*Nucleic Acids Research, 2005, Vol. 33, Database issue* ***D135–D138***
*doi:10.1093/nar/gki015*

# The Small Subunit rRNA Modification Database

**James A. McCloskey and Jef Rozenski[1],***

Departments of Medicinal Chemistry and Biochemistry, University of Utah, Salt Lake City, UT 84112, USA and [1]Rega Institute for Medical Research, Katholieke Universiteit Leuven, Minderbroedersstraat 10, B-3000 Leuven, Belgium

Received August 17, 2004; Accepted September 17, 2004

## ABSTRACT

**The Small Subunit rRNA Modification Database provides a listing of reported post-transcriptionally modified nucleosides and sequence sites in small subunit rRNAs from bacteria, archaea and eukarya. Data are compiled from reports of full or partial rRNA sequences, including RNase T1 oligonucleotide catalogs reported in earlier literature in studies of phylogenetic relatedness. Options for data presentation include full sequence maps, some of which have been assembled by database curators with the aid of contemporary gene sequence data, and tabular forms organized by source organism or chemical identity of the modification. A total of 32 rRNA sequence alignments are provided, annotated with sites of modification and chemical identities of modifications if known, with provision for scrolling full sequences or user-dictated subsequences for comparative viewing for organisms of interest. The database can be accessed through the World Wide Web at http://medlib.med.utah.edu/SSUmods.**

## INTRODUCTION

Post-transcriptional nucleoside modification in rRNA was recognized and reported in the early literature (1,2), but has received far less attention than that for tRNA because of limited sequence studies carried out at the RNA level (which was historically a frequent means of discovery of the new nucleosides in tRNA), coupled with considerably less understanding of modification structure–function relationships compared with those of tRNA. Knowledge of the presence, chemical identities and sequence locations of rRNA modifications has steadily accumulated, even prior to the more recent availability of corresponding rRNA gene sequences. Although rRNA gene sequences are now accumulating at a rapid rate (http://www.rna.icmb.utexas.edu/) they unfortunately do not address nucleotide modifications, which are performed by the actions of specific modification enzymes after RNA transcription. Nevertheless, the relevant modification data when extracted from a variety of literature sources, particularly with regard to RNA from the small ribosomal subunit due to the availability of extensive RNase T1 oligonucleotide catalogs (3), have reached a point at which assembly into a single database is useful to workers in multiple developing fields. These include (4) studies of RNA modification enzymes, phylogenetic relationships between modification and evolution of the organisms that contain them, and functions of highly conserved modifications, for example through influence on rRNA secondary and tertiary structure. Interest in the details of rRNA structure/function is further propelled by reports of the high-resolution crystal structures of small (5) and large (6) ribosomal subunits.

The structures of presently known modified nucleosides in rRNA are shown in Figure 1. Further information on each nucleoside, including systematic chemical name, Chemical Abstracts registry number useful for computer-based searching and citations for original structure determination and chemical synthesis, are given in Ref. (7) and in updated form at http://medlib.med.utah.edu/RNAmods.

## DATABASE DESCRIPTION

The Small Subunit (SSU) rRNA Modification Database is accessible on the World Wide Web at http://medlib.med.utah.edu/SSUmods. The site is divided into four main areas: (i) OVERVIEW, which provides a description of the database, sources of modification data and comments on numbering systems, and the use of sequence alignments; (ii) MAP, in which modifications can be displayed for selected organisms in the three primary evolutionary domains either as a simple list or as a full sequence map; (iii) ALIGN, which provides modification-annotated aligned sequences for comparisons among multiple sources, presently including 17 bacteria, 5 archaea, 8 eukarya, 1 chloroplast and 1 mitochondrion; and (iv) BROWSE, providing two search functions for the user, reported rRNA sequence locations for specific modifications of interest and retrieval of literature citations from the database files either for selected authors or publication names or years.

Comments on subdivisions and information in each of these categories are as follows.

