3756–3762 *Nucleic Acids Research, 1996, Vol. 24, No. 19* © 1996 Oxford University Press

Identification of new RNA modifying enzymes by iterative genome search using known modifying enzymes as probes

Claes Gustafsson, Ralph Reid[1], Patricia J. Greene and Daniel V. Santi*

Departments of Pharmaceutical Chemistry and of Biochemistry and Biophysics, University of California, San Francisco, CA 94143-0448, USA and [1]Biomolecular Resource Center, University of California, San Francisco, CA 94143-0541, USA

Received June 14, 1996; Revised and Accepted August 12, 1996

ABSTRACT

The complete nucleotide sequences of the *Haemophilus influenzae* and *Mycoplasma genitalium* genomes and the partially sequenced *Escherichia coli* chromosome were analyzed to identify open reading frames (ORFs) likely to encode RNA modifying enzymes. The protein sequences of known RNA modifying enzymes from three families—m^5U methyltransferases, Ψ synthases and 2′-*O* methyltransferases—were used as probes to search sequence databases for homologs. ORFs identified as homologous to the initial probes were retrieved and used as new probes against the databases in an iterative manner until no more homologous ORFs could be identified. Using this approach, we have identified two new m^5U methyltransferases, seven new Ψ synthases and four new 2′-*O* methyltransferases in *E.coli*. Many of the ORFs found in *E.coli* have direct genetic counterparts (orthologs) in one or both of *H.influenzae* and *M.genitalium*. Since there is a near-complete knowledge of RNA modifications in *E.coli*, functional activities of the proteins encoded by the identified ORFs were proposed based on the level of conservation of the ORFs and the modified nucleotides.

INTRODUCTION

As of April 1996, high throughput genomic sequencing has provided hundreds of viral genome sequences, ~20 organellar sequences and the complete nucleotide sequences of two free living organisms: *Haemophilus influenzae* (1.8 Mbp; 1) and *Mycoplasma genitalium* (0.6 Mbp; 2). Also, 74% of the *Escherichia coli* chromosome (4.7 Mbp) has been reported in the *E.coli* database collection release 25 (January 1996; 3). The numerous open reading frames (ORFs) now identified brings the genome projects to the next level of analysis: to identify the functions of the uncharacterized ORFs. A practical first approach towards this objective involves prediction of function through homology comparisons of known proteins to uncharacterized ORFs. Indeed, such comparisons of *E.coli* ORFs have led to assignments of many general functions (4).

RNA modifications have been well characterized in *E.coli*. Mature RNA contains many modified nucleotides, of which three are 5-methyluridines (m^5U), seventeen are pseudouridines (5-ribosyluridine, Ψ) and seven are 2′-*O* methylated nucleosides (Nm, where N denotes A, C, G or U) (reviewed in 5). Six of the enzymes that catalyze formation of the m^5U, Ψ and Nm nucleotide modifications have been identified in *E.coli*.

In this paper, we attempt to identify ORFs likely to encode RNA modifying enzymes. We used the amino acid sequence of eight known RNA modifying enzymes, six from *E.coli* and one each from *Streptomyces azureus* and *Saccharomyces cerevisiae*, as probes to search the databases for homologous ORFs. The probes used represent enzymes which catalyze three types of RNA modifications; uracil m^5U methyltransferases, pseudouridine synthases and 2′-*O* methyltransferases. By iterative homology searches, we have identified ORFs in *E.coli*, *H.influenzae* and *M.genitalium* likely to encode enzymes with similar function. These ORFs, together with knowledge of RNA modifications in *E.coli*, allowed us to predict specific substrates for many of the ORF encoded proteins. In *E.coli*, eleven ORFs which are likely to code for RNA modifying enzymes were found in addition to the six previously characterized. These seventeen enzymes could account for most or all of the three m^5Us, seventeen Ψs and seven Nms present in *E.coli* rRNA and tRNA. We also assigned the direct genetic counterparts, or orthologs, of the ORFs found in *E.coli* to ORFs present in *H.influenzae* and, where applicable, *M.genitalium*.

The genomic search procedure described here exploited the following information: (i) knowledge of a set of related endproducts, the production of which requires a set of potentially related unknown enzymes, (ii) the amino acid sequence of one or more related enzymes to use as initial probes. This procedure should be generally applicable to other situations which meet similar criteria.

* To whom correspondence should be addressed

+Present address: Kosan Biosciences, Inc., 1450 Rollins Road, Burlingame, CA 94010, USA

MATERIAL AND METHODS

Amino acid sequences of proteins of known function (Table 1) were used as probes in searches for homologous sequences present in the GenBank database (National Center for Biotechnology Information), the SwissProt database (EBI EMBL) and The Institute of Genomic Research (TIGR) databases for *Haemophilus influenzae* Rd and *Mycoplasma genitalium.* The initial searches were carried out using the BLAST program (6) or the GRASTA program [a modification of FASTA (7)]. These programs are part of the network service provided by GenBank and TIGR respectively. ORFs having a probability of an accidental match of 10^{-4} or less to the respective probe were retrieved from the databases and further analyzed.

Table 1. Genes encoding known RNA modifying enzymes used as probes in the database searches

Organism	Gene	Activity[a]	Acc. number[b]
E.coli	*trmA*	tRNA U54→m^5U	P23003
E.coli	*truA*	tRNA U38, U39, U40→Ψ	P07649
E.coli	*truB*	tRNA U55→Ψ	P09171
E.coli	*rluA*	23S U746, tRNA U32→Ψ	P39219
E.coli	*rsuA*	16S U516→Ψ	P33918
E.coli	*spoU*	tRNA G18→Gm	P19396
St.azureus[c]	*tsr*	23S A1067→Am	P18644
Sa.cerevisiae[c]	*PET56*[d]	23S G2251→Gm	431760

[a]Numbering and designation of RNA according to *E.coli*.
[b]SwissProt accession number.
[c]*St.* denotes *Streptomyces* and *Sa.* denotes *Saccharomyces*.
[d]The PET56 protein methylates the yeast mitochondrial 23S equivalent.

