4264–4274 *Nucleic Acids Research, 2007, Vol. 35, No. 13* *Published online 18 June 2007*
doi:10.1093/nar/gkm411

Functional specialization of domains tandemly duplicated within 16S rRNA methyltransferase RsmC

S. Sunita[1], Elzbieta Purta[2,3], Malgorzata Durawa[2], Karolina L. Tkaczuk[2,4], J. Swaathi[1], Janusz M. Bujnicki[2] and J. Sivaraman[1,*]

[1]Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543, [2]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Trojdena 4, 02-109 Warsaw, [3]Institute of Biochemistry and Biophysics PAS, Pawinskiego 5a, 02-106 Warsaw and [4]Institute of Technical Biochemistry, Technical University of Lodz, B. Stefanowskiego 4/10, 90-924 Lodz, Poland

Received February 14, 2007; Revised May 3, 2007; Accepted May 5, 2007

ABSTRACT

RNA methyltransferases (MTases) are important players in the biogenesis and regulation of the ribosome, the cellular machine for protein synthesis. RsmC is a MTase that catalyzes the transfer of a methyl group from S-adenosyl-L-methionine (SAM) to G1207 of 16S rRNA. Mutations of G1207 have dominant lethal phenotypes in *Escherichia coli*, underscoring the significance of this modified nucleotide for ribosome function. Here we report the crystal structure of *E. coli* RsmC refined to 2.1 Å resolution, which reveals two homologous domains tandemly duplicated within a single polypeptide. We characterized the function of the individual domains and identified key residues involved in binding of rRNA and SAM, and in catalysis. We also discovered that one of the domains is important for the folding of the other. Domain duplication and subfunctionalization by complementary degeneration of redundant functions (in particular substrate binding versus catalysis) has been reported for many enzymes, including those involved in RNA metabolism. Thus, RsmC can be regarded as a model system for functional streamlining of domains accompanied by the development of dependencies concerning folding and stability.

INTRODUCTION

S-adenosyl-L-methionine (SAM)-dependent methyltransferases (MTases) transfer the methyl group from SAM to carbon, oxygen or nitrogen atoms of the target. Substrates for SAM-dependent MTases include DNA, RNA, proteins, polysaccharides, lipids or small molecules, implying their key importance in most biological processes (1,2) Methylation of DNA has crucial roles in DNA damage repair, regulation of expression and embryonic development (3). In prokaryotes, DNA MTases are part of restriction–modification systems, which protect the cells from viral invasion (4). Protein methylation is a post-translational process that typically occurs on arginine or lysine residues and is found in prokaryotic and eukaryotic signal transduction pathways, and a role in intracellular signaling has been identified (5). Recently, numerous RNA MTases have been discovered and their function studied (6).

RNA plays a central role in the flow of biological information. Ribosomal RNA undergoes modifications by a number of enzymes during the maturation of ribosome. To date, over 100 different types of nucleotide modifications have been identified, out of which about one-third are present in rRNA (7). Of all species, rRNA modifications are the best characterized in *Escherichia coli* (7). There are three basic types of post-transcriptional modifications found in rRNA, namely base methylation, ribose methylation and pseudouridylation. Base methylation is the simplest and most common type of modification found in rRNA of prokaryotes. It occurs at the final stages of ribosome maturation and the rRNA sequences in which it occurs are highly conserved. Ribose methylation, occurring at the 2′ hydroxyl position on the sugar backbone, is more common in eukaryotes, and is less frequent in bacteria (8).

rRNA MTases play a crucial role in the assembly, maturation and regulation of the protein synthetic cellular machinery (9). In spite of the wealth of literature on rRNA methylation, the structural information currently available for RNA MTases is insufficient to elucidate their mechanism of action. The structure of an rRNA MTase in complex with the substrate RNA is available only for the ternary complex of *E. coli* 23S rRNA m^5U MTase RlmD (previously called RumA) with the ribosomal substrate and SAH (S-adenosyl-L-homocysteine) (10). RNA MTases are thus relatively less well-understood compared to DNA MTases, whose structure–function relationships

*To whom correspondence should be addressed. Tel: +65 65161163; Fax: +65 67792486; Email: dbsjayar@nus.edu.sg

are well established (4).

As a continuation of our efforts to understand the structure and function of rRNA modifying enzymes, we have undertaken structure determination and functional analysis of *E. coli* RsmC that specifically methylates the N2 atom of G1207 in 16S rRNA. The modified residue m^2G1207 occurs in a region of the rRNA that is involved in the recognition of peptide chain termination codons. *In vivo*, transversion mutants of G1207, namely C1207 and U1207, were shown to have dominant lethal phenotypes (11).

Here we report the crystal structure of RsmC from *E. coli* refined at 2.1 Å resolution. RsmC is the first structurally characterized MTase, which exhibits the phenomenon reported earlier for many enzymes, including those involved in RNA modification: presence of duplicated, mutually homologous domains, which preserved the ancestral 3D fold, but accumulated divergent mutations in different regions, leading to the complementary loss of conserved motifs and selective retention of different aspects of function present in the ancestral non-duplicated enzyme. Thus, we combined computational and experimental analyses to identify the key amino acids involved in different functions and to assign the roles to the two domains of RsmC.

MATERIALS AND METHODS

Recombinant DNA techniques

The *rsmC* gene, cloned into pCA24N vector with a non-cleavable N-terminal His6 tag and corresponding strain with the knocked-out *rsmC* gene, were obtained as a gift from the ASKA recloned library [NBRP (NIG, Japan): *E. coli*]. Site-directed mutagenesis of the rsmC gene was performed by a PCR-based technique according to the QuikChange site-directed mutagenesis strategy (Stratagene) following the manufacturer's instructions. NTD- and CTD-RsmC variants were constructed by recloning single domains into the pET28 vector by removing single domains in the PCR reaction. The mutant genes were sequenced and found to contain only the desired mutations.

Expression and purification

For the native protein, plasmid DNA carrying *rsmC* gene was transformed into *E. coli* BL21 (DE3) and grown in 1 l of LB media at 37°C till the OD_{600} reached 0.5–0.6. Induction of the culture was then carried out with 100 μM IPTG after cooling it down to room temperature. The cells were continuously grown overnight at 25°C in a shaking flask at 180 rpm. The next day, cells were harvested by centrifugation (9000 *g* for 30 min, 4°C) and pelleted. The cell pellet was first washed with pre-binding buffer (10 mM Na-Hepes pH 7.9, 0.17 M NaCl), and resuspended in 20 ml of binding buffer (20 mM Na–Hepes pH 7.9, 0.5 M NaCl), 5 mM imidazole (pH 7.0), 5%(v/v) glycerol, 10 mM BME, 0.5% (v/v) Triton-X-100 and 1 tablet of Complete™ EDTA-free protease inhibitor cocktail (Roche diagnostics). The buffer conditions were slight modifications to the ones mentioned in an earlier work describing the purification, cloning and characterization of RsmC (33). For the selenomethionine (SelMet) substituted RsmC, the cells were grown in Le-Master medium (36), using the DL41 strain of *E. coli* (methionine auxotroph).

The purification of RsmC was carried out at room temperature. Both native and SelMet RsmC were purified using the same two-step protocol: DEAE sepharose (Amersham biosciences) column followed by Ni-NTA beads (Qiagen) purification. After binding the protein to the Ni-NTA resin for 30–40 min, the beads were washed with binding buffer (without Triton-X100). The protein was then eluted with 20 mM Na-Hepes pH 7.9, 0.5 M NaCl, 5 mM BME, 0.5 M imidazole, 5%(v/v) glycerol. Furthermore, RsmC was passed through a Superdex-200 gel filtration column using an AKTA-FPLC UPC-900 system (Amersham Biosciences). The gel filtration buffer was the same as the final protein storage buffer: 20 mM HEPES pH 7.9, 0.5 M NaCl, 5 mM BME, 10 mM $MgCl_2$ and 5%(v/v) glycerol. The protein eluted as a monomer (~40 kDa). The peak fractions were pooled together and concentrated to 4.5 mg/ml by ultra filtration, using a Centriprep centrifugal filter device from Millipore, with a molecular weight cut-off of 10 kDa.

