Nucleic Acids Research, 2006, Vol. 34, No. 9 ***2483–2494***
doi:10.1093/nar/gkl145

THUMP from archaeal tRNA:m^2_2G10 methyltransferase, a genuine autonomously folding domain

Guillaume Gabant[1], Sylvie Auxilien[2], Irina Tuszynska[3], Marie Locard[2], Michal J. Gajda[3,4], Guylaine Chaussinand[1], Bernard Fernandez[1], Alain Dedieu[1], Henri Grosjean[2], Béatrice Golinelli-Pimpaneau[2], Janusz M. Bujnicki[3] and Jean Armengaud[1,*]

[1]CEA VALRHO, DSV-DIEP—SBTN, Service de Biochimie post-génomique & Toxicologie Nucléaire, F-30207 Bagnols-sur-Cèze, France, [2]Laboratoire d'Enzymologie et Biochimie Structurales, CNRS, Bld 34, avenue de la Terrasse 1, F-91198 Gif-sur-Yvette, France, [3]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Trojdena 4, 02-109, Warsaw, Poland and [4]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

Received January 13, 2006; Revised February 13, 2006; Accepted March 16, 2006

ABSTRACT

The tRNA:m^2_2G10 methyltransferase of *Pyrococus abyssi* (PAB1283, a member of COG1041) catalyzes the N^2,N^2-dimethylation of guanosine at position 10 in tRNA. Boundaries of its THUMP (THioUridine synthases, RNA Methyltransferases and Pseudo-uridine synthases)—containing N-terminal domain [1–152] and C-terminal catalytic domain [157–329] were assessed by trypsin limited proteolysis. An inter-domain flexible region of at least six residues was revealed. The N-terminal domain was then produced as a standalone protein (THUMPα) and further characterized. This autonomously folded unit exhibits very low affinity for tRNA. Using protein fold-recognition (FR) methods, we identified the similarity between THUMPα and a putative RNA-recognition module observed in the crystal structure of another THUMP-containing protein (ThiI thiolase of *Bacillus anthracis*). A comparative model of THUMPα structure was generated, which fulfills experimentally defined restraints, i.e. chemical modification of surface exposed residues assessed by mass spectrometry, and identification of an intramolecular disulfide bridge. A model of the whole PAB1283 enzyme docked onto its $tRNA^{Asp}$ substrate suggests that the THUMP module specifically takes support on the co-axially stacked helices of T-arm and acceptor stem of tRNA and, together with the catalytic domain, screw-clamp structured tRNA. We propose that this mode of interactions may be common to other THUMP-containing enzymes that specifically modify nucleotides in the 3D-core of tRNA.

INTRODUCTION

After their transcription, tRNA precursors are subjected to numerous post-transcriptional modifications. To date, out of 102 chemically distinct modified nucleosides presently known in all types of RNA, 87 have been identified in tRNA [(1), see also http://medlib.med.utah.edu/RNAmods/ and http://genesilico.pl/modomics/]. Methylation at different positions of bases and/or of the 2′-hydroxyl group of riboses are the most frequently encountered ones. These modifications can alter the tRNA's codon specificity or stabilize the tRNA tertiary structure [reviewed in (2) and several chapters in (3)], but significance of many tRNA modifications only begins to be understood [see several chapters in (4)]. Concerning tRNA modification enzymes, several have been identified over the last decade, mostly from *Escherichia coli*, *Saccharomyces cerevisiae* and also a few from Archaea (5–7). Remaining uncharacterized enzymes are now systematically studied after identification by comparative genomics and/or structural genomics approaches [see for examples (8–12)].

Most enzymes involved in RNA metabolism are conserved multi-domain proteins: beside a catalytic domain carrying out the enzymatic reaction, they often require one or several other domains to recognize and possibly help to bind the RNA substrate. These domains also may help to bind other

*To whom correspondence should be addressed at CEA VALRHO, DSV-DIEP—SBTN, Service de Biochimie post-génomique & Toxicologie Nucléaire, Marcoule, BP 17171, F-30207 Bagnols-sur-Cèze cedex, France. Tel: 33 4 66 79 68 02; Fax: 33 4 66 79 19 05; Email: armengaud@cea.fr

macromolecules thus forming multi-subunits ribonucleoprotein complexes (13). Among these maturation enzymes, RNA methyltransferases (MTases) have been reported to comprise a catalytic domain that belongs to a limited number of well-studied enzyme superfamilies (10) and one or several additional variable domains, such as domain S4, PUA, TRAM, THUMP, NusB, OB-fold and wHTH (14), which are supposed, but not always experimentally tested, to be involved in binding of nucleic acids. While most studies focused on the catalytic domains of tRNA modification enzymes [see for examples refs (15–18)], only in rare cases their putative RNA-binding domains have been characterized biochemically and structurally (19,20).

Recently, we reported the characterization of PAB1283 protein from the archaeon *Pyrococcus abyssi*, a prototype member of COG1041. This enzyme catalyzes the N^2,N^2-dimethylation of guanosine at position 10 in archaeal tRNAs and was previously called Trm-G10 (9,21). It is hereby designated TrMet(m2,2G10) according to a newly developed, uniform nomenclature (7). Enzymes belonging to COG1041 are ubiquitous in Eukaryota and Archaea but are not present in Bacteria (9,22). In their C-terminus, they all exhibit a characteristic Rossmann fold, S-adenosylmethionine-dependent MTase domain (pfam01170) and in their N-terminus, a predicted RNA-binding THUMP domain (abbreviated after THioUridine synthases, RNA MTases and Pseudo-uridine synthases (23); pfam02926). As implied by its name, THUMP domain is present in many families of enzymes that catalyze very diverse reactions on tRNA. For example, ThiI (COG0301) from *E.coli* is involved in 4-thiouridine formation at position 8 of some bacterial tRNAs (24–26) and Tan1 (KOG3943, related to COG1818) from *S.cerevisiae* was reported to be required for N^4-acetylcytidine formation at position 12 in tRNAs harbouring a long extra arm (27). Other THUMP-containing proteins, COG1258 (a tRNA pseudouridylate synthase; Martine Roovers and Louis Droogmans, personal communication) and COG0116 (predicted as MTase), are still uncharacterized. Therefore, THUMP that is present in proteins of the three domains of life, is an ancient module, which probably has been recruited during evolution to act in different processes related to tRNA modification.

Recently, the structures of two ThiI orthologs (members of COG0301) have been solved: that of PH1313 from *Pyrococcus horikoshii* [1vbk in the Protein Data Bank; unpublished analysis by M. Sugahara, and N.Kunishima, the RIKEN structural genomics initiative) and BA4899 from *Bacillus anthracis* strain Ames (2c5s, (28)]. Both structures are composed of a C-terminal PP-loop domain that contains the thiolase active site (probably degenerated in PH1313) and a N-terminal predicted RNA-binding module comprising the THUMP domain [as defined in its minimal form by Aravind and Koonin (23)], closely linked with a N-terminal ferredoxin-like domain (NFLD) (28). Waterman *et al*. (28) suggested that the NFLD domain may be specific to ThiI.

Despite the widespread occurrence and presumed importance of the THUMP domain, no experimentally proven function has yet been assigned to it. Recently, we proposed that THUMP may interact with a specific region of tRNA and target the catalytic domains of various enzymes towards the central 3D-core of the tRNA molecule (9). This hypothesis was based on the fact that three well characterized THUMP-containing proteins, ThiI, Tan1 and TrMet(m2,2G10), are all involved in site-specific modification within the same region of the L-shaped tRNA substrate: positions 8, 12 and 10, respectively. In order to obtain better insight into the structural and functional features of the THUMP domain and the TrMet(m2,2G10) variant of its N-terminal extension, we first delineated the boundaries of domains in PAB1283 and then purified the N-terminal region as a standalone protein [1–155 aa, including the THUMP domain (59–139 aa) as initially defined by Aravind and Koonin (23)]. We found that this N-terminal fragment of PAB1283 (here termed THUMPα) is autonomously folded and exhibits only a very low affinity for tRNA. We propose a structural model based on protein fold-recognition analysis, which we then validate experimentally using chemical modification of surface-exposed residues and identification of an intramolecular disulfide bridge. Our results suggest that the THUMPα fragment of PAB1283 assumes a similar structure to that of the N-terminal fragment of ThiI, i.e. that it contains two interlinked α/β subdomains [the NFLD (sub)domain and the classical THUMP domain]. Finally, we constructed a docking model of the whole PAB1283 enzyme onto yeast $tRNA^{Asp}$ (a genuine substrate of PAB1283) based on experimental restraints from our previous work on the elucidation of the identity elements in tRNA required for dimethylation of G10 in tRNA (21). Our model suggests a potential binding mode for the THUMP domain, which may be common to various RNA modification enzymes that specifically modify nucleotides in the 3D-core of the tRNA molecule.

MATERIALS AND METHODS

Limited trypsin proteolysis of PAB1283 and THUMPα proteins

Proteolysis reactions were carried out in 25 mM TRIS/HCl buffer, pH 8.0, containing 200 mM NaCl and 1 mM DTT at 25°C for 60–180 min using different protease/polypeptide ratio [1:5 and 1:2 (w/w)]. The protein concentrations were 0.54 mg/ml for PAB1283 and 1.23 mg/ml for THUMPα. The reactions were stopped by addition of 4 mM 4-(2-aminoethyl)-benzene-sulfofluoride (a Ser-protease inhibitor, Pefabloc SC from Pentafarm). The reaction mixtures were directly analyzed by Matrix Assisted Laser Desorption Ionization-Time Of Flight (MALDI-TOF) mass spectrometry or resolved by SDS–PAGE prior fingerprint identification and membrane blotting for Edman sequencing.

