FEBS Letters 569 (2004) 18–26 FEBS 28487

Novel Sm-like proteins with long C-terminal tails and associated methyltransferases ☆

Mario Albrecht*, Thomas Lengauer

Max-Planck-Institute for Informatics, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany

Received 23 December 2003; revised 11 March 2004; accepted 15 March 2004

Available online 31 May 2004

Edited by Robert B. Russell

Abstract **Sm and Sm-like proteins of the Lsm (like Sm) domain family are generally involved in essential RNA-processing tasks. While recent research has focused on the function and structure of small family members, little is known about Lsm domain proteins carrying additional domains. Using an integrative bioinformatics approach, we discovered five novel groups of Lsm domain proteins (Lsm12-16) with long C-terminal tails and investigated their functions. All of them are evolutionarily conserved in eukaryotes with an N-terminal Lsm domain to bind nucleic acids followed by as yet uncharacterized C-terminal domains and sequence motifs. Based on known yeast interaction partners, Lsm12-16 may play important roles in RNA metabolism. Particularly, Lsm12 is possibly involved in mRNA degradation or tRNA splicing, and Lsm13-16 in the regulation of the mitotic G2/M phase. Lsm16 proteins have an additional C-terminal YjeF_N domain of as yet unknown function. The identification of an additional methyltransferase domain at the C-terminus of one of the Lsm12 proteins also led to the recognition of three new groups of methyltransferases, presumably dependent on *S*-adenosyl-L-methionine. Further computational analyses revealed that some methyltransferases contain putative RNA-binding helix–turn–helix domains and zinc fingers.**

Keywords: Lsm domain; Methyltransferase; RNA metabolism; Fold recognition; Function prediction

1. Introduction

Sm and Sm-like proteins of the RNA-binding Lsm (like Sm) domain family are found in all domains of life and are generally involved in important RNA-processing tasks [1,2]. Some of them are main components of the spliceosomal small nuclear ribonucleoproteins (snRNPs) and form cyclic hetero- or homo-oligomers, which preferentially bind uridine-rich, single-stranded snRNA. Lsm domain proteins have also been observed to function in hypermethylation of snRNA caps and mRNA splicing, decapping, and degradation [3]. Lsm proteins can be part of large complexes for distinct processes in the cytoplasm and nucleus. For instance, the cytosolic Lsm1-7 complex functions in mRNA decay, while a nuclear Lsm2-8 complex works in pre-mRNA splicing [4].

While current research has focused on the function and structure of small Sm and Sm-like proteins (not more than about 150 residues), little is known about large Lsm domain proteins carrying additional domains. A recently described Lsm11 protein possesses an N-terminal extension and has been found to be involved in histone mRNA processing [5]. Two very long Lsm domain proteins with both N- and C-terminal sequence prolongations are human ataxin-2 (1312 residues) and its yeast homolog PBP1 (PAB1-binding protein 1) [6,7]. Ataxin-2 is still of unknown function, but a polyglutamine expansion in the N-terminal region of ataxin-2 is causative of the inherited neurodegenerative disease spinocerebellar ataxia type 2 (SCA2) [8]. The C-terminal tail of PBP1 has been observed in experiment to bind to the C-terminal PABC domain of poly(A)-binding protein (PABP) and to regulate polyadenylation in mRNA splicing [9]. The absence of PBP1 leads to incomplete poly(A) tails, although the 3′-end of pre-mRNAs is properly cleaved. Interestingly, in contrast to PBP1, ataxin-2 contains a conserved PABP interacting motif in the C-terminal tail, which is, however, not yet investigated experimentally [10].

Based on a comprehensive bioinformatics analysis, we found at least five novel and evolutionarily conserved Lsm domain proteins with long C-terminal tails, containing further domains and highly conserved sequence motifs (Fig. 1). The additional methyltransferase domain of one of the Lsm domain homologs also led to the identification of three as yet uncharacterized groups of methyltransferases (MTases), presumably dependent on *S*-adenosyl-L-methionine (AdoMet) [11]. Using yeast interaction data, we could assign putative functions to some of the new Lsm domain proteins.

2. Materials and methods

We employed an integrative bioinformatics approach combining sequence and domain database searches with protein interaction data and the consensus of prediction results from fold recognition servers. We obtained protein sequences from the SWISS-PROT/TrEMBL (SPTrEMBL) [12] and the NCBI non-redundant databases [13] and searched the domain databases Pfam [14], SMART [15], and CDD [16]. The GRID resource [33] and the GeneDB database [34] provided information on yeast protein interactions for *Saccharomyces cerevisiae* and *Schizosaccharomyes pombe*, respectively. Species names are abbreviated by first letters, see Web Table I. We searched the sequence databases by means of the (PSI-)BLAST [17] and HMMER [18] suites of programs (using the *E*-value cut-offs 0.005 and 0.5, respectively). We

☆ Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2004.03.126.

* Corresponding author.
E-mail addresses: mario.albrecht@mpi-sb.mpg.de (M. Albrecht), thomas.lengauer@mpi-sb.mpg.de (T. Lengauer).

0014-5793/$22.00
doi:10.1016/j.febslet.2004.03.126

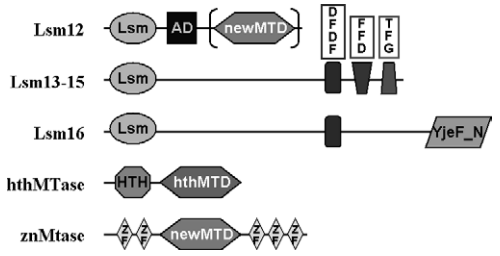

Fig. 1. Domain architectures of newly identified Lsm domain proteins and associated methyltransferases.

computed multiple sequence alignments using CLUSTAL W [19] and T-COFFEE [20] and improved them manually in some relevant regions. We clustered homologous sequences manually into groups of proteins based on full-length sequence similarity and common domains and sequence motifs. The multiple alignments shown in figures were prepared in the SEAVIEW editor [21] and illustrated by the ESPript online service [22].

We also explored the structure prediction summaries provided by the BioInfo.PL meta-server [23], which collects and evaluates the results of a dozen fold recognition servers [24]. Additionally, we investigated the suggested PDB template structures [25] with the help of the SCOP database [26] and compared the DSSP secondary structure assignment [27] to the secondary structure predicted for newly identified proteins by the PSIPRED server [28]. Alternative PDB identifiers and corresponding SPTrEMBL accession nos. are given in Web Table II. A single capital letter appended to the PDB identifier denotes the chosen structure chain. For the prediction of intrinsically unstructured, that is, natively unfolded, regions in proteins, we used the consensus of the DisEMBL [29], GlobProt [30], NORSp [31] and PONDR [32] servers. Note that the online version of this article contains supplementary material, which also provides further details on sequence fragments and omitted SPTrEMBL sequences related to the sequences shown in Figs. 2–4, 7.