Purification and refolding of C-RsmC from inclusion bodies

Inclusion bodies were collected from the cell extract by centrifugation at 20 000 rpm and resuspended in buffer B (10 mM Tris, 50 mM NaCl, 10 mM imidazole) supplemented with 6 M urea. Dissolved pellet was then centrifuged, followed by addition of buffer B and Ni-NTA resin was equilibrated with buffer B. After 1 h incubation, Ni-NTA resin with the bound protein was washed three times with buffer B. The deletion mutant protein RsmC-CTD was eluted with elution buffer (10 mM Tris, 50 mM NaCl, 10 mM imidazole) supplemented with 6 M urea. Refolding of the purified RsmC-CTD was achieved by sequential dialysis with reducing urea concentrations from 6 M to 4 M, 2 M, 1 M, 0 M against RF buffer (100 mM Tris pH 8.8 400 mM L-arginine, 10% glycerol, 0.5% TritonX-100, 1 mM EDTA,1 mM DTT). The dialysis buffer was exchanged every 24 hours. The composition of the dialysis buffer (suitable for the subsequent ITC analyses) was 20 mM Na-Hepes pH 7.0, 300 mM NaCl, 5%(v/v) glycerol, 10 mM $MgCl_2$, 10 mM BME.

In vitro methylation assay

30S ribosomal subunits were isolated as described previously (37). Quantitation of subunits was determined by absorbance at 260 nm (1 A_{260} unit is equivalent to 34.5 pmol of 30S ribosomes). *In vitro* methylation reactions were carried out using 2 μg pure RsmC protein or its variants, 6 μM [methyl-14C]-SAM (52.8 mCi/mmol, NEN) and 3 μM 30S RNA ribosome subunit isolated from the *rsmC*_knockout (K.O.) strain in the total volume 60 μl of the buffer (50 mM PIPES [piperazine-*N*,*N*'-*bis*(2-ethanesulfonic acid)]-Na (pH 7.0), 4 mM $MgCl_2$). After 60 min incubation at 37°C, methylation was stopped by heating the reaction mixture to 70°C for 10 min. The RNA was precipitated with 10% TCA onto Whatman GF/C filter

disks. The disks were washed twice with 5% TCA, once with 5 ml ethanol and air-dried. The filter-bound radioactivity was determined by liquid scintillation counting.

MALDI-TOF analysis

The native and SelMet substituted RsmC was further analyzed for the incorporation of selenium on a Voyager-STR MALDI-TOF mass spectrometer (Applied Biosystems) by comparing the experimentally measured molecular weight of the native protein with that of the SelMet protein.

Dynamic light scattering (DLS)

DLS measurements were performed at room temperature on a DynaPro (Protein Solutions) DLS instrument. The homogeneity of the protein samples was monitored during the various stages of concentration in order to avoid aggregation. The percentage of polydispersity was below 16% and the SOS error was less than 10 for all protein samples at various concentrations.

Isothermal titration calorimetry (ITC)

SAM was procured from MP biomedicals. For the titration experiments, the protein (both native and variants), was extensively dialyzed against a 500-fold excess volume of the buffer containing 20 mM Na-Hepes pH 7.0, 5%(v/v) glycerol, 10 mM $MgCl_2$, 0.3 M NaCl, 10 mM BME, for ~14 h. SAM solutions were prepared by weight, in the same dialysis buffer. The ITC experiments were carried out using VP-ITC calorimeter (Microcal, LLC) at 20°C using 0.02–0.06 mM of the protein in the sample cell and 1–2 mM of SAM in the injector. All samples were thoroughly degassed and then centrifuged to get rid of precipitates. Injection volumes of 4–5 µl per injection were used for the different experiments and for every experiment, the heat of dilution for each ligand was measured and subtracted from the calorimetric titration experimental runs for the protein. Consecutive injections were separated by at least 4 min to allow the peak to return to the baseline. The ITC data was analyzed using a single site fitting model using Origin 7.0 (OriginLab Corp.) software.

Crystallization and data collection

RsmC was crystallized using the hanging drop vapor diffusion method. Initial crystals were obtained from a Jena Biosciences (Jena, Germany) screen and further optimized. The best crystals were obtained when a volume of 1 µl of reservoir solution containing 25% (w/v) PEG MME 5000, 0.1 M Tris–HCl pH 8.5, 0.2 M ammonium sulfate was mixed with 1 µl of protein (Hanging drop). Diffraction quality crystals formed in 3 days, with the smallest dimension measuring ~0.14 mm. RsmC crystals belonged to the space group C2 with one molecule in the asymmetric unit. The cell parameters were $a = 123.94$, $b = 51.50$, $c = 73.33$, ß = 121.52. The Matthew's co-efficient was 2.49 Å^3/Da and the solvent content 50.7% (38).

The crystals were directly taken from the drop, and flash cooled in a N_2 cold stream at 100°K. The native crystals diffracted up to 2.5 Å resolution using an *R*-axis 1V++ image plate detector mounted on a RU-H3RHB rotating anode generator (Rigaku Corp., Tokyo, Japan). Synchrotron data were collected at beam lines X12C and X29, NSLS, Brookhaven National Laboratory for the SelMet protein. A complete SAD dataset was collected (Table 1) using Quantum 4-CCD detector (Area Detector Systems Corp., Poway, CA, USA) to 2.1 Å resolution. Data were processed and scaled using the program HKL2000 (39).

Structure solution and refinement

Of the expected seven selenium sites in the asymmetric unit, five were located by the program SOLVE (40). The N-terminal, as well as the C-terminal methionine, was disordered. The initial phases were further improved by density modification using Sharp (v 3.0.15) (41) that improved the overall figure of merit (FOM) to 0.73. The ARP/wARP (42) built ~65% of the molecule. The remaining parts of the model were built manually using the program O (43). Further cycles of model building alternating with refinement using the program CNS (44) resulted in the final model, with an *R*-factor of 0.21 ($R_{free} = 0.26$) to 2.1 Å resolution with no sigma cutoff used during refinement. The final model comprises 334 residues (Ala3-Met336) and 231 water molecules. The N-terminal His-tag and the linker residues were not visible in the electron density map. PROCHECK (13) analysis shows no residues in the disallowed regions of the Ramachandran plot. A simulated annealing *Fo*–*Fc* omit map of the putative SAM-binding site of RsmC is shown (Figure 1c).

Bioinformatics analyses

Sequence searches were carried out with PSI-BLAST (45), and multiple sequence alignment was constructed with MUSCLE (46). Sequence conservation was calculated from the sequence alignment and mapped onto the protein structure using ConSurf (47). Structure manipulations and modeling was carried out with SwissPDBViewer and PyMol. Structure database searches and superpositions were done with DALI (16).

Protein Data Bank accession code

Coordinates and structure factors have been deposited with RCSB Protein Data Bank with code 2PJD.

RESULTS

Overall structure

The structure of RsmC from *E. coli* was solved by the single-wavelength anomalous dispersion (SAD) (12) method from synchrotron data using SelMet-labeled protein and was refined to a final *R*-factor of 0.21 ($R_{free} = 0.26\%$) at 2.1 Å resolution. The asymmetric unit contains one RsmC molecule comprising 334 residues from Ala3 to Met336 and a total of 232 water molecules.