Detailed analysis of the ORFs retrieved in the GRASTA and BLAST search was performed using sequence analysis programs that are part of the Wisconsin sequence analysis package (8). PILEUP was used to create multiple sequence alignments of related sequences. LINEUP was used as screen editor to edit the multiple sequence alignments, and PRETTY was used to display the sequence alignments. The level of similarity and identity between the probe used and the identified sequences were determined using the GAP program. The gap creation penalty was set to 3.0 and gap extension penalty was set to 0.1. Before the final sequence analysis, the compiled amino acid sequences were truncated uniformly at the C- and N-terminal ends to provide sequences of similar overall length as shown in the figures (Figs 2, 3 and 4). The PILEUP program was also used to generate and plot a graph drawn in unrooted tree format (dendrogram) which shows the clustering relationships used to create the alignments.

The multiple sequence alignments were used to identify conserved amino acid sequence motifs. ORFs having the conserved motifs were considered true homologs and were used as probes in an iterative manner, i.e. the identified ORFs were used to search the databases for additional homologs using the same criteria as with the initial probes. As more ORFs were added to the separate sequence alignments (one alignment for each family of enzymes), the identity of conserved sequence motifs was further improved. The identities of the motifs were further refined by aligning homologous sequences from organisms other than *E.coli*, *H.influenzae* and *M.genitalium* found in the GenBank

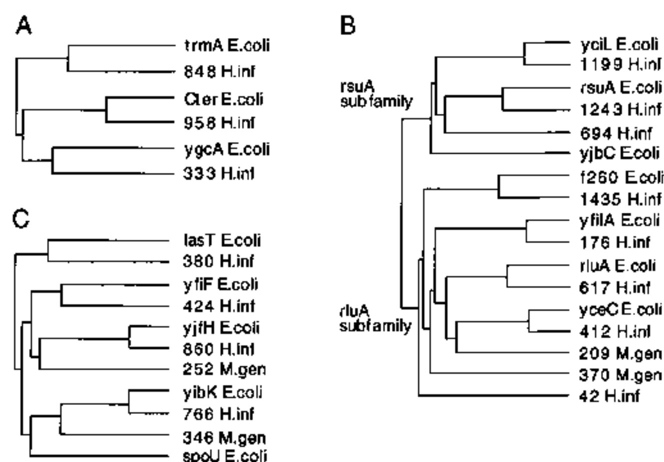

Figure 1. Dendrograms representing the relationship of gene homologs. Distance along the linear axis of the dendrograms is proportional to the difference between sequences. (**A**) Dendrogram representing the relationship of *trmA* homologs found in *E.coli* and *H.influenzae*. (**B**) Dendrogram representing the relationship of *rluA* and *rsuA* homologs found in *E.coli*, *H.influenzae* and *M.genitalium*. (**C**) Dendrogram representing the relationship of *spoU*, *tsr* and PET56 homologs found in *E.coli*, *H.influenzae* and *M.genitalium.*

database. These sequences, many of which are incomplete ORFs, have not been included in the alignment figures of this paper.

Escherichia coli and *H.influenzae* ORFs were considered as orthologous gene pairs when the ORFs paired together strongly in the dendrogram generated by PILEUP with branch points well separated from the next most distant branch points. In the comparisons presented here, the strongly paired ORFs had an identity score >30% and a similarity score >50% as determined by the GAP program. *E.coli* and *M.genitalium* ORFs were considered as orthologous gene pairs when the ORFs paired together strongly in the dendrogram and had an identity score >20% and a similarity score >40%. In this context identity means having the same amino acid at the same position whereas similarity is having a similar amino acid as defined by the GAP program at the same position in the two ORFs. The homology and similarity analyses were based on the truncated ORFs as noted above.

All rRNA designations and nucleotide numbering reflect the *E.coli* equivalent rRNA and nucleotide position respectively.

RESULTS AND DISCUSSION

We have searched genomic databases in an iterative manner using as initial probes the deduced amino acid sequences of eight genes encoding RNA modifying enzymes (Table 1). First we searched the *H.influenzae* and *M.genitalium* databases for homologous ORFs and retrieved sequences having a probability of an accidental match of 10^{-4} or less to the probe. The retrieved ORFs were aligned with the probe(s), analyzed for the presence of conserved motifs and subsequently used again as probes against the GenBank database to find additional homologous ORFs. This was repeated until no more homologous ORFs could be identified. ORFs were considered as orthologous gene pairs (i.e. encoding 'the same protein' in different organisms) when the ORFs (i) paired together in a dendrogram, (ii) shared conserved motifs, and (iii) showed homology as defined in the Materials and Methods section.

m^{5}U Methyltransferases

The single m^{5}U methyltransferase characterized to date (the TrmA protein encoded by the *trmA* gene) (9) methylates U54 of all *E.coli* tRNAs. We used the deduced amino acid sequence of *trmA* (10) to identify other ORFs that are likely to encode additional m^{5}U methyltransferases. A homology search of the *H.influenzae* and *M.genitalium* genome databases employing the GRASTA program identified three ORFs in the *H.influenzae* genome as *trmA* homologs with probabilities of an accidental match of 10^{-7} or less (Table 2). The minimal genome of *M.genitalium* did not encode any *trmA* homologs; notably, *Mycobacteria* sp. are the only eubacterial organisms which do not have m^{5}U54 in the tRNA (11). One ORF of *H.influenzae*, HI848, was identified as ortholog of *trmA*. The other *trmA* homologs identified in *H.influenzae* were HI333 and HI958. Reverse searching of the GenBank database using HI333 and HI958 as probes identified apparent orthologs for each in *E.coli*. The ortholog of HI333 is the *E.coli* ORF *ygcA*, and the ortholog of HI958 is a C-terminal portion of an *E.coli* ORF which we designated *Cter*. The presence of an AdoMet binding site in *ygcA* has previously been noted (4). The three pairs of orthologs from *E.coli* and *H.influenzae* within the m^{5}U methyltransferase family are displayed as a dendrogram in Figure 1A.

