J Mol Model (2011) 17:1863–1875
DOI 10.1007/s00894-010-0888-0

ORIGINAL PAPER

Prediction of a new class of RNA recognition motif

Núria Cerdà-Costa · Jaume Bonet ·
M. Rosario Fernández · Francesc X. Avilés ·
Baldomero Oliva · Sandra Villegas

Received: 30 April 2010 / Accepted: 21 October 2010 / Published online: 17 November 2010

Abstract The observation that activation domains (AD) of procarboxypeptidases are rather long compared to the pro-regions of other zymogens raises the possibility that they could play additional roles apart from precluding enzymatic activity within the proenzyme and helping in its folding process. In the present work, we compared the overall pro-domain tertiary structure with several proteins belonging to the same fold in the structural classification of proteins (SCOP) database by using structure and sequence comparisons. The best score obtained was between the activation domain of human procarboxypeptidase A4 (ADA4h) and the human U1A protein from the U1 snRNP. Structural alignment revealed the existence of RNP1- and RNP2-related sequences in ADA4h. After modeling ADA4h on U1A, the new structure was used to extract a new sequence pattern characteristic for important residues at key positions. The new sequence pattern allowed scanning protein sequences to predict the RNA-binding function for 32 sequences undetected by PFAM. Unspecific RNA electrophoretic mobility shift assays experimentally supported the prediction that ADA4h binds an RNA motif similar to the U1A binding-motif of stem-loop II of U1 small nuclear RNA. The experiments carried out with ADA4h in the present work suggest the sharing of a common ancestor with other RNA recognition motifs. However, the fact that key residues preventing activity within the proenzyme are also key residues for RNA binding might have induced the activation domains of procarboxypeptidases to evolve from the canonical RNP1 and RNP2 sequences.

Keywords RNA recognition motif · Ribonucleoprotein · Electrophoretic mobility shift assay · Remote homology · Fold conservation

Núria Cerdà-Costa and Jaume Bonet contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00894-010-0888-0) contains supplementary material, which is available to authorized users.

N. Cerdà-Costa · M. R. Fernández · F. X. Avilés · S. Villegas (✉)
Departament de Bioquímica i Biologia Molecular, Unitat de Biociències, Universitat Autònoma de Barcelona,
08193, Cerdanyola del Vallès, Spain
e-mail: sandra.villegas@uab.cat

J. Bonet · B. Oliva
Structural Bioinformatics Group (GRIB), Barcelona Research Park of Biomedicine (PRBB), Universitat Pompeu Fabra,
Doctor Aiguader 88,
Barcelona 08003 Catalonia, Spain

B. Oliva
e-mail: boliva@imim.es

F. X. Avilés
Institut de Biotecnologia i Biomedicina,
Universitat Autònoma de Barcelona,
08193, Cerdanyola del Vallès, Spain

Introduction

Carboxypeptidases (CPs) are proteases that hydrolyze C-terminal peptide bonds from peptides and proteins. A relatively large number of different CPs exist, covering a variety of specificities and complementing each other when acting in the same location [1].

The most used classification of proteases can be found at the MEROPS database [2], in which metallocarboxypeptidases belong to the M14 family. The M14 family can be divided into two main different subfamilies according to their sequence homology: the N/E subfamily and the A/B subfamily.

The A/B subfamily has been the most studied one, with the bovine carboxypeptidase A (CPA) as the reference model. It comprises enzymes produced as zymogens, having an N-terminal activation segment (or pro-region), about 95 residues long, which inhibits the enzyme's activity until it is released due to limited proteolysis. They are usually synthesized in the pancreas and transported to the duodenum where they participate in food processing; nonetheless, some members, like CPU and CPA3, originate in other tissues and act in alternative processes such as blood coagulation [3], or anaphylactic and inflammatory responses [4], respectively.

The activation segment, or pro-region, found in the A/B subfamily is remarkably long, constituting one-fourth of the entire proenzyme. It is composed of an activation domain (AD), with a globular fold independent of the catalytic domain, and a connecting α-helix that covalently holds them together.

The sequence similarity among different CPs pro-regions is quite low (20% to 50%), the residues involved in the enzyme inhibition being the most conserved; however, they share an almost identical α+β topology formed by two α-helices and four β-strands arranged in an open-sandwich conformation. As a result, all of the activation domains with a resolved three-dimensional (3D) structure belong to the same family in the structural classification of proteins (SCOP) database (pancreatic carboxypeptidase, activation domain) [5]; http://scop.mrc-lmb.cam.ac.uk/scop/).

The fact that the activation segments are relatively long compared to the pro-regions of other zymogens, brought up some questions about the possibility that they could play other roles apart from precluding enzymatic activity within the proenzyme. Folding studies performed in ADA2h showed an extremely fast two-state kinetics [6], suggesting that ADs could assist in proenzyme folding. Supporting this hypothesis, recombinant expression of various forms of CPs in the methylotrophic yeast *Pichia pastoris* have proved unsuccessful, whereas the proenzyme forms can be produced at high yields.

One of the first sets of experiments studied the possible Ca^{2+} binding properties, since some structural similarity with the EF-hand protein family had been found [7]. No significant results regarding Ca^{2+} binding were obtained and, subsequently, the similarity was also discarded using bioinformatic sequential and structural approaches, although some homologies with complement factor B were proposed [8].

In the present work, the overall pro-domain tertiary structure was compared with several proteins belonging to the same fold in SCOP. The sequence alignment derived from the structure comparison was used to extract a sequence pattern characteristic of important residues at key positions. The sequence pattern was further used to scan protein sequences to detect RNA-binding proteins. Subsequently, the RNA-binding capability of ADA4h was tested experimentally in order to confirm our predictions.

Materials and methods

Bioinformatics

The search for remote homologs of ADA4h was performed with three iterations of PSI-BLAST [9] searching in Uniprot [10], using the BLOSUM45 matrix and standard parameters. After the first iteration, the position-specific scoring matrix (PSSM) included the first hit sequence not belonging to the procarboxypeptidase family in order to find remote homologs of ADA4h. A second search was performed on PDB [11] (using the same parameters) searching remote homologs with solved structure.

The 3D structure of ADA4h was extracted from the structure of the human procarboxypeptidase A4 (code 2BOA [10]) retrieved from the PDB [11]. ADA4h belongs to the pancreatic carboxypeptidase, activation domain family, in the ferredoxin-like fold, from SCOP [5]. All proteins with a similar fold were retrieved from PDB, and all protein domains with these folds were extracted from these proteins. The structural time series analyser modeller and predictor (STAMP) program [12] was used to superimpose AD folds and to compare their structures by means of the root mean square deviation (RMSD).

The capability of ADA4h to change the orientation of some of its loops to simulate the function of a different family of the same ferredoxin-like fold was tested via comparative modeling using MODELLER [13] and the selected templates from the ferredoxin-like fold.

This fold contains many proteins with functions unrelated to the CP precursor. Among them, many belong to several superfamilies with nucleotide-recognition functions. Particularly interesting is one superfamily whose function is related to RNA binding and that can be identified by an RNA recognition motif (RRM). This motif is defined in PROSITE [14] by means of two sequence signatures called as RNP1 and RNP2, separated by about 30 residues, and is thought to bind single-stranded RNA. Due to the structural similarity among ADA4h and the proteins of families containing this motif, we were able to identify similar sequence patches (RNP1 and RNP2) in ADA4h, also separated by approximately 30 residues. Previous work had shown the possibility of inferring functional relationships within the same SCOP fold by distant homology [15] and the importance of functional loops to imprint a particular function [16]. Therefore, we hypothesize a new motif based on the RNP1 and RNP2 sequences of RRM that could also include the sequence patches of ADA4h. The new motif was tested by searching proteins containing the new motif in the Uniprot database with ScanProsite [17].