Table 1. Crystallographic data and refinement statistics

Data set	Peak	High resolution
Cell parameters and Space group	a = 123.94, b = 51.50, c = 73.33, β = 121.52 C2	a = 123.94, b = 51.50, c = 73.33, β = 121.52 C2
Data collection		
Resolution range (Å)	20–2.4	50–2.04
Wavelength (Å)	0.9792	0.9792
Observed reflections > 1σ	110250	170781
Unique reflections	15187	46172
Completeness (%)	96.9	93.5
Overall (I/σI)	12.9	15.2
R_{sym}[a] (%)	7.9	7.4
Refinement and quality		
Resolution range (Å)		15–2.1
R_{work}[b] (no. of reflections)		0.204 (37909)
R_{free}[c] (no. of reflections)		0.256 (3223)
RMSD bond lengths (Å)		0.01
RMSD bond angles (°)		1.78
Average B-factors[d] (Å^2)		
Main chain		25.03
Side chain		27.68
B-rmsd main chain (Å^2)		1.33
B-rmsd side chain (Å^2)		2.25
Ramachandran plot		
Most favored regions (%)		88.6
Additional allowed regions (%)		10
Generously allowed regions (%)		1.3
Disallowed regions (%)		0

[a]$R_{sym} = \Sigma|Ii - \langle I\rangle|/|Ii|$where Ii is the intensity of the *i*th measurement, and ⟨I⟩ is the mean intensity for that reflection.
[b]$R_{work} = |F_{obs} - F_{calc}|/|F_{obs}|$ where F_{calc} and F_{obs} are the calculated and observed structure factor amplitudes, respectively.
[c]R_{free} = as for R_{work}, but for 8% of the total reflections chosen at random and omitted from refinement.
[d]Individual B-factor refinements were calculated.

Neither the N-terminal His-tag nor the C-terminal residues Thr337-Gly343 had interpretable density and were not modeled. The RsmC molecules eluted as a monomer from the gel filtration column. This was consistent with observations in the dynamic light scattering experiments as well as the analysis of intermolecular contacts in the crystal (data not shown). Analysis of the Ramachandran plot using the program PROCHECK (13) showed 88.6% of all residues within the most favored regions and no residues in the disallowed regions. The crystallographic statistics are given in Table 1.

The structure of the full-length RsmC with overall dimensions of ~35 × 40 × 60 Å reveals the presence of two homologous domains of a mixed α/β fold, characteristic for SAM-dependent MTases (Figure 1 ribbon diagram). The existence of intramolecular homology in RsmC has

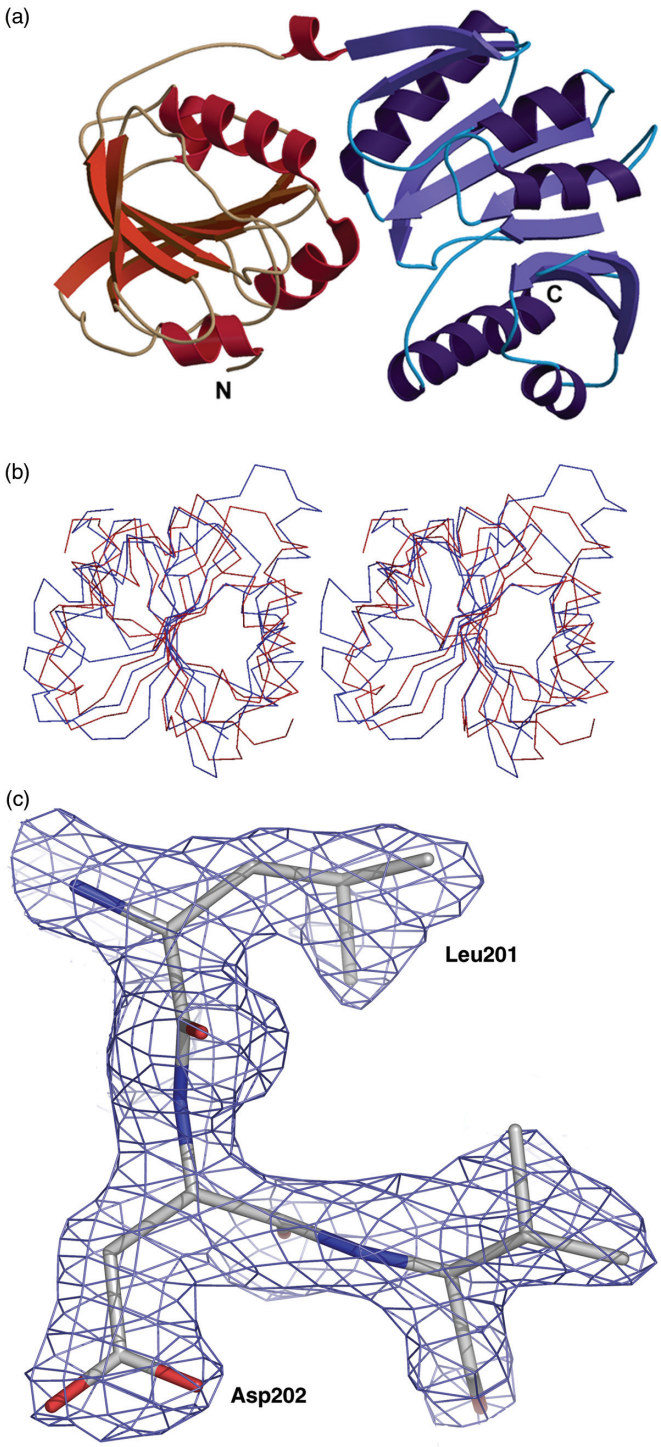

Figure 1. Ribbon diagram showing the domain duplication in the RsmC structure. (**a**) Full-length protein. The N-terminal domain (putative RNA-binding domain: residues 3–150) is depicted in red and the C-terminal domain (SAM-binding domain: residues 179–336) in blue. The N- and C-termini are labeled. (**b**) Superposition of the N-terminal and C-terminal domains (blue and red, respectively) in stereo. These figures were prepared using the programs MOLSCRIPT (48) and Raster3D (49). (**c**) Simulated annealing *Fo*–*Fc* omit map in the putative SAM-binding site of RsmC. The key residue Asp202 and all atoms within 3.5 Å of Asp202 were omitted prior to refinement and map calculation. The map is contoured at a level of 3.0σ. This figure was prepared using PyMol (www.pymol.org).