Table 2. RNA m^{5}U methyltransferases

E.coli		*H.influenzae*				
Gene	Accession number	Gene[a]	Identity[b] (%)	Similarity[b] (%)	Function	Activity
trmA (probe)	P23003	HI848	64	78	known	tRNA U54→m^{5}U
ygcA	U29580	HI333	47	66	predicted	23S U747→m^{5}U or 23S U1939→m^{5}U
Cter[c]	X69108	HI958	(52)	(74)	predicted	23S U747→m^{5}U or 23S U1939→m^{5}U

[a]The *H.influenzae* ORFs are denoted according to Fleischmann *et al*. (1).

[b]The percentages of identity and similarity of the corresponding amino acid sequence to the *E.coli* ortholog are calculated using the GAP program from the GCG package, which aligns sequences using the Needleman–Wunsch algorithm. The ORFs were truncated at the N- and C-terminal ends as shown in Figure 2 before determining the identity and similarity.

[c]The 5′ half of the DNA sequence encoding *Cter* is not currently available, therefore only the C-terminal was aligned. The identity and similarity values thus only reflect the C-terminal part of the alignment.

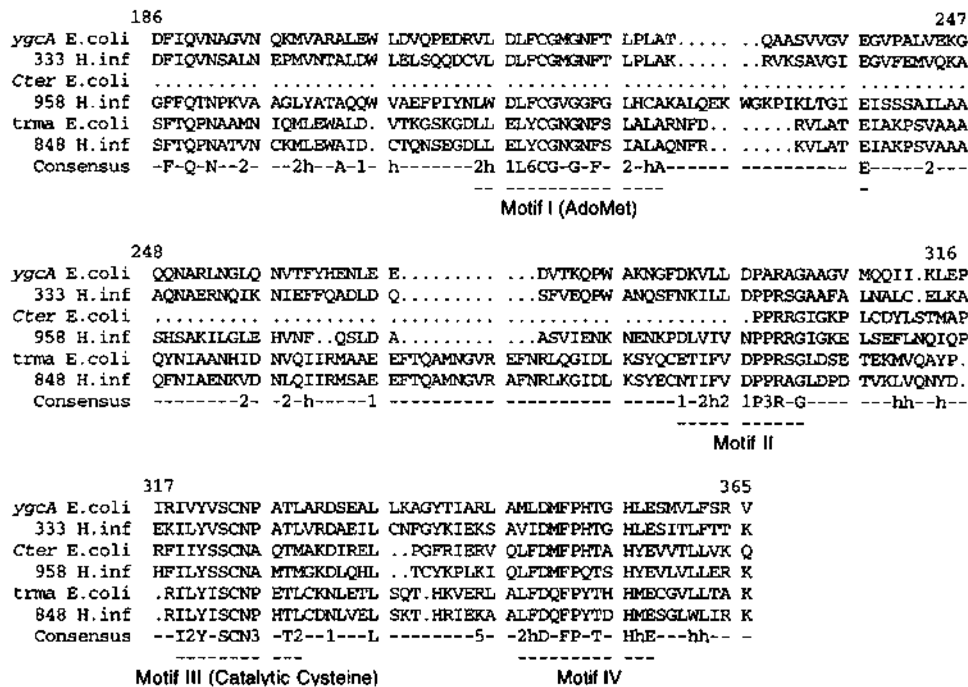

Figure 2. Alignment of *trmA* homologs found in *E.coli* and *H.influenzae*. The 5′ half of the DNA sequence encoding *Cter* is not currently available, therefore only the C-terminal part was used in the alignment. Otherwise, the sequences shown have been truncated at the N- and C-terminal ends to produce ORFs of uniform length. These are the truncated sequences used in the sequences analysis. Residues were considered as consensus only when present in five of the five sequences. The numbering of residues reflects the *trmA* sequence. Consensus abbreviations: h, hydrophobic; 1, EDQN; 2, VLIM; 3, GPA; 4, ST; 5, KR; 6, FYW.

```
                 19                                                                          76
rluA E.coli  QDDHIMVV..  NKPSGLLSV  PGRLEEHKD.  .SVMTRIQR.  .DY...PQ..  .AESVHRLDM  ATSGVIVVAL
 617 H.inf   QDNHLCVV..  NKPSGLLSV  PGNQPQYYD.  .SAMSRVKE.  .KF...GF..  .CEPAHRLDM  ATSGIIVFAL
yceC E.coli  EDDHILVL..  NKPSG.TAV  HGGSGLSFGV  IEGLRALRP.  .EA...RF..  .LELVHRLDR  DTSGVLLVAK
 412 H.inf   EDDCLIIL..  NKPSG.IAV  HGGSGLNFGV  IEALRALRP.  .EA...RF..  .LELVHRLDR  DTSGILLIAK
yfiI E.coli  EDEDIIII..  NKPRDLVVH  PGAGNPDGTV  LNALLHYYP.  .PIADVPR..  .AGIVHRLDK  DTTGLMVVAK
 176 H.inf   EDDDIIVI..  NKPKDLVVH  PGAGNPNGTV  LNALLYHYP.  .PIVEVPR..  .AGIVHRLDK  DTTGLMVVAK
 209 M.gen   EDKDLMVI..  NKPSGLLTH  PTTFNEKASL  LAACIFH...  ...NNKNP..  .VYLVHRLDR  DTSGAIVVCK
f260 E.coli  QDEWLVAV..  NKPSGWL.V  HRSWLDRDEK  VVVMQTVRD.  .QIG..QH..  .VFTAHRLDR  PTSGVLLMGL
1435 H.inf   QDGFLVAV..  NKPAGML.V  HRSWLDPHET  QFVMQTLRD.  .QIG..QH..  .VFPIHRLDR  PTSGVLLFAL
 370 M.gen   EDENIIVV..  DKPTGIVCQ  PDKKHSIINL  SNMLLKHCGY  RQFDSNKLNF  YPQFAHRIDR  NTSGIVIGAK
  42 H.inf   QHSDFIII..  NKSEGISVH  KDQEEQ....  ..GLTELVA.  .KQLNVPK..  .VWLVHRLDK  VTSGLLILAL

rsuA E.coli  HGPRYFML..  NKPQGYVCS  T.DDPDHPTV  LYFLDE....  ....PVAW..  KLHAAGRLDI  DTTGLVLMTD
1243 H.inf   EEGQYFML..  NKPQGCVCS  N.DDGDYPTI  YQFFDY....  ....PLAG..  KLHSAGRLDV  DTTGLVLLTD
yciL E.coli  AEQICRVLAY  YKPEGELCT  RNDPEGRPTV  FDRLPK....  .....LRGAR  WI.AVGRLDV  NTCGLLLFTT
1199 H.inf   QKEICRVLMY  YKPEGELCT  RSDPEGRATV  FDRLPR....  .....LTGSR  WI.AVGRLDI  NTSGLLLFTT
yjbC E.coli  EDLVLIAL..  NKPVGIVST  TEDG.ERDNI  VDF.......  .....VNHSK  RVFPIGRLDK  DSQGLIPLTN
 694 H.inf   KVVLF.....  NKPFDVLTQ  FTDEQGRATL  KDFISI....  ....P.....  NVYAAGRLDR  DSEGLLILTN
 Consensus   ----h--2--  1KP---2--  ----------  ---h------  ----------  ------RLD-  -T-G222h--
                         --------                                                ---- --------