Materials

The MEGAscript® T7 Kit for transcription, the RNase inhibitor SuperaseIn and the RNase decontamination solution RNaseZAP® were obtained from Ambion (http://www.ambion.com/). Micro Bio-Spin RNA purification columns were purchased from Bio-Rad (http://www.bio-rad.com), [α-^{32}P]UTP from GE Healthcare (http://www.gehealthcare.com), and the C-UVette® DNase RNase Protein-free from Eppendorf (http://www.eppendorf.com). Primers were synthesized by Roche Molecular Biochemicals (Basel, Switzerland). All other chemicals were obtained from Sigma (St. Louis, MO).

Cloning and expression

ADA4h was cloned into the pET-30 Xa/LIC vector from Novagen (http://www.merck-chemicals.com/life-science-research/novagen) resulting in an N-terminal His-tagged protein. The recombinant vector was then transformed into *Escherichia coli* strain BL21 (DE3), and the obtained clones were screened for positives.

Cells were grown and induced as previously described [18] and ADA4h was purified from the intracellular soluble fraction using nickel sepharose. A 60-mM imidazole wash was applied once the sample was loaded and the column was re-equilibrated, followed by an elution step with 300 mM imidazole. The elution fractions containing His-tagged ADA4h were dialyzed overnight against binding buffer, and then were again applied to nickel sepharose in a batch manner. Nickel sepharose-bound protein was digested overnight at room temperature with the necessary amount of Factor Xa, in order to eliminate the His-tag. The protein was recovered in the supernatant and kept at 0°C in 50 mM Tris-HCl 0.5 M NaCl, pH 8.0. The identity of the protein was confirmed using MALDI-TOF mass spectrometry and N-terminal sequencing.

In vitro transcription of RNA

An RNA probe carrying two copies of the U1A-binding motif from stem-loop II of U1 snRNA hairpin [19] was synthesized by in vitro transcription and named bipolar-stem-loop: gggAgggUUAACAUUgCACUCCgUUgUCCAUCCCAAAAAAAAAAgggUUAACAUUgCACUCCgUUgUCC.

The online program OligoAnalyzer 3.0 from Integrated DNA Technologies® (http://www.idtdna.com/analyzer/Applications/OligoAnalyzer/Default.aspx) was used to test the secondary structure.

In order to obtain the afore-mentioned RNA probe, a direct primer carrying the T7 RNA polymerase promoter was designed: 5′-TAATACGACTCACTATAGGGAGGGTTAACATTGCACTCCGTTGTCCATCCCAAAAAAAAA-3′ partially overlapping with the reverse primer, 5′-GGACAACGGAGTGCAATGTTAACCCTTTTTTTTTTGGGATGGACAACGG-3′. Both primers (1 nmol of each) were incubated with 4 U Klenow polymerase and 0.6 mM dNTPs at 37 °C for 1 h, and the enzyme was inactivated afterwards at 65 °C for 10 min. The DNA was then analyzed in a 2% agarose gel and quantified.

In vitro transcription reactions were performed essentially as indicated in the MEGAshortscript® T7 Kit. The reaction was performed in 10 µl, using a 1:10 ratio for UTP with respect to other NTPs, adding 40 µCi [α-^{32}P]UTP, and incubating at 37 °C for 2 h. A second in vitro transcription in the absence of [α-^{32}P]UTP was carried out as indicated in the kit instruction manual and used as the cold probe. The amount of DNA template used in both reactions was 125 nM. In both cases, after the transcription reaction, 10 U RNase-free DNaseI were added and the mixture incubated for 15 min at 37 °C. The RNA obtained was then purified using Bio-Rad Micro Bio-Spin columns P-30 Tris RNase-Free. The unlabeled RNA was quantified by UV-spectrometry using C-UVette® DNase RNase Protein-free and the radioactive probe by scintillation counting.

RNA electrophoretic mobility shift assays

For RNA electrophoretic mobility shift assays (EMSA), 25 fmol labeled RNA was incubated in a final volume of 10 µl for 20 min at room temperature with the indicated amounts of protein in a buffer containing 10 mM Hepes, pH 7.4, 25 mM potassium acetate, 2.5 mM magnesium acetate, 0.5% Igepal CA 630, 5% glycerol, 1 mM dithiothreitol and 1 U/µl of SuperaseIn [20]. Yeast tRNA was also added in competition assays (0.5 µg tRNA; probe:competitor ratio 1:2×10^7, w:w).

RNA probes were previously heated at 95°C in order to disrupt their secondary structure and left 5 min on ice so that their proper conformation was finally achieved. EMSA were performed using one-phase 5% polyacrylamide gels with TBE buffer (90 mM Tris-HCl, pH 8.5, 110 mM boric acid, 2 mM EDTA) as running buffer and were left at 60–70 V for 90 min. The gel was then fixed with 20% ethanol, 10% acetic acid, for 30 min, introduced into a plastic-bag and visualized using a Fluor-S™ MultiImager (Bio-Rad).

Results

ADA4h particularities

First of all, a computational characterization of the available activation domains (ADs) was carried out using online tools such as ProtParam [21] from the Expasy webpage (http://www.expasy.org/tools/protpar-ref.html). All domains show a conserved length, although their composition is quite different. The pI for all domains is below neutrality; nonetheless, ADA4h presents a fairly basic pI (9.30) compensated by a highly acidic connecting helix. Thus, the entire activation segment retains an acidic pI, similar to the other procarboxypeptidases. Apart from ADA4h, only the ADs of the recently described procarboxypeptidases A5 and A6 and the AD from PCPU approach neutrality [22, 23].

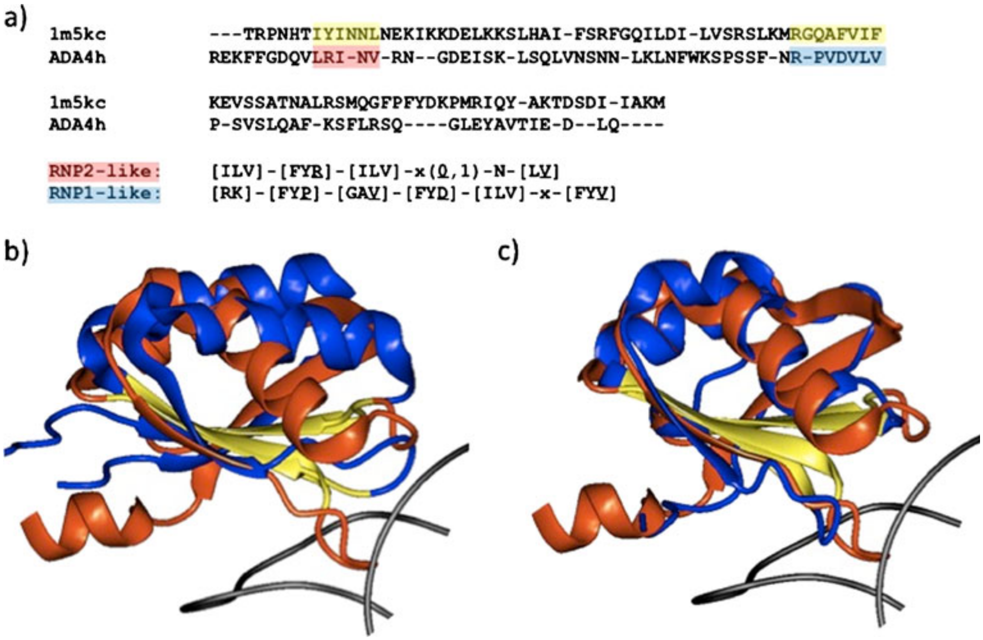
Structural search

A significant sequence alignment was found (e-value 1e-17) using Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) between ADA4h and the U1 SNP1-associating protein (Q03714), a U1-type yeast-protein involved in mRNA splicing.