3. Results and discussion

3.1. Novel Lsm domain protein Lsm12

The GRID database indicates that the yeast proteins PBP1/YGR178C and YHR121W interact. While iterated PSI-BLAST searches for YHR121W did not return any significant

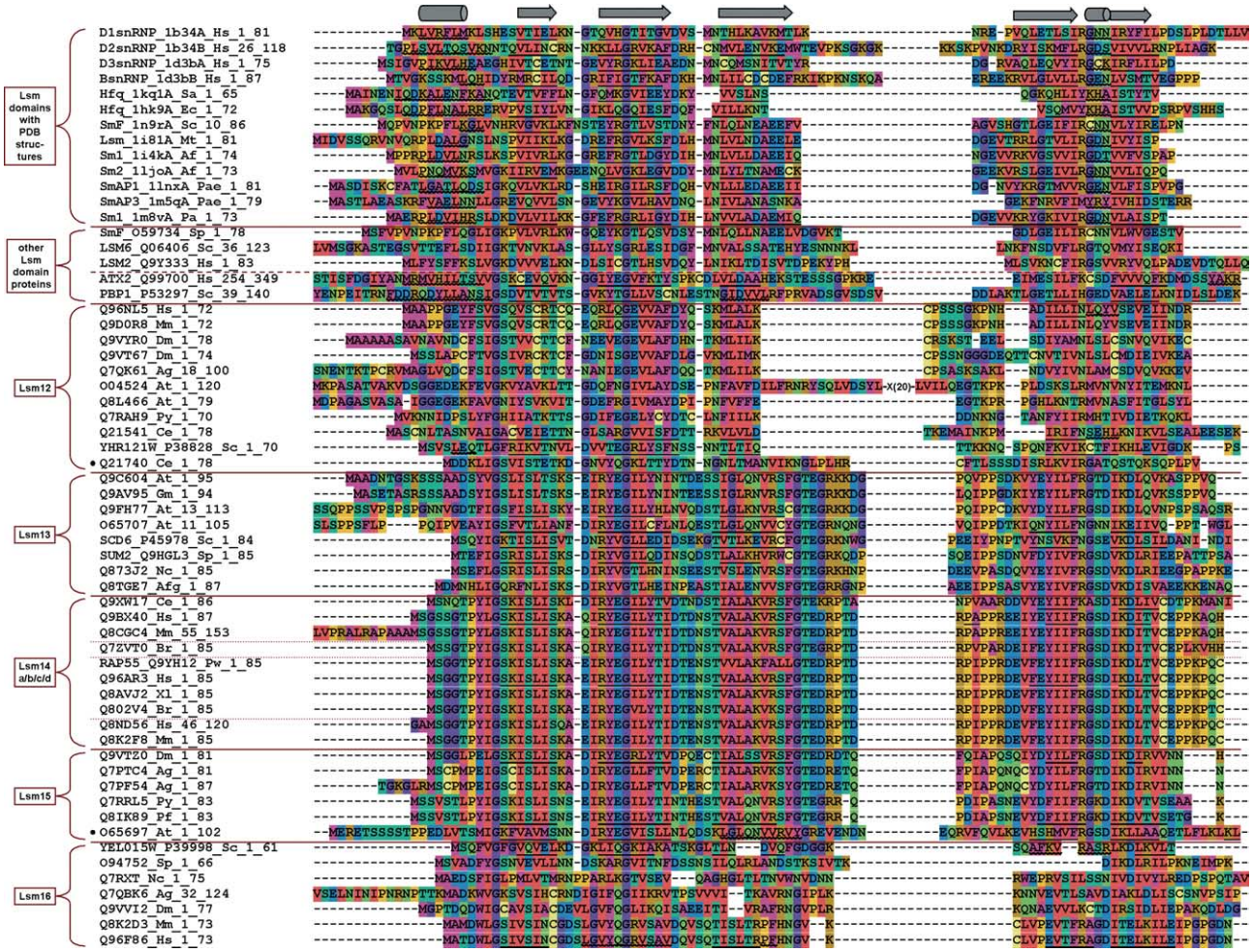

Fig. 2. Multiple sequence alignment of the Lsm12–16 protein families and other Lsm domain proteins such as the homologs ataxin-2 and PBP1 and Sm and Sm-like proteins whose crystallographic structures have been determined. As an example, the Lsm14 group is further subdivided into four putative subgroups a–d by dotted horizontal lines. The known DSSP secondary structure assignment of the snRNP protein D_1 is shown at the top of the alignment (cylinder for α-helix, arrow for β-strand), and the amino acid sequences of PDB structures are underlined accordingly (curled line for α-helix, straight line for β-strand). The corresponding PSIPRED secondary structure predictions for selected Lsm domain proteins is also underlined. The depicted second α-helix of the snRNP D_1 is actually a less conserved 3_{10} helical turn. Physico-chemically similar amino acids are colored identically. The highly conserved glycines characteristic of Lsm domains are in violet. Alternative PDB identifiers and corresponding SPTrEMBL accession nos. are given in Web Table II. The *C. elegans* protein R05D11.8 and the *A. thaliana* protein T5K18.40 are marked by solid circles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

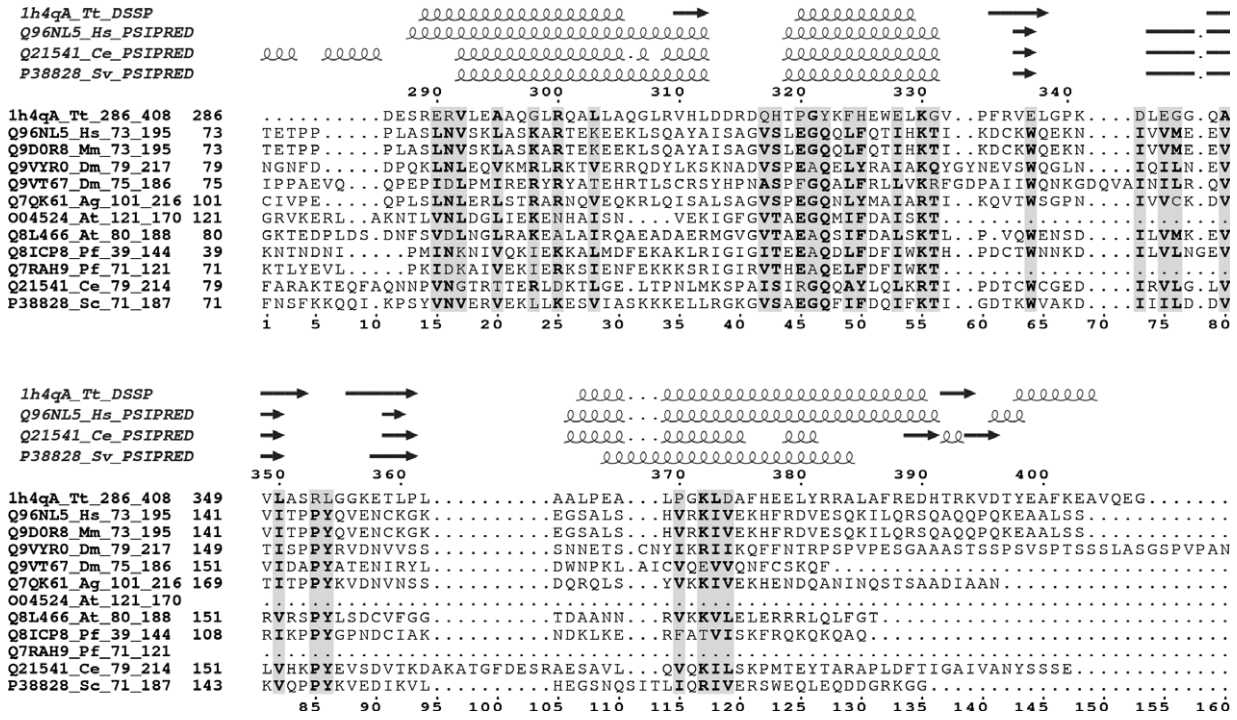

Fig. 3. Multiple sequence alignment of the domain following the N-terminal Lsm domain of Lsm12 proteins. The alignment includes the anticodon-binding domain of a prolyl-tRNA synthetase, whose PDB structure is available under the identifier 1h4q. The DSSP secondary structure of 1h4q and the corresponding PSIPRED predictions for selected Lsm12 members are drawn on top of each alignment row (curled lines for α-helices, arrows for β-strands). Alignment columns in which more than 65% of the residues are physico-chemically equivalent are highlighted in light gray boxes. The residues of 1h4q (chain A) are numbered at the top, the alignment columns at the bottom of each row.