                         Motif I                                                   Motif II

                 77                                                                          140
rluA E.coli  TKAAERELKR  QFREREPKKQ  YVARVWGHPS  PAEGLVDLPL  I......CDW  PNRPKQKVCY  ETGKPAQTEY
 617 H.inf   SKAADRELKR  QFREREPKKH  YQAIVWGHLE  NDYGEVNLPM  I......CDW  ENRPRQRLDF  VLGKRAVTKF
yceC E.coli  KRSALRSLHE  QLREKGMQKD  YLALVRGQWQ  SHVKSVQAPL  LKNILQSGER  IVRVSQE...  ..GKPSETRF
 412 H.inf   KRSALRNLHE  QLRVKTVQKD  YLALVRGQWQ  SHIKVIQASL  LKNELSSGER  IVRVSEQ...  ..GKPSETRF
yfiI E.coli  TVPAQTRLVE  SLQRREITRE  YEAVA.IGHM  TAGGTVDEPI  SRHPTKRTH.  .MAVHPM...  ..GKPAVTHY
 176 H.inf   TIPAQTKLVR  DLQKRKITRE  YEAVA.SGIM  TKGGTVDQPM  ARHATKRTL.  .MAVHPM...  ..GKPAVTHY
 209 M.gen   NQASLLNLQN  QLQNRTLKRY  YVALVHFPFN  ALTGSINAPL  ARVNNNKVMF  KIAQTAK...  ..AKQAITKF
f260 E.coli  SSEAGRLLAQ  QFEQHQIQKR  YHAIVRGWLM  EE.AVLDYPL  VEELDKIADK  FAREDKG...  ..PQPAVTHY
1435 H.inf   SSEIANLMCE  QFEQKYVQKS  YLAVVRGYLQ  GK.ERIDYPL  KIQLDKIADK  FSQEDKE...  ..PQEAITDY
 370 M.gen   TNKALKELNK  VFKNNHLTKR  YKGLVFGQFN  HLG..LQTAY  WKKDNNNG..  .IVTVKWKPF  PEAKKISTFF
  42 H.inf   LFSEHKIHKT  YLALSNQKPK  NAESAAEFSR  KKQGLIIGDM  KK..AREGAW  KLCQTK....  ..DNPAITRF

rsuA E.coli  DGQWSHRITS  ..PRHHCEKT  YLVTLESPVA  DDTAEQFAKG  VQLHNEKDL.  ..........  .TKPAVLEVI
1243 H.inf   DGQWSHRITS  ..PKHHCEKT  YLVTLADPVE  ENYSAACAEG  ILLRGEKEP.  ..........  .TKPAKLEIL
yciL E.coli  DGELANRLMH  ..PSREVERE  YAVRVFGQVD  DAKLRDLSRG  VQLEDGPAA.  ..........  .FKTIK..FS
1199 H.inf   DGELANRLMH  ..PSREVERE  YSVRVFGQVD  DAMLARLRKG  VQLEDGLAN.  ..........  .FKEIK..FT
yjbC E.coli  HGDLVNKILR  ..AGNDHEKE  YLVTVDKPIT  EEFIRGMSAG  VPILGTV...  ..........  .TKKCKVKKE
 694 H.inf   NGELQHRLAD  ..PKFKTEKT  YWVQVEGIPE  ETDLAQLRKG  VELKDGV...  ..........  .TKSAKVRLI
 Consensus   -------2--  --------5-  Y---------  ----------  ----------  ----------  ----------

                141                                                      187
rluA E.coli  E....VVEYA  ADNT......  ARVVLKPITG  RSHQLRVHML  ALGHPI...L  GDRFYASPEA
 617 H.inf   E....VLARL  PNNS......  TRVKLTPVTG  RSHQLRLHML  ALGHPI...L  GDKFYSHPQA
yceC E.coli  K....VEERY  A.FA......  TLVRCSPVTG  RTHQIRVHTQ  YAGHPI..AF  DDRYGDREFD
 412 H.inf   S....IEERY  I.NA......  TLVKASPVTG  RTHQIRVHTQ  YAGHPI..AL  DDKYGDKDFD
yfiI E.coli  RIMEHFRV..  .........H  TRLRLRLETG  RTHQIRVHMA  HITHPL..VG  DPVYGGRPRP
 176 H.inf   RIMENYRN..  .........Y  TRLRLRLETG  RTHQIRVHMA  HIAHPL..LG  DQTYGGRPRP
 209 M.gen   KVINQNEK..  .........A  ALISLELLTG  RTHQIRVHLK  FIQHPV..YN  DPLYGIKSEK
f260 E.coli  RGLATVEMPV  ATGRYPTTRY  GLVELEPKTG  RKHQLRRHLA  HLRHPI..IG  DSKHGDLRQN
1435 H.inf   EGLKIVEMPY  PAGRYQTARY  SLVKLIPHTG  QKTP......  ..........
 370 M.gen   ENSSYIAQK.  ........DM  SLITIRLISG  RTHQIRACLN  LFSNQL..VG  DKKYSLIQFK
  42 H.inf   ESVSCEPN..  .........L  RLFILKPQTG  KTHQLRVAMK  SLGSPI..LG  DGLYGKNTEK

rsuA E.coli  TPTQ......  ..........  ..VRLTISEG  RYHQVKRMFA  AVGNHVVELH  RERIGGITLD
1243 H.inf   DDYN......  ..........  ..VNLTISEG  RYHQVKRMFA  ALGNKVVGLH  RWKIGDVVLD
yciL E.coli  GGEG......  .....INQ..  .WYNVTLTEG  RNREVRRLWE  AVGVQVSRLI  RVRYGDIPLP
1199 H.inf   GGVG......  .....INQ..  .WYDVTLMEG  RNREVRRLWE  SQGIQVSRLI  RIRYGNIKLM
yjbC E.coli  APF.......  ..........  .VFRITLVQG  LNRQIRRMCE  HFGYEVKKLE  RTRIMNVSLS
 694 H.inf   SEPNLWERNP  PIRERKNIPT  SWLEIKISEG  RNRQVRRMTA  HIGFPTLRLV  RVSMGLLSI.
 Consensus   ----------  ----------  --h-2----G  5--12R----  -----2----  ---h------
                                             - ------
                                             Motif III
```