The structure of ADA4h was assigned to the SCOP family "pancreatic carboxypeptidase, activation domain within ferredoxin-like fold". However, the structure of U1 SNP1-associating protein (Q03714) is not available in the Protein Data Bank (PDB). Therefore, a PSI-BLAST search on PDB was performed in order to find a structural template for Q03714. We found a significant alignment between Q03714 and several domains from the canonical RNA-binding-domain (RBD) family of SCOP (psi-blast e-values between 7e-33 and 3e-19), suggesting some evolutive relationship between Q03714 and this SCOP family. Consequently, the domain of U1 SNP1-associating protein was assigned to the canonical RNA-binding-domain (RBD) SCOP family, also in the ferredoxin-like fold. From the members of this family, we selected as a template for Q03714 the human U1 small nuclear ribonucleoprotein A (P09012), with code 1M5Kc [24] in the PDB. This protein has been crystallized in interaction with an RNA molecule. However, instead of modeling the structure of Q03714 with 1M5Kc, we compared ADA4h directly with 1M5Kc. The structural superimposition of ADA4h pro-domain and 1M5Kc produces a sequence alignment that proves a 14% identity and covers 58% of sequence similarities according to the BLOSUM45 matrix (Fig. 1a).

This superimposition allowed us to allocate the RNA molecule co-crystallized with U1A in 1M5Kc, and to assume the putative binding between RNA and the AD of human procarboxypeptidase A4 (Fig. 1b). Although the superimposition shows a putative binding region between ADA4h and RNA, with two sequence patterns similar to the RRM of U1A (Fig. 1a), there are conformational restrictions to accommodate the backbone chain of ADA4h to perform similar binding. The major differences between both backbone conformations are found in loops 1 and 3. We postulate that the orientation of these loops would be dependent on the partner molecule that interacts with ADA4h, i.e., that the actual conformation of ADA4h binds the catalytic domain within procarboxypeptidase A4, while upon interaction with RNA it would adopt a conformation similar to U1A. In agreement with this hypothesis, the main motions of the binding region studied in a previous work [25] shows the correlation between the RNP1 and RNP2 sequence patches and loops 1 and 3. In order to test this hypothesis, ADA4h was modeled using the conformation of U1A as a template and the structural alignment of the superimposition where the putative RNA-binding regions found on ADA4h corresponded to the RRM patterns of U1A (Fig. 2). The possibility that the interaction with RNA was due to the positive charge of ADA4h was also assessed. Comparison with U1A suggested that ADA4h has less positive charges on its surface than U1A (Fig. 3).

The new conformation of ADA4h (ADA4hm) was evaluated with the pseudo-energies of ProsaII [26], the Z-score based on split-potentials developed recently in our group [27], QMEAN [28], Gromos [29] and ANOLEA [30] (see Electronic supplementary material). According to ProsaII statistical potentials, the modified orientation of loops 1 and 3, whose flexibility has been described before [31], was energetically possible but not favorable (Fig. 4a, b). First, the comparison with the energies of the template U1A (Fig. 4a) shows the conflicting regions of the model with respect to the conformation of the original sequence that produces this fold; second, the comparison between modified and unmodified conformations of ADA4h (Fig. 4b) shows that the conformation can endure the re-orientation of the first helix and of loops 1 and 3. Furthermore, to check if the energetic conflicts of this conformational change were due to side-chain packing or to the backbone conformation, we used the split-potentials developed recently in our group [27]. Z-scores based on split-potentials of ADA4h and ADA4hm were calculated using pair-residue distances between Cβ carbons (Fig. 4c) and the minimum distance between the atoms of each two residues (Fig. 4d). The location of the energy differences between the conformations of ADA4h and ADA4hm, calculated with energies that take into account the side-chain packing (Fig. 4d) or only the backbone conformation (Fig. 4c), were similar. This indicates that the conformational conflicts detected in the modified conformation of ADA4hm are due mainly to the change of the backbone, but not to the new residue packing. We further compared the disposition of the side-chains of ADA4h in the binding region with the RRM residues of U1A to check whether they might perform similar binding with RNA (Table 1, Fig. 5). Among the key residues for RNA binding by U1A, the most important are Arg52, Gln54 and Phe56 belonging to RNP1 and Tyr13 belonging to RNP2. On ADA4h, these residues are substituted by Arg45, Pro46 and Asp48 for RNP1, and Arg11 for RNP2 (Table 1).



```
1m5kc        ---TRPNHTIYINNLNEKIKKDELKKSLHAI-FSRFGQILDI-LVSRSLKMRGQAFVIF
ADA4h        REKFFGDQVLRI-NV-RN--GDEISK-LSQLVNSNN-LKLNFWKSPSSF-NR-PVDVLV

1m5kc        KEVSSATNALRSMQGFPFYDKPMRIQY-AKTDSDI-IAKM
ADA4h        P-SVSLQAF-KSFLRSQ----GLEYAVTIE-D--LQ----

RNP2-like:   [ILV]-[FYR]-[ILV]-x(Q,1)-N-[LV]
RNP1-like:   [RK]-[FYP]-[GAV]-[FYD]-[ILV]-x-[FYV]
```

Fig. 1a–c Structural superimposition of human procarboxypeptidase A4 (ADA4h) and human U1A protein from the U1 snRNP. **a** Sequence alignment derived from the structural superimposition. The RNP regions of U1A are emphasized in *yellow*. RNP1 and RNP2 regions stand out in *red* and *blue* for ADA4h. The PROSITE pattern description is shown below the alignment and modifications from the original pattern are *underlined*. **b** Superimposition of structures obtained from the structural alignment of the activation domain of human procarboxypeptidase A4 (in *blue*) with the RNA-binding protein U1A (in *red*). RNP regions of both structures are highlighted in *yellow* (RMSD 2.16). (**c**) Structural alignment of RNA-binding protein U1A (in *red*) and the structural model of ADA4h (in *blue*) using the structure of U1A in 1M5Kc as template (RMSD 1.70). RNP regions of both structures are highlighted in *yellow*

The interactions of Arg52 in U1A are produced mostly by Arg45 in ADA4h (Fig. 5b). Also, Arg11 (ADA4h) produces a π interaction similar to that of Tyr13 (U1A) with C40 of the RNA chain. This interaction is improved in the modeled structure (Fig. 5d).

The aromatic stacking of Phe56 (U1A) with A41 of the RNA [32] would be weakened if the interaction were carried out by Asp48 of ADA4h (Fig. 5c). Finally, the Gln54 orientation cannot be emulated by Pro46 of ADA4h (Fig. 5a). Both interactions could occur with a different RNA substrate (see Discussion).