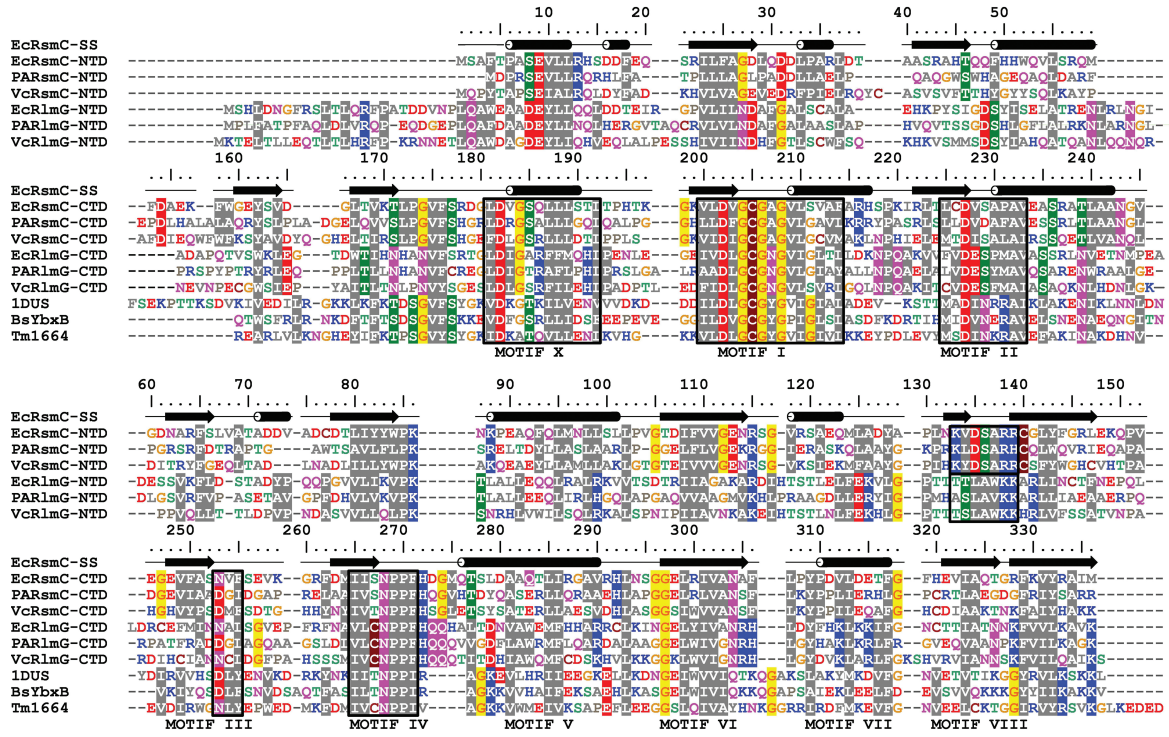

Figure 2. Structure-based sequence alignment of two domains of RsmC, RlmG, together with their closest homolog MJ0882 (1dus). The superposition of RsmC-NTD, RsmC-CTD and MJ0882 was performed with O program (43). For RsmC and RlmG families three representative members are shown: *E. coli* (Ec), *P. aeruginosa* (Pa) and *V. cholerae* (Vc). For MJ0882, homologs are from *B. subtilis* (Bs) and *T. maritima* (Tm). Residues that are conserved within families are highlighted. For EcRsmC, the sequence ruler and the secondary structural elements are shown in the upper panel. Common motifs (conservation at the 3D level) are indicated below the alignment and conserved regions of functional importance (RNA-binding in the NTD and SAM-binding and catalysis in the CTD) are boxed.

been earlier predicted by bioinformatics analysis (14). The N-terminal domain (NTD) consists of seven β-strands and five α-helices and the C-terminal domain (CTD) has nine β-strands and six α-helices. The NPPF (N269-F271) tetrapeptide motif, which is conserved in m^2G MTases (15) is located in a loop between β4 and α5 of the CTD (Figure 2). This motif is absent from the NTD.

The DALI search (16) shows that there is no structure in the PDB with global similarity to the entire RsmC. However, the isolated NTD and CTD show expected similarity to SAM-dependant MTases from the RFM superfamily (17) as well as to each other. In particular, the NTD shows higher similarity to the CTD than to any other structure: RMSD 2.4 Å for 135 Cα atoms, DALI *Z*-score of 13.1. As predicted by bioinformatics analyses (14), among other proteins of known structure, MJ0882, a putative MTase from *Methanococcus jannaschii* (PDB code 1dus) is the closest homolog of both NTD and CTD: it superimposes onto the NTD with 2.5 Å RMSD over 138 Cα atoms, DALI *Z*-score of 13.1 and onto the CTD with RMSD 2.0 Å over 173 Cα atoms, DALI *Z*-score of 23.3. Other MTases from the large RFM superfamily show significant, but lower structural similarity (data not shown).

Although the structures of the NTD and CTD of RsmC are highly similar to each other, the structure-based sequence alignment of the two domains indicates that there is only 12% amino acid identity between them (Figure 2). Clearly noticeable is the preservation of a non-polar character of the residues forming the ß-sheet core of both domains and the lack of conservation of residues at the surface. These features suggest that both domains of RsmC originated by intragenic tandem duplication from a primitive single-domain ancestor similar to MJ0882, and that they accumulated divergent mutations that made them dissimilar on the surface, while preserving the structural scaffold. It is important to note that the NTD appears to have accumulated more sequence and structural changes than the CTD with respect to MJ0882: while the CTD exhibits 22.7% amino acid sequence identity to MJ0882, the NTD shows 11.4% identity both to the CTD and to MJ0882 (see also the aforementioned DALI Z-scores, 23.3 versus 13.1).

Bioinformatics analyses

Although the sequence analysis of RsmC had been reported (14), thus far no high-resolution structure was available to provide a 3D framework for sequence-function considerations. Both domains of RsmC are members of the RFM superfamily of MTases, which is characterized by the presence of a series of motifs conserved at the structural level, and typically also at the sequence level (17). Motifs I, II and III form a SAM-binding pocket, while motifs X and IV usually form the 'floor' and the 'roof' of the catalytic site and may be important for the methyl group donor SAM and substrate

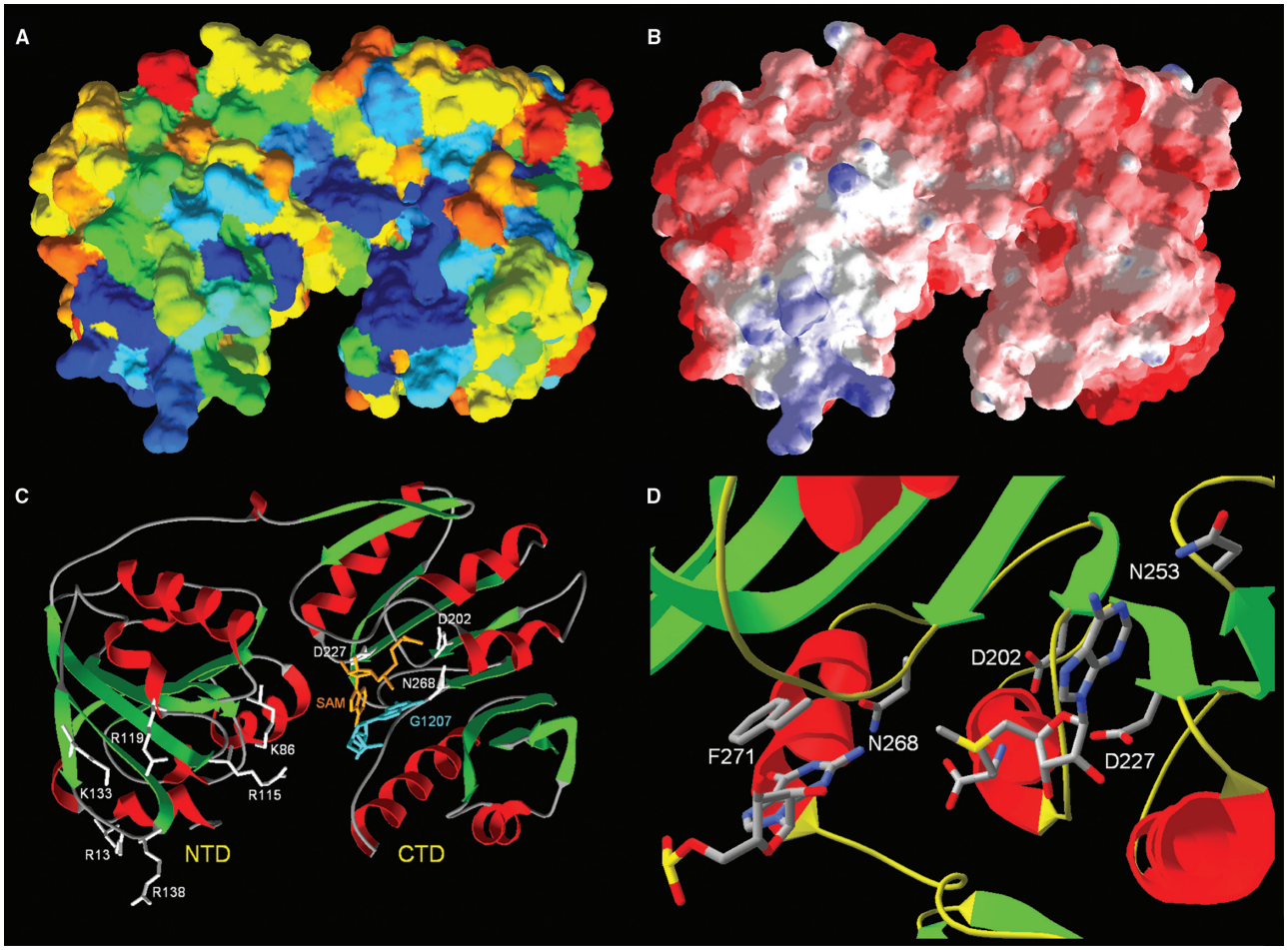

Figure 3. (**A**) Amino acid sequence conservation in the RsmC family mapped onto the RsmC surface using ConSurf (from red: no conservation, to blue: identity). (**B**) Electrostatic potential mapped onto the RsmC surface (from red −5 kT to blue, +5 kT): showing the positively charged protuberance in the NTD and the dominantly negatively charged CTD (left and right, respectively) separated by a cleft. (**C**) Ribbon diagram of RsmC with residues studied by mutagenesis shown in white. The docked SAM molecule is shown in orange, and the docked substrate guanosine 1207 in cyan. Ligands were docked manually, to visualize the active site, in analogy to other MTase structures. (**D**) A detailed view of the predicted ligand-binding/active site pocket in RsmC. Atoms of docked ligands and predicted ligand-binding residues are colored using the following scheme: C, gray; O, red; N, blue; S (in SAM) and P (in guanosine), yellow. The figures were produced with SwissPDBViewer.

