**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 rRNA modification in bacteria and eukaryotes. The clustering of sites of modification in similar regions of the rRNA is, therefore, predicted to be the result of convergent rather than divergent evolution.

Three sites of Ψ 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-ACA snoRNA U19 is predicted to direct Ψ formation at sites equivalent to E. coli Ψ1915 and Ψ1917, while U65 (snK34 in yeast) is predicted to select a position equivalent to E. coli Ψ 2457 (Refs 11, 43). The human box C+D snoRNA U31 (snR87 in yeast) is predicted to direct the 2'-O-methylation equivalent to E. coli Cm2251, 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
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 ψ or 2′-O-methylation in the eukaryotic 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 such sites. 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 m6A6A 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. coli46. 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 clearances in eukaryotes are also tightly 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 trans 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 eukaryotic evolution gave rise to the RNA components of both enzymes43, which retain common structural features and are associated with eight common proteins44. In most bacteria and Archaea, RNase P can process the pre-tRNA because of the conserved presence of a tRNA in the internal transcribed-spacer region. Following gene duplication, co-evolution of RNase MRP and the pre-tRNA might have led to the loss of the spacer tRNA but retention of the RNase MRP cleavage site in the pre-tRNA.

When did the snoRNAs arise?
The gene duplication that gave rise to ChlPp and Pus4p could, in principle, have occurred in an Archaron rather than a eukaryote. However, inspection of the complete genomic sequences of Methanobacterium thermoautotrophicum, Archaeoglobus fulgidus and Methanococcus jannaschii reveals only single TruII-like open-reading frames. Moreover, no clear homologue of the other common box HACA 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 C/D snoRNAs but not the box H/ACA snoRNAs (Table I).

In agreement with this model, there is a dramatic discrepancy in the numbers of Ψ residues and 2′-O-methyl groups in the Archaea *Sulfolobus*. The number of Ψ 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. *S. solfataricus* has a total of 67 2′-O-methyl modifications in the tRNA, 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. 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 tRNA. Recent evolutionary models propose that the eukaryotes arose from the thermophilic Archaea and they would, therefore, have inherited this system of RNA modification.

Interestingly, while Chbp-depleted strains are not impaired in the formation of Ψ in U snRNAs (C. Branlant, pers. commun.), a potential 2′-O-methyl-ation-guide snoRNA has been identified for the U6 spliceosomal snoRNA. 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 Ψ synthase that previously recognized a target sequence in *cis* within pre-snoRNAs acquired the ability to recognize an interaction in *trans* between an snoRNA base-paired to the tRNA. 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 snoRNP-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. Kuss, 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 Nune, K. et al. (1995) EMBO J. 14, 8027–8034
4 Mikkola, J. et al. (1997) EMBO J. 16, 4015–4024
7 Simos, G. et al. (1999) EMBO J. 18, 9198–9214
20 Laterra, D. J. et al. (1998) Genes Dev. 12, 527–537
21 Giraud, J. P. et al. (1992) EMBO J. 11, 673–682
22 Bourouza-Antonielli, C. et al. (1997) EMBO J. 16, 4770–4776
26 Schramm, T. et al. (1989) EMBO J. 8, 4015–4024

Table I. Birth of the snoRNPs

<table>
<thead>
<tr>
<th>Bacteria</th>
<th>Archaea</th>
<th>Eukaryotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>T sites</td>
<td>10 (Escherichia coli)</td>
<td>9 (Sulfolobus)</td>
</tr>
<tr>
<td>H+ACA snoRNPs</td>
<td>Single Trf homologues; no clear homologue of Gar1p</td>
<td>Single Trf homologues; no clear homologue of Gar1p</td>
</tr>
<tr>
<td>2′-O-methyl sites</td>
<td>4 (E. coli)</td>
<td>67 (Sulfolobus)</td>
</tr>
<tr>
<td>C+D snoRNPs</td>
<td>No clear homologues of Nop1p or Hops88p</td>
<td>Homologues of Nop1p and Hops88p – in some species</td>
</tr>
</tbody>
</table>

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

FREE SUPPLEMENT NEXT MONTH
The Trends Guide to Bioinformatics

Bioinformatics and genomics are becoming accepted in the research community as vital strategies for genetic analysis and elucidation of the molecular relationships between proteins. The Trends Guide to Bioinformatics demystifies the concepts involved in this cutting-edge area with a series of tutorials from respected authors in the field.

Contents:
Introduction by Mark Boguski
Text-based database searching by Fran Lewitter
Fundamentals of database searching by Stephen Altschul
Practical database searching by Steve Brenner
Computational genethinking by David Haussler
Multiple-alignment and -sequence searches by Sean Eddy
The future of bioinformatics by Janet Thomson

Protein classification and functional assignment by Kay Hofmann
Phylogenetic analysis and comparative genomics by Jim Lake and Jonathan Moore
Databases of biological information by Minoru Kanekawa
Functional genomics by Mike Brownstein, J. Trent and Mark Boguski