Pattern discovery

From the structural alignment, a pattern performing the RNA-recognition function was identified. The RRM identified in 1M5Kc has been described as a combination of two regions (RNP1 and RNP2) in PROSITE. According to

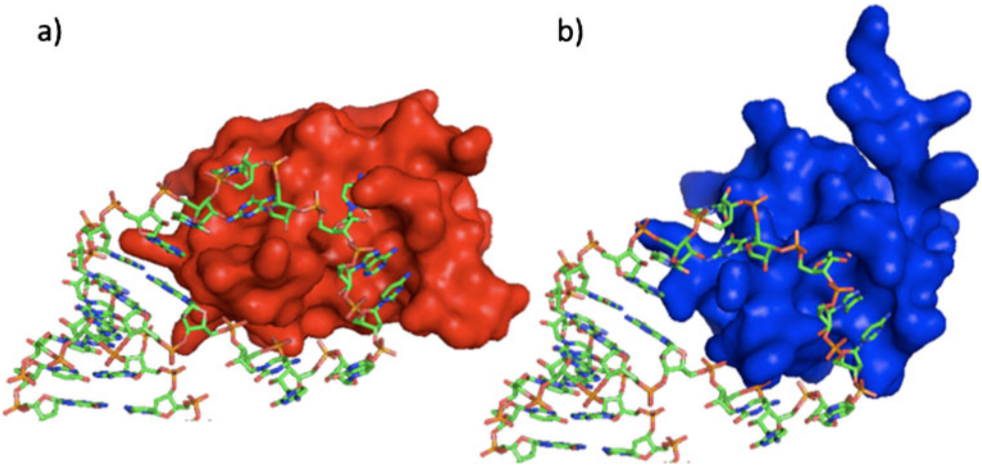


Fig. 2a,b RNA surface interaction. Interacting surface of U1A (**a**, in *red*) and the proposed conformational change of ADA4h (**b**, in *blue*) when binding RNA

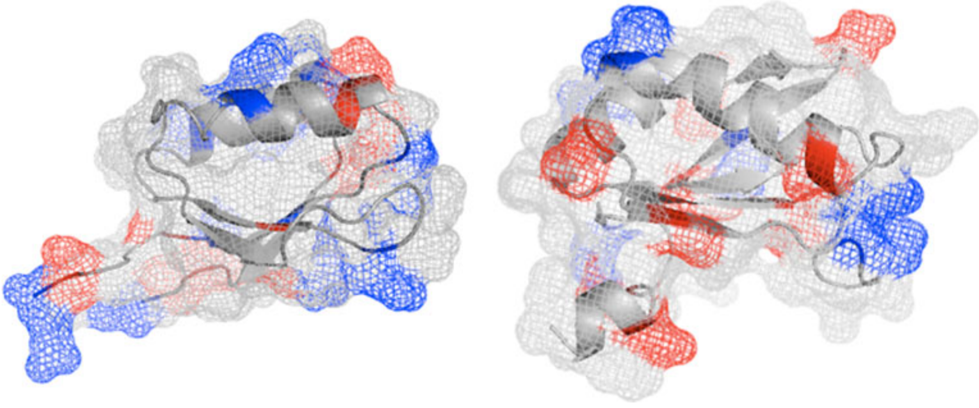


Fig. 3 Electrostatic representations of the original ADA4h (*left*) and 1M5Kc (*right*) are shown here. In both proteins positively charged amino acids are represented in *blue* while negatively charged are represented in *red*. Globally, ADA4h has 10 positively charged amino acids while 1M5Kc has 18

sequence comparison from the structural alignment, these two regions present variants of the original sequences of U1A in the ADA4h sequence (Fig. 1a). Furthermore, both regions are structurally located in the same position according to the structural alignment (Fig. 1b, c). When structurally compared to other known RRM domains, the new conformation of ADA4h displayed the same disposition for its RNP motifs (Fig. 6). Consequently, we propose a new RRM pattern that would characterize a putative binding of RNA.

The new RRM pattern was defined by variations in the sequences of the RNP1 and RNP2 regions (Fig. 1a). A PROSITE SCAN [17] search was performed in the UniProt database to find all sequences matching the new motif. Table 2 shows the distribution of functions of the 70 new sequences found in Uniprot using the new RRM motif but not the original PROSITE RRM motif.

A total of 48 sequences out of these 70 were aligned with the RRM_1 motif from PFAM [33] under an e-value threshold of 10^{-5}. Moreover, the RRM_1 motif aligned with our motif only for 35 of these 48, while for the remaining 13 the alignment was located on a different position of the sequence. Consequently, the new pattern here described predicts the RNA-binding function for 32 sequences undetected by PFAM and may imply a different binding behavior for 13 known RNA-binding proteins, while 50% confirmed the expected function of our new pattern.

Additionally, 52 sequences out of these 70 perform functions involving nucleic-acid binding according to the information extracted from Swissprot. This means that we have found an explanation for four sequences that could not be described by using PFAM. Finally, of the remaining 18 proteins, 5 contributed to functions unrelated to nucleotide binding and 13 had an unknown function. ADA4h was among the 5 proteins unrelated to RNA binding, according to Uniprot labeled functions. The corroboration of our prediction would imply that: (1) a new RRM-motif type for RNA binding would be possible, and (2) ADA4h would have a moonlighting, undetected function.

RNA electrophoretic mobility shift assays

The purified ADA4h protein was tested for RNA-binding activity using the U1A binding-motif of stem-loop II of U1 small nuclear RNA (snRNA) [19]. A 70-mer RNA probe, called the bipolar-stem-loop, was designed to carry two copies of the target RNA motif. The presence of the designed secondary structure was checked using the online program OligoAnalyzer 3.0 from Integrated DNA Technologies®. A free energy of $-18.9\ \mathrm{kcal\,mol^{-1}}$ is calculated for this secondary structure.

The recently found yeast ζ-crystalline (ZTAIp), an RNA-binding protein [34], and bovine serum albumin (BSA) were used as positive and negative controls, respectively (Fig. 7). A small amount of radioactivity was found in the shifted band (Fig. 7a), which increases concomitantly with the concentration of protein (lanes 3–5) and was present neither in the negative control (lane 2) nor when the cold probe was added as competitor in a ten-fold excess (lane 1).

Another experiment with competitor tRNA was performed in order to check the ADA4h specificity for the bipolar-stem-loop (Fig. 7b), and, although the shifted band is not as clear as in the EMSA without competitor tRNA (Fig. 7), some faint shift can still be observed (see lanes 3–4). The shifting also decreases for yeast ζ-crystalline as expected (lane 5).

These results confirm that ADA4h has RNA-binding capability, although it does not show a strong specificity for the bipolar-stem-loop designed in this work.

Discussion

Definition of RRM

The RNA recognition motif (RRM) is one of the most abundant motifs in eukaryotic proteins, and is estimated to be present in 2% of human gene products [35]. Initial biochemical studies on poly(A)-binding protein (PABP) and hnRNP protein C defined a consensus RNA-binding