binding, positioning them in optimal orientation for the methyl group transfer to occur. Motif VI often participates in the formation of the active site from the substrate side, motifs V and VII are typically important for the structural stability and motif VIII can participate in substrate binding. On the sequence level, motif I is strongly conserved among nearly all members of the RFM superfamily and typically assumes the pattern similar to (D/E)XGXGXG. Motif IV typically contains the key substrate-binding and/or catalytic residues and assumes very different sequence patterns in MTase families acting on different molecules. In MTases acting on exocyclic amino groups of nucleic acid bases (those yielding m^6A, m^4C and m^2G modifications), the typical pattern of conservation is (N/D/S)PP(Y/F/W/H) (15).

To identify the potential functionally important sites in both domains of RsmC, we calculated the sequence conservation in the RsmC family and mapped it onto the protein surface. This analysis reveals two conserved patches: the larger one lining up a deep pocket in the CTD formed by motifs: X, I, II, III, IV and VI, and the smaller one on the exposed protuberance of the NTD formed by motifs VII and VIII. Importantly, the conservation is asymmetric across the domains—neither the NTD pocket nor the CTD protuberance shows any significant conservation (Figure 3A). On the other hand, mapping of the electrostatic potential on the surface of RsmC reveals that the protein is almost uniformly negatively charged with the exception of a small positive patch on the conserved NTD protuberance (Figure 3B). We carried out analogous analyses for a comparative model of RlmG (YgjO), a MTase closely related to RsmC and also exhibiting two domains, but specific for m^2G modification at the G1835 another position of 23S rRNA (18). The distribution of conservation in the RlmG family is similar to that in the RsmC family, with high conservation in the CTD pocket and on the NTD protuberance (Figure 2 and Supplementary Figure 1). Interestingly, while the CTD pocket is conserved between RsmC and RlmG, the NTD protuberance is not, i.e. motifs VII and VIII in both

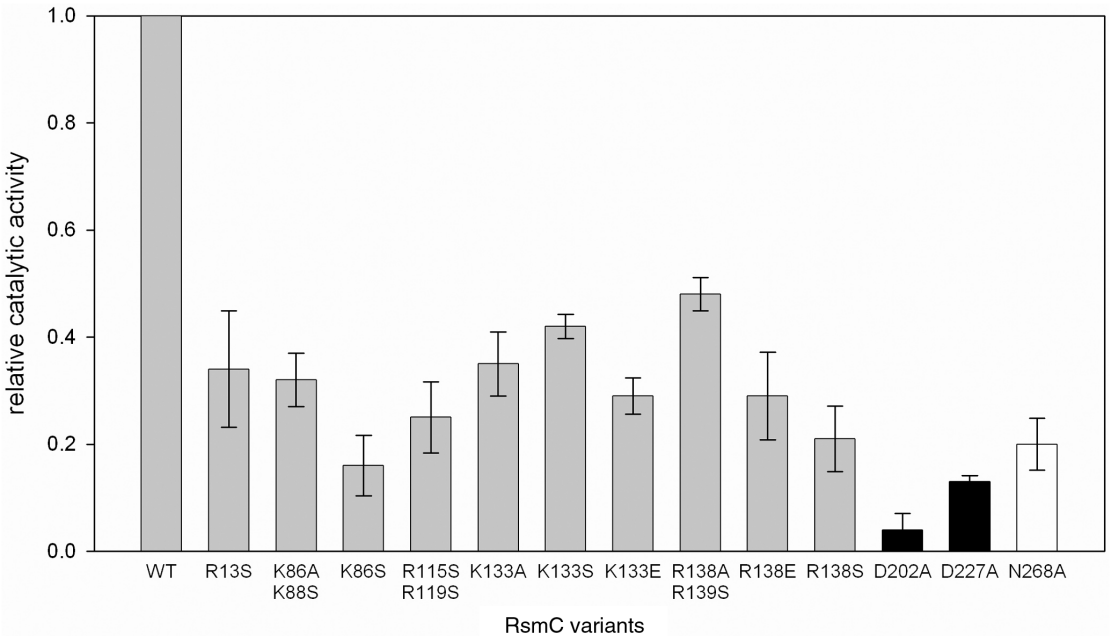

Figure 4. *In vitro* MTase activity of the mutant RsmC variants, measured on the 30S RNA ribosome subunits isolated from the *rsmC* K.O. strain. The activity is shown as the percentage of the wild-type MTase activity. Double and single substitutions in the presumed RNA-binding site are indicated in gray, substitutions in the SAM-binding site are indicated in black, and the substitution in the active site is indicated in white.

families exhibit different conserved amino acids (Figure 2). RlmG is also negatively charged, with positive patches on both NTD and CTD protuberances (Supplementary Figure 2). Conservation of the pocket with motifs I and IV suggest that the CTD of RsmC and RlmG is important for binding of the SAM cofactor and the catalysis of the methyl transfer reaction. On the other hand, a positively charged protuberance that shows differential conservation in MTase families of different specificity is likely to be important for the recognition and binding of their different rRNA substrates. This prediction is further supported by bioinformatics methods for prediction of RNA-binding sites RNABindR (19) and BindN (20) that identify region 130–145 (encompassing motif VIII in the NTD) as a likely RNA-binding site (data not shown).

Structure–function relationships in RsmC

To characterize the function of each domain of RsmC and to confirm the predicted role of individual residues, we designed and constructed two deletion mutants corresponding to the isolated NTD and CTD (amino acids 1–158 and 159–336, respectively), and a series of point mutants of conserved residues in the full-length RsmC that mapped to the predicted SAM-binding site, guanosine-binding/catalytic site and the RNA-binding site. For the potential RNA-binding site we constructed three double mutants in the neighboring positively charged residues (Figure 3C). The NTD as well as the point mutants expressed and purified easily using procedure optimized for the wild-type protein, while the isolated CTD turned out to be very difficult to purify in these conditions and only the purification and refolding from inclusion bodies enabled us to obtain sufficient amounts of the deletion mutant protein for further experiments. It is known that a maltose-binding protein (MBP) can act as a 'passive chaperone' to improve the solubility and promote the proper folding of their fusion partners (21). Thus, we constructed two variants of the RsmC CTD, fused to the MBP either in the N- or C-terminus of the isolated domain (i.e. MBP–CTD or CTD–MBP). We found that the MBP–CTD fusion protein purifies well, similar to the wild-type RsmC (NTD–CTD), while the CTD–MBP fusion protein purifies poorly, similar to the isolated CTD (data not shown). In the MBP–CTD fusion, the MBP domain physically replaces the NTD of the wt RsmC and has the opportunity to fold before the CTD, as it leaves the ribosome earlier. On the other hand, in the CTD–MBP fusion CTD leaves the ribosome first, and it is likely that it starts to fold before it has a chance to interact with the MBP domain. Our results suggest that the RsmC CTD has lost the ability to fold on its own and requires a pre-folded 'intramolecular chaperone' localized at its N-terminus, be it the NTD or another well-folded domain such as MBP.