Figure 3. Alignment of *rluA* and *rsuA* homologs found in *E.coli*, *H.influenzae* and *M.genitalium*. The C-terminal part of ORF HI1435 has been deleted due to poor nucleotide sequence in the corresponding database entry. Otherwise, the sequences shown have been truncated at the N- and C-terminal ends to produce ORFs of uniform length. These are the truncated sequences used in the sequences analysis. The *rluA* and the *rsuA* subfamilies are found on the upper and lower half of the alignment, respectively. Residues were considered as consensus only when present in 15 of the 17 sequences. The numbering of residues reflects the *rluA* sequence. Consensus abbreviations as above.

The alignment of the deduced amino acid sequence of *trmA* and its five homologs showed four conserved sequence motifs (Fig. 2). Motif I, 2-h-1-L-6-C-G-x-G-x-F-x-2-x-h-A-x_{10}-E, (abbreviations used: h, hydrophobic; 1, EDQN; 2, VLIM; 3, GPA; 4, ST; 5, KR; 6, FYW; x, any residue) shares considerable homology to the established consensus of the *S*-adenosyl-L-methionine (AdoMet) binding motif identified in other methylases that use this cofactor (12). This motif contributes directly to the binding pocket for AdoMet in the three dimensional structure of DNA m^5C methyltransferase *Hha*I (13). Conserved motif III (I-2-Y-x-

```
                    31                                                                             95
yibK E.coli  NIVLYEPEIP PNTGNIIRLC ANTGFRLHII EPMGFAWDDK RLR...RAGL DYHEFTAVTR HHDYRAFLEA
 766 H.inf   DIVLYEPEIP QNTGNIIRLC ANTGFRLHLI EPLGFTWDDK RLR...RSGL DYHEFAEIKR HKTFEAFLES
 346 M.gen   NIVLFCPEIP NNTGNIVRSC TAFKANLHLI KPYGFFLNDK RMV...RAGL NCWDKIQLFE HKSWEHFLQA
spoU E.coli  TVCMEQVHKP HNVSAIIRTA DAVG..VHEV HAV...WPGS RMRTMASAAA GSNSWVQVKT HRTIGDAVAH
 380 H.inf   RIVLIETSHS GNIGSAARAM KTMGLTQLCL VSPK.SVDEQ SYALSA.GAE NIVKNARVV. .DSFDEAVDD
lasT E.coli  TIILVAPARA ENIGAAARAM KTMGFSDLRI VDSQ.AHLEP ATRWVAHGSG DIIDNIKVF. .PTLAESLHD
 424 H.inf   LVLLDNVNNA QNIGGVLRTC AYFGVKNIVA ..DNVENLYS AASM..RVAE GGAEYIRVLE ADYIDSALMQ
yfiF E.coli  VLALENESNP HNLGGMMRSC AHFGVKGVVV ..QDAALLES GAAI..RTAE GGAEHVQPIT GDNIVNVLDD
 860 H.inf   LLVLDGVTDP HNLGACLRTA DAAGAVAVIV PKDKSAQLTS IAR...KVAC GAAETVPLIR VTNLSRTLRD
yjfH E.coli  LLILDGVTDP HNLGACLRSA DAAGVHAVIV PKDRSAQLNA TAK...KVAC GAAESVPLIR VTNLARTMRM
 252 M.gen   .VMLDEIQDP YNFGAILRTC LASEVDGIIF KKNNQVPINN TVM...KTSM GSVFYQNLVQ VANLSYTITK
Consensus    -2hL-----3 -N-G---R-- ---------- ---------- ---------- -------2-- ---h---2--
                      - --------
                        motif I

                    96                                                                            148
yibK E.coli  .ENPQRLFAL TTKGTPAHSA .......... .......VSY QDGDYLMFGP ETRGLPASIL DALPAEQKIR
 766 H.inf   .EKPKRLFAL TTKGCPAHSQ .......... .......VKF KLGDYLMFGP ETRGIPMSIL NEMPMEQKIR
 346 M.gen   TTENKTIWLL TKSGDKTPDQ .......... ...ICMTNKL PNELYFVFGQ ETKGLPKTIM DN.FKQNQIR
spoU E.coli  LKGQGMQILA THLSDNA... .......... ..VDFREIDY TRPTCILMGQ EKTGITQEAL A..LADQDII
 380 H.inf   CP......LV IGTSARLRHL QNTLIEP... RECAEKVVAY KGKIAIVFGR ERIGLTNEEL L..KCHYHLN
lasT E.coli  VD......FT VATTARSRAK YHYYATPVEL VPLLEEKSSW MSHAALVFGR EDSGLTNEEL A..LADVLTG
 424 H.inf   LRKSG..YQI IHVSHNKQGD PLDK...... .......VRL KNKVVFVLS. ESSTESLATP EDTQARLTLA
yfiF E.coli  FRQAG..YTV VTTS.SEQGK PLFK...... .......TSL PAKMVLVLGQ EYEGLPDAAR D..PNDLRVK
 860 H.inf   LQQNHNIW.V VGTAGEATET IYQ....... .......SKL TGPLALVMGA EGEGMRRLTR E..HCDQLIS
yjfH E.coli  LQEE.NIW.I VGTAGEADHT LYQ....... .......SKM TGRLALVMGA EGEGMRRLTR E..HCDELIS
 252 M.gen   LKEI.GFWTV VSTLDPIWKP IDY....... ......RKVD FAKKILIVGN EDRGVNQLIT K..NADCRIK
Consensus    ---------- ---------- ---------- ---------h -----h2hG- E--G2----- --------2-
                                                            ----- -----
                                                               motif II

                   149                  168
yibK E.coli  IPMVPDSRSM NLSNAVSVVV
 766 H.inf   IPMTANSRSM NLSNSVAVTV
 346 M.gen   IPIWNSVRSI NLANAVVCIL
spoU E.coli  IPMIGMVQSL NVSVASALIL
 380 H.inf   IPANPDYSSL NLAMAVQLVS
lasT E.coli  VPMVADYPSL NLGQAVMVYC
 424 H.inf   SPI..K.SGL NIAVNAGVLL
yfiF E.coli  IDGTGNVAGL NISVATGVLL
 860 H.inf   IPMAGSVSSL NVSVATGVCL
yjfH E.coli  IPMAGSVSSL NVSVATGICL
 252 M.gen   IPMNNKINSL NVSVALGIIL
Consensus    2P------S2 N2-----2hh
             ---------- --
                motif III
```