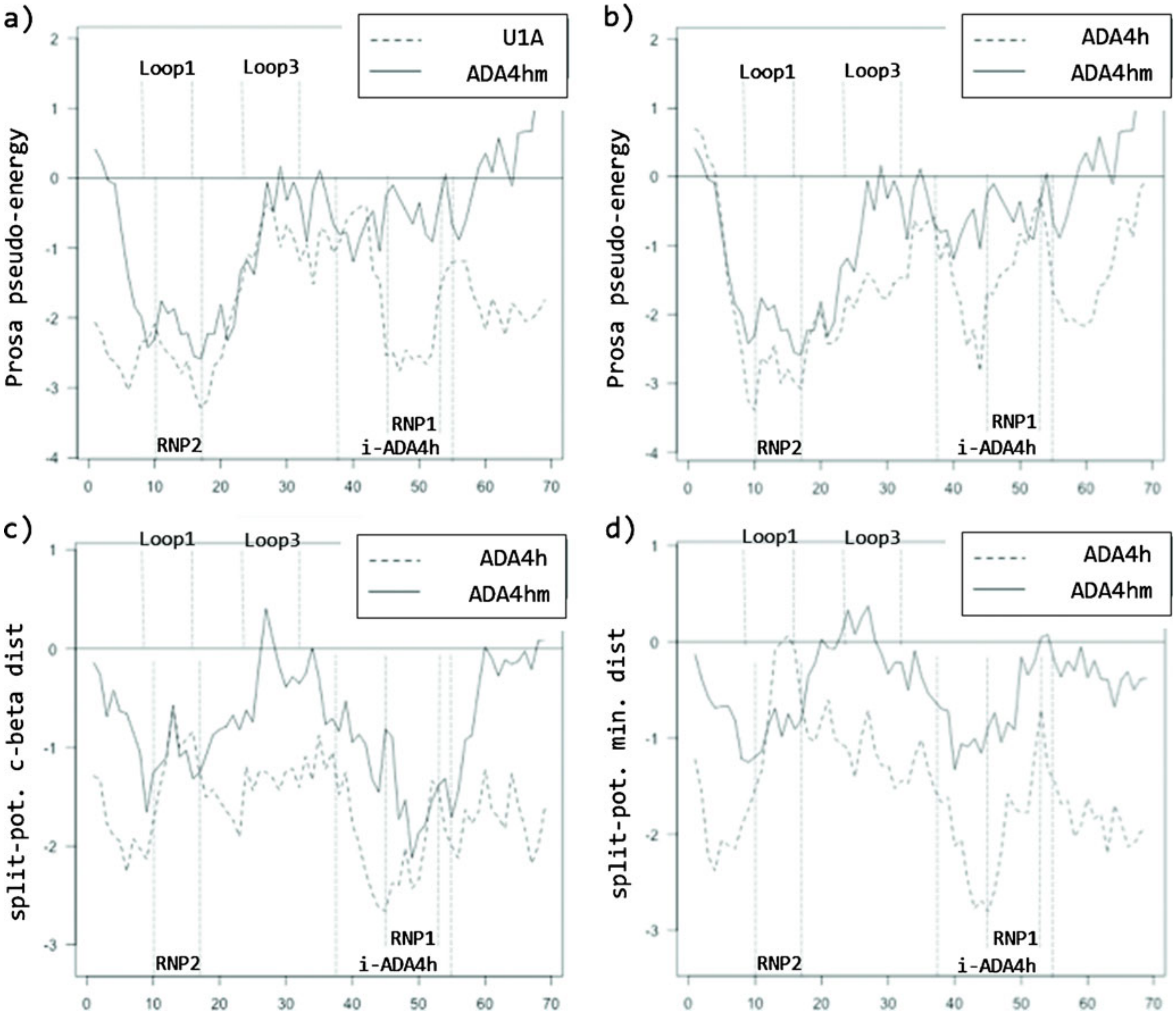


Fig. 4a–d Evaluation of the structure of ADA4h modeled with the conformation of U1A. Energy profiles of the sequence of ADA4h and U1A are calculated with knowledge-based potentials of ProsaII [26] and Z-scores based on split potentials [27]. Energy profiles were smoothed using a window of ten residues. Regions of loop1, loop3, and RNA binding motifs are shown in all plots. **a** Comparison between the ProsaII pseudo-energy of U1A (*dashed line*) and the sequence of ADA4h thread on the structure of U1A (*solid line*). **b** Comparison between the ProsaII pseudo-energy of the original structure of ADA4h (*dashed line*) and the sequence of ADA4h thread in the structure of U1A (*solid line*). In **c** and **d**, the energies of Z-scores based in split-potentials in ADA4h conformation (*dashed line*) and ADA4hm conformation (*solid line*), are compared using pair-residue distances calculate with Cβ atoms (**c**) or the minimum distance between the atoms of the residues (**d**). i-ADA4h stands for the region interacting within the proenzyme

domain of about 90 residues containing a central sequence of eight conserved residues [36, 37]. The [RK]-G-[FY]-[GA]-[FY]-[ILV]-X-[FY] sequence was called RNP1 when a second consensus sequence, [ILV]-[FY]-[ILV]-X-N-L, was found in the N-terminus [38]. RNP2 is less conserved and shorter than RNP1. RNP1 contains mainly aromatic and basic residues, whereas RNP2 is rather hydrophobic.

The search for structures related to the activation domains of procarboxypeptidases, but not to CP function, yielded as the best score the U1A RNP (1M5Kc) with the activation domain of human procarboxypeptidase A4. RNP2 of U1A fits perfectly in the consensus with the IYINNL sequence. In the case of ADA4h, the RNP2-aligned sequence corresponds to: L$\underline{\text{R}}$I-N$\underline{\text{V}}$; differences with respect to the consensus sequence are underlined. Overall, the sequence conserves four out of six residues and is quite similar to the RNP2 consensus sequence.

The RNP1 sequence of U1A is RG$\underline{\text{Q}}$AFVIF, where Gln does not correspond either to Phe or Tyr. This particular Gln, Gln54, stacks on G42 of the polyadenylation inhibi-

Table 1 Comparison of RNA interactions by U1A and ADA4h. Comparison between the interactions performed by U1A and ADA4h, and the proposed conformational change of ADA4h (ADA4hm). The differential distance with which they perform the interaction is detailed in Angstroms. The arrow indicates the corresponding residue in ADA4h

	ADA4h	ADA4hm	U1A	→	
Stacking	X	X	Gln45	Pro46	RNP1
H-bond	1	1.5	Arg52	Arg45	
Stacking	X	X	Phe56	Asp48	
Π interact	0.5	−1	Tyr13	Arg11	RNP2

tion element (PIE) RNA in the crystal structure of the complex, performing the same function as any Phe or Tyr [19]. In the case of ADA4h, the corresponding sequence is R-PVDVLV, showing the lack of residue 2. Although just three residues out of eight are strictly conserved, it could be considered that Val47 and Asp48 are not so different from [GA]-[FY] of the consensus sequence. In fact, an equivalent Val is found in the PABP and in the U2AF35 RNP1 sequences, and an equivalent Asp is also found in the ALY RNP1 sequence [35]. Finally, the residue at position 8 is not conserved, as is the case of the polypyrimidine-tract-binding protein (PTB) RNP1 sequences A and C, the Nucleolin RNP1 sequence B, the LA and the PABP