hit to known Sm and Sm-like proteins, the consensus of fold recognition servers contacted through the Bioinfo.PL meta-server predicted an N-terminal Lsm domain for YHR121W with strong confidence scores. This finding is in agreement with the predicted secondary structure and highly conserved glycines that are characteristic of Lsm domain structures (Fig. 2) [1,2,35–45]. The amide groups of the glycines stabilize the protein fold by forming hydrogen bonds to neighboring β-strands [36]. Like many other Lsm domain proteins such as the snRNP proteins B and D_3 [36], PBP1 and YHR121W may also form heteromers. A subsequent PSI-BLAST search with the complete sequence of YHR121W revealed that it is an evolutionarily conserved protein in eukaryotes (Fig. 2), which we named Lsm12. We also found that other PBP1 interaction partners contained in the GRID database appear not to have Lsm domains.

Lsm12 has an N-terminal Lsm domain followed by another domain termed AD (Fig. 3) that may function in RNA-binding based on the predicted anticodon-binding fold of tRNA synthetases [46]. In agreement with these observations, Lsm12/YHR121W has been found in association with the transcription factor STB5/YHR178W, the mRNA guanylyltransferase CEG1/YGL130W, and the mRNA-binding protein PUF3/YLL013C. STB5 interacts with the transcriptional repressor SIN3/YOL004W [47]. CEG1 forms a heteromeric mRNA capping enzyme complex with the triphosphatase CET1/YPL228W. This enzyme complex removes the γ-phosphate from the 5′-end of mRNA and transfers GMP from GTP to the resulting mRNA diphosphate end before an RNA methyltransferase finally modifies this guanine cap [48]. PUF3 contains an RNA-recognition motif RRM and is a homolog of PUF1, which is functionally related to the reverse ORF of another newly identified Lsm domain protein SCD6/Lsm13 as detailed below. In particular, PUF3 works as a transcript-specific regulator of mRNA degradation by enhancing its rate of deadenylation [49].

We also found that an uncharacterized *Caenorhabditis elegans* protein R05D11.8 (SPTrEMBL: Q21740) has an N-terminal Lsm domain very similar to the Lsm12 domain, but its long C-terminal tail of almost 500 residues does not exhibit significant homology to other proteins or known domains.

3.2. New methyltransferases

Another *C. elegans* homolog (SPTrEMBL: Q21541) in the Lsm12 group carries a new C-terminal methyltransferase (MTase) domain (newMTD) consistently predicted by fold recognition servers (Fig. 4). This newMTD domain is included in the Pfam-B_5894 domain cluster, which revealed further MTases that consist solely of the newMTD domain. Domain databases searches also returned weak, but significant hits (*E*-values between 0.015 and 0.62) of the *C. elegans* ortholog to the RNA methyltransferase families MT-A70, UPF0020, and COG1568, and less significant hits to other RNA methylase families [11,50]. The closest search hit to an as yet uncharacterized archaebacterial Pfam protein family DUF43 had the significant *E*-value 0.009. Fold recognition servers also predict the methyltransferase domain fold for the C-terminal part of DUF43 proteins (Fig. 5).

Presumably, the DUF43 members possess an N-terminal 'winged' helix–turn–helix (HTH) RNA-binding domain instead of the Lsm domain, because structure prediction servers proposed MarR/SlyA-like transcriptional regulators as closest