Figure 4. Alignment of *spoU*, *tsr* and PET56 homologs found in *E.coli*, *H.influenzae* and *M.genitalium*. The sequences shown have been truncated at the N- and C-terminal ends to produce ORFs of uniform length. These are the truncated sequences used in the sequences analysis. Residues were considered as consensus only when present in 10 of the 11 sequences. The numbering of residues reflects the *spoU* sequence. Consensus abbreviations as above.

S-C-N-3-x-T-2) contains the catalytic cysteine residue of the TrmA protein which forms a covalent adduct to C6 of U during catalysis (14). Two additional conserved motifs, II and IV, were found. Motif II (1-x-2-h-2-1-P-3-R-x-G) is located directly upstream from the catalytic cysteine motif, and motif IV (2-h-D-x-F-P-x-T-x-H-h-E) reflects the extensive amino acid sequence similarity at the C-terminal. The presence of conserved motifs I–IV among the *trmA* homologs supports the prediction that the ORFs identified encode m^5U methyltransferases.

Other than the TrmA-dependent m^5U54 in tRNA, the only other m^5Us that have been found in *E.coli* RNA are m^5U747 and m^5U1939 in 23S rRNA. The methyltransferases that catalyze these latter two modifications have not been identified. Since we have identified only two other ORFs in *E.coli* that code for m^5U methyltransferases, we believe that the *E.coli* ORFs *ygcA* and *Cter* encode the two putative 23S rRNA m^5U-methyltransferases. Although the presence of these modifications have not been established in *H.influenzae*, we further believe that HI333 and HI958 gene products catalyze the corresponding modifications in *H.influenzae*.

Ψ Synthases

Four *E.coli* Ψ synthases have been identified (Table 1). The product of the *truA* gene, TruA (also known as HisT or Ψ synthase I), converts U residues to Ψ in the anticodon arm of some tRNAs (15). The *truB* gene product, TruB, forms Ψ at U55 in the T-arm of all *E.coli* tRNAs (16). The product of the *rsuA* gene, RsuA, introduces the only Ψ found in 16S rRNA (17). The product of the *rluA* gene, RluA, has two enzymatic activities; it catalyzes Ψ formation at U746 in domain II of 23S rRNA and also catalyzes Ψ formation at U32 in some tRNAs (18).

The amino acid sequences of the four known Ψ synthase genes were used as probes for iterative searching of the genome sequences of *E.coli*, *H.influenzae* and *M.genitalium*. The search using the *truA* and *truB* probes identified an ortholog for each in *H.influenzae*, and an ortholog for *truA*, but not *truB*, in *M.genitalium*; the search did not identify any non-orthologous homologous ORFs. Searches using the *rluA* and *rsuA* probes yielded two families of homologs, one to each probe. A distant, but distinct homology exists between the *rluA* and *rsuA* families (Figs 1B and 3).

Table 3. RNA Ψ synthases

E.coli		*H.influenzae*			*M.genitalium*				
Gene	Acc. no.	Gene[a]	Identity[b] (%)	Similarity[b] (%)	Gene[c]	Identity[b] (%)	Similarity[b] (%)	Function	Activity
rsuA (probe)	P33918	HI1243	58	74	–			known	16S U516→Ψ
yciL	P37765	HI1199	72	82	–			predicted	tRNA or 23S U→Ψ
yjbC[d]	P32684	–			–			predicted	tRNA or 23S U→Ψ
		HI694						predicted	tRNA or 23S U→Ψ
rluA (probe)	P39219	HI617	62	75	–			known	23S U746, tRNA U32→Ψ
yfiI	P33643	HI176	72	85	MG370[d]	(26)	(51)	predicted	23S U2580 or 23S U(s) domain IV →Ψ(s)
yceC	P23851	HI412	72	85	MG209[d]	(30)	(55)	predicted	23S U2580 or 23S U(s) domain IV →Ψ(s)
f260	CO05945	HI1435	59	70	–			predicted	tRNA or 23S U→Ψ
		HI42			–			predicted	tRNA or 23S U→Ψ

[a]The *H.influenzae* ORFs are denoted according to Fleischmann *et al.* (1).
[b]The percentages of identity and similarity of the corresponding amino acid sequence to the *E.coli* ortholog are calculated using the GAP program from the GCG package, which aligns sequences using the Needleman–Wunsch algorithm. The ORFs were truncated at the N- and C-terminal ends as shown in Figure 2 before determining the identity and similarity.
[c]The *M.genitalium* ORFs are denoted according to (2).
[d]The assignment of ORFs MG209 and MG370 as an orthologs of *yceC* and *yfiI* is speculative (see text).