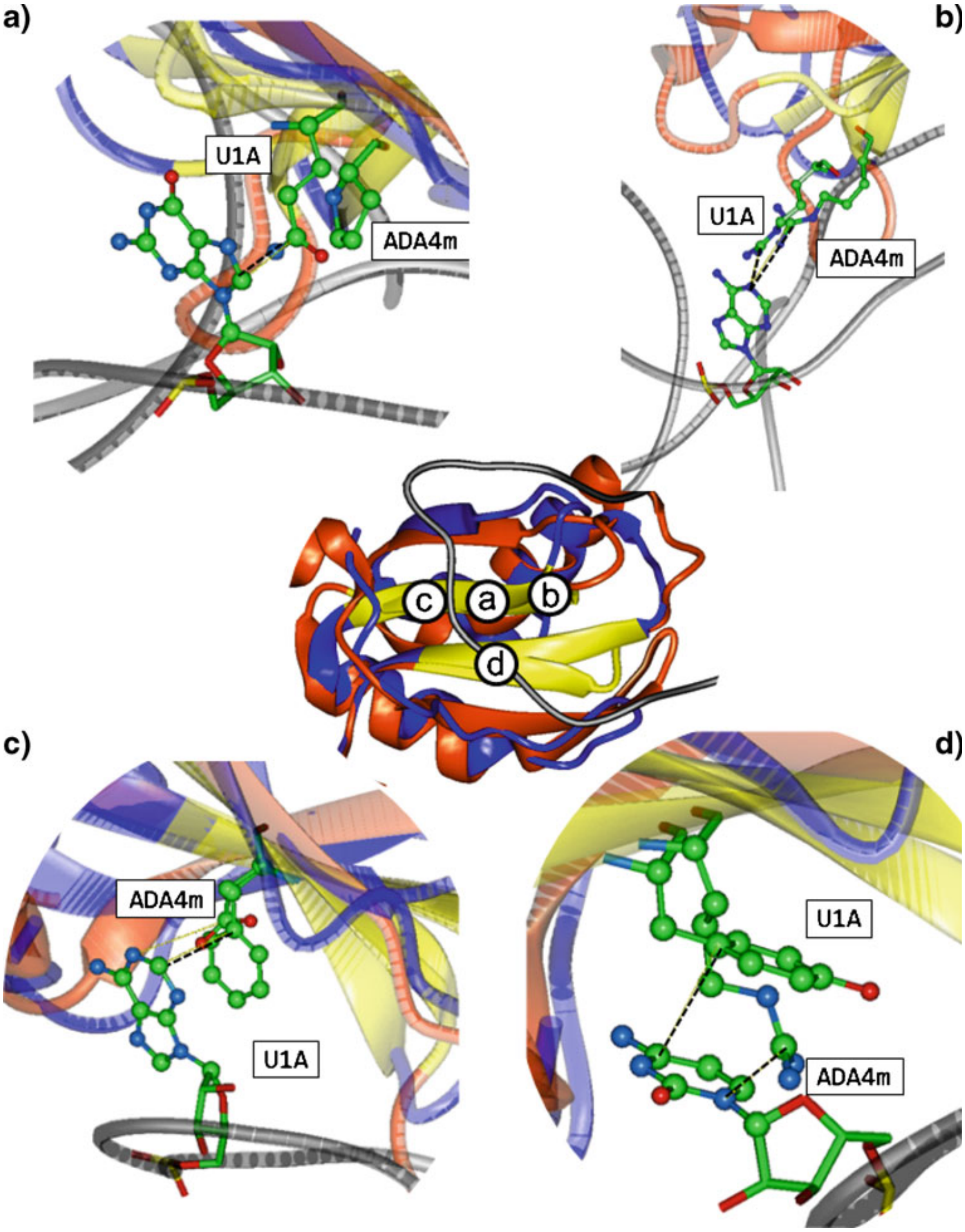
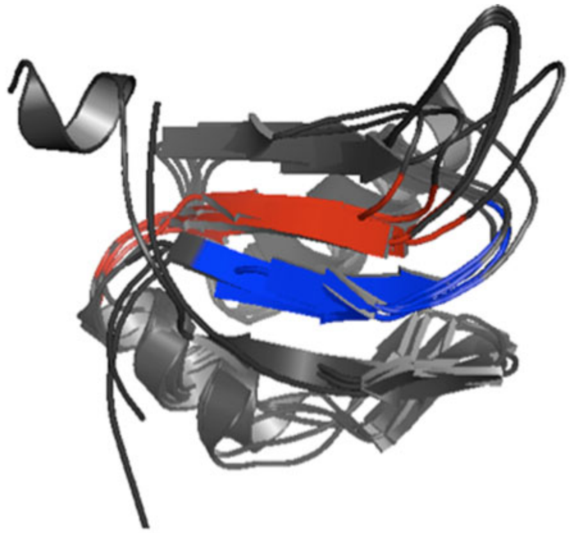


Fig. 5 Comparison between U1A and ADA4h according to their ability to bind RNA. **a** Detail of RNP1 motif. The stacking interaction between Gln54 (U1A) and G39 of the RNA (3.15 Å) cannot be mimicked by Pro46 (ADA4h). **b** Detail of RNP1 motif. The hydrogen bond interaction between Arg52 (U1A) and A36 (3.67 Å) is also performed by Arg45 in ADA4h (5.06 Å). **c** Detail of RNP1 motif. The interaction between Phe56 (U1A) and A41 of the RNA (3.68 Å) cannot be performed by Asp48 in ADA4h (4.01 Å. **d** Detail of RNP2 motif. The π interaction between Tyr13 and C40 of the RNA (3.52 Å) can still be carried out by Arg11 of ADA4h (2.23 Å)



```
>ADA4h
REKFFGDQVLRI-NV-RN---GDEISKLSQLVNSNN-LKLNFWKSPSS----F--N-RPVDVL-VPSVSLQAF-KSFLRS----QGLEYA-VTIED--LQ----
>1UA
---TRPNHTIYINNLNEKIKKDELKKSLHAI--FSRFGQILDI-LV-SR---SLKMRGQAFVIFKEVSSATNALRSMQGFPFYDKPMRIQYAK-TDSDI-IAKM
>1CVJ
-------ASLYVGDLHPDVTEAMLY----EK--FSPAGPILSI-RV-CRDMITRRSLGYAYVNFQQPADAERALDTMNFDVIKGKPVRI--------------
>1FXL
---------LIVNYLPQNMTQEEFR----SL--FGSIGEIESC-KL-VRDKITGQSLGYGFVNYIDPKDAEKAINTLNGLRLQTKTIKV--------------
>1HD0
---------MFIGGLSWDTTKKDLK----DY--FSKFGEVVDC-TL-KLDPITGRSRGFGFVLFKESESVDKVMD-QKEHKLNGKVIDP--------------
>1L3K
---------IFVGGIKEDTEEHHLR----DY--FEQYGKIEVI-EI-MTDRGSGKKRGFAFVTFDDHDSVDKIVI-QKYHTVNGHNCEV--------------
```

Fig. 6 Structural alignment of ADA4h with other RNA recognition motif (RRM) structures. ADA4h is aligned with U1A and all the non-repetitive hits of the PFAM motif of RRM_1 over sequences with known structure with an e-value lower than 1e-23. Although the sequence homology with those sequences is low for both ADA4h and 1UA, both locate their RNPs (RNP1 in *blue* and RNP2 in *red*) in the same orientation as the rest of the RRM structures

sequences [35]. Taking all of these criteria into account, the sequence of RNP1 of ADA4h approaches the consensus except for the lack of Gly at position 2.

It is important to note three different general features of RNA-binding domains. Firstly, although some conserved aromatic residues are always found at the interface of the complex, the topology of the bound RNA is characteristic of each complex solved and the sequence-specificity cannot easily be predicted. Secondly, RRMs that bind RNA show a wide spectrum of affinities. And, finally, some RRMs are specific for protein binding rather than for RNA binding, suggesting that some RRM might have evolved from RNA- to protein-recognition [39].

Table 2 Functional distribution of the new sequences found with the modified RNA recognition motif. The first six functions, excluding the procarboxypeptidase precursor, are related to nucleotide recognition. It must be noted that 75% of the new sequences found with the modified RNA recognition motif (RRM) are related to this function

Function found	Frequency (%)
Carboxypeptidase precursor	2
Nucleolin	34
Splicing factor	10
RNA-binding	16
Nucleotide binding	13
DNA helicase	2
Polysaccharide biosynthesis	1
Dioxygenase	4
Transmembrane component	1
Hypothetical protein	17

Previous works have shown that Gly is a relevant residue for several nucleotide recognition motifs such as the RRM [40, 41]. Despite its importance, it is clear from the fact that Gly has no side-chain that this residue has to be involved in the flexibility and length of the pattern, improving interaction or facilitating specificity. Although our proposed RNP1-like motif lacks this Gly, it aligned several times (50% correctly aligned) with the RRM PFAM domain, thus showing that the interaction with RNA could still be possible through the conservation of other key residues.