*To whom correspondence should be addressed. Tel: +32 16 33 73 90; Fax: +32 16 33 73 40; Email: jef.rozenski@rega.kuleuven.ac.be

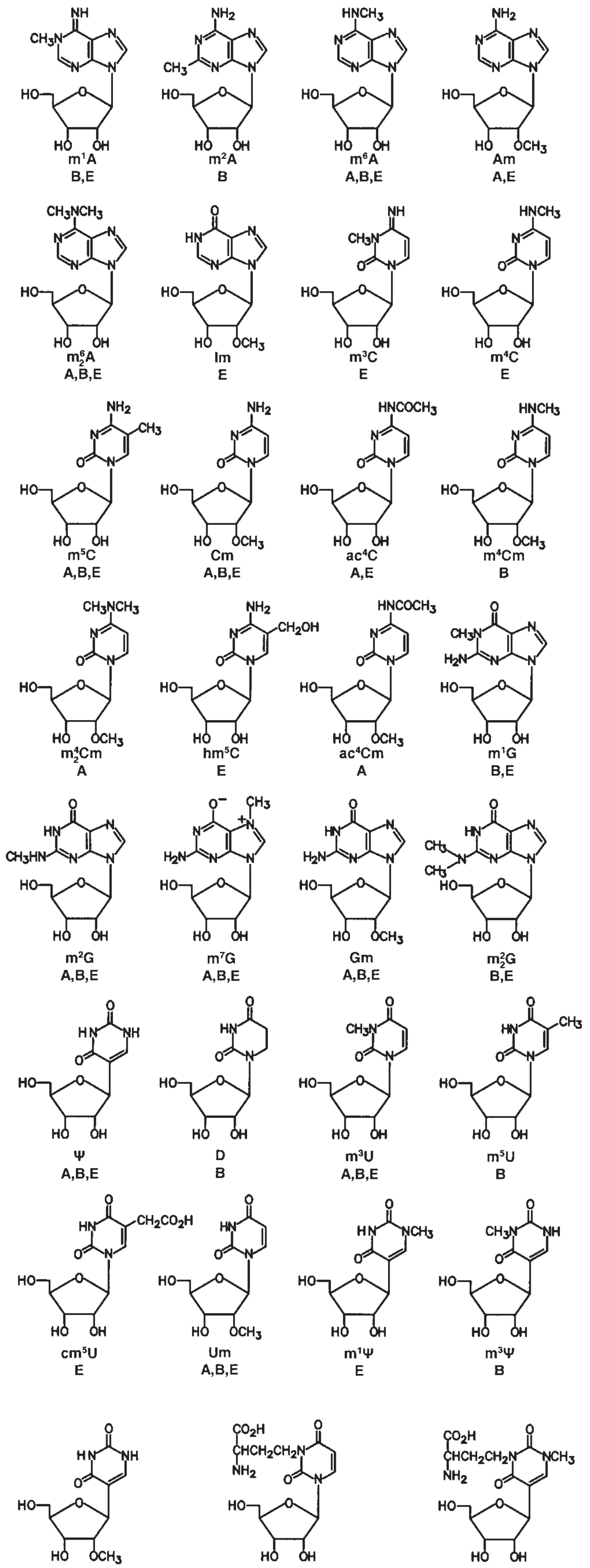

**Figure 1.** Chemical structures of presently known post-transcriptionally modified nucleosides from rRNA. A, archaea; B, bacteria; and E, eukarya. *Symbols*: m$^1$A, 1-methyladenosine; m$^2$A, 2-methyladenosine; m$^6$A, *N*$^6$-methyladenosine; Am, 2′-*O*-methyladenosine; m$^6_2$A, *N*$^6$,*N*$^6$-dimethyladenosine; Im, 2′-*O*-methylinosine; m$^3$C, 3-methylcytidine; m$^4$C, *N*$^4$-methlycytidine; m$^5$C, 5-methlycytidine; Cm, 2′-*O*-methylcytidine; ac$^4$C, *N*$^4$-acetylcyti dine; m$^4$Cm, *N*$^4$,2′-*O*-dimethylcytidine; m$^4_2$Cm, *N*$^4$,*N*$^4$,2′-*O*-trimethylcytidine; hm$^5$C, 5-hydroxymethylcytidine; ac$^4$Cm, *N*$^4$-acetyl-2′-*O*-methylcytidine; m$^1$G, 1-methylguanosine; m$^2$G, *N*$^2$-methylguanosine; m$^7$G, 7-methylguanosine; Gm, 2′-*O*-methylguanosine; m$^2_2$G, *N*$^2$,*N*$^2$-dimethylguanosine; Ψ, pseudouridine; D, dihydrouridine; m$^3$U, 3-methyluridine; m$^5$U, 5-methyluridine; cm$^5$U, 5-carboxymethyluridine; Um, 2′-*O*-methyluridine; m$^1$Ψ, 1-methylpseudouridine; m$^3$Ψ, 3-methylpseudouridine; Ψm, 2′-*O*-methylpseudouridine; acp$^3$U, 3-(3-amino-3-carboxypropyl)uridine; and m$^1$acp$^3$Ψ, 1-methyl-3-(3-amino-3-carboxypropyl)pseudouridine.