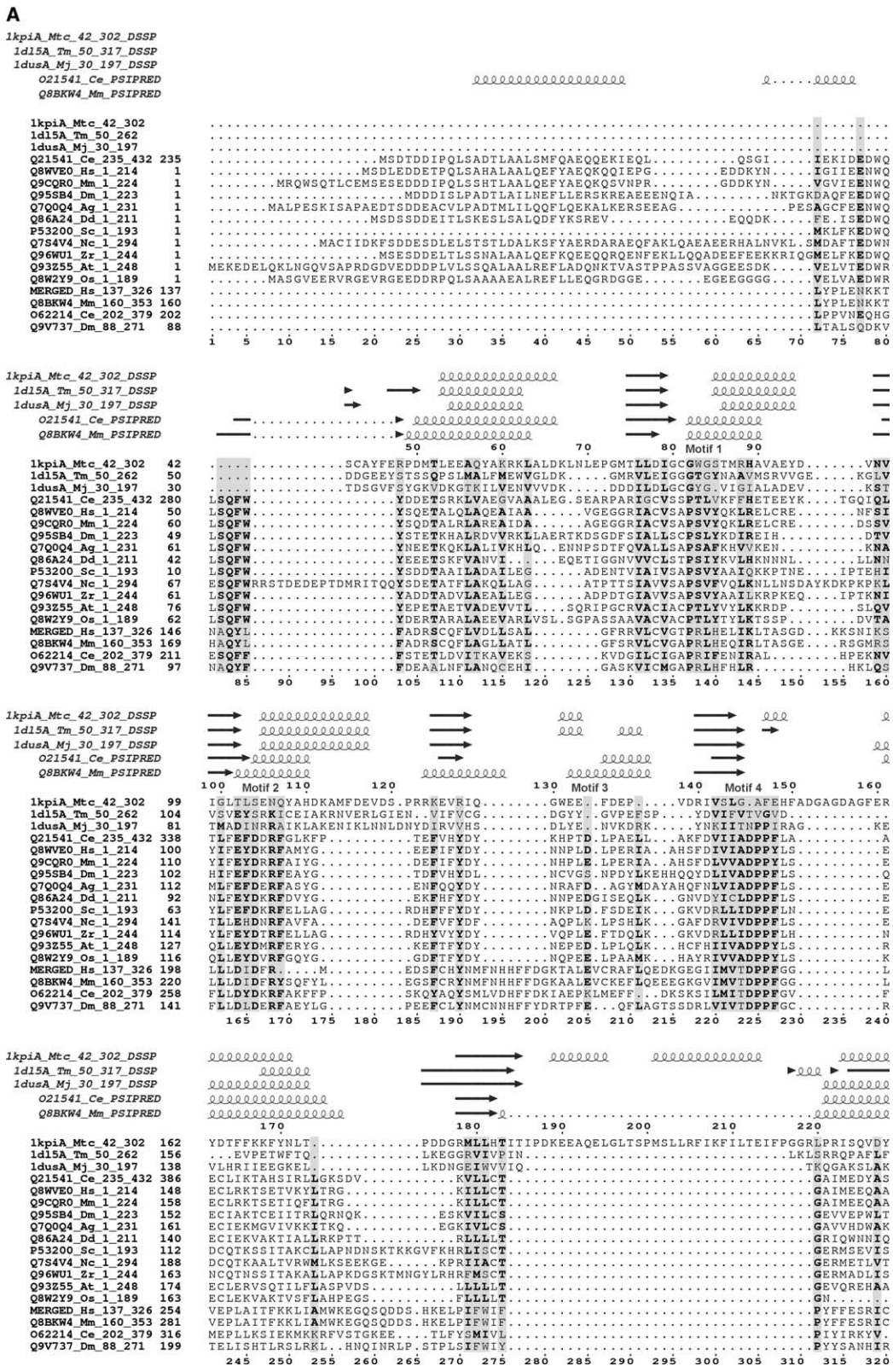

Fig. 4. Multiple sequence alignment of the C-terminal domain of a *C. elegans* Lsm12 protein (SPTrEMBL: Q21541) and other AdoMet-dependent methyltransferase domains, including the novel newMTases and znMTases (the latter including MERGED_Hs and all sequences below) and related MTase domain sequences of three representative PDB structures 1kpi, 1dl5 and 1dus. (A) and (B) depict the N- and C-terminal alignment parts, respectively. The DSSP secondary structure assignments and the corresponding PSIPRED predictions for selected proteins are drawn on top of each alignment row (curled lines for α-helices, arrows for β-strands). Alignment columns in which more than 65% of the residues are physico-chemically equivalent are highlighted in light gray boxes. Text labels point to the positions of four conserved sequence motifs 1–4 of functional relevance. The residues of 1kpi (chain A) are numbered at the top, the alignment columns at the bottom of each alignment row. Note that the human sequence fragments Q96AN7 (205 residues) and Q9H5U6 (279 residues) have been connected in the sequence MERGED_Hs based on the complete mouse ortholog Q8BKW4.

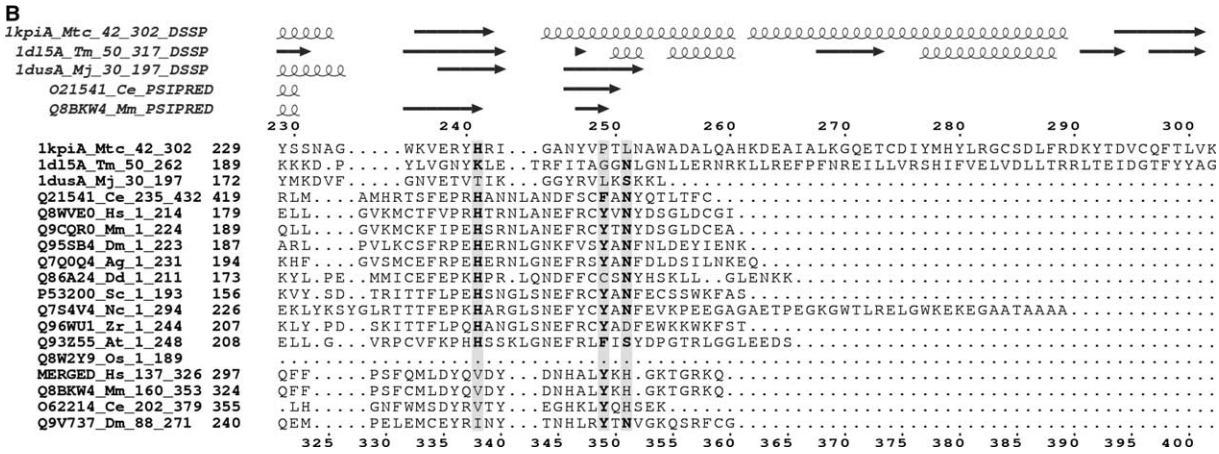

Fig. 4 (*continued*)

homologs [51] for this N-terminal domain (Fig. 6). Accordingly, we named the DUF43 members of this novel HTH-methyltransferase family hthMTases. This family includes the yeast protein YGR001C (SPTrEMBL: P53200) of unknown function. One particularly conserved sequence motif, a TEGK box, is shared by hthMTases and MarR/SlyA family members and is assumed to be important to position the HTH-domain onto the bound nucleic acids [51]. In contrast, two conspicuous cysteines found in all hthMTases seem not to be a feature of MarR/SlyA-like proteins, but may form a disulfide bond or a binding site for zinc or another molecule together with additional histidines next to them.

Both discovered newMTases and hthMTases share four very conserved sequence motifs 1–4 of functional relevance (Figs. 4 and 5). Particularly, sequence motif 4 is very similar to the N6-adenine-specific DNA-methylase Prosite signature N6_MTASE: [LIVMAC]–[LIVFYWA]–x–[DN]–PP–[FYW]. In addition, the best structural templates for both new methylase protein families are predicted to be members of the AdoMet-dependent methyltransferase (MTase) superfamily such as a mycolic acid cyclopropane synthase (PDB: 1kpi, SPTrEMBL: Q11196) [52], a protein isoaspartyl methyltransferase (PDB: 1dl5, SPTrEMBL: Q56308) [53], and another AdoMet-dependent methyltransferase named MJ0882 (PDB: 1dus, SPTrEMBL: Q58292) [54].

The described MTase superfamily includes the above-mentioned MT-A70 family [50], which shares the conserved AdoMet-binding site characterized by motifs 1–4 with Lsm- and hthMTases (Figs. 4 and 5) [50]. Motif 1 usually facilitates AdoMet-binding and contains more than one conserved glycine, but this appears not to be the case for our novel MTases. Acidic residues in motifs 2 and 3 form hydrogen bonds to the ribose hydroxyl and the adenine ring of AdoMet, respectively. Generally, methylation of adenosines in pre-mRNA may affect the efficiency of RNA processing and splicing reactions [55].

A PSI-BLAST search with the described *C. elegans* methyltransferase domain newMTD as seed sequence uncovered another as yet uncharacterized group of proteins with newMTD as centrally located domain containing the N6_MTASE Prosite motif (Fig. 4). An RNA/DNA-processing function is strongly supported by the fact that this protein group has additional C-terminal zinc fingers (ZFs) with DHHC- and CHHC-type zinc-binding motifs. Similarly, further sequence and domain searches as well as structure predictions for the N-terminal cysteine-rich domain also reported zinc-binding sites as found, for instance, in DNA topoisomerases. Thus, we call the related proteins znMTases (for zinc-binding MTases).

3.3. More Lsm domain proteins Lsm13–16

Using another HMM constructed from sequence profiles of ataxin-2 and Lsm12 homologs, we detected many known short Sm and Sm-like proteins from SPTrEMBL and at least four novel groups of Lsm domain proteins with long C-terminal tails. We numbered the corresponding yeast proteins Lsm13 to Lsm16 (Fig. 2), but some of them have more than one homolog in the same species due to putative gene duplication events, giving rise to subgroups as demonstrated exemplarily for at least four identifiable subgroups Lsm14a–d (Figs. 2 and 7). We call all Lsm13–16 homologs DFDF box Lsm domain proteins because of the common C-terminal consensus motif DFDF–x(7)–F of unknown function with predicted helical structure, closely preceded and followed by further phenylalanines and charged aspartates/glutamates and arginines/lysines/histidines (Fig. 7). The variable seven-residue tract of this consensus motif usually contains an asparagine at the third or fourth position except of one sequence where the asparagine is replaced by a glycine. In few other sequences, the DFDF box is replaced by a DYDF or EFDF box. We could not find DFDF box proteins without an N-terminal Lsm domain in repeated HMM searches using a growing sequence profile of proteins containing the DFDF box motif.