In order to characterize the function of individual residues, we carried out the functional, biochemical and biophysical characterization of the point mutants. The biochemical assay involving the *in vitro* methylation of ribosomes isolated from the rsmC△ strain (see 'Materials and methods' section for details) revealed that all mutant proteins exhibit reduced activities compared to the wild-type RsmC (Figure 4). In particular, alanine substitution of residues predicted to be important for SAM binding showed the most severe loss of activity (D202A in motif I to 4% and D227A in motif II to 13%). The alanine substitution of the Asn residue in the predicted catalytic NPPF motif IV (N268A) has reduced the activity to 20% of the wild-type. On the other hand, substitutions of

Table 2. ITC data for titration of RsmC variants with SAM

Protein Binding to SAM	Ka (10^5/M)	−ΔH (kcal/mol)	−ΔS (cal/deg.mol)	−ΔG (kcal/mol)	Number of binding sites
Native	2.09	11.52	14.9	7.15	0.94
R138A	1.78	11.84	16.4	7.03	0.99
K133A	1.96	8.92	6.19	7.07	0.97
K133E	1.91	12.08	17	7.1	0.91
K133S	1.74	6.03	−3.42	7.03	1.2
R138S, R139S	1.89	5.6	−5.06	7.09	1.11
K86A	2.1	16.39	31.6	6.97	1.04
R13S	1.74	11.68	15.8	7.05	0.95
R115S, R119S	1.89	10.4	11.3	7.09	0.93
K86S, K88S	1.94	11.33	14.5	7.08	0.99
N268A	0.35	11.66	19	6.09	1.03
D202A	NB (no binding)				
D227A	NB (no binding)				
NTD	NB (no binding)				
MBP-CTD	NB (no binding)				

Ka, binding affinity.
ΔH, ΔS, ΔG—change in enthalpy, entropy and Gibb's free energy, respectively.

individual residues in the predicted RNA-binding site had relatively mild effects on the RsmC activity—their activity was typically reduced only to 30–50% of the wt enzyme (Figure 4). Double mutants exhibited further reduction of activity, e.g. K86S/K88S to 16%. These results are very similar to those obtained in the course of mutagenesis of the rRNA:m^6A methyltransferase ErmC' (22), where it was also impossible to obtain a mutant that would be completely inactive *in vitro* even with multiple substitutions in the predicted RNA-binding site.

The interactions between RsmC (and its mutant variants) and the methyl group donor SAM were studied by the Isothermal Titration Calorimetry. The thermodynamics of binding is given in Table 2. The mutants D202A and D227A in the potential SAM-binding site in the CTD showed complete inability to bind the cofactor (Figure 5), while the N268A mutant in the predicted catalytic motif NPPF that coordinates interactions between SAM and the target guanosine showed almost 5-fold reduction in the SAM-binding affinity (Supplementary Figure 1 and Table 2). On the other hand, mutants in the predicted RNA-binding site in the NTD could still bind SAM with wild-type-like affinities (Table 2) indicating that their reduced activity is not due to the compromised cofactor binding.

ITC experiments on the NTD of RsmC with SAM indicated that this domain is not capable of binding SAM (Supplementary Figure 1). We failed to obtain a preparation of the isolated CTD (with MBP cleaved off) that would be suitable for ITC measurements. Interestingly, the entire MBP–CTD fusion protein that could be purified, was unable to bind SAM, which indicates one of the three possibilities (i) the CTD is misfolded (despite the presence of MBP) or (ii) some portion of MBP blocks the access to the SAM-binding site on a properly folded CTD or (iii) SAM binding by RsmC requires the presence of both NTD and CTD. Our crystal-structure-based docking model suggests that the NTD does not make direct contacts with SAM. Thus, based on our analysis of the calorimetric studies on the wild type RsmC and point mutants in the CTD that are incapable of SAM binding, we propose that the NTD has evolved to be an essential intramolecular chaperone of the CTD that promotes the formation of the SAM-binding site.

The presented data allows us to conclude that the CTD of RsmC is involved in SAM binding and catalysis of the N2-guanosine methylation reaction, while the NTD is important for the folding of CTD and contains residues that are important (but not essential) for the RNA MTase activity, not by direct involvement in cofactor binding, but most likely by RNA binding. We were unable to measure the binding of RsmC and its variants to the ribosome; however the analysis of protein structure and sequence conservation strongly suggests that the NTD is the principal substrate-recognition and binding module of the RsmC. Thus, despite the homology between NTD and CTD they appear to perform completely different and complementary roles.

DISCUSSION

Domain duplication and functional specialization is a common evolutionary process. The duplication of a gene encoding a primitive multifunctional protein yields two independent proteins or one protein with two similar domains, which may experience relaxation of functional constraints and increased rate of mutations [review: (23)]. A number of primitive homooligomeric enzymes have been reported to possess heterooligomeric counterparts with specialized subunits, the best known example being probably the proteasome [review: (24)]. Among enzymes involved in RNA metabolism, the most frequent specialization in enzymes composed of two or more homologous domains concerns substrate-binding, catalysis or structural stability, accompanied by the degeneration of ancestral activities. Examples include heterodimeric tRNA deaminases (25) and heterotetrameric tRNA:m^1A58 MTases (26,27). Similar mechanisms have been postulated for other MTases, including the protein-modifying enzyme PRMT7 comprising two domains in the single polypeptide (28) and eukaryotic DNA MTases Dnmt3a/Dnmt3b/Dnmt3L, where the 'degenerated' Dnmt3L is a regulatory subunit in the heterodimeric complex with Dnmt3a or Dnmt3b (29,30). However, thus far no structural information existed to analyze this phenomenon in detail.

The structure of RsmC provides the first atomic-level picture of an RNA-modification enzyme as well as of an MTase which comprises two domains apparently derived from a common ancestor, which underwent differential functional specialization. According to ITC measurements, RsmC binds only one SAM molecule, and mutational analyses clearly demonstrate that conserved residues in the CTD are responsible for SAM binding. The direct involvement of the NTD in rRNA binding remains to be established, nonetheless mutational analyses of