Binding specificity of U1A

The U1A protein tertiary structure has been obtained in solution in complex with one of its specific RNA targets—

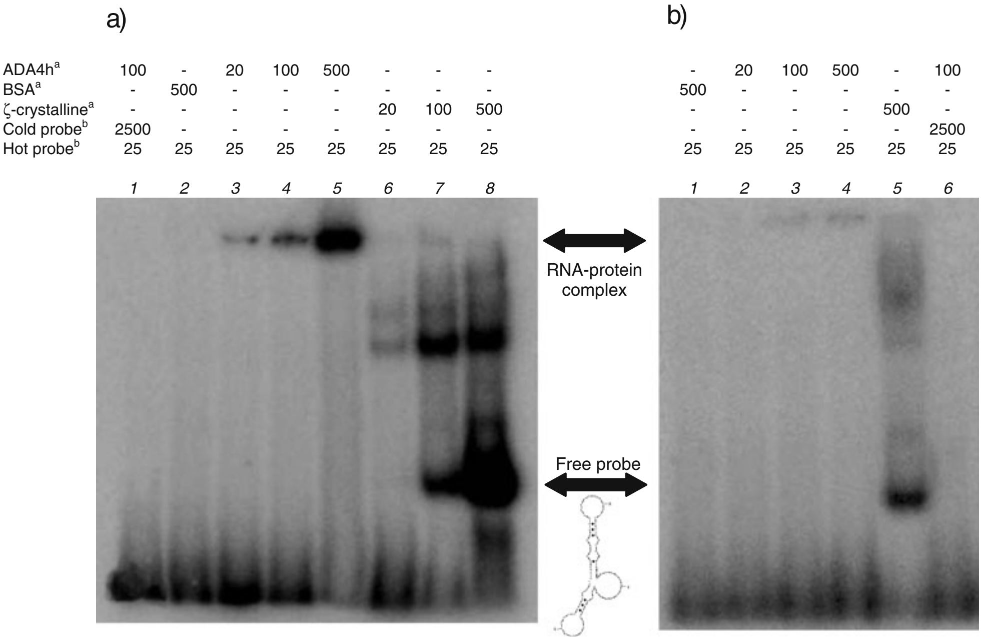


Fig. 7 Unspecific (**a**) and specific (**b**) electrophoretic mobility shift assay (EMSA; *a* ng, *b* fmol). **a** Lanes: *1* ADA4h + cold probe; *2* protein negative control [bovine serum albumin (BSA)]; *3–5* increasing amounts of ADA4h + hot probe; *6* increasing amounts of positive control protein (yeast ζ-crystalline) + hot probe. **b** All lanes contain 0.5 μg competitor tRNA. Lanes: *1* protein negative control (BSA); *2–4* increasing amounts of ADA4h + hot probe; *5* protein positive control (yeast ζ-crystalline) + hot probe; *6* ADA4h + cold probe. In both cases, part of the free probe was left to escape the gel in order to be able to visualize faint retarded bands. The secondary structure of the bipolar-stem-loop is indicated between the panels. The structure was generated with OligoAnalyzer 3.0 from Integrated DNA Technologies®

the PIE from its pre-mRNA [19, 42]. The determinants for binding have been analyzed in detail and compared with those of the previously known crystal complex with stem-loop II of U1-snRNA [43]. Afterwards, the complex with stem-loop II was also solved by nuclear magnetic resonance (NMR). Intermolecular recognition of both RNAs required not only conformational changes in the protein, shifting loop 3 and the C-terminal extension, but also in the RNA, kinking it by induced fitting.

The general fold for U1A is a four-stranded β-sheet flanked on one side by two α-helices. The complex for PIE shows the β-sheet surface interacting with the bases, whereas the phosphate backbone points towards the solvent, away from the protein. This implies that electro-

Table 3 Comparison between U1A RNP and ADA4h RNP-like motifs

RNP	U1A	ADA4h	Observation
RNP2	Ile12	Leu10	Substitution fits the standard motif
	Tyr13	Arg11	Stacking with C40 is kept
	Ile14	Ile12	Substitution fits the standard motif
	Asn15	–	Represents the loss of a wildcard in the motif
	Asn16	Asn13	Fits the standard motif
	Leu17	Val14	Substitution does not affect main properties of the motif position
RNP1	Arg52	Arg45	Fits the standard motif
	Gly53	–	This loss might reduce affinity with RNA
	Gln54	Pro46	The interaction with G39 is lost
	Ala55	Val47	Substitution does not affect main properties of the motif position
	Phe56	Asp48	Stacking with A41 is kept
	Val57	Val49	Fits the standard motif
	Ile58	Leu50	Wildcard position
	Phe59	Val51	This loss might reduce affinity with RNA

static interactions are not determinant for binding and that hydrophobic contact must be regarded as the main force fitting together both molecules. This is also the case for stem-loop II.

PIE and stem-loop II have a seven-nucleotide sequence that interacts with the β-sheet and the C-terminal extension of U1A. The RNP1 and RNP2 sequences of the β-sheet interact with this ssRNA in loop conformation. These interactions are also shown by other RNA protein complexes, and are mainly responsible for unspecific contacts. The fact that ADA4h contains RNP1-like and RNP2-like sequences could explain the experimentally observed binding to the bipolar-stem-loop constructed in this work. Differences between U1A RNP and ADA4h RNP-like motifs are summarized in Table 3. Finally, two hydrophobic interactions are important for U1A, those of Ser46 and Lys88; Pro40 and Glu75 of ADA4h align to them, respectively.

Concerning the contacts with the dsRNA structure, loops 1 and 3 are the interacting loops and have been pointed out as the key residues for specificity. An important region in U1A is an unusual cluster of six basic residues. Lys20 is found aligned to Asn16 of ADA4h but is preceded by Arg15, a residue more basic than Lys. Lys22 and Lys23 have no basic counterpart in ADA4h. Although neither Arg47 nor Lys50 align with Lys38 of ADA4h, they are all located in loop 3 near the RNA-binding interface. Finally, the most important basic residue in U1A, and characteristic of RNP1, Arg52, aligns perfectly to Arg45 in ADA4h. Specificity in these loops is not a common feature of all RRMs, since their size and distribution of basic amino-acids is quite variable [44]. Thus, it is fairly possible that ADA4h, with a cluster of three basic residues in loops 1 and 3, would use this mechanism for molecular discrimination to a lesser extent than does U1A.

Loop 3 is the most variable segment between RRMs, and determines specificity in the case of the U1A complex. However, important residues in U1A have their counterparts in ADA4h: Arg52 (Arg45 in ADA4h) contributes with several hydrogen bonds and packs against Leu49 (Phe43 in ADA4h) and the CG pair closing the structure of the RNA. The residues positioning the dsRNA portion are Ser46 (Pro40 in ADA4h), Ser48 (Ser42 in ADA4h), Leu49 (Phe43 in ADA4h) and Arg52 (Arg45 in ADA4h). This could account for the inability of 0.5 μg competitor tRNA to totally abolish the complex between ADA4h with 25 fmol bipolar-stem-loop RNA, a ratio of 2×10^7:1 (w:w), in the EMSA.

Proposed moonlighting function for ADA4h

Apart from their relevance in assisting the folding process, a known function for the activation domains of CPs is inhibition of the enzyme in the proenzyme form. The contacts in the crystal structure between ADA4h and its enzyme moiety involve one residue of RNP2 (Arg11), three residues corresponding to β-strand 2 (Asn35, Phe36, and Trp37) and one located in loop 3 (Lys38) of the domain [10]. The fact that important residues for inhibition are located in the same region determinant for RNA binding might induce the ADA4h domain to evolve from the canonical RNP1 and RNP2 sequences. This hypothesis is based on the suggestion that some RRMs might have evolved from RNA- to protein-recognition [39], as previously mentioned. It is likely that, among the activation domain of procarboxypeptidases, ADA4h is the one that best conserves RNA-binding properties. Apart from the structural approach presented in this work pointing out ADA4h as being related to RNA-binding proteins, it is noteworthy that ADA4h shares about a 20% homology with U1A, whereas ADA2h and ADB only share 11% and 3%, respectively, as determined by using the CLUSTALW multiple sequence alignment program [45].

hPCPA4 is a procarboxypeptidase expressed mainly in cancerous prostate, brain and during development. Its gene is subject to genomic imprinting and its regulation controlled by histone hyperacetylation, as observed in the sodium butyrate treatment of androgen-independent prostate-cancer PC-3 cell lines [46], which lead to growth inhibition. Very low expression has been found in the pancreas or other normal tissues. Recently, some proteins involved in different processes such as metabolism have been proved to bind RNA, participating in regulatory mechanisms [47]. Although further experiments should be performed, this could also be the case of ADA4h, as pointed out in this work, where fold homology has been revealed as a suitable tool for predicting protein function.