## Accuracy and completeness of database entries

Very few SSU rRNA modification maps are believed to be complete, for example as is *Escherichia coli* [leading citations in (8)] and *Saccharomyces cerevisiae* (9), generally due at the time of the study, to the unavailability of the corresponding gene sequence for use as a template, or by using inadequate methods for nucleoside structure assignments. [Further comments on reports of modification in the early rRNA literature are given in Ref. (7).]

For some organisms in the database there are three major sources of incomplete modification listings. The first is that only portions of the rRNA molecule may have been sequenced, thus giving the (possibly incorrect) appearance in the comparative alignments that portions of the rRNA are unmodified. For example in the case of *Bacillus stearothermophilus* 16S rRNA only the 52 3′-terminal nucleotides were sequenced (10), as noted in the database tabular listing, which produces an incomplete picture when the *B. stearothermophilus* sequence is aligned against that from *E.coli*, which is complete.

An additional source of incomplete listings was provided in the data derived from RNase T1 catalogs in the earlier literature. Most such studies were made for the purpose of establishing patterns of phylogenetic relatedness (11), in which most reported catalog data omitted short sequences (whether modified or not) because they were of less use in establishing characteristic ‘signature’ sequences. The extent of this problem is difficult to judge unless a total nucleoside modification assay (as from HPLC or LC/MS analysis) had been carried out to determine the identities and approximate numbers of modifications.

A third source of incompleteness may result from the method used for the detection of modification. For instance, pseudouridine residues [abundant in eukarya (12)] would be missed if the $^{14}$C-methionine method (9), indicating only methylated nucleosides, had been used. Similarly, maps reporting locations of pseudouridine have in recent years often been based on the derivatization-reverse transcriptase method of Ofengand (13), which is not intended to detect ribose methylation. References to the original citations will usually be required to establish whether the methodology used would have influenced the reported modified nucleoside distribution.

## Identification of the modified nucleoside

In many instances, most notably in the RNase T1 catalogs, the presence—but not the identity—of modification at a given site was established. In some cases the designation 'N' was used in the original literature for reporting oligonucleotide modification, perhaps reflecting unusual modification, or uncertainty in making structure assignments based solely on unusual electrophoretic or chromatographic mobility of nuclease fragments. Examples of such designations are found in the T1 catalogs from methanogens (14).

As discussed (7), the identification of modified nucleosides in the early rRNA literature is somewhat problematic partly due to the use of inadequate analytical methods of identification. A related problem encountered during compilation of this database arises when a conserved site of modification was encountered during sequencing of the RNA, and the structure of the modification then assigned without reported experimental evidence as to how the nucleoside was actually identified. The most common example arises in assignment of the very highly conserved modified sequence **aa**CCUG as $m^6_2$A $m^6_2$ACCUG in stem–loop 45, an assumption which is likely to be correct in this particular case.

### Sequence locations of modified nucleosides

Literature reports can be considered in terms of three likely levels of accuracy.

(i) The modification position has unequivocally been derived and the reported location was supported by correlation with the corresponding gene sequence. Examples are in *E.coli* (15), *Sulfolobus solfataricus* (16) and human (17) rRNAs. Such modifications are color-coded in the database.