Construction of a N-terminal 6His-tagged THUMPα overexpressing plasmid

Two synthetic oligonucleotide primers were designed in order to amplify a truncated version of the *PAB1283* gene from *P.abyssi* using pSBTN-AC18 plasmid (Armengaud *et al*. (19) as template. These primers are oAJ01 (5′-caccATGTTCTACGTTGAAATCCTAGGTTTGC-3′) and oAJ02 (5′-atcaATCGGCCTTCCTCTCGTCAAACTCC-3′). Nucleotides in lower cases were not present in the original coding sequence. PCR performed with Pwo polymerase (Roche Diagnostics) gave a 473 bp homogeneous product that was resolved on a 1.5% GTG agarose gel, purified by

means of a QiaexII agarose gel extraction kit (Qiagen) and cloned into pET200 D-TOPO (Invitrogen). The resulting plasmid was sequenced in order to ascertain the integrity of the nucleotide sequence and was named pSBTN-AD55. Thirty-six amino acid residues (MRGSHHHHHHGM-ASMTGGQQMGRDLYDDDDKDHPFT) were introduced with the N-terminal 6His/Xpress-tag from pET200.

Purification of recombinant THUMPα module

Large scale liquid cultures of *E.coli* Rosetta(DE3)pLysS strain (Novagen) transformed with pSBTN-AD55 were set up at 30°C and induced with 1 mM Isopropyl-β-D-thiogalacto-pyranoside (IPTG) as described earlier (29). Cells (63 g of wet material) were resuspended in 315 ml of cold 50 mM Tris/HCl buffer (pH 8.0 at 20°C) containing 500 mM KCl and 10% (w/w) glycerol, disrupted and centrifuged. THUMPα was purified from 80 ml of this cell extract (corresponding to 16 g of wet cells). The sample was subjected to a 20 min heat treatment at 60°C. After centrifugation, the supernatant was diluted with 60 ml of 50 mM Tris/HCl buffer (pH 8.0) containing 500 mM KCl, 10% glycerol (w/w) and 50 mM imidazole (buffer A) and applied onto a 5 ml HiTrap Chelating HP column (Amersham Biosciences) at a flow rate of 1.5 ml/min. After wash with buffer A, the 6His-tagged THUMPα protein was eluted over a 45 ml linear gradient comprising 50–300 mM imidazole. The major peak, which eluted at about 180 mM imidazole, was desalted by gel filtration on a G25SF gel (Amersham Biosciences) previously equilibrated with 50 mM Tris/HCl buffer (pH 8.0) containing 20 mM KCl and 10% (w/w) glycerol. Protein concentrations were determined using the molar absorption coefficient of 17 900 M^{-1} cm^{-1} at 280 nm. Determination of native molecular mass and tRNA binding assay by gel filtration were done essentially as described earlier (9).

Circular dichroïsm

Far- and near-ultraviolet (UV) circular dichroïsm spectra were recorded at 25°C on a J-810 Jasco spectropolarimeter equipped with a PTC-424S Jasco Peltier, using a quartz cuvette of 1 mm path length, with a 20 nm/min scanning speed and a band-width of 1 nm. For each sample, three spectra were averaged and corrected from the baseline for buffer solvent contribution. Experimental data were analyzed using the program K2D (http://www.embl-heidelberg.de/~andrade/k2d/).

tRNA gel retardation assay

tRNAs used for band shift assays were transcribed *in vitro* in presence of [α-^{32}P]CTP to label tRNA as described previously (30). PAB1283 (15 nM to 4 μM) or THUMPα (98 nM to 25 μM) were incubated with ^{32}P-labeled tRNA (10 fmol) in 25 mM Tris-HCl buffer (pH 7.5) containing 50 mM NaCl, 5 mM $MgCl_2$, 10% glycerol, 0.1 mg/ml RNase free BSA, 2 mM DTT in a final volume of 20 μl. After incubation at 25°C for 20 min, the mixture was placed on ice and bromophenol blue was added to a final concentration of 0.05% before loading on a 6% polyacrylamide gel (mono/bis, 37.5:1) containing 5% glycerol and 1 mM EDTA in 45 mM Tris/Boric acid buffer (pH 8.0) at 4°C. After electrophoresis, the gel was dried and analyzed by using a Storm PhosphorImager (Molecular Dynamics) to quantify free and bound tRNA.

Lysine, tyrosine and serine labeling by three NHS ester reagents

Sulfo-N-hydroxysuccinimide-biotin (Sulfo-NHS-Biotin), Sulfo-N-hydroxysuccinimyl-6-(biotin-amido)-hexanoate (Sulfo-NHS-LC-Biotin) and Sulfo-N-hydroxysuccinimyl-6-(biotin-amido)-6-hexanamidohexanoate (Sulfo-NHS-LC-LC-Biotin) were obtained from Pierce. Labeling of one residue by Sulfo-NHS-Biotin, Sulfo-NHS-LC-Biotin or Sulfo-NHS-LC-LC-Biotin, should result in a mass increase of 226.293/226.078 ($C_{10}H_{14}O_2N_2S_1$), 339.452/339.162 ($C_{16}H_{25}O_3N_3S_1$) and 452.611/452.246 ($C_{22}H_{36}O_4N_4S_1$) (average/monoisotopic masses in a.m.u.), respectively. THUMPα modification was performed by incubating 1.66 nmol of protein in 50 mM K_2HPO_4/KH_2PO_4 buffer (pH 7.5), 50 mM NaCl, with various amounts of freshly prepared chemical dissolved at 2 mM in the same buffer with a constant protein concentration of 33.2 μM. After 30 min of incubation at room temperature, samples were dialyzed against the same reaction buffer. All samples were desalted by means of $ZipTip_{C18}$ (Millipore) prior MALDI-TOF analysis. Trypsin proteolysis was carried out for 5 h at 37°C with a trypsin/protein ratio of 1:50 (w/w).

Mass spectrometry

Mass spectra were recorded on a MALDI-TOF Biflex IV mass spectrometer (Bruker Daltonics) in positive ionization mode. The matrix solutions for desalted protein or peptide samples were sinapinic acid prepared as saturated solution in 30% acetonitrile, 70% milli-Q water and 0.1% trifluoroacetic acid and α-cyano-4-hydroxycinnamic acid prepared as one-fourth diluted satured solution in 50% acetonitrile containing 0.1% trifluoroacetic acid, respectively. Spectra of proteins and peptides were acquired in linear mode (150–250 laser shots) and reflectron mode (90–180 laser shots), respectively. A pepmix calibration kit (Bruker Daltonics) or internal peaks were used for calibration. MALDI mass spectra were processed using the Xmass 5.1.5 software from Bruker Daltonics. Peptide assignment and identification of labeled residues were carried out using the FindMod package from ExPaSy (http://www.expasy.org/tools/findmod/).

Bioinformatic methods

The multiple sequence alignment was carried out using MUSCLE (31) and refined based on the results of structure predictions, to place insertions and deletions in the regions of solvent-exposed loops. Structure prediction was carried out via the GeneSilico metaserver gateway [(32), and references therein]. Access to PONDR was kindly provided by Molecular Kinetics (Indianapolis, IN). A combination of various methods for prediction of secondary structure, protein disorder and residue accessibility was used to predict the local structure, along with a number of protein FR servers to detect the best structural templates in the Protein Data Bank and to align them to the target sequence (i.e. the sequence of the protein to be modeled). The modeling of the 3D structure was done independently for each domain, using the 'FRankenstein's monster' approach (33,34). It consists of iterative cycles comprising the following steps: automatic model-building using

MODELLER (35), assessment of the local model quality using the VERIFY3D scoring system (36) via the COLORADO3D engine (37), creation of hybrid models by recombination of consensus fragments with best-scoring non-consensus fragments, inference of a new alignment by superposition of the hybrid model with the template structure, and local modification of the alignment for regions with poor score. The initial pool of models was generated based on raw FR alignments and the cycles of realignment and remodeling were continued until the VERIFY3D score of resulting models could not be improved. 100 models with the best score were retained for experimental validation.

Mapping of the sequence conservation onto the protein sequence was carried out via COLORADO3D (37), using the Rate4Site method (38), with the JTT matrix and the Bayesian model of sequence substitution, based on the alignment of the N-terminal regions of PAB1283 orthologs.

Protein–RNA docking was performed using the Global RAnge Molecular Matching (GRAMM) method (39). In the absence of the scoring function specific for protein–RNA interactions we resorted to the low-resolution docking option of GRAMM that optimizes only the geometric fit of molecular surfaces between the two molecules, without taking into account the energy of interactions, e.g. from electrostatics. Thousand docking solutions were retained for each domain and tested for agreement with the experimental data using the FILTREST3D method for discrimination of models that fulfill distance restraints (J. M. Bujnicki, M. J. Gajda, M. Kaczor and A. Bakulina, manuscript in preparation).

To facilitate reading and future discussions about amino acids of TrMet(m2,2G10) from *P.abyssi*, position of residues in purified proteins (6HIS-tagged PAB1283 and 6HIS-Xpress-tagged THUMPα proteins) and models refers to the native PAB1283 sequence as defined in (9).