Conclusions

Bioinformatic tools have proven useful in the prediction of both the possible function of ADA4h as an RNA-interacting protein and the affinity of such interactions. Our results seem to indicate that activation domains of procarboxypeptidases would have lost some of their ability to bind RNA during evolution as a consequence of the optimization of the inhibition of the catalytic domain.

Acknowledgments N.C.C. was supported by an AGAUR FI fellowship (2003 FI 00388, Generalitat de Catalunya). J.B. and B.O. acknowledge support from MEC (BIO08-0206). F.X.A. acknowledges financial support from the CICYT (BIO04-2046). S.V. acknowledges funding from the Fondo de Investigación Sanitaria (FIS-PI07-0148), the Fundación Mutua Madrileña (FMM-08) and the Generalitat de Catalunya (SGR-2009-00761).

References

1. Vendrell J, Aviles FX, Fricker LD (2004) Metallocarboxypeptidases. In: Messerschmidt A, Bode W, Cygler M (eds) Handbook of metalloproteins, vol 3. Wiley, New York, pp 176–189
2. Rawlings ND, Morton FR, Barrett AJ (2006) MEROPS: the peptidase database. Nucleic Acids Res 34(Database issue):D270–D272
3. Hendriks DF (1998) Carboxypeptidase. In: Barrett AJ, Rawlings ND, Woessner JF (eds) Handbook of proteolytic enzymes. Academic, London, pp 1328–1330
4. Springman EB, Dikov MM, Serafin WE (1995) Mast cell procarboxypeptidase. A molecular modeling and biochemical characterization of its processing within secretory granules. J Biol Chem 270:1300–1307
5. Murzin AG et al (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540
6. Villegas V, Vendrell J, Aviles X (1995) The activation pathway of procarboxypeptidase B from porcine pancreas: participation of the active enzyme in the proteolytic processing. Protein Sci 4:1792–1800
7. Vilanova M et al (1988) Analysis of the conformation and ligand-binding properties of the activation segment of pig procarboxypeptidase A. Biochem J 251:901–955
8. Vendrell J et al (1989) Procarboxypeptidase A activation segment compared to structures of other proteins. Protein Seq Data Anal 2:461–462
9. Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
10. Garcia-Castellanos R et al (2005) Detailed molecular comparison between the inhibition mode of A/B-type carboxypeptidases in the zymogen state and by the endogenous inhibitor latexin. Cell Mol Life Sci 62:1996–2014
11. Berman HM et al (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242
12. Russell RB, Barton GJ (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 14:309–323
13. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815
14. Hulo N et al (2006) The PROSITE database. Nucleic Acids Res 34(Database issue):D227–D230
15. Aloy P et al (2002) Structural similarity to link sequence space: new potential superfamilies and implications for structural genomics. Protein Sci 11:1101–1116
16. Espadaler J et al (2006) Identification of function-associated loop motifs and application to protein function prediction. Bioinformatics 22:2237–2243
17. de Castro E et al (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34(Web Server issue):W362–W365
18. Viguera AR et al (1997) Favourable native-like helical local interactions can accelerate protein folding. Fold Des 2:23–33
19. Allain FH et al (1997) Structural basis of the RNA-binding specificity of human U1A protein. EMBO J 16:5764–5772
20. Tang A, Curthoys NP (2001) Identification of zeta-crystallin/NADPH:quinone reductase as a renal glutaminase mRNA pH response element-binding protein. J Biol Chem 276:21375–21380
21. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the ExPASy server.In: Walker (JM) The proteomics protocols handbook. Humana, Totowa, NJ, pp 571–607
22. Barbosa Pereira PJ et al (2002) Human procarboxypeptidase B: three-dimensional structure and implications for thrombin-activatable fibrinolysis inhibitor (TAFI). J Mol Biol 321:537–547
23. Wei S et al (2002) Identification and characterization of three members of the human metallocarboxypeptidase gene family. J Biol Chem 277:14954–14964
24. Rupert PB, Ferre-D'Amare AR (2001) Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. Nature 410:780–786
25. Gargallo R et al (2003) Molecular dynamics simulation of highly charged proteins: comparison of the particle-particle particle-mesh and reaction field methods for the calculation of electrostatic interactions. Protein Sci 12:2161–2172
26. Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17:355–362
27. Aloy P, Oliva B (2009) Splitting statistical potentials into meaningful scoring functions: testing the prediction of near-native structures from decoy conformations. BMC Struct Biol 9:71
28. Benkert P, Schwede T, Tosatto SC (2009) QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information. BMC Struct Biol 9:35
29. van Gunsteren W (1996) Simulations: The GROMOS96 Manual and User Guide WdF Hochschulverlag ETHZ
30. Melo F, Feytmans E (1998) Assessing protein structures with a non-local atomic interaction energy. J Mol Biol 277:1141–1152
31. Mittermaier A et al (1999) Changes in side-chain and backbone dynamics identify determinants of specificity in RNA recognition by human U1A protein. J Mol Biol 294:967–979
32. Nagai K et al (1995) The RNP domain: a sequence-specific RNA-binding domain involved in processing and transport of RNA. Trends Biochem Sci 20:235–240
33. Finn RD et al (2006) Pfam: clans web tools and services. Nucleic Acids Res 34(Database issue):D247–D251
34. Fernandez MR et al (2007) Human and yeast zeta-crystallins bind AU-rich elements in RNA. Cell Mol Life Sci 64:1419–1427
35. Maris C, Dominguez C, Allain FH (2005) The RNA recognition motif a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J 272:2118–2131
36. Adam SA et al (1986) mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence. Mol Cell Biol 6:2932–2943
37. Swanson MS et al (1987) Primary structure of human nuclear ribonucleoprotein particle C proteins: conservation of sequence and domain structures in heterogeneous nuclear RNA mRNA and pre-rRNA-binding proteins. Mol Cell Biol 7:1731–1739
38. Dreyfuss G, Swanson MS, Pinol-Roma S (1988) Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. Trends Biochem Sci 13:86–91
39. Kielkopf CL, Lucke S, Green MR (2004) U2AF homology motifs: protein recognition in the RRM world. Genes Dev 18:1513–1526
40. DeAngelo DJ et al (1995) The embryonic enhancer-binding protein SSAP contains a novel DNA-binding domain which has homology to several RNA-binding proteins. Mol Cell Biol 15:1254–1264
41. Kim S et al (1997) Identification of N(G)-methylarginine residues in human heterogeneous RNP protein A1: Phe/Gly-Gly-Gly-Arg-Gly-Gly-Gly/Phe is a preferred recognition motif. Biochemistry 36:5185–5192
42. Allain FH et al (1996) Specificity of ribonucleoprotein interaction determined by RNA folding during complex formulation. Nature 380:646–650

43. Oubridge C et al (1994) Crystal structure at 192 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. Nature 372:432–438
44. Burd CG, Dreyfuss G (1994) Conserved structures and diversity of functions of RNA-binding proteins. Science 265:615–621
45. Chenna R et al (2003) Multiple sequence alignment with the clustal series of programs. Nucleic Acids Res 31:3497–3500
46. Huang H et al (1999) Carboxypeptidase A3 (CPA3): a novel gene highly induced by histone deacetylase inhibitors during differentiation of prostate epithelial cancer cells. Cancer Res 59:2981–2988
47. Ciesla J (2006) Metabolic enzymes that bind RNA: yet another level of cellular regulatory network? Acta Biochim Pol 53:11–32