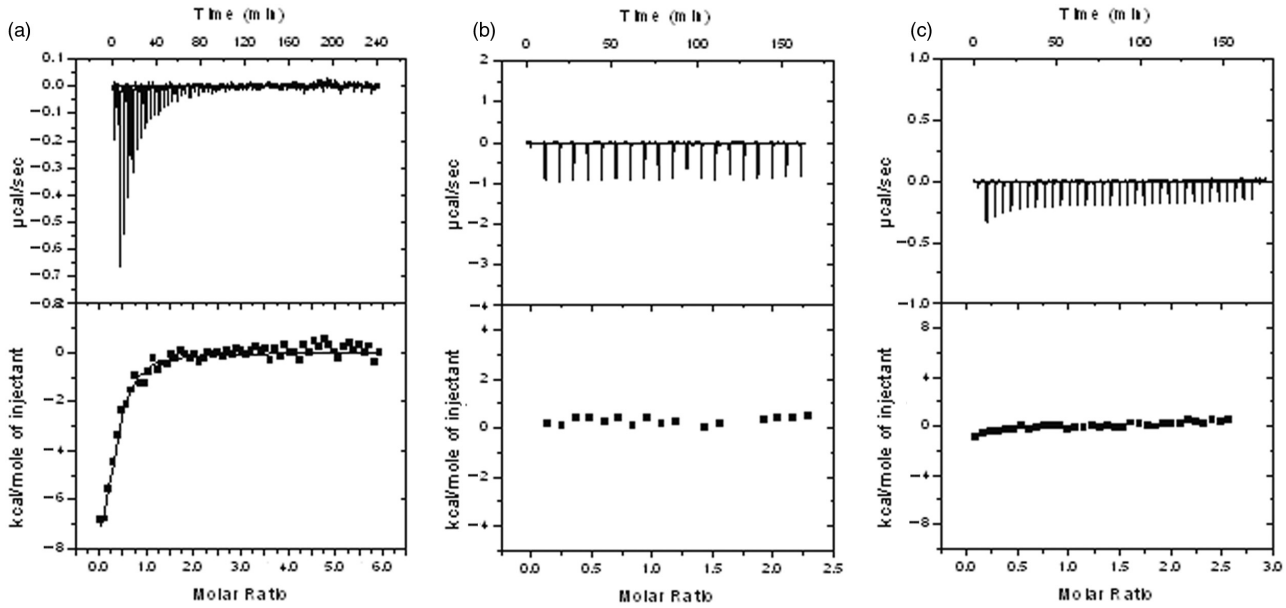

Figure 5. ITC spectra for RsmC wild type and mutants. Baseline subtracted raw ITC data for injections of SAM (ligand) is indicated in the upper panels of each of the ITC profiles shown (for the wild-type as well as the variants of RsmC). The peaks normalized to the ligand/protein molar ratio were integrated as is shown in the bottom panels. The solid dots indicate the experimental data and the best fit to the experimental data were obtained from a non-linear least squares method of fitting using a one-site binding model depicted by a solid line. (**a**) Wild type (**b**) D202A (**c**) D227A.

residues in the conserved charged patch on the NTD surface, predicted to be involved in RNA binding by the RNABindR and BindN methods, give strong support for this prediction. Substitutions of these residues significantly affected the MTase activity, while they had no effect on the SAM-binding ability of the enzyme. Despite the conservation of the structural 'MTase-like' scaffold, two RFM domains of RsmC exhibit complementary pattern of sequence loss or conservation in motifs implicated in substrate-binding (NTD) versus cofactor-binding and catalysis (CTD). Not surprisingly, the isolated domains are unable to carry out the methylation reaction. Moreover, even when the two isolated domains of RsmC are mixed together, they fail to form a catalytically active complex, suggesting that cooperation between the domains requires physical linkage or begins already at the stage of protein synthesis. Indeed, we found that the CTD requires a well-folded N-terminal partner to fold correctly. It is also possible that the peptide linker between the NTD and the CTD plays a role in coordinating binding and catalysis. This specialization of complementary functions and resulting mutual dependence of domains (concerning both protein stability and enzymatic activity) are likely to be common to other 'pseudodimeric' MTases with partially degenerated motifs, such as the protein-arginine MTase PRMT7 and in RNA modification enzymes composed of several homologous domains.

Recently, Dontsova and coworkers characterized experimentally three *E. coli* rRNA:m^2G MTases: RlmL that modifies G2445 in 23S rRNA (31), RlmG that modifies G1835 in 23S rRNA (18) and RsmD that modifies G66 in 16S rRNA (32). They have demonstrated that RsmD is encoded by the YhhF open reading frame (ORF), and that the YgjO ORF encodes not the RsmD enzyme as previously believed (14,33), but RlmG. They have also determined the structure of YhhF/RsmD, which revealed a single catalytic domain (32). Based on these findings, Dontsova and coworkers proposed a hypothesis that *E. coli* rRNA:m^2G MTases can be divided into two categories based on the domain structure and substrate specificity: MTases composed of multiple domains would recognize protein-free ribosomal RNA *in vitro* and most probably, unfolded early assembly intermediates *in vivo*, while MTases comprising only the catalytic domain would recognize only late assembly intermediates resembling the completed 30S particle and not the free RNA (34). They predicted that RlmG and RlmL (whose structures remain unknown) are composed of multiple domains, and that RsmC closely resembles RsmD in that it is composed only of a single domain (32). On the other hand, our results clearly show that RsmC is composed of two domains and is closely related to RlmG (YgjO) rather than RsmD (YhhF). Besides, RsmD has been shown to require the presence of proteins S7 and S19 with the 16S rRNA to be recognized by the enzyme (35). Thus, it appears that the relationship between structure and substrate specificity in rRNA:m^2G MTases is more complex and cannot be inferred simply from the number of domains in different proteins.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We thank Professor Hirotada Mori, Nara Institute of Science and Technology, Japan for the RsmC

clone (50,51). We thank Dr Anand Saxena for assistance in data collection. Data for this study were measured at beam lines X12C and X29 of the National Synchrotron Light Source, BNL. J. Sivaraman acknowledges full research support from the Academic Research Fund (ARF), National University of Singapore (NUS). E.P., M.D., K.L.T. and J.M.B. were supported by MNiSW (grant PBZ-KBN-088/P04/2003). The authors also acknowledge Ms Cherlyn Ng's assistance during the ITC experiments. We thank the Protein and Proteomics Center, Department of Biological Sciences, NUS for providing mass spectrometry facilities. S. Sunita is a graduate scholar in receipt of a research scholarship from the National University of Singapore (NUS). Funding to pay the Open Access publication charges for the article was provided by ARF, NUS.

Conflict of interest statement. None declared.