Two other strongly conserved FFD box and TFG box sequence motifs Y–x–K–x(3)–FFD–x–[IL]–S and [RKH]–x(2-5)–E–x(0-2)–[RK]–x(3-4)–[DE]–TFG contained in Lsm13–15, but not Lsm16, homologs succeed the DFDF–x(7)–F motif and are also predicted to be of helical nature (Fig. 7). An exception to the DFDF/FFD/TFG boxes is an uncharacterized *Arabidopsis thaliana* Lsm13 protein (SPTrEMBL: Q65707), whose corresponding motifs appear to be DFEA/SYK/AFG. All three described motifs may be linked structurally and functionally to the always co-occurring N-terminal Lsm domain. In addition, the N-terminal region of the DFDF–x(7)–F motif and the C-terminus of Lsm13–15 consist of RG(G) repeats that are indicative of RNA/DNA-binding [56] and are also found at the C-terminus of short Lsm proteins such as Lsm4 and snRNP core subtypes D1 and D3.

Apparently, the sequence regions between the Lsm domain and the highly conserved motifs are of quite variable length

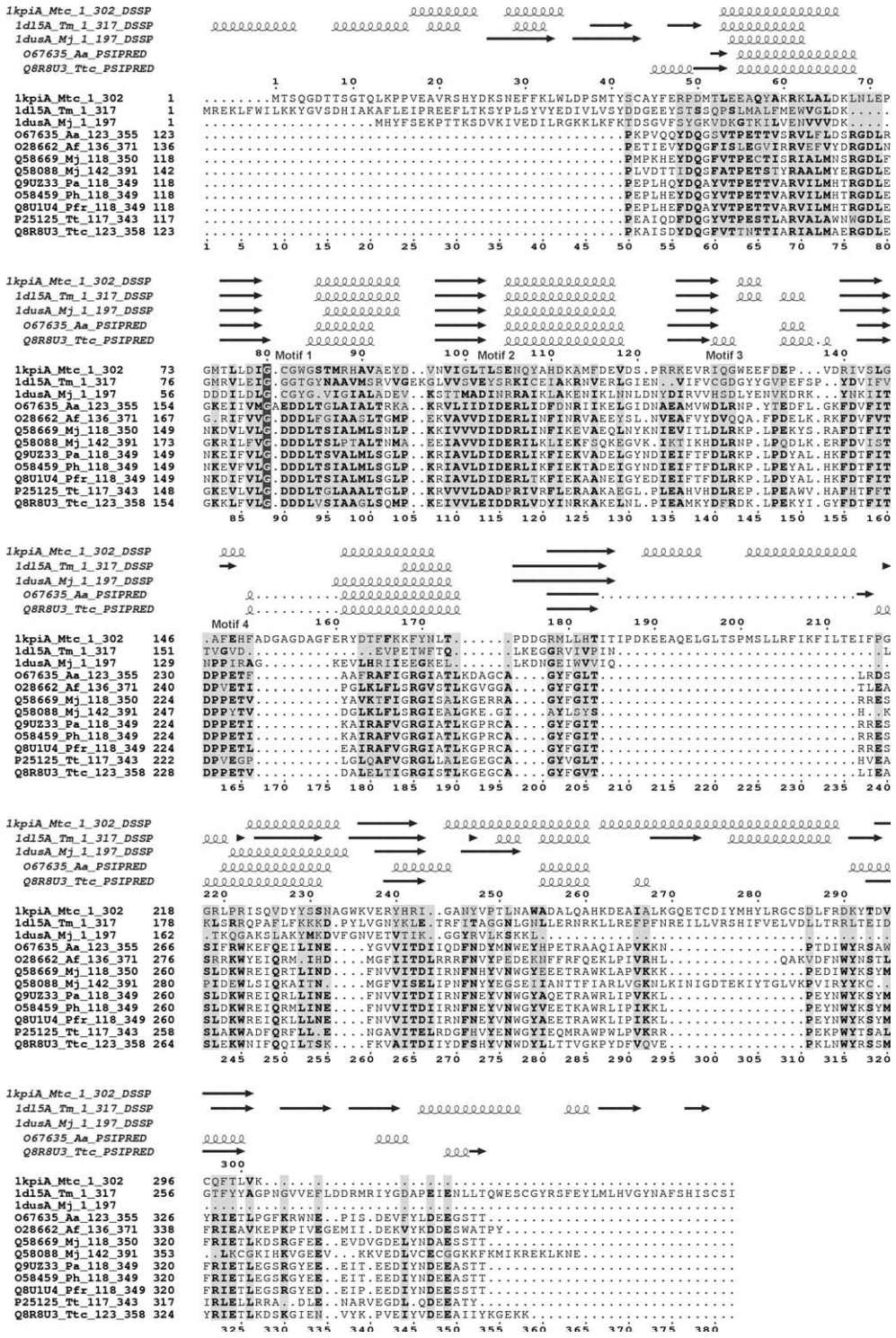

Fig. 5. Multiple sequence alignment of the C-terminal methyltransferase domains of the archaebacterial DUF43/hthMTase family and the MTase domain sequences of three representative PDB structures also used in Fig. 4. The DSSP secondary structure assignments and the corresponding PSIPRED predictions for selected proteins are drawn on top of each alignment row (curled lines for α-helices, arrows for β-strands). Alignment columns with strictly conserved residues are highlighted in dark gray boxes, those in which more than 65% of the residues are physico-chemically equivalent are shown in light gray boxes. Text labels point to the positions of four conserved sequence motifs 1–4 of functional relevance. The residues of 1kpi (chain A) are numbered at the top, the alignment columns at the bottom of each row.

and diverged, containing numerous short poly-P/-R/-S (1–3 residues) and long poly-G/-N/-Q sequence stretches (Web Fig. A). Thus, the C-terminal conserved motif boxes appear as stable islands in a large sea of intrinsically unstructured sequence regions [31]. Indeed, solely coil is predicted as secondary structure for the diverged regions, for which four prediction servers of disordered structure return similar results. Such observations can also be made for regions outside of the