(ii) The modification position has been determined by sequencing of a nuclease fragment, which in turn has a unique sequence context when compared with the gene sequence, and has been thereby placed in the total rRNA sequence by the database curators. This method of placement is indicated in the 'table of modifications' listings under MAP. Examples are found in the tables for *Methanococcus jannaschii* and *Rattus norvegicus* (Novikoff hepatoma) and are also color-marked.

(iii) The modification site was determined by sequencing of a nuclease fragment, but the gene sequence permits placement of the fragment in more than one site in the rRNA molecule. In a limited number of cases, the database curators have nonetheless placed the modified oligonucleotide in the overall sequence because the modification site is very highly conserved in rRNAs of related organisms (within bacteria, archaea or eukarya). This circumstance is uniquely color-marked in ALIGN presentations and in sequence and tabular listings under MAP. Good examples of this fall under listings derived from nuclease catalogs from *R.norvegicus* 18S rRNA (18), which was an extensive and carefully carried out study in which modified nucleotides were separately identified in the sequenced RNA fragments.

### Choice of SSU rRNA numbering systems

Two parallel numbering systems have been used throughout the database: 'unaligned numbering', in which the absolute numbering for a reported RNA sequence has been used; and 'aligned numbering', in which a 'universal' numbering system from the European ribosomal RNA database (19) (http://www.psb.ugent.be/rRNA/) has been adopted. The universal system takes advantage of the generally conserved nature of rRNA secondary structures within all three evolutionary domains, by adopting residues numbers sufficiently broad so as to include all presently reported SSU RNA sequences so that meaningful alignments between RNAs from distantly related organisms can be visualized.

### Modification display options under MAP

The data compiled in the MAP section are derived primarily from RNase T1 catalog data sets and are presented in the tabular form, showing those oligonucleotide sequences that were reported to be modified. These tabular listings have been somewhat arbitrarily divided into two categories as follows.

*Modifications lists and maps* show organisms for which more extensive lists of modification sites are available, derived from both T1 fragments and other types of data. Many of these sites can be referenced to the full rRNA sequence and may therefore be displayed as a full sequence map by using either the aligned (universal) or unaligned (absolute) numbering system. If modification sites were originally determined solely from the T1 catalogs, they are likely to be incomplete because only 5mer or 6mer (and longer) sequences were reported usually (14). If the modification site is ambiguous, often as a consequence of multiple possibilities for placement within the rRNA molecule, the location is shown as 'unknown'.

*RNase T1 catalog data* show listings of modified T1 oligonucleotides, usually from groups of related organisms, and usually without reference to overall sequence of the rRNA. For example, under Actinomycetes, modification data from 25 different organisms are presented. Therefore sequence maps, which are generally not available in such cases, are not provided. The data are always presented in this subsection rather than under 'lists and maps' when fewer than three types of nucleoside modifications were originally reported.

In most instances, the position of the modification in the T1 fragment was reported but the identity of the modified nucleoside was not established. In the present database these modified sites are indicated by a lower case a, g, c or u. The designation 'N' is used if listed in that fashion in the original report, except when contemporary availability of the gene sequence permits assignment as a, g, c or u.

### Alignment of modification-annotated full SSU rRNA sequences or subsequences

A total of 32 SSU rRNA sequences have been aligned by the database curators, with reference in part to alignments shown in the European ribosomal RNA database (19). In some instances, reported T1 catalog data were placed in full sequence context by the curators when made possible by later reports of the gene sequences. Examples are in sequence maps for *Haloferax volcanii* (20) and *S.solfataricus* 16S rRNAs (16). Color-coded annotations of individually modified nucleotides as reported in the literature have been performed, with modified nucleoside symbols (as in Figure 1) included if known. Full sequences or user-dictated subsequences can be scrolled for comparative viewing for any organism of interest.

## ACKNOWLEDGEMENTS

The authors thank P. F. Crain for assistance and advice, G. E. Fox and other colleagues for discussions, and S. C. Pomerantz for structural drawings. We are grateful to N. Lombardo in the Eccles Health Sciences Library for assistance in setting up the database. Compilation of the database was supported by grant GM29812 from the National Institute of General Medical Sciences (J.A.M.) and grant 1.5.014.02 from the Fonds voor Wetenschappelijk Onderzoek—Vlaanderen (J.R.).