REFERENCES

1. Loenen,W.A. (2006) S-adenosylmethionine: jack of all trades and master of everything? *Biochem. Soc. Trans.*, **34**, 330–333.
2. Cheng,X. and Blumenthal,R.M. (1999) *S-Adenosylmethionine-dependent Methyltransferases: Structures and Functions*. World Scientific Publishing, New Jersey.
3. Marinus,M.G. (1996) In Neihardt,F.C. (ed.), *Escherichia coli and Salmonella typhimurium, 2nd edn*. ASM Press, Washington DC, pp. 782–791.
4. Cheng,X. and Roberts,R.J. (2001) AdoMet-dependent methylation, DNA methyltransferases and base flipping. *Nucleic Acids Res.*, **29**, 3784–3795.
5. Pahlich,S., Zakaryan,R.P. and Gehring,H. (2006) Protein arginine methylation: cellular functions and methods of analysis. *Biochim. Biophys. Acta.*, **1764**, 1890–1903.
6. Bujnicki,J.M., Droogmans,L., Grosjean,H., Purushothaman,S.K. and Lapeyre,B. (2004) In Bujnicki,J.M. (ed.), *Practical Bioinformatics*., Springer, Berlin, Vol. 15, pp. 139–168.
7. Rozenski,J., Crain,P.F. and McCloskey,J.A. (1999) The RNA modification database: 1999 update. *Nucleic Acids Res.*, **27**, 196–197.
8. Maden,B.E. and Hughes,J.M. (1997) Eukaryotic ribosomal RNA: the recent excitement in the nucleotide modification problem. *Chromosoma*, **105**, 391–400.
9. Lapeyre,B. (2005) In Grosjean,H. (ed.), *Fine-tuning of RNA Functions by Modification and Editing*., Springer, Berlin, Heidelberg, Vol. 12.
10. Lee,T.T., Agarwalla,S. and Stroud,R.M. (2005) A unique RNA fold in the RumA-RNA-cofactor ternary complex contributes to substrate selectivity and enzymatic function. *Cell*, **120**, 599–611.
11. Jemiolo,D.K., Taurence,J.S. and Giese,S. (1991) Mutations in 16S rRNA in Escherichia coli at methyl-modified sites: G966, C967, and G1207. *Nucleic Acids Res.*, **19**, 4259–4265.
12. Hendrickson,W.A. and Teeter,M.M. (1981) Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur. *Nature*, **290**, 107–113.
13. Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.*, **26**, 283–291.
14. Bujnicki,J.M. and Rychlewski,L. (2002) RNA:(guanine-N^2) methyltransferases RsmC/RsmD and their homologs revisited - bioinformatic analysis and prediction of the active site based on the uncharacterized Mj0882 protein structure. *BMC Bioinformatics*, **3**, 10.
15. Bujnicki,J.M. (2000) Phylogenomic analysis of 16S rRNA: (guanine-N^2) methyltransferases suggests new family members and reveals highly conserved motifs and a domain structure similar to other nucleic acid amino-methyltransferases. *Faseb J.*, **14**, 2365–2368.
16. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
17. Kozbial,P.Z. and Mushegian,A.R. (2005) Natural history of S-adenosylmethionine-binding proteins. *BMC Struct. Biol.*, **5**, 19.
18. Sergiev,P.V., Lesnyak,D.V., Bogdanov,A.A. and Dontsova,O.A. (2006) Identification of Escherichia coli m2G methyltransferases: II. The ygjO gene encodes a methyltransferase specific for G1835 of the 23S rRNA. *J. Mol. Biol.*, **364**, 26–31.
19. Terribilini,M., Lee,J.H., Yan,C., Jernigan,R.L., Honavar,V. and Dobbs,D. (2006) Prediction of RNA binding sites in proteins from amino acid sequence. *RNA*, **12**, 1450–1462.
20. Wang,L. and Brown,S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.*, **34**, W243–248.
21. Nallamsetty,S. and Waugh,D.S. (2006) Solubility-enhancing proteins MBP and NusA play a passive role in the folding of their fusion partners. *Protein Expr. Purif.*, **45**, 175–182.
22. Maravic,G., Bujnicki,J.M., Feder,M., Pongor,S. and Flogel,M. (2003) Alanine-scanning mutagenesis of the predicted rRNA-binding domain of ErmC' redefines the substrate-binding site and suggests a model for protein-RNA interactions. *Nucleic Acids Res.*, **31**, 4941–4949.
23. Roth,C., Rastogi,S., Arvestad,L., Dittmar,K., Light,S., Ekman,D. and Liberles,D.A. (2006) Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J. Exp. Zool. B Mol. Dev. Evol.*, **308**, 58–73.
24. Gille,C., Goede,A., Schloetelburg,C., Preissner,R., Kloetzel,P.M., Gobel,U.B. and Frommel,C. (2003) A comprehensive view on proteasomal sequences: implications for the evolution of the proteasome. *J. Mol. Biol.*, **326**, 1437–1448.
25. Gerber,A.P. and Keller,W. (1999) An adenosine deaminase that generates inosine at the wobble position of tRNAs. *Science*, **286**, 1146–1149.
26. Bujnicki,J.M. (2001) In silico analysis of the tRNA:m^1A58 methyltransferase family: homology-based fold prediction and identification of new members from Eubacteria and Archaea. *FEBS Lett.*, **507**, 123–127.
27. Roovers,M., Wouters,J., Bujnicki,J.M., Tricot,C., Stalon,V., Grosjean,H. and Droogmans,L. (2004) A primordial RNA modification enzyme: the case of tRNA (m^1A) methyltransferase. *Nucleic Acids Res.*, **32**, 465–476.
28. Gros,L., Renodon-Corniere,A., de Saint Vincent,B.R., Feder,M., Bujnicki,J.M. and Jacquemin-Sablon,A. (2006) Characterization of prmt7alpha and beta isozymes from Chinese hamster cells sensitive and resistant to topoisomerase II inhibitors. *Biochim. Biophys. Acta*, Epub 2006 Sep 14, 1646–1656.
29. Gowher,H., Liebert,K., Hermann,A., Xu,G. and Jeltsch,A. (2005) Mechanism of stimulation of catalytic activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L. *J. Biol. Chem.*, **280**, 13341–13348.
30. Kareta,M.S., Botello,Z.M., Ennis,J.J., Chou,C. and Chedin,F. (2006) Reconstitution and mechanism of the stimulation of de novo methylation by human DNMT3L. *J. Biol. Chem.*, Epub 2006 Jul 7, 25893–2590.
31. Lesnyak,D.V., Sergiev,P.V., Bogdanov,A.A. and Dontsova,O.A. (2006) Identification of Escherichia coli m2G methyltransferases: I. the ycbY gene encodes a methyltransferase specific for G2445 of the 23 S rRNA. *J. Mol. Biol.*, **364**, 20–25.
32. Lesnyak,D.V., Osipiuk,J., Skarina,T., Sergiev,P.V., Bogdanov,A.A., Edwards,A., Savchenko,A., Joachimiak,A. and Dontsova,O.A. (2007) Methyltransferase that modifies guanine 966 of the 16S rRNA: functional identification and tertiary structure. *J. Biol. Chem.*, **282**, 5880–5887.
33. Tscherne,J.S., Nurse,K., Popienick,P. and Ofengand,J. (1999) Purification, cloning, and characterization of the 16S RNA m2G1207 methyltransferase from Escherichia coli. *J. Biol. Chem.*, **274**, 924–929.
34. Sergiev,P.V., Bogdanov,A.A. and Dontsova,O.A. (2007) Ribosomal RNA guanine-(N2)-methyltransferases and their targets. *Nucleic Acids Res.*, **35**, 2295–2301.
35. Weitzmann,C., Tumminia,S.J., Boublik,M. and Ofengand,J. (1991) A paradigm for local conformational control of function in the ribosome: binding of ribosomal protein S19 to Escherichia coli 16S rRNA in the presence of S7 is required for methylation of m2G966 and blocks methylation of m5C967 by their respective methyltransferases. *Nucleic Acids Res.*, **19**, 7089–7095.

36. LeMaster,D.M. and Richards,F.M. (1985) 1H-15N heteronuclear NMR studies of Escherichia coli thioredoxin in samples isotopically labeled by residue type. *Biochemistry*, **24**, 7263–7268.
37. Daigle,D.M. and Brown,E.D. (2004) Studies of the interaction of Escherichia coli YjeQ with the ribosome in vitro. *J. Bacteriol.*, **186**, 1381–1387.
38. Matthews,B.W. (1968) Solvent content of protein crystals. *J. Mol. Biol.*, **33**, 491–497.
39. Otwinowski,Z. and Minor,W. (1997) Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.*, **276**, 307–326.
40. Terwilliger,T.C. and Berendzen,J. (1999) Automated MAD and MIR structure solution. *Acta Crystallogr. D Biol. Crystallogr.*, **55**, 849–861.
41. Bricogne,G., Vonrhein,C., Flensburg,C., Schiltz,M. and Paciorek,W. (2003) Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. *Acta Crystallogr. D Biol. Crystallogr.*, **59**, 2023–2030.
42. Perrakis,A., Morris,R. and Lamzin,V.S. (1999) Automated protein model building combined with iterative structure refinement. *Nat. Struct. Biol.*, **6**, 458–463.
43. Jones,T.A., Zou,J.Y., Cowan,S.W. and Kjeldgaard,M. (1991) Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A*, **47**(Pt 2), 110–119.
44. Brunger,A.T., Adams,P.D., Clore,G.M., DeLano,W.L., Gros,P., Grosse-Kunstleve,R.W., Jiang,J.S., Kuszewski,J., Nilges,M. *et al.* (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.*, **54**(Pt 5), 905–921.
45. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
46. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
47. Glaser,F., Pupko,T., Paz,I., Bell,R.E., Bechor-Shental,D., Martz,E. and Ben-Tal,N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.
48. Kraulis,J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.*, **24**, 946–950.
49. Merritt,E.A. and Murphy,M.E. (1994) Raster3D Version 2.0. A program for photorealistic molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.*, **50**, 869–873.
50. Kitagawa,M., Ara,T., Arifuzzaman,M., Ioka-Nakamichi,T., Inamoto,E., Toyonaga,H. and Mori,H. (2006) Complete set of ORF clones of Escherichia coli ASKA library (A Complete Set of E. coli K-12 ORF Archive): Unique Resources for Biological Research. *DNA Res.*, **12**, 291–299.
51. Baba,T., Ara,T., Hasegawa,M., Takai,Y., Okumura,Y., Baba,M., Datsenko,K.A., Tomita,M., Wanner,B.L. and Mori,H. (2006) Construction of Escherichia coli-K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol*, **2**, 2006.0008.