Table 4. 2′-*O* methyltransferases

E.coli		*H.influenzae*			*M.genitalium*				
Gene	Acc. no.	Gene[a]	Identity[b] (%)	Similarity[b] (%)	Gene[c]	Identity[b] (%)	Similarity[b] (%)	Function	Activity
spoU (probe)	P19396	–			–			known	tRNA G18→Gm
yfiF	P33635	HI424	38	60	–			predicted	tRNA, 16S or 23S N→Nm
yjfH	P39290	HI860	72	85	MG252	27	47	predicted	23S G2251→Gm
yibK	P33899	HI766	76	83	MG346	39	60	predicted	23S U2552→Um
lasT	P37005	HI380	32	52	–			predicted	tRNA, 16S or 23S N→Nm

[a]The *H.influenzae* ORFs are denoted according to Fleischmann *et al.* (1).
[b]The percentages of identity and similarity of the corresponding amino acid sequence to the *E.coli* ortholog are calculated using the GAP program from the GCG package, which aligns sequences using the Needleman–Wunsch algorithm. The ORFs were truncated at the N- and C-terminal ends as shown in Figure 2 before determining the identity and similarity.
[c]The *M.genitalium* ORFs are denoted according to Fraser *et al.* (2).

The *rluA* subfamily has five homologs in the completely sequenced *H.influenzae* genome, four of which have orthologs in *E.coli*. Although an *E.coli* ortholog of the fifth *rluA* homolog found in *H.influenzae* has not been found, it may exist in the 26% of the *E.coli* chromosome that remains to be sequenced. Two *rluA* homologs were identified in *M.genitalium*; MG209 and MG370. The ORF MG209 is the more conserved of these two and is slightly more related to the *E.coli* ORF *yceC* than to the other *rluA* homologous ORFs (Figs 1B and 3; Table 3); however the sequence conservation is not strong enough to assign clear orthology.

The *rsuA* sub-family has three homologs in *E.coli*, three in *H.influenzae* and none in *M.genitalium*. Two of the *E.coli* ORFs, *yciL* and *rsuA*, have apparent orthologs in *H.influenzae*. The remaining *rsuA* homolog in *E.coli* (*yjbC*) and in *H.influenzae* (HI42) are not orthologous to each other and have no apparent orthologs in other sequences examined here (Figs 1B and 3; Table 3). The similarity of *yceC*, *yfiI* and *yjbC* to *rluA* and *rsuA* has previously been noted (4).

An alignment was made of the amino acid sequences of the six *rsuA* homologs from *E.coli* and *H.influenzae* and the eleven *rluA* homologs from *E.coli*, *H.influenzae* and *M.genitalium*. Upon aligning the two subgroups of Ψ synthases, three conserved sequence motifs (motif I: 1-K-P-x_3-2, motif II: R-L-D-x_2-T-x-G-2-2-2-h and motif III: G-5-x_2-1-2-R) were found in both sets of Ψ synthases (Fig. 3).

A total of 17 Ψs are known to be present in *E.coli* RNA. Mature tRNA has seven Ψ nucleotides. Three enzymes—RluA, TruA and TruB—which catalyze five of the seven tRNA modifications have been characterized (Table 1). The enzymes catalyzing the two remaining modifications in tRNA have not been identified. 16S RNA has a single Ψ at nucleotide 516 which is formed by RsuA (17). 23S rRNA has nine Ψs which, with one exception (Ψ955) are located at the peptidyl transferase center (19). RluA catalyzes formation of Ψ746, but enzymes for the eight remaining Ψs in 23S rRNA (Ψ955, Ψ1911, m^3Ψ1915, Ψ1917, Ψ2457, Ψ2504, Ψ2580 and Ψ2605) have not been identified. Thus, there are a total of ten Ψ nucleotides for which the modifying enzyme have not been identified. We have identified five new ORFs in *E.coli* predicted to encode RNA Ψ synthases. The difference in numbers—ten Ψ nucleotides versus five Ψ synthases—may be explained by (a) enzymes having multiple substrates as with TruA and RluA, (b) *E.coli* genes not yet sequenced, or (c) genes not related to the major *rsuA*/*rluA* branch of Ψ synthases as is the case for *truA* and *truB*.

We assume that ortholog pairs of enzymes from *E.coli* and *H.influenzae* having the highest homology as well as having close homologs present in *M.genitalium*, will catalyze those modified nucleotides present at the most conserved locations. The most conserved Ψ nucleotides found in ribosomal RNA are Ψ2580 located in domain V of 23S rRNA, and two Ψs clustered in domain IV (Ψ1915 and Ψ1917) (20,21). Thus, we predict that one of the two *E.coli* ORFs with the most conserved orthologs (*yceC* and *yfiI*) encodes the enzyme that catalyzes formation of Ψ2580 and that the other Ψ synthase encodes the enzyme that catalyzes formation of one or both of the conserved Ψ1915 and Ψ1917 in domain IV (Table 3). It is noteworthy that TruA can modify up to three closely spaced Us in the anticodon arm of some tRNAs, thus providing precedent for multiple Ψ modifications at closely spaced positions by a single enzyme.

2′-*O* Methyltransferases

The sequences of three genes encoding 2′-*O* methyltransferases are available: *spoU* in *E.coli*, encoding the tRNA Gm18 methyltransferase (C. Gustafsson, unpublished), *tsr* in *Streptomyces azureus* encoding the thiostreptone resistance marker 23S rRNA Am1067 methyltransferase (22), and PET56 in yeast, encoding the 23S rRNA Gm2251 methyltransferase (23) (Table 1). The amino acid sequences of these three gene products were used as probes for iterative searching of the genome sequences of *E.coli*, *H.influenzae* and *M.genitalium* to identify other ORFs encoding enzymes that catalyze methylation of the 2′ hydroxyl of the ribose in RNA.