RESULTS

FR analyses for THUMP modeling

COG1041 proteins are predicted to be organized in at least two structural domains. Figure 1 shows a sequence alignment of the N-terminal half of some archaeal COG1041 members. As the 60 N-terminal amino acids are notably less conserved (7 identical residues in at least 2/3 of the sequences shown in Figure 1 over 60) than the next 90 aa (14 identical residues over 90), it is difficult to assess if they really belong to the THUMP domain or if they should be considered as a distinct structural domain. Protein FR methods reported with high confidence that the best template for modeling the 155 N-terminal amino acid of PAB1283 is the N-terminal region of PH1313 (1vbk). The recently solved structure of *B.anthracis* ThiI ortholog (2c5s) has not yet been included in the template libraries of most FR servers and, when reported, gave similar scores compared to 1vbk. The aligned region spanned both subdomains of ThiI orthologs: 1–76 aa matched the NFLD domain, while 77–165 aa matched the 'minimal' THUMP domain. The β−α−β−β−α−β*−α−β−α−β−β−β pattern predicted for the N-terminus of PAB1283 was perfectly aligned with the pattern observed in the crystal structure of PH1313 ('*' indicates the β-strand shared by both subdomains). In the sequence alignments between PAB1283 and PH1313 reported by FR servers (compare first and last sequences in Figure 1B), the only regions that were not always similarly aligned are 25–65 aa of PAB1283 (the predicted junction between the NFLD and THUMP domains) and 135–155 aa (the predicted C-terminus of the THUMP domain, which in PH1313 folds back onto the NFLD). As expected, no similarity was detected between the catalytic domain of PAB1283 and the PP-loop pyrophosphatase domain of ThiI orthologs. We have also independently submitted 1–58 aa and 59–139 of PAB1283 to FR analysis to confirm the presence of two domains in the N-terminal region of PAB1283. While region 77–165 was unambiguously aligned to the THUMP domain of PH1313, region 1–58 could not be confidently aligned to any particular fold, although nearly all templates proposed by FR servers displayed the two-layer α/β structure, with the ferredoxin-like or a similar fold with the same pattern of secondary structures β−α−β−β−α−β. This result suggests that, despite the absence of significant sequence similarity (28), 1–58 aa of PAB1283 are most likely a considerably diverged version of the N-terminal ferredoxin-like fold present in PH1313 and also that this (sub)domain could require the presence of the THUMP core domain for correct folding.

Experimental delineation of domain boundaries of PAB1283

In order to discriminate the function of the THUMP domain within the whole PAB1283 methyltransferase, we intended to purify it as a structurally stable standalone polypeptide. We first attempted to identify the precise domain boundaries in PAB1283, using bioinformatic methods available via the GeneSilico metaserver (32). All methods confidently identified 145–165 aa as the inter-domain linker region. In particular, this region was predicted to be disordered by PONDR (40) as indicated in Figure 1. Besides, FR methods found residues 150–160 to be at the N-terminus of the conserved core of the MTase domain, in agreement with our previously published model (9), while residues 140–150 were found to mark the C-terminus of the THUMP domain (see below).

The domain boundaries were then delineated experimentally by subjecting PAB1283 to limited proteolysis. Trypsin was found eminently suitable for pinpointing sites of chain flexibility or local unfolding in between the two domains because 12 arginines and 4 lysines are scattered along the region (128 to 180 aa) encompassing the predicted linker. Figure 2 (left part) shows the fragmentation pattern generated by limited trypsin proteolysis as revealed by SDS–PAGE. PAB1283 appears quite refractory to proteolysis as one-third of the sample was still not digested after 180 min incubation, even with 50% (w/w) trypsin, a unusually high ratio (Figure 2, lane 5). Two protein fragments of ∼20 and 19 kDa (polypeptides C and D in Figure 2, lane 3) were generated after 60 min incubation with 20% (w/w) trypsin. Edman sequencing and MALDI-TOF mass spectrometry showed that PAB1283 is divided into two polypeptides: the almost full-length tagged N-terminus domain (Tag+[1–152], 19 kDa) and the full-length C-terminal domain ([157–329], 20 kDa) [see Figure 2 (right part) and Supplementary Data]. The trypsin cleavage sites delineate a short interdomain tetrapeptide

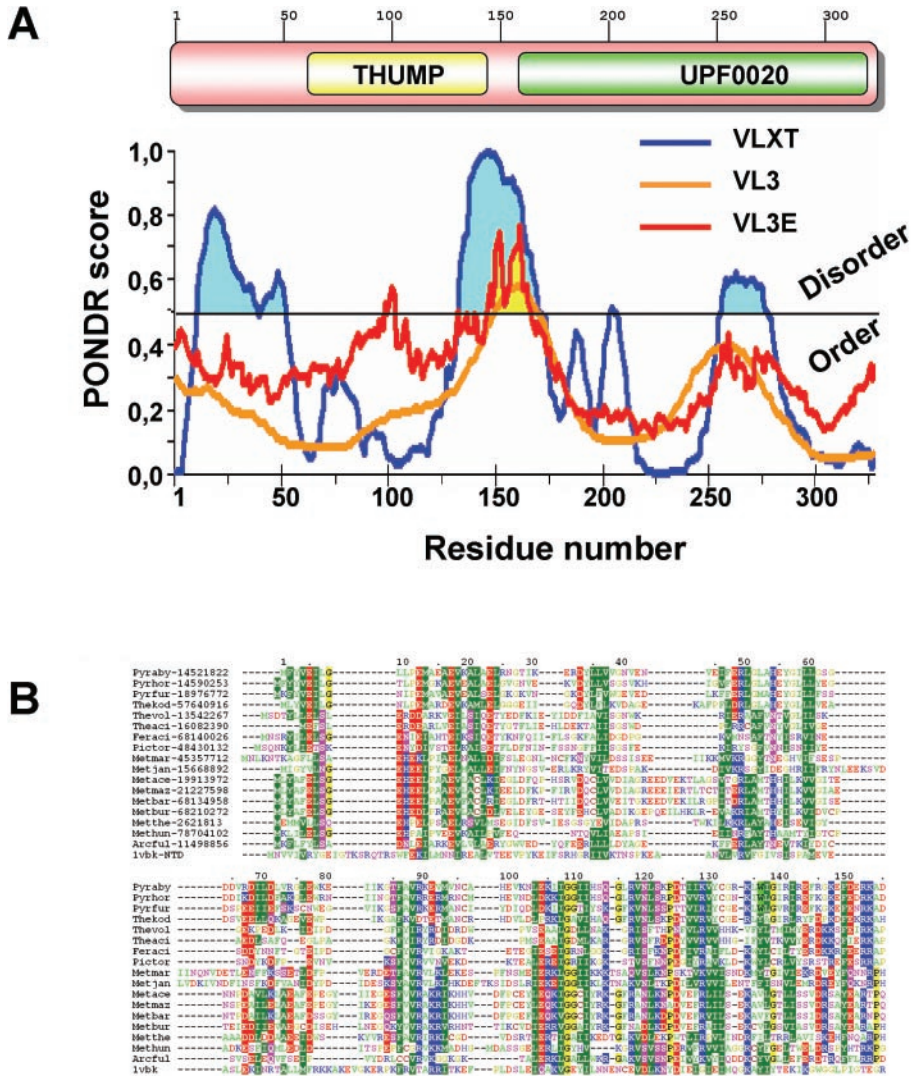

Figure 1. Domain organization of COG1041 sequences and archaeal THUMP sequences alignement. Domains of COG1041 orthologs identified by the NCBI Conserved Domain search tool (http://www.ncbi.nlm.nih.gov) are indicated in (**A**). Residue-by-residue PONDR score for VLXT, VL3 and VL3E algorithms are shown in blue, orange and red, respectively. Predicted disorder regions (score >0.5) are indicated with blue and yellow areas (40). A multiple alignments of 17 archaeal COG1041 sequences with the PH1313 sequence (PDB accession number: 1vbk) is presented in (**B**). Accession numbers (gi) are indicated beside each organism name: *P.abyssi* (Pyraby), *P.horikoshii* (Pyrhor), *Pyrococcus furiosus* (Pyrfur), *Thermococcus kodarensis* (Thekod), *Thermoplasma volcanium* (Thevol), *Thermoplasma acidophilum* (Theaci), *Ferroplasma acidarmanus* (Feraci), *Picrophilus torridus* (Pictor), *Methanococcus maripaludis* (Metmar), *Methanocaldococcus jannaschii* (Metjan), *Methanosarcina acetivorans* (Metace), *Methanosarcina mazei* (Metmaz), *Methanosarcina barkeri* (Metbar), *Methanococcoides burtonii* (Metbur), *Methanothermobacter thermautotrophicus* (Metthe), *Methanospirillum hungatei* (Methun), *Archaeoglobus fulgidus* (Arcful). Residues are colored according to the conservation groups: positively charged (blue), negatively charged (red), hydrophilic (magenta), hydrophobic and aromatic (green), Gly and Pro (yellow), Cys (brown).

linker region: [153–156]. Under harsh proteolysis conditions, no other stable intermediate could be detected (Figure 2, lane 5). The experimentally determined domain boundaries fit with those predicted from sequence comparisons and disorder predictions. These results confirm the existence of at least two well-defined structural domains in PAB1283 connected by a short solvent-exposed linker. This linker is flexible as it is sensitive to trypsin proteolytic cleavage.