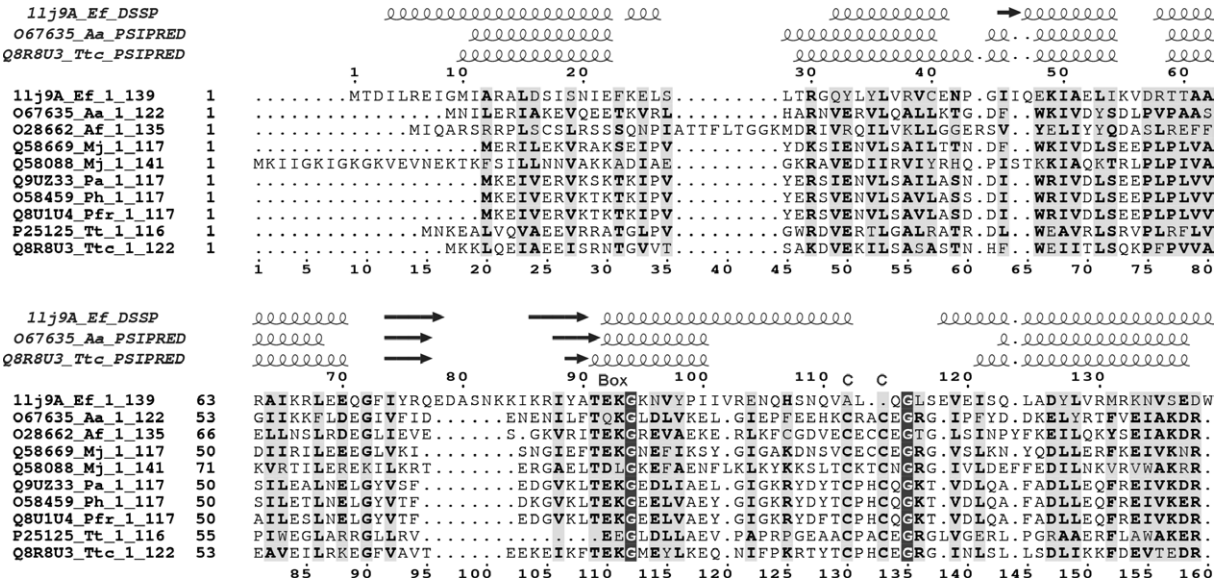

Fig. 6. Multiple sequence alignment of the N-terminal helix-turn-helix domains of the archaebacterial DUF43/hthMTase family and of the PDB structure 1lj9 of a MarR/SlyA-like transcriptional regulator. The DSSP secondary structure assignment of 1lj9 and the corresponding PSIPRED predictions for selected proteins are drawn on top of each alignment row (curled lines for α-helices, arrows for β-strands). Alignment columns with strictly conserved residues are highlighted in dark gray boxes, those in which more than 65% of the residues are physico-chemically equivalent are shown in light gray boxes. Text labels point to the positions of a conserved sequence motif box of functional relevance and of two conspicuous cysteines. The residues of 1lj9 (chain A) are numbered at the top, the alignment columns at the bottom of each row.

Lsm domain of ataxin-2 and PBP1. We hypothesize that the long C-terminal tail of Lsm domain proteins serves as flexible lasso, whose conserved sequence motifs have evolved as specific binding sites to catch nucleic acids and other interaction partners of Lsm proteins.

3.4. The YjeF_N domain of Lsm16 proteins

In contrast to Lsm13–15 proteins, all Lsm16 members possess an additional C-terminal YjeF_N domain of unknown function. According to the Pfam database, this domain is also found as YjeF-related N-terminal domain (YjeF_N) in bacterial carbohydrate kinases such as the *Escherichia coli* protein YjeF and in several putative plant pyridoxamine 5′-phosphate oxidases (EC 1.4.3.5). The domain also occurs singly in eukaryotes and bacteria, for instance, in the human apolipoprotein A-I (ApoA-I) binding protein AI-BP [57] and the yeast protein YNL200C of unknown function. The YjeF_N domain structure has been determined as a novel fold (PDB identifier 1jzt), but this did not elucidate its function [58]. According to the SCOP annotation, the YjeF_N domain may be a circular permutation of the ribokinase-like fold of the YjeF C-terminal carbohydrate kinase domain (PDB identifier 1kyh) [59].

Consistent with these observations, the sequence region between the DFDF box and the C-terminal YjeF_N domain in Lsm16 proteins shows some sequence similarity to an ATP-dependent DNA-ligases in a NCBI conserved domain search and is predicted by some fold recognition servers to adopt a protein kinase-like fold (data not shown). The latter fold shares functional and structural similarities with the ATP-grasp fold of both the ATP-dependent adenylation domain of DNA-ligases and of the homologous GTP-binding domain of mRNA-capping guanylyltransferases [60,61]. Both domains show the same circularly permutated topology in contrast to the normal ATP-grasp fold. Unfortunately, the confidence values for these prediction results are quite low and it remains difficult to assign a reliable functional hypothesis to the C-terminal domains of Lsm16 proteins, but an experimental investigation of NTP-binding may be worthwhile.

3.5. Functional observations on Lsm13–16 proteins

Few DFDF box Lsm domain proteins have already real names assigned. The *S. cerevisiae* yeast Lsm13 protein SCD6/YPR129W has been observed as multicopy suppressor of clathrin deficiency (Gelperin et al., unpublished, personal communication [62]), while its putative *S. pombe* ortholog SUM2 is required in G2/M phase checkpoint control during mitosis [62]. SUM2 functions in the same pathway as the DEAD/H-box RNA-helicase SUM3, which is involved in translation initiation as suppressor of uncontrolled mitosis [62]. The *S. cerevisiae* homologs of SUM3 are the DEAD/H-box RNA-helicases DBP1 and DED1. The Lsm14 protein RAP55 (RNA-associated protein of 55 kDa) has been observed as oocyte-specific constituent of mRNP particles [63].

Interestingly, the ORF of SCD6 overlaps with the reverse ORF of YPR130C, whose gene product interacts with the RRM domain protein PUF1/YJR091C. PUF1 belongs to the family of pumilio-like Puf repeat containing proteins, some of which regulate the G2/M stage as RNA-binding translational repressors by binding the 3′-UTR of target mRNA [64]. The GRID interaction database shows that the Lsm16 yeast protein DCP3/EDC3/YEL015W associates with 14 other proteins with functions related to RNA-processing such as the Lsm1 and Lsm8 proteins, the exportin CRM1 [65], the mRNA decapping enzymes DCP1 and DCP2 [66,67], and the RNA-helicase DHH1 [68], which regulates the G1/S-checkpoint after DNA damage [69] and also interacts with PBP1, DCP1–DCP2 [70], and the 5′-to-3′ exonuclease XRN1 [71,72].

Some Lsm proteins do not seem to be members of a group and carry other additional C-terminal domains. For instance, a hypothetical *A. thaliana* protein T5K18.40 (SPTrEMBL:

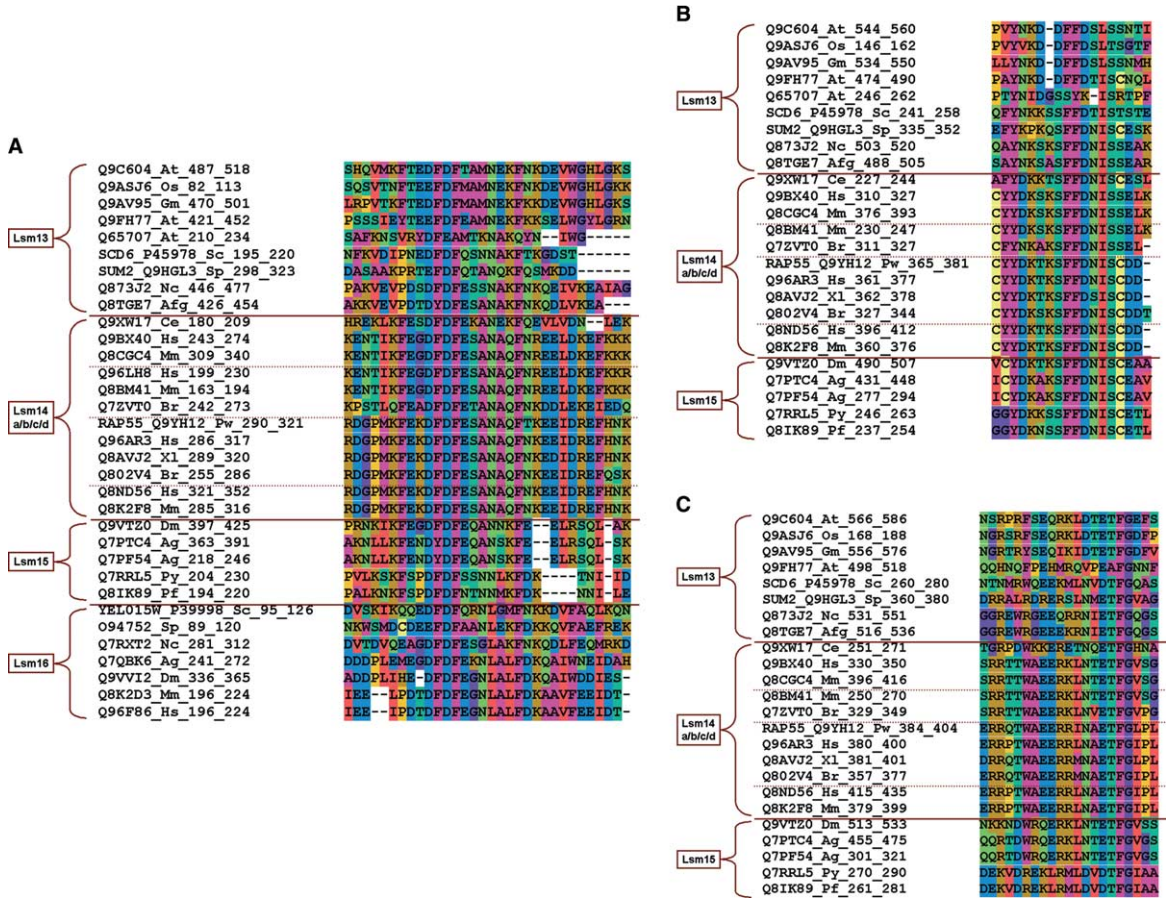

Fig. 7. Multiple sequence alignments of sequence regions in Lsm13-16 proteins, which contain DFDF boxes (A), FFD boxes (B), or TFG boxes (C). The latter two FFD and TFG boxes are not found in Lsm16 proteins. The Lsm14 group is exemplarily further subdivided into four putative subgroups a–d by dotted horizontal lines. Physico-chemically similar amino acids are colored identically. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

O65697) has an N-terminal Lsm domain very similar to Lsm13 and carries a Kelch domain at the C-terminus (NCBI conserved domain search *E*-value 0.004). The Kelch domain consists of a β-propeller, which occurs in diverse protein families as protein–protein interaction module implicated in diverse functions [73,74].

4. Conclusions

Starting with the as yet uncharacterized yeast Lsm domain protein YHR121W, we discovered five evolutionarily conserved groups of homologous Lsm domain proteins Lsm12–16 with an N-terminal Lsm domain and additional C-terminal domains and as yet uncharacterized sequence motifs. The Lsm12 group possesses a C-terminal domain with a predicted tRNA-binding fold and its interaction partners are involved in mRNA degradation. The identification of an additional methyltransferase domain at the C-terminus of one of the Lsm12 group members may point to a function in RNA methylation, but also led to the recognition of three new groups of AdoMet-dependent methyltransferases (new/hth/znMTases).

The Lsm13–15 groups of proteins possess three highly conserved DFDF/FFD/TFG motif boxes of unknown function near the C-terminus, while large central sequence parts appear to be rather variable in orthologs and are predicted to be natively unfolded. We also found some functional and experimental evidence that some of the Lsm13–16 homologs may be involved in RNA metabolism like other Lsm proteins and the control of the mitotic G2/M phase. The Lsm16 group has an additional YjeF_N domain of as yet unknown function, but may bind NTP.

Generally, our findings on the novel Lsm domain proteins and methyltransferases suggest experiments on RNA-binding and methylase activity and on the interaction with other Lsm and RNA-processing proteins. Although the function of the discovered Lsm domain protein families points to RNA-binding, note that they may also bind DNA as shown recently [75]. Furthermore, it is now straightforward to submit the sequence-structure alignments contained in our multiple sequence alignments to modeling web servers in order to obtain 3D model structures bound to RNA or AdoMet for detailed functional analyses.

Note added in proof

Recently, yeast Lsm16 was reported as enhancer of mRNA decapping (EDC3) [76].

Acknowledgements: Part of the research has been funded by the German Research Foundation (DFG) under contract number LE 491/14-1.

