3592–3601 *Nucleic Acids Research, 2002, Vol. 30 No. 16*

Metallo-β-lactamase fold within nucleic acids processing enzymes: the β-CASP family

Isabelle Callebaut*, Despina Moshous[1], Jean-Paul Mornon and Jean-Pierre de Villartay[1]

Systèmes moléculaires et Biologie structurale, LMCP, CNRS UMR 7590, Universités Paris 6 et Paris 7, case 115, 4 place Jussieu, F-75252 Paris Cedex 05, France and [1]Développement Normal et Pathologique du Système Immunitaire, INSERM U429, Hôpital Necker Enfants Malades, 149 rue de Sèvres, F-75015 Paris, France

Received April 11, 2002; Revised and Accepted June 20, 2002

ABSTRACT

A separate family of enzymes within the metallo-β-lactamase fold comprises several important proteins acting on nucleic acid substrates, involved in DNA repair (Artemis, SNM1 and PSO2) and RNA processing [cleavage and polyadenylation specificity factor (CPSF) subunit]. Proteins of this family, named β-CASP after the names of its representative members, possess specific features relative to those of other metallo-β-lactamases, that are concentrated in the C-terminal part of the domain. In this study, using sensitive methods of sequence analysis, we identified highly conserved amino acids specific to the β-CASP family, some of which were unidentified to date, that are predicted to play critical roles in the enzymatic function. The identification and characterisation of all the extant, detectable β-CASP members within sequence databases and genome data also allowed us to unravel particular sequence features which are likely to be involved in substrate specificity, as well as to describe new but as yet uncharacterised members which may play critical roles in DNA and RNA metabolism.

INTRODUCTION

Metallo-β-lactamase fold proteins constitute a large superfamily of proteins possessing a wide variety of substrates, most of them having in common an ester linkage and a negative charge (1). Besides class B β-lactamases hydrolysing lactams, this superfamily includes among others, glyoxalase II, aryl sulfatases, cytidine monophosphate-*N*-acetyl neuraminic acid (CMP-NeuAc) hydrolases, cAMP phosphodiesterases and the phnP protein, involved in alkylphosphonate uptake (1–3). A separate group within the metallo-β-lactamase fold superfamily comprises proteins with nucleic acid substrates. Among these, the 73 kDa subunit of cleavage and polyadenylation specificity factor (CPSF) and its yeast orthologue Ysh1p, are involved in RNA processing whereas mouse SNM1 and yeast PSO2 are implicated in DNA crosslink repair (1). Artemis, a novel member of this group, has recently been identified as involved in V(D)J recombination/DNA repair and mutations of which cause human severe combined immune deficiency with increased radiosensitivity (RS-SCID) (4). Following this first characterisation, Artemis was also shown to possess intrinsic single-strand-specific 5′ to 3′ exonuclease activity which is modified to an endonuclease activity when Artemis forms a complex with the DNA-dependent protein kinase DNA-PKcs (5), consistent with its presumed catalytic activity (4).

Metallo-β-lactamase fold consists of a four-layered β-sandwich with two mixed β-sheets flanked by α-helices, with the metal-binding sites located at one edge of the β-sandwich (Fig. 1) (6). The dinuclear Zn(II) centre, used to perform the cleavage reaction, is located at the bottom of a wide shallow groove (Fig. 1) (6). Five sequence motifs, consisting mostly of histidine and aspartic acid residues, are highly conserved in active enzymes of the superfamily and participate in zinc coordination and hydrolysis reaction. The first two motifs are located at the end of two β-strands of the first β-sheet (red and yellow in Fig. 1). Motif 2 (yellow in Fig. 1) is typical of the entire superfamily and is typified by the highly conserved HxHxDH signature, in which the first histidine and the aspartic acid are invariant. The third and fifth motifs, ending strands of the second β-sheet (pink and green in Fig. 1), each contain a conserved histidine whereas the fourth one, also located at the end of an in-between β-strand, contains an acidic residue or a cysteine (violet in Fig. 1). However, as noticed by Aravind (1) and Daiyasu *et al.* (3), the length between motifs 4 and 5 is predicted to be extremely variable, particularly important for metallo-β-lactamases acting on nucleic acids, but these authors disagree on the location of the catalytic histidine of motif 5 within this particular family.

Prompted by our interest in the Artemis function in DNA repair, we undertook a detailed analysis of its sequence C-terminal to motif 4, up to which the similarity with proteins of the metallo-β-lactamase superfamily can be significantly detected against domain databases. Indeed, motif 5 could not be easily identified in Artemis. Instead, this C-terminal sequence shares obvious similarity with yeast PSO2 and mouse SNM1 and, to a lower extent, with the 73 kDa subunit of CPSF (4). The conserved sequences lying after the four typical metallo-β-lactamase motifs (motifs 1–4) therefore constitute a hallmark of proteins of this family specifically acting on nucleic acids. They are not restricted to a limited

*To whom correspondence should be addressed. Tel: +33 1 44 27 45 87; Fax: +33 1 44 27 37 85; Email: isabelle.callebaut@lmcp.jussieu.fr