## REFERENCES

1. Littlefield,J.W. and Dunn,D.B. (1958) Natural occurrence of thymine and three methylated adenine bases in several ribonucleic acids. *Nature*, **181**, 254–255.
2. Smith,J.D. and Dunn,D.B. (1959) An additional sugar component of ribonucleic acids. *Biochim. Biophys. Acta*, **31**, 573–575.
3. Fox,G.E., Pechman,K.R. and Woese,C.R. (1977) Comparative cataloguing of 16S ribosomal RNA: molecular approach to prokaryotic systematics. *Int. J. Syst. Bacteriol.*, **27**, 44–57.
4. Grosjean,H. and Benne,R. (eds) (1998) *Modification and Editing of RNA*. ASM Press, Washington, DC.
5. Wimberly,B.T., Brodersen,D.E., Clemons,W.M.,Jr, Morgan-Warren,R.J., Carter,A.P., Vonrhein,C., Hartsch,T. and Ramakrishnan,V. (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327–339.
6. Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.
7. Limbach,P.A., Crain,P.F. and McCloskey,J.A. (1994) Summary: the modified nucleosides of RNA. *Nucleic Acids Res.*, **22**, 2183–2196.
8. Bakin,A., Kowalak,J.A., McCloskey,J.A. and Ofengand,J. (1994) A single pseudouridine residue in *E. coli* 16S RNA is located at position 516. *Nucleic Acids Res.*, **22**, 3681–3684.
9. Maden,B.E.H. (1990) The numerous modified nucleosides in eukaryotic ribosomal RNA. *Progr. Nucleic Acids Res. Mol. Biol.*, **39**, 241–303.
10. Van Charldorp,R., Van Kimmende,A.M.A. and Van Knippenberg,P.H. (1981) Sequence and secondary structure of the colicin fragment of *Bacillus stearothermophilus* 16S ribosomal RNA. *Nucleic Acids Res.*, **9**, 4909–4918.
11. Woese,C.R. and Fox,G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA*, **74**, 5088–5090.
12. Ofengand,J. and Fournier,M.J. (1998) The pseudouridine residues of rRNA: number, location, biosynthesis, and function. In Grosjean,H. and Benne,R. (eds), *Modification and Editing of RNA*. ASM Press, Washington, DC, pp. 229–253.
13. Bakin,A. and Ofengand,J. (1993) Four newly located pseudouridylate residues in *Escherichia coli* 23S ribosomal RNA are all at the peptidyl transferase center: analysis by the application of a new sequencing technique. *Biochemistry*, **32**, 9754–9762.
14. Balch,W.E., Fox,G.E., Magrum,L.J., Woese,C.R. and Wolfe,R.S. (1979) Methanogens: reevaluation of a unique biological group. *Microbiol. Rev.*, **43**, 260–296.
15. Brosius,J., Palmer,M.L., Kennedy,P.J. and Noller,H.F. (1978) Complete nucleotide sequence of a 16S ribosomal RNA gene from *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **75**, 4801–4805.
16. Olsen,G.J., Pace,N.R., Nuell,M., Kaine,B.P., Gupta,R. and Woese,C.R. (1985) Sequence of the 16S rRNA gene from the thermoacidophilic archaebacterium *Sulfolobus solfataricus* and its evolutionary implications. *J. Mol. Evol.*, **22**, 301–307.
17. McCallum,F.S. and Maden,D.E.H. (1985) Human 18S ribosomal RNA sequence inferred from DNA sequence. Variations in 18S sequences and secondary modification patterns between vertebrates. *Biochem. J.*, **232**, 725–733.
18. Choi,Y.C. and Busch,H. (1978) Modified nucleotides in $T_1$ RNase oligonucleotides of 18S ribosomal RNA of the Novikoff hepatoma. *Biochemistry*, **17**, 2551–2560.
19. Wuyts,J., Perrière,G. and Van de Peer,Y. (2004) The European ribosomal RNA database. *Nucleic Acids Res.*, **32**, D101–D103.
20. Gupta,R., Lanter,J.M. and Woese,C.R. (1983) Sequence of the 16S ribosomal RNA from *Halobacterium volcanii*, an archaebacterium. *Science*, **221**, 656–659.