References

[1] Mura, C., Phillips, M., Kozhukhovsky, A. and Eisenberg, D. (2003) Proc. Natl. Acad. Sci. USA 100, 4539–4544.
[2] Sauter, C., Basquin, J. and Suck, D. (2003) Nucleic Acids Res. 31, 4091–4098.
[3] He, W. and Parker, R. (2000) Curr. Opin. Cell. Biol. 12, 346–350.
[4] Tharun, S., He, W., Mayes, A.E., Lennertz, P., Beggs, J.D. and Parker, R. (2000) Nature 404, 515–518.
[5] Pillai, R.S., Grimmler, M., Meister, G., Will, C.L., Lührmann, R., Fischer, U. and Schumperli, D. (2003) Genes Dev. 17, 2321–2333.
[6] Satterfield, T.F., Jackson, S.M. and Pallanck, L.J. (2002) Genetics 162, 1687–1702.
[7] Neuwald, A.F. and Koonin, E.V. (1998) J. Mol. Med. 76, 3–5.
[8] Pulst, S.M. et al. (1996) Nat. Genet. 14, 269–276.
[9] Mangus, D.A., Amrani, N. and Jacobson, A. (1998) Mol. Cell. Biol. 18, 7383–7396.
[10] Albrecht, M. and Lengauer, T. (2004) Biochem. Biophys. Res. Commun. 316, 129–138.
[11] Cheng, X. and Roberts, R.J. (2001) Nucleic Acids Res. 29, 3784–3795.
[12] Boeckmann, B. et al. (2003) Nucleic Acids Res. 31, 365–370.
[13] Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2003) Nucleic Acids Res. 31, 34–37.
[14] Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M. and Sonnhammer, E.L. (2002) Nucleic Acids Res. 30, 276–280.
[15] Letunic, I., Goodstadt, L., Dickens, N.J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R.R., Ponting, C.P. and Bork, P. (2002) Nucleic Acids Res. 30, 242–244.
[16] Marchler-Bauer, A. et al. (2003) Nucleic Acids Res. 31, 383–387.
[17] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[18] Eddy, S.R. (1996) Curr. Opin. Struct. Biol. 6, 361–365.
[19] Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. and Thompson, J.D. (2003) Nucleic Acids Res. 31, 3497–3500.
[20] Poirot, O., O'Toole, E. and Notredame, C. (2003) Nucleic Acids Res. 31, 3503–3506.
[21] Galtier, N., Gouy, M. and Gautier, C. (1996) Comput. Appl. Biosci. 12, 543–548.
[22] Gouet, P., Robert, X. and Courcelle, E. (2003) Nucleic Acids Res. 31, 3320–3323.
[23] Bujnicki, J.M., Elofsson, A., Fischer, D. and Rychlewski, L. (2001) Bioinformatics 17, 750–751.
[24] Ginalski, K. and Rychlewski, L. (2003) Nucleic Acids Res. 31, 3291–3292.
[25] Westbrook, J., Feng, Z., Chen, L., Yang, H. and Berman, H.M. (2003) Nucleic Acids Res. 31, 489–491.
[26] Lo Conte, L., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2002) Nucleic Acids Res. 30, 264–267.
[27] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.
[28] McGuffin, L.J., Bryson, K. and Jones, D.T. (2000) Bioinformatics 16, 404–405.
[29] Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J. and Russell, R.B. (2003) Structure (Camb.) 11, 1453–1459.
[30] Linding, R., Russell, R.B., Neduva, V. and Gibson, T.J. (2003) Nucleic Acids Res. 31, 3701–3708.
[31] Liu, J. and Rost, B. (2003) Nucleic Acids Res. 31, 3833–3835.
[32] Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z. and Dunker, A.K. (2002) J. Mol. Biol. 323, 573–584.
[33] Breitkreutz, B.J., Stark, C. and Tyers, M. (2003) Genome Biol. 4, R23.
[34] Hertz-Fowler, C. et al. (2004) Nucleic Acids Res. 32 (Database issue), D339–D343.
[35] Achsel, T., Brahms, H., Kastner, B., Bachi, A., Wilm, M. and Lührmann, R. (1999) EMBO J. 18, 5789–5802.
[36] Kambach, C., Walke, S., Young, R., Avis, J.M., de la Fortelle, E., Raker, V.A., Lührmann, R., Li, J. and Nagai, K. (1999) Cell 96, 375–387.
[37] Collins, B.M., Harrop, S.J., Kornfeld, G.D., Dawes, I.W., Curmi, P.M. and Mabbutt, B.C. (2001) J. Mol. Biol. 309, 915–923.
[38] Mura, C., Cascio, D., Sawaya, M.R. and Eisenberg, D.S. (2001) Proc. Natl. Acad. Sci. USA 98, 5532–5537.
[39] Törö, I., Thore, S., Mayer, C., Basquin, J., Séraphin, B. and Suck, D. (2001) EMBO J. 20, 2293–2303.
[40] Urlaub, H., Raker, V.A., Kostka, S. and Lührmann, R. (2001) EMBO J. 20, 187–196.
[41] Schumacher, M.A., Pearson, R.F., Møller, T., Valentin-Hansen, P. and Brennan, R.G. (2002) EMBO J. 21, 3546–3556.
[42] Törö, I., Basquin, J., Teo-Dreher, H. and Suck, D. (2002) J. Mol. Biol. 320, 129–142.
[43] Collins, B.M., Cubeddu, L., Naidoo, N., Harrop, S.J., Kornfeld, G.D., Dawes, I.W., Curmi, P.M. and Mabbutt, B.C. (2003) J. Biol. Chem. 278, 17291–17298.
[44] Mura, C., Kozhukhovsky, A., Gingery, M., Phillips, M. and Eisenberg, D. (2003) Protein Sci. 12, 832–847.
[45] Thore, S., Mayer, C., Sauter, C., Weeks, S. and Suck, D. (2003) J. Biol. Chem. 278, 1239–1247.
[46] Yaremchuk, A., Tukalo, M., Grøtli, M. and Cusack, S. (2001) J. Mol. Biol. 309, 989–1002.
[47] Kasten, M.M. and Stillman, D.J. (1997) Mol. Gen. Genet. 256, 376–386.
[48] Takagi, T., Cho, E.J., Janoo, R.T., Polodny, V., Takase, Y., Keogh, M.C., Woo, S.A., Fresco-Cohen, L.D., Hoffman, C.S. and Buratowski, S. (2002) Eukaryot. Cell 1, 448–457.
[49] Olivas, W. and Parker, R. (2000) EMBO J. 19, 6602–6611.
[50] Bujnicki, J.M., Feder, M., Radlinska, M. and Blumenthal, R.M. (2002) J. Mol. Evol. 55, 431–444.
[51] Wu, R.Y., Zhang, R.G., Zagnitko, O., Dementieva, I., Maltzev, N., Watson, J.D., Laskowski, R., Gornicki, P. and Joachimiak, A. (2003) J. Biol. Chem. 278, 20240–20244.
[52] Huang, C.C., Smith, C.V., Glickman, M.S., Jacobs Jr., W.R. and Sacchettini, J.C. (2002) J. Biol. Chem. 277, 11559–11569.
[53] Skinner, M.M., Puvathingal, J.M., Walter, R.L. and Friedman, A.M. (2000) Struct. Fold Des. 8, 1189–1201.
[54] Huang, L., Hung, L., Odell, M., Yokota, H., Kim, R. and Kim, S.H. (2002) J. Struct. Funct. Genomics 2, 121–127.
[55] Narayan, P. and Rottman, F.M. (1992) Adv. Enzymol. Relat. Areas Mol. Biol. 65, 255–285.
[56] Dreyfuss, G., Matunis, M.J., Pinol-Roma, S. and Burd, C.G. (1993) Annu. Rev. Biochem. 62, 289–321.
[57] Ritter, M. et al. (2002) Genomics 79, 693–702.
[58] Chance, M.R. et al. (2002) Protein Sci. 11, 723–738.
[59] Zhang, R., Grembecka, J., Vinokour, E., Collart, F., Dementieva, I., Minor, W. and Joachimiak, A. (2002) J. Struct. Biol. 139, 161–170.
[60] Grishin, N.V. (1999) J. Mol. Biol. 291, 239–247.
[61] Doherty, A.J. and Suh, S.W. (2000) Nucleic Acids Res. 28, 4051–4058.
[62] Forbes, K.C., Humphrey, T. and Enoch, T. (1998) Genetics 150, 1361–1375.
[63] Lieb, B., Carl, M., Hock, R., Gebauer, D. and Scheer, U. (1998) Exp. Cell Res. 245, 272–281.
[64] Parisi, M. and Lin, H. (2000) Curr Biol. 10, R81–R83.
[65] Moroianu, J. (1998) J. Cell Biochem. 70, 231–239.
[66] Piccirillo, C., Khanna, R. and Kiledjian, M. (2003) RNA 9, 1138–1147.
[67] She, M., Decker, C.J., Sundramurthy, K., Liu, Y., Chen, N., Parker, R. and Song, H. (2004) Nat. Struct. Mol. Biol. 11, 249–256.
[68] Tseng-Rogenski, S.S., Chong, J.L., Thomas, C.B., Enomoto, S., Berman, J. and Chang, T.H. (2003) Nucleic Acids Res. 31, 4995–5002.
[69] Westmoreland, T.J., Olson, J.A., Saito, W.Y., Huper, G., Marks, J.R. and Bennett, C.B. (2003) J. Surg. Res. 113, 62–73.
[70] Fischer, N. and Weis, K. (2002) EMBO J. 21, 2788–2797.
[71] Ingelfinger, D., Arndt-Jovin, D.J., Lührmann, R. and Achsel, T. (2002) RNA 8, 1489–1501.
[72] Long, R.M. and McNally, M.T. (2003) Mol. Cell 11, 1126–1128.
[73] Adams, J., Kelso, R. and Cooley, L. (2000) Trends Cell Biol. 10, 17–24.
[74] Prag, S. and Adams, J.C. (2003) BMC Bioinform. 4, 42.
[75] Collins, B.M., Cubeddu, L., Naidoo, N., Harrop, S.J., Kornfeld, G.D., Dawes, I.W., Curmi, P.M. and Mabbutt, B.C. (2003) J. Biol. Chem. 278, 17291–17298.
[76] Kshirsagar, M. and Parker, R. (2004) Genetics 166, 729–739.