Purification of a standalone THUMPα protein

To confirm that THUMPα is a structurally independent unit, an expression plasmid (pSBTN-AD55) encoding a 6HIS/Xpress-tagged version of the N-terminal module (NFLD+THUMP, residues [1–155]) was constructed. The 22 kDa soluble protein, hereby designated THUMPα, was obtained with a high degree of purity (Figure 2, lane 6). THUMPα behaves as a monomeric protein on a Superdex75 column (data not shown), like the native PAB1283 protein (9). The near UV-CD spectrum of this protein (Figure 3) shows typical negative ellipticity signals with minima at 208–218 nm. Deconvolution of the CD signal leads to an estimation of the content of secondary structure elements of about 30% of α-helices and 20% of β-sheets. To further confirm that the protein was not unfolded, it was subjected to limited trypsin proteolysis. As shown in Figure 2 (lanes 6 and 7), a truncated polypeptide of ∼19 kDa (band F) was identified on SDS–PAGE. Edman sequencing and mass

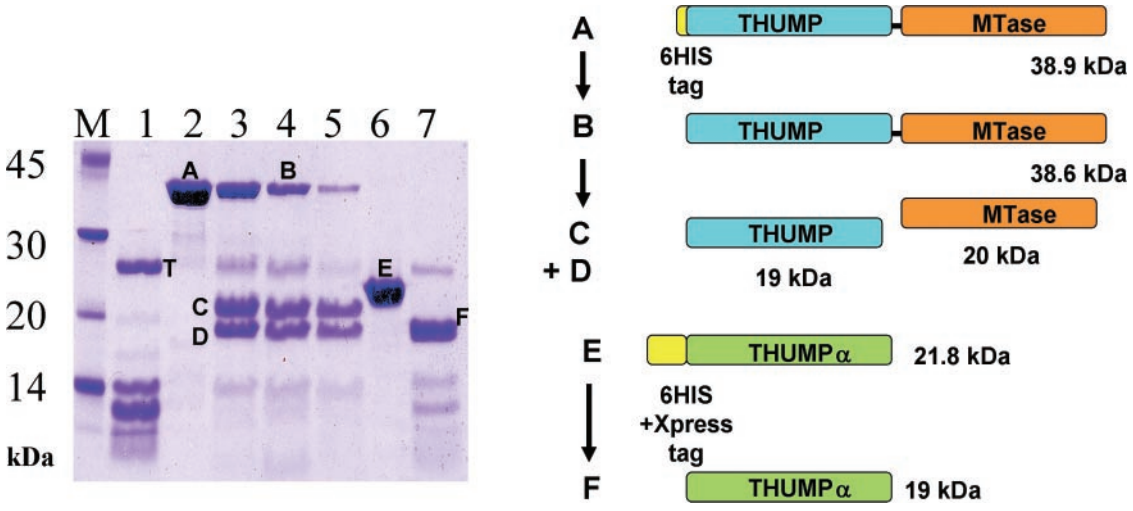

Figure 2. 7SDS–PAGE analysis of limited proteolysis products. SDS–PAGE was performed on a 15% polyacrylamide gel and stained with Coomassie blue. Lane M, molecular weight markers; Lane 1, trypsin; Lane 2, purified PAB1283; Lanes 3–5, proteolyzed PAB1283, Lane 6, purified THUMPα; Lane 7, proteolyzed THUMPα. Samples consisted of 5 μg of proteins. Incubation time is 60 min for all samples except Lane 5 where it was extended to 180 min. Ratios trypsin/protein are: 1:5 (Lanes 3 and 7) or 1:2 (Lanes 4–5). Bands corresponding to trypsin (T) and polypeptides of interest (A–F) are indicated. See Supplementary Data for the delineation of each products by mass spectrometry.

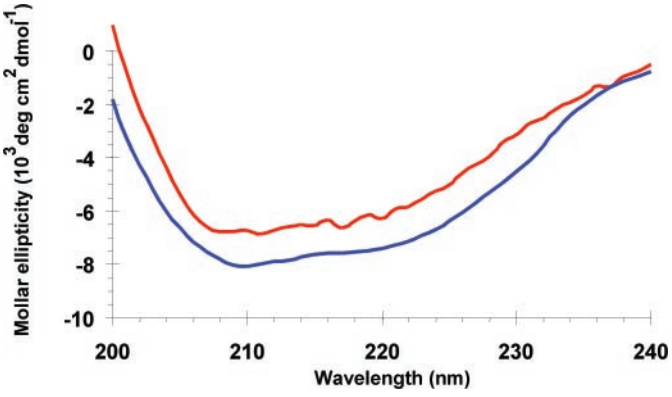

Figure 3. Evaluation of THUMPα secondary structure elements by circular dichroïsm. The molar ellipticity was calculated on the basis of exact amino acid composition of recombinant THUMPα product. Signal of 8.48 μM purified protein in 10 mM Tris/HCl buffer (pH 8.0) is shown in red while k2D estimation is shown in blue. Secondary structures predictions by different methods available via the GeneSilico metaserver (32) gave on the average 44% of α-helices and 21% of β-sheets. The content of α-helices is overestimated by the predictions (or alternatively, the CD measurements may underestimate it).

measurement of this entity confirmed that only the N-terminal tag is proteolyzed (Supplementary Data). The THUMPα core as defined above, remains quite resistant to further proteolysis, even under the harshest conditions used, as shown in Figure 2 (see lanes 5 and 7). Such proteolysis results are clearly in favor of a well structured and folded THUMPα domain. Thus, we conclude that the N-terminal domain of TrMet(m2,2G10) can fold autonomously.

The affinity of THUMPα for tRNA is very low

Since THUMP has been predicted to be a RNA-binding domain (23), we analyzed whether purified recombinant THUMPα could interact with tRNAs. THUMPα was incubated with *E.coli* bulk tRNAs (or yeast bulk tRNAs) in the same experimental conditions where a 1/1 complex with the full-length recombinant PAB1283 (9) was previously identified. The mixture was subjected to gel filtration on Superdex75 and the elution profile compared to those obtained under identical conditions with the protein or bulk tRNAs alone (data not shown). A new peak of compound eluting faster than that of either THUMPα or tRNAs alone was not observed, indicating no formation of a complex between THUMPα and tRNAs. In order to estimate the affinity of THUMPα for tRNA, we performed tRNA gel retardation assays. Radiolabeled *in vitro* transcribed tRNA was incubated with increasing amounts of PAB1283 or THUMPα. Similar results were obtained with the two tRNA substrates that were tested: *P.abyssi* $tRNA^{Asp}$ (data not shown) and yeast $tRNA^{Asp}$ (Figure 4). As shown in Figure 4, one band shift was observed with PAB1283 protein, a result consistent with the formation of a 1/1 complex as indicated by gel filtration assays (9). The dissociation constant (K_d) was estimated at about 1 μM. With THUMPα, the complex is barely detectable even at 25 μM of protein (Figure 4) and this weak interaction can be considered as unspecific. Therefore, THUMPα is not responsible *per se* for the affinity of TrMet(m2,2G10) for tRNA, although its contribution in the interaction with tRNA in the context of the whole protein can be anticipated.

Modeling the 3D-structure of THUMPα

We modeled the structure of THUMPα using the 'FRankenstein's Monster' approach. Briefly, we generated alternative models based on different alignments between the region 1–155 of PAB1283 and the region 1–175 of PH1313 from *P.horikoshii* and BA4899 from *B.anthracis* strain Ames. Then, we assessed the local sequence-structure fit in the models using VERIFY3D (41), and iteratively recombined the models and locally shifted the alignments to generate new models with improved VERIFY3D score. The shifts in the alignments were constrained to maintain

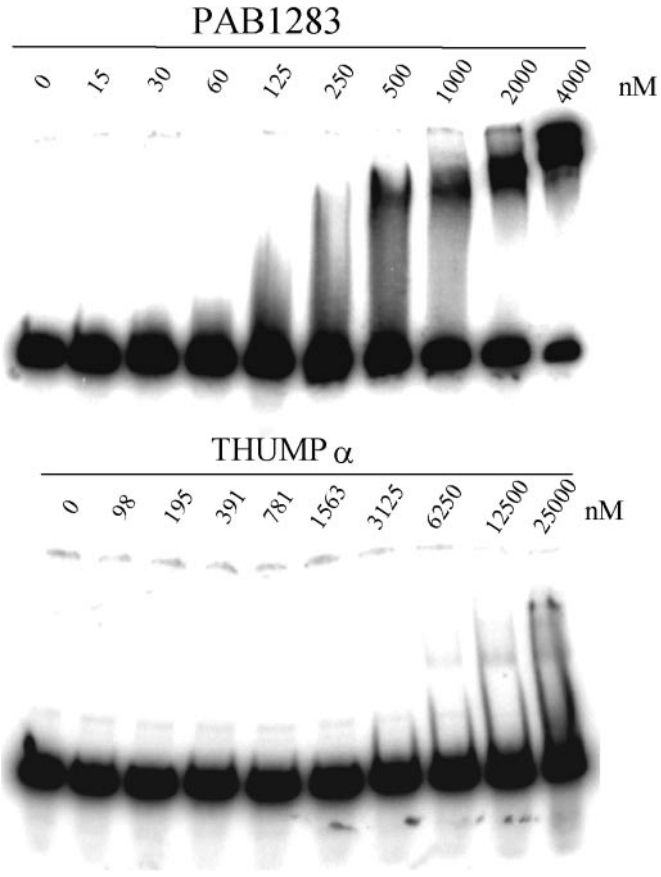

Figure 4. Gel retardation assay of PAB1283 and THUMPα with tRNA substrate. Radiolabeled *in vitro* transcribed yeast $tRNA^{Asp}$ was incubated with PAB1283 protein (upper panel) or THUMPα protein (lower panel).

the overlap of identical secondary structure patterns in PAB1283 and PH1313/BA4899 (ThiI). As a result, we obtained a number of relatively similar models of THUMPα, which differed mainly in the conformation of the variable loops (e.g. residues 7–11, 80–84, 92–97 and 113–117) and orientation of the side-chains (see below). The root mean square deviation (RMSD) between the models ranged from <1 Å (models based on very similar alignments, with differences limited to loops) to 4.5 Å (models based on significantly different alignments).

Validation of the THUMPα 3D-structural model by chemical modification and mass spectrometry

To validate the models and identify the variant with the best conformation, we attempted to determine, which residues are accessible to labeling reagents using mass spectrometry measurements. The amino groups of Lys and the N-terminal Met, as well as Ser and Tyr hydroxyl groups, were specifically labeled with Sulfo-NHS-biotin. After reaction with the chemical reagent, samples were subjected to trypsin proteolysis and compared to untreated samples by MALDI-TOF mass spectrometry. The coverage of the THUMPα sequence was estimated to be 80% with two lysine residues (Lys19 and one amino acid from the tag) not included in this coverage. To unequivocally confirm the assignment of modified peptides, we labeled in separate assays THUMPα with two other reagents (Sulfo-NHS-LC-biotin and Sulfo-NHS-LC-LC-biotin), exhibiting the same overall reactivity but introducing mass differences (+113.084 a.m.u.) due to their spacer arm lengths. As shown in Table 1, peptide assignment of a majority of ions is clearly redundant and confirmed by the theoretical mass increment observed for each modified peptides. Besides three residues from the tag, nine reactive residues were unambiguously labeled and are therefore solvent accessible: Lys30, Lys79, Lys83, Lys90, Lys105, Lys122, Tyr130, Lys134, Lys147 and Lys153 (Table 1). The labeling of Tyr34, Tyr56, Lys101 and Lys128 was not observed under these experimental conditions, although the native peptides encompassing these residues were clearly detected.

During the biochemical study of THUMPα, we observed the presence of an additional discrete band on SDS–PAGE at 22 kDa under non-reducing conditions, migrating slightly faster than the main protein band (See Supplementary Data). This additional band was not detected when the sample was treated with DTT prior to electrophoresis. This indicates that an intrapolypeptide disulfide bridge can be formed in THUMPα. Hence, the two cysteines (Cys96 and Cys131) present in THUMPα are likely to be close in the 3D structure. These two cysteines are conserved among the four TrMet(m2,2G10) sequences from *Thermococcales* (See Figure 1). As disulfide bridges are often present in hyperthermophilic proteins because of their stabilizing effects (42,43), the disulfide bridge is probably formed in PAB1283.

Identification of the THUMPα model that minimizes the violation of experimental constraints

The accessibility characteristics of several amino acids determined by NHS-biotin labeling as well as the predicted close distance of Cys96 and Cys131 (Cβ atoms <7 Å) were used as constraints to select between 100 alternative models. However, no model was found to fully agree with all data. Only one constraint was violated in all models: the ε-amino group of Lys101 was always exposed to the solvent, despite it was never found to be labeled. Besides, in all but two models (very similar to each other, RMSD only 1 Å), the ε-amino group of Lys90 was buried in the protein core, despite we found that it can be labeled. Figure 5 shows one of the two best models. Its RMSD to the template structures 1vbk and 2c5s is 1.8 and 3.9 Å over all alignable pairs or residues. The larger value of RMSD to 2c5s results from a slightly different angle between the NFLD and THUMP domains in 1vbk and 2c5s. Inspection of these two models reveals that the side-chain of Lys101 is close to the side-chain of Glu104 (and in fact, the orientation of these residues is very similar in all models). Perhaps the formation of a salt bridge between these residues prevents the labeling of Lys101 (44,45). In the immediate neighborhood of these residues, the side-chains of Lys90 and Lys105 are partially solvent-exposed, and do not form any salt bridges, in agreement with their labeling. In the rejected models, Lys90 was found at the same position, but with the side-chain inserted into the hydrophobic core. Finally, it is satisfying to find that the distance constraint is respected as the side-chains of Cys96 and Cys131 are close to each other (Cβ atoms at 4.9 Å) allowing them to form an intramolecular disulfide bridge.

Table 1. Monoisotopic $[M+H]^+$ labeled THUMPα peptides generated by trypsin[a]

NHS-Biotin		NHS-LC-Biotin		NHS-LC-LC-Biotin		Theoretical peptides and labeling assignment				
m/z observed	ΔMass	*m/z* observed	ΔMass	*m/z* observed	ΔMass	*m/z* expected	Sequence	Position start-end	Number of labels	Modified residues
532.255	−34	645.312	14	758.446	−55	306.159	MR	1–2	1	M^1
548.226	11	661.257	89	774.401	−1	322.154		1–2*	1	M^1
2504.071	−10	2617.219	−34	2730.249	−13	2277.967	GSHHHHHHGMASM TGGQQMGR	3–23	1	S^4 or S^{14}
2520.045	−2	2633.081	16	2746.209	0	2293.962		3–23*	1	S^4 or S^{14}
2730.079	16	2956.306	−5	3182.497	−12	2277.967		3–23	2	S^4 and S^{14}
2746.050	24	2972.168	39	3198.465	−4	2293.962		3–23*	2	S^4 and S^{14}
2260.004	−21	2259.950	3	2259.939	8	2277.967		3–23	0[b]	S^4 or S^{14}
2486.037	−1	2599.077	16	2712.221	−7	2277.967		3–23	1[b]	S^4 and S^{14}
1043.523	7	1156.590	21	1269.630	54	817.453	NGTIKER	26–32	1	K^{30}
1341.737	−10	1454.720	60	1567.765	80	1115.646	GLEWKEIIK	75–83	1	K^{79}
1359.773	−21	1472.800	20	1585.807	67	1133.668	EIIKGTFAVR	80–89	1	K^{83}
1487.839	1	1600.878	29	1713.901	62	1261.763	EIIKGTFAVRK	80–90	1	K^{83}
1612.781	−11	1725.792	32	1838.844	48	1386.687	KEVMVNCAHEVK	90–101	1	K^{90}
1974.137	−21	2087.169	5	2200.258	2	1748.018	NLEKIIGGIIHSQGLR	102–117	1	K^{105}
1453.784	17	1566.795	62	1679.840	81	1227.731	VNLSKPDTIIK	118–128	1	K^{122}
823.291	83	936.386	61	1049.486	39	597.281	VYCGR	129–133	1	Y^{130}
nd[c]	—	1224.705	19	1337.740	54	885.567	KLWIGIR	134–140	1	K^{134}
1106.504	−10	1219.560	14	1332.590	53	880.416	GKEFDER	146–152	1	K^{147}
nd[c]	—	nd[c]	—	785.431	−11	333.177	KAD	153–155	1	K^{153}

[a]Reagent/polypeptide molar ratio: 26:1; NHS-biotin label (+226.078 a.m.u.); NHS-LC-biotin label (+339.162 a.m.u.); NHS-LC-LC-biotin label (+452.246 a.m.u.); *m/z* and mass tolerance are expressed in a.m.u. and p.p.m., respectively. Only assignments with Δmass tolerance below 100 p.p.m. are shown; peptide positions refer to PAB1283 sequence except for the 6HIS-Xpress-tag shown in grey shading.
[b]Labeling on serine removed (−18 a.m.u.).
[c]'nd' denotes peptides not detected in one of the samples.
*Oxydized methionine.

Summarizing, the overall structure of THUMPα in solution is predicted to resemble the N-terminal region of PH1313 in the crystal, i.e. to contain two closely interlinked α/β domains with the ferredoxin-like and THUMP fold, respectively. The interface of both subdomains includes a common hydrophobic β-strand that spans both domains (Figure 5A). Thus, in agreement with the data from the limited proteolysis experiments, both subdomains of THUMPα are predicted to be linked very rigidly and to form a single independently folded unit. Mapping of the sequence conservation onto the surface of the THUMPα model reveals specific clustering of conserved residues from both subdomains (e.g. Glu5, Leu7 (conserved as Ser in most members of COG1041), His54, Arg89, Lys128, Arg152) on one side of the molecule, which we predict to be involved in tRNA binding (Figure 5B). On the other hand, mapping of the electrostatic potential reveals no particular concentration of positive charge (data not shown).

Prediction of the PAB1283:tRNA complex structure

The availability of experimentally validated models of the C-terminal catalytic domain (9) and the N-terminal domain (this work) of PAB1283, as well as data concerning the regions of tRNA that are involved in the enzyme recognition (21), allows us to speculate about a reasonable model of interactions between PAB1283 and tRNA, using as substrate the bona fide yeast $tRNA^{Asp}$. This tRNA was shown to form a 1:1 complex with the archaeal enzyme and is efficiently methylated in the presence of S-adenosylmethionne (9). The field of computational protein–RNA docking is in its infancy and there are no established algorithms to carry out this type of modeling. In particular, there are no confident methods to account for the flexibility of the RNA molecule, as well as to precisely calculate the energy of RNA–protein interactions, which are often dominated by electrostatics. Thus, we decided to use the GRAMM software for low-resolution docking (39) to calculate independently for each domain 1000 alternative docked models to the yeast $tRNA^{Asp}$ structure [PDB entry 2tra, (46)] that simply exhibit surface complementarities but no requirement for electrostatic compatibility was imposed. We assumed that no major conformational rearrangements occur in the RNA or the protein, except for possible base-flipping of G10/m^2G10 or conformational changes of the flexible linker between THUMPα and the catalytic domain of PAB1283. Subsequently, we screened all the models for the following constraints, based on experimental data. (i) The C-terminal MTase domain must be positioned in such a way that the methylated nucleoside G10 can be accommodated in the catalytic pocket. Thus, we screened for docking solutions in which the N2 group of G10 in tRNA was within 15 Å from the C-β atom of the catalytic side-chain D254 (9), the distance threshold being deliberately large to account for possible base-flipping of G10). (ii) The T-arm is the key determinant of specificity of interactions between PAB1283 and tRNA, and the amino acid acceptor stem is required for the second round of methylation (to $m^2{}_2$G10) to occur (21). Thus, we looked for such docking solutions, in which any of the atoms from either domain of PAB1283 is within 5 Å from both any atoms of nucleosides 49–65 (T-arm) and 1–7 or 66–72 (acceptor arm) in the $tRNA^{Asp}$. (iii) The C-terminus of THUMPα is covalently joined to the N-terminus of the catalytic domain. Thus, we looked for such combinations of

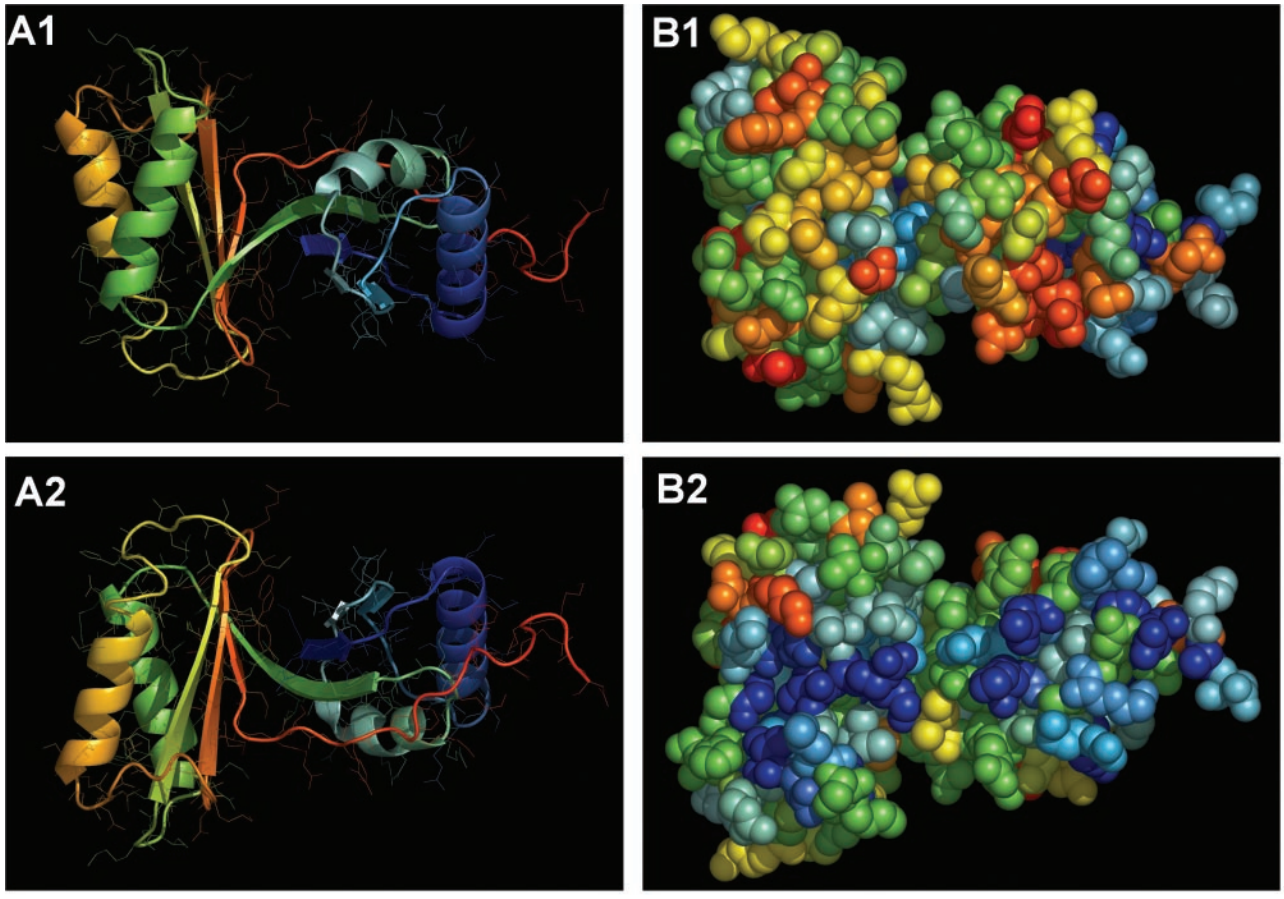

Figure 5. THUMPα 3D-structural model. Four panels showing the THUMPα structure in different orientations (1 and 2), either (**A**) with the backbone shown as a cartoon and colored by progression of sequence (from blue N-terminus, to red C-terminus), or (**B**) with all atoms in the space-filled representation and colored according to the sequence conservation (from dark blue highly conserved, to red highly variable).

docking solutions that passed conditions 1 and 2, in which aa 153 in the N-terminal domain was within 50 Å from aa 167 in the C-terminal domain (the distance threshold was deliberately large to account for the flexibility of the linker). (iv) Last, but not least, THUMPα and the catalytic domain must not overlap spatially.

Among the 1000 docking models obtained for each domain (100 000 0 possible combinations), only five models of the N-terminal domain and two models of the C-terminal domain passed all criteria. Interestingly, in the case of the C-terminal domain, conditions 1 and 2 turned out to be mutually exclusive, i.e. we found no models, which would simultaneously bind to the T-arm and position the catalytic pocket in the vicinity of the nucleoside to be methylated. Hence, we predict that binding of the T-arm occurs through the N-terminal THUMPα domain. Among the five docking solutions for the N-terminal region, only two similar models (RMSD 5.7 Å) interact with the T-arm using the conserved side of the molecule, while the three others (dissimilar to each other) do not make any contacts with the conserved protein residues. Thus, we introduced a 5th, *ad hoc* criterion to retain only the models that interact with the tRNA using the conserved residues. The two selected models of the C-terminal domain approach tRNA from the same side, albeit they are rotated with respect to each other by nearly 180°.

We analyzed the conservation of nucleotides in Archaeal tRNAs with the $m^2{}_2$G10 modification in close contact with THUMPα and found no common features except the G53–C61 pair (which is important for the tRNA folding) and the CCA extension. Analysis of the docking model reveals that the THUMPα module of TrMet(m2,2G10) makes contacts with C61 through residues from the region 145–148 (Arg145 and Lys147 in PAB1283, see Supplementary Data). This region is not strongly conserved on the level of the amino acid sequence in TrMet(m2,2G10) orthologs, but always contains at least one positively and one negatively charged residue that may be important for binding either to the base or to the backbone (the phosphate is exposed). The stacking of the G53–C61 base pair over the U/T54-A58 reverse-Hoogsteen base pair *in trans* induces a characteristic bulged-out conformation of the 59 and 60 nt (47–49) that are not found here in contacts with the THUMPα module. As no other conserved nucleotides belong to the tRNA region predicted to be in contact with the THUMPα module, the conserved geometry of the T-loop and acceptor stem may be the most important element for recognition by the enzyme.

Figure 6 shows the superposition of $tRNA^{Asp}$ and the docking solutions that fulfill the imposed constraints. Together, they form a 'fuzzy' model of PAB1283-tRNA interactions, which should not be interpreted at the atomic level, but

Figure 6. A docking model of the two domains from PAB1283 onto $tRNA^{Asp}$. The tRNA backbone is shown as a white tube, while the functionally important nucleosides are shown in the wireframe representation and colored: the methylation target G10 in cyan, the T-arm in yellow and the acceptor stem in violet. The conserved G53–C61 pair is in magenta. The AdoMet molecule is shown in white. The models of PAB1283 domains are shown in the cartoon representation, colored by the progression of sequence—from blue (N-terminus of the N-terminal domain) to red (C-terminus of the C-terminal domain). In the current model the C-terminus of the N-terminal domain and the N-terminus of the C-terminal domain (in green) are not connected—this region is predicted to be flexible and its conformation should be regarded as undefined. Only one variant of each domain is shown for the clarity of presentation, the superposition of all selected solutions is available as a Supplementary Data. The alternative orientation of the THUMPα module corresponds to a very minor rotation, while the alternative orientation of the MTase domain corresponds to ~180° rotation around the axis defined by the methylated base and the methyl group donor.

provides the first educated guess of the mutual localizations of domains from the enzyme and the substrate molecule. This docking model is compatible with the scheme that the two structural domains of TrMet(m2,2G10) are sideways the central 3D-core of tRNA.

DISCUSSION

The existence of a new putative RNA-binding domain, abbreviated THUMP, was deduced from sequence comparison of various proteins known or predicted to be involved in RNA metabolism (23). For a long time, the THUMP domain remained essentially uncharacterized. Recently, new experimental data have been collected on three different THUMP-containing enzymes: ThiI from *E.coli*, Tan1 from *S.cerevisiae* and TrMet(m2,2G10) from *P.abyssi*. All three proteins are involved in tRNA modification and target nucleosides in the central 3D-core of the tRNA molecule. Thus, we proposed that THUMP may interact with a specific region of tRNA and target the catalytic domains of these various enzymes towards the common region of the substrate (9). Recently, two structures of ThiI family members have been solved: PH1313 from *P.horikoshii* (without any published analysis) and BA4899 from *B.anthracis* strain Ames (28). However, none of these proteins have been biochemically characterized and actually their activity remains putative [PH1313 is even predicted to be catalytically inactive, (28)]. Thus, the experimental validation that THUMP interacts with tRNA and can be called a RNA-binding domain was needed.

In this paper, we first defined by limited proteolysis and mass spectrometry the boundaries of domains in PAB1283, a prototype member of the TrMet(m2,2G10) family. We then purified the N-terminal region [1–155] of PAB1283, which contains the THUMP domain, as a standalone protein. This autonomously folding unit was characterized as a soluble monomeric protein showing only a very weak affinity for tRNA in contrast to the entire PAB1283 protein.

In order to gain insight into the structure of this domain and understand how it functions concomitantly with the C-terminal catalytic domain, we proposed an original modeling approach. It is now well established that *in silico* modeling can be a fast approach to predict the structure of a protein or a macromolecular complex. However, theoretical methods generate models that need experimental validation, and also the number and diversity of the proposed solutions is often considerable. A promising strategy is the use of experimental data for model discrimination or refinement (50,51). Thus, we combined the experimental and theoretical analyses at three stages, corresponding to characterization of the primary structure (domain boundaries in the PAB1283 sequence), tertiary structure (3D fold of the THUMPα module), and quaternary structure (interactions between the two protein domains and the tRNA). First, it is satisfying to find that the experimentally determined boundaries of both domains from PAB1283 parallel those predicted by our FR analysis. Second, to distinguish among various threading models of THUMPα, we determined which residues are solvent accessible with a set of three labeling reagents and identified a disulfide bridge between the two cysteines found in the THUMPα sequence. These constraints allowed us to validate the hypothesis that the N-terminal structural module of PAB1283 is related to the equivalent module observed in the structure of ThiI (28), and that both proteins share not only the easily detectable THUMP domain, but also the strongly diverged NFLD domain (Figure 5). Therefore the role of the NFLD domain may not be specific to ThiI, as previously suggested (28). Our model shows that THUMPα by itself does not present any particular concentration of positively charged residues, although its conserved residues map on the surface corresponding to the ThiI homolog (28). These results suggest that the THUMP domain is not

necessarily responsible *per se* for the affinity of TrMet(m2,2G10) for tRNA but rather may be used to target the catalytic domain to a particular region of the tRNA structure. The results of our experiments show that the linker between THUMPα and the catalytic domain of PAB1283 is flexible. As it is highly basic, it may participate in nucleic acid binding.

Finally, we constructed a docking model of the two domains of PAB1283 onto the tRNA substrate (Figure 6 and Supplementary Data), based on experimental restraints from interactions between different elements of both molecules. This docking model confidently places the catalytic MTase domain of PAB1283 near the G10 nucleoside, at the junction between the anticodon stem and the D-stem, and the THUMPα domain at the co-axially stacked helices of the T-arm and the acceptor stem. Most of the interactions between the THUMPα domain and the tRNA occur through the phosphate backbone, suggesting that it is the RNA structure rather than the sequence that is recognized by THUMPα. Recently, it was found that the T-arm is an essential specificity determinant for PAB1283 (21). Likewise, the minimal substrate for the tRNA:s^4U8 synthase ThiI from *E.coli* is an RNA mini-helix comprising the acceptor and T-stems (26). Thus, the specificity determinants of *E.coli* ThiI and TrMet(m2,2G10) appear to be strikingly similar, despite the fact that these enzymes use unrelated catalytic domains to carry out completely different reactions. If the C-terminal domain is removed from our docking model of PAB1283, the site of U8 is relatively exposed, so that the catalytic domain of the thiouridine synthase could be modeled to bind tRNA in a similar manner to that predicted for the MTase domain of PAB1283. In order that both enzymes carry out their respective reactions, some conformational rearrangement, such as base flipping (52) in the substrate tRNA must occur, to expose the target nucleoside and place it in the active site. Likewise, the Tan1 protein required for N^4-acetylcytidine formation at position 12 appears to contain not only the THUMP domain, but also the N-terminal extension that exhibits a similar structure to the NFLD domain (J. M. Bujnicki, unpublished data). Thus, we postulate that ThiI thiolase, TrMet(m2,2G10) and the enzymes responsible for N^4-acetylcytidine formation at position 12 possess a common THUMPα module [comprising the NFLD and THUMP (sub)domains] that should bind tRNA in a very similar manner.

In conclusion, our docking model provides a useful platform for future studies and can be tested experimentally. In particular, footprinting and cross-linking analyses could provide specific restraints for inter-residue distances that could be used to refine the current model. THUMP-domain containing protein Tan1 from *S.cerevisiae* is required for the formation of N4-acetylcytidine at position 12 (27), a nucleoside, which is located just between G10 and U8. It is of interest to determine what could be the minimal substrate for this enzyme. Interestingly, it remains to be determined if other tRNA-modifying enzymes that contain the THUMP domain interact with the substrate in a similar way as TrMet(m2,2G10), i.e. structured tRNA as target and the co-axially stacked helices of the T-arm and the acceptor stem as unmodified support. Such hypothesis may be the basis of original strategies to search for the function of these still uncharacterized THUMP-containing proteins.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Jaunius Urbonavicius (CNRS-LEBS, Gif-sur-Yvette, France) for stimulating discussions on THUMP-containing proteins and sharing information on TrMet(m2,2G10) minimal substrate prior to publication. We are indebted to Martine Roovers and Louis Droogmans (both from Université Libre de Bruxelles, Belgium) for sharing information on PsuX prior to publication. The authors thank the following colleagues (both from CEA VALRHO, DSV-DIEP-SBTN): Charles Marchetti for operating fermentor facilities and Jean-Charles Gaillard for assistance with Edman sequencing. J.M.B., M.J.G. and I.T. were supported by the Polish Ministry of Science (grant PBZ-KBN-088/P04/2003). J.M.B. is an EMBO/HHMI Young Investigator. H.G. was supported by a research grant from the CNRS (Programme Interdépartemental de Géomicrobiologie des Environnements Extrêmes GEOMEX 2002–2004). Funding to pay the Open Access publication charges for this article was provided by Commissariat à l'Energie Atomique, DSV-DIEP-SBTN.

Conflict of interest statement. None declared.

REFERENCES

1. Rozenski,J., Crain,P.F. and McCloskey,J.A. (1999) The RNA Modification Database: 1999 update. *Nucleic Acids Res*., **27**, 196–197.
2. Agris,P.F. (2004) Decoding the genome: a modified view. *Nucleic Acids Res*., **32**, 223–238.
3. Grosjean,H. and Benne,R. (1998) *Modification and editing of RNA*. ASM Press, Washington, DC.
4. Grosjean,H. (2005) *Fine-Tuning of RNA Functions by Modification and Editing*. Springer Verlag, Berlin-Heidelberg, NY.
5. Hopper,A.K. and Phizicky,E.M. (2003) tRNA transfers to the limelight. *Genes Dev*., **17**, 162–180.
6. Björk,G.R. and Hagervall,T.G. (2005) Transfer RNA modification. In Curtiss,R.,IIIrd, Böck,A., Ingrahan,J.L., Kaper,J.B., Maloy,S., Neidhardt,F.C., Riley,M.M., Squires,C.L. and Wanner,B.L. (eds), *Escherichia coli and Salmonella. Cellular and Molecular Biology*. ASM Press, Washington DC, pp. 4.6.2.
7. Dunin-Horkawicz,S., Czerwoniec,A., Gajda,M.J., Feder,M., Grosjean,H. and Bujnicki,J.M. (2006) MODOMICS: a database of RNA modification pathways. *Nucleic Acids Res*., **34**, D145–D149.
8. Reader,J.S., Metzgar,D., Schimmel,P. and de Crecy-Lagard,V. (2004) Identification of four genes necessary for biosynthesis of the modified nucleoside queuosine. *J. Biol. Chem*., **279**, 6280–6285.
9. Armengaud,J., Urbonavicius,J., Fernandez,B., Chaussinand,G., Bujnicki,J.M. and Grosjean,H. (2004) N2-methylation of guanosine at position 10 in tRNA is catalyzed by a THUMP domain-containing, S-adenosylmethionine-dependent methyltransferase, conserved in Archaea and Eukaryota. *J. Biol. Chem*., **279**, 37142–37152.
10. Bujnicki,J.M., Droogmans,L., Grosjean,H., Purushothaman,S.K. and Lapeyre,B. (2004) Bioinformatics-guided identification and experimental characterization of novel RNA Methyltransferases. In Bujnicki,J.M. (ed.), *Practical Bioinformatics*. Springer-Verlag, Berlin Heidelberg, Vol. 15, pp. 139–168.
11. Urbonavicius,J., Skouloubris,S., Myllykallio,H. and Grosjean,H. (2005) Identification of a novel gene encoding a flavin-dependent tRNA:m5U methyltransferase in bacteria–evolutionary implications. *Nucleic Acids Res*., **33**, 3955–3964.
12. Crecy-Lagard,V. (2004) Finding missing tRNA modification genes: a comparative genomics goldmine. In Bujnicki,J.M. (ed.), *Nucleic Acids and Molecular Biology Series*, *Vol. 15*, *Practical Bioinformatics*. Springer-Verlag, Berlin Heidelberg, pp. 169–190.