Ten previously uncharacterized ORFs were found in the search which, after assignment of orthologous pairs, corresponded to four new 2′-*O* methyltransferases. The four previously uncharacterized ORFs in *E.coli* all had orthologs present in *H.influenzae* and two of which (*yjfH* and *yibK*) also had orthologs present in *M.genitalium*. The two ORFs in *M.genitalium* were orthologs of the two most homologous gene pairs in *E.coli* and *H.influenzae*. The probe *spoU* did not have an ortholog in either *H.influenzae* or *M.genitalium* (Table 4; Fig. 1C).

An alignment made of the three known and four newly identified 2′-*O* methyltransferases revealed three motifs found in all of the ORFs (Fig. 4). One of the motifs, motif II (h-2-h-G-x-E-x_2-G-2), consists of a series of bulky aliphatic amino acid residues followed by two conserved glycines, resembling an AdoMet binding motif (12). Two additional conserved motifs were found, motif I (3-x-N-x-G-x_3-R) located at the N-terminus of the sequences and motif III (2-P-x_6-S-2-N-2) located at the C-terminus (Fig. 4).

A total of seven 2′-*O* modified nucleotides have been found in *E.coli* RNA. One is in 16S rRNA (m^4Cm1402) and three are in 23S rRNA (Gm2251, Cm2498 and Um2552). There are also three 2′-*O* modified nucleotides in tRNA; however, two of the three, Um32 and Cm32, are both pyrimidine nucleotides and occur at the same position in different tRNAs, and are likely to be catalyzed by the same enzyme. Thus, we propose that there are six 2′-*O* methyltransferases in *E.coli* which catalyze the seven RNA modifications. One, *spoU*, has been previously identified and we have here identified four previously uncharacterized ORFs as putative 2′-*O* methyltransferases. The remaining ORF may be (a) in the part of the *E.coli* genome not sequenced yet, (b) attributed to one enzyme having multiple target substrates, or (c) part of another 2′-*O* methyltransferase family. The similarity of *lasT*, *yibK* and *yfiF* to *spoU* has previously been noted (4).

Since the two most highly conserved putative 2′-*O* methyltransferases, *yibK* and *yjfH*, are the only ORFs within this family present in *M.genitalium*, we suggest they encode the enzymes catalyzing the 2′-*O* methylations 23S Gm2251 and Um2552, which are the only modified nucleotides found in all organisms so far analyzed (20). Since the *yibK* ortholog set is phylogenetically more closely related to the guanosine methylase *spoU* (Fig. 1C), it probably encodes a guanosine methylase. Thus, we propose that *yibK* encodes the 23S rRNA Gm2251 methyltransferase and *yjfH* consequently encodes the 23S rRNA Um2552 methyltransferase.

We are currently experimentally testing the functional predictions described in this paper. So far, we have cloned and expressed three of the *E.coli* ORFs described above; *ygcA, yceC* and *yfiF.* Although the specific bases modified have not yet been identified, we have determined that each of the three enzymes encoded by these ORFs does indeed catalyze the formation of the predicted RNA modifications.

ACKNOWLEDGEMENT

This work was supported by grant GM-51232 from the National Institute of Health.

REFERENCES

1 Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) *Science*, **269**, 496–512.
2 Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M. *et al.* (1995) *Science*, **270**, 397–403.
3 Wahl,R., Rice,P., Rice,C.M. and Kröger,M. (1994) *Nucleic Acids Res.*, **22**, 3450–3455.
4 Koonin,E.V., Tatusov,R.L. and Rudd,K.E. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 11921–11925
5 Björk,G.R. (1996) In Neidhardt,F.C. (ed.), *Escherichia coli and Salmonella. Cellular and Molecular Biology.* American Society for Microbiology, Washington, DC, pp. 861–886.
6 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
7 Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
8 Genetics Computer Group, (1994). Version **8**, Madison, Wisconsin, USA.
9 Björk,G.R. and Isaksson,L.A. (1970) *J. Mol. Biol.*, **51**, 83–100.
10 Gustafsson,C., Lindström,P.H., Hagervall,T.G., Esberg,K.B. and Björk,G.R. (1991) *J. Bacteriol.*, **173**, 1757–1764.
11 Sprinzl,M., Dank,N., Nock,S. and Schön,A. (1991) *Nucleic Acids Res.* **19**, 2127–2171.
12 Kagan,R.M. and Clarke,S. (1994) *Arch. Biochem. Biophys.*, **310**, 417–427.
13 Cheng,X., Kumar,S., Posfai,J., Pflugrath,J.W. and Roberts,R.J. (1993) *Cell*, **74**, 299–307.
14 Kealey,J.T. and Santi,D.V. (1991) *Biochemistry*, **30**, 9724–9728.
15 Bruni,C.B., Colantuoni,V., Sbordone,L., Cortese,R. and Blasi,F. (1977) *J. Bacteriol.*, **130**, 4–10.
16 Nurse,K., Wrzesinski,J., Bakin,A., Lane,B.G. and Ofengand,J. (1995) *RNA*, **1**, 102–112.
17 Wrzesinski,J., Bakin,A., Nurse,K., Lane,B.G. and Ofengand,J. (1995) *Biochemistry*, **34**, 8904–8913.
18 Wrzesinski,J., Nurse,K., Bakin,A., Lane,B.G. and Ofengand,J. (1995) *RNA*, **1**, 437–448.
19 Brimacombe,R., Mitchell,P., Osswald,M., Stade,K. and Bochkariov,D. (1993) *FASEB J.*, **7**, 161–166.
20 Sirum-Connolly,K., Peltier,J.M., Crain,P.F., McCloskey,J.A. and Mason,T.L. (1995) *Biochimie*, **77**, 30–39.
21 Bakin,A., Lane,B.G. and Ofengand,J. (1994) *Biochemistry*, **33**, 13475–13483.
22 Bibb,M.J., Bibb,M.J., Ward,J.M. and Cohen,S.N. (1985) *Mol. Gen. Genet.*, **199**, 26–36.
23 Sirum-Connolly,K. and Mason,T.L. (1993) *Science*, **262**, 1886–1889.