13. Aravind,L. and Koonin,E.V. (1999) Novel predicted RNA-binding domains associated with the translation machinery. *J. Mol. Evol.*, **48**, 291–302.
14. Anantharaman,V., Koonin,E.V. and Aravind,L. (2002) Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res.*, **30**, 1427–1464.
15. Liu,Y. and Santi,D.V. (2000) m5C RNA and m5C DNA methyl transferases use different cysteine residues as catalysts. *Proc. Natl Acad. Sci. USA*, **97**, 8263–8265.
16. Watanabe,K., Hori,H. and Endo,Y. (2001) Identification of essential amino acid residues of tRNA (Gm18)methyltransferase for methyl-transfer activity. *Nucleic Acids Res.*, Suppl. **1**, 33–34.
17. Watanabe,K., Nureki,O., Fukai,S., Ishii,R., Okamoto,H., Yokoyama,S., Endo,Y. and Hori,H. (2005) Roles of conserved amino acid sequence motifs in the SpoU (TrmH) RNA methyltransferase family. *J. Biol. Chem.*, **280**, 10368–10377.
18. Purta,E., van Vliet,F., Tricot,C., De Bie,L.G., Feder,M., Skowronek,K., Droogmans,L. and Bujnicki,J.M. (2005) Sequence-structure-function relationships of a tRNA (m7G46) methyltransferase studied by homology modeling and site-directed mutagenesis. *Proteins*, **59**, 482–488.
19. Hoang,C., Hamilton,C.S., Mueller,E.G. and Ferre-D'Amare,A.R. (2005) Precursor complex structure of pseudouridine synthase TruB suggests coupling of active site perturbations to an RNA-sequestering peripheral protein domain. *Protein Sci.*, **14**, 2201–2206.
20. Sabina,J. and Soll,D. (2006) The RNA-binding PUA domain of archaeal tRNA-guanine transglycosylase is not required for archaeosine formation. *J. Biol. Chem.*, **281**, 6933–7001.
21. Urbonavicius,J., Armengaud,J. and Grosjean,H. (2006) Identity elements required for enzymatic formation of N2, N2-dimethylguanosine and its role in avoiding alternative conformations in archaeal tRNAs. *J. Mol. Biol.*, **357**, 387–399.
22. Purushothaman,S.K., Bujnicki,J.M., Grosjean,H. and Lapeyre,B. (2005) Trm11p and Trm112p are both required for the formation of 2-methylguanosine at position 10 in yeast tRNA. *Mol. Cell. Biol.*, **25**, 4359–4370.
23. Aravind,L. and Koonin,E.V. (2001) THUMP-a predicted RNA-binding domain shared by 4-thiouridine, pseudouridine synthases and RNA methylases. *Trends Biochem. Sci.*, **26**, 215–217.
24. Palenchar,P.M., Buck,C.J., Cheng,H., Larson,T.J. and Mueller,E.G. (2000) Evidence that ThiI, an enzyme shared between thiamin and 4-thiouridine biosynthesis, may be a sulfurtransferase that proceeds through a persulfide intermediate. *J. Biol. Chem.*, **275**, 8283–8286.
25. Mueller,E.G., Palenchar,P.M. and Buck,C.J. (2001) The role of the cysteine residues of ThiI in the generation of 4-thiouridine in tRNA. *J. Biol. Chem.*, **276**, 33588–33595.
26. Lauhon,C.T., Erwin,W.M. and Ton,G.N. (2004) Substrate specificity for 4-thiouridine modification in *Escherichia coli*. *J. Biol. Chem.*, **279**, 23022–23029.
27. Johansson,M.J. and Byström,A.S. (2004) The *Saccharomyces cerevisiae* TAN1 gene is required for N(4)-acetylcytidine formation in tRNA. *RNA*, **10**, 712–719.
28. Waterman,D.G., Ortiz-Lombardia,M., Fogg,M.J., Koonin,E.V. and Antson,A.A. (2006) Crystal structure of *Bacillus anthracis* ThiI, a tRNA-modifying enzyme containing the predicted RNA-binding THUMP domain. *J. Mol. Biol.*, **356**, 97–110.
29. Armengaud,J., Fernandez,B., Chaumont,V., Rollin-Genetet,F., Finet,S., Marchetti,C., Myllykallio,H., Vidaud,C., Pellequer,J.L., Gribaldo,S. *et al.* (2003) Identification, purification, and characterization of an eukaryotic-like phosphopantetheine adenylyltransferase (coenzyme A biosynthetic pathway) in the hyperthermophilic archaeon *Pyrococcus abyssi*. *J. Biol. Chem.*, **278**, 31078–31087.
30. Auxilien,S., Crain,P.F., Trewyn,R.W. and Grosjean,H. (1996) Mechanism, specificity and general properties of the yeast enzyme catalysing the formation of inosine 34 in the anticodon of transfer RNA. *J. Mol. Biol.*, **262**, 437–458.
31. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
32. Kurowski,M.A. and Bujnicki,J.M. (2003) GeneSilico protein structure prediction meta-server. *Nucleic Acids Res.*, **31**, 3305–3307.
33. Kosinski,J., Cymerman,I.A., Feder,M., Kurowski,M.A., Sasin,J.M. and Bujnicki,J.M. (2003) A 'FRankenstein's monster' approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation. *Proteins*, **53**, 369–379.
34. Kosinski,J., Gajda,M.J., Cymerman,I.A., Kurowski,M.A., Pawlowski,M., Boniecki,M., Obarska,A., Papaj,G., Sroczynska-Obuchowicz,P., Tkaczuk,K.L. *et al.* (2005) FRankenstein becomes a cyborg: the automatic recombination and realignment of Fold-Recognition models in CASP6. *Proteins*, **61**, 106–113.
35. Fiser,A. and Sali,A. (2003) Modeller: generation and refinement of homology-based protein structure models. *Meth. Enzymol.*, **374**, 461–491.
36. Luthy,R., Bowie,J.U. and Eisenberg,D. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.
37. Sasin,J.M. and Bujnicki,J.M. (2004) COLORADO3D, a web server for the visual analysis of protein structures. *Nucleic Acids Res.*, **32**, W586–W589.
38. Pupko,T., Bell,R.E., Mayrose,I., Glaser,F. and Ben-Tal,N. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**, S71–S77.
39. Katchalski-Katzir,E., Shariv,I., Eisenstein,M., Friesem,A.A., Aflalo,C. and Vakser,I.A. (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl Acad. Sci. USA*, **89**, 2195–2199.
40. Romero,P., Obradovic,Z., Li,X., Garner,E.C., Brown,C.J. and Dunker,A.K. (2001) Sequence complexity of disordered protein. *Proteins*, **42**, 38–48.
41. Bowie,J.U., Luthy,R. and Eisenberg,D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.
42. Mallick,P., Boutz,D.R., Eisenberg,D. and Yeates,T.O. (2002) Genomic evidence that the intracellular proteins of archaeal microbes contain disulfide bonds. *Proc. Natl Acad. Sci. USA*, **99**, 9679–9684.
43. 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 (m1A) methyltransferase. *Nucleic Acids Res.*, **32**, 465–476.
44. Kaplan,H., Stevenson,K.J. and Hartley,B.S. (1971) Competitive labelling, a method for determining the reactivity of individual groups in proteins. *Biochem. J.*, **124**, 289–299.
45. Bresciani,D. (1977) Different reactivities of free and bound amino groups in deoxy-and liganded haemoglobin. *Biochem. J.*, **163**, 393–395.
46. Westhof,E., Dumas,P. and Moras,D. (1988) Restrained refinement of two crystalline forms of yeast aspartic acid and phenylalanine transfer RNA crystals. *Acta Crystallogr. A.*, **44**, 112–123.
47. Westhof,E., Dumas,P. and Moras,D. (1983) Loop stereochemistry and dynamics in transfer RNA. *J. Biomol. Struct. Dyn.*, **1**, 337–355.
48. Romby,P., Carbon,P., Westhof,E., Ehresmann,C., Ebel,J.P., Ehresmann,B. and Giege,R. (1987) Importance of conserved residues for the conformation of the T-loop in tRNAs. *J. Biomol. Struct. Dyn.*, **5**, 669–687.
49. Becker,H.F., Motorin,Y., Sissler,M., Florentz,C. and Grosjean,H. (1997) Major identity determinants for enzymatic formation of ribothymidine and pseudouridine in the T psi-loop of yeast tRNAs. *J. Mol. Biol.*, **274**, 505–518.
50. Ye,X., O'Neil,P.K., Foster,A.N., Gajda,M.J., Kosinski,J., Kurowski,M.A., Bujnicki,J.M., Friedman,A.M. and Bailey-Kellogg,C. (2004) Probabilistic cross-link analysis and experiment planning for high-throughput elucidation of protein structure. *Protein Sci.*, **13**, 3298–3313.
51. Armengaud,J., Dedieu,A., Solques,O., Pellequer,J.L. and Quemeneur,E. (2005) Deciphering structure and topology of conserved COG2042 orphan proteins. *BMC Struct. Biol.*, **5**, 3.
52. Roberts,R.J. and Cheng,X. (1998) Base flipping. *Annu. Rev. Biochem.*, **67**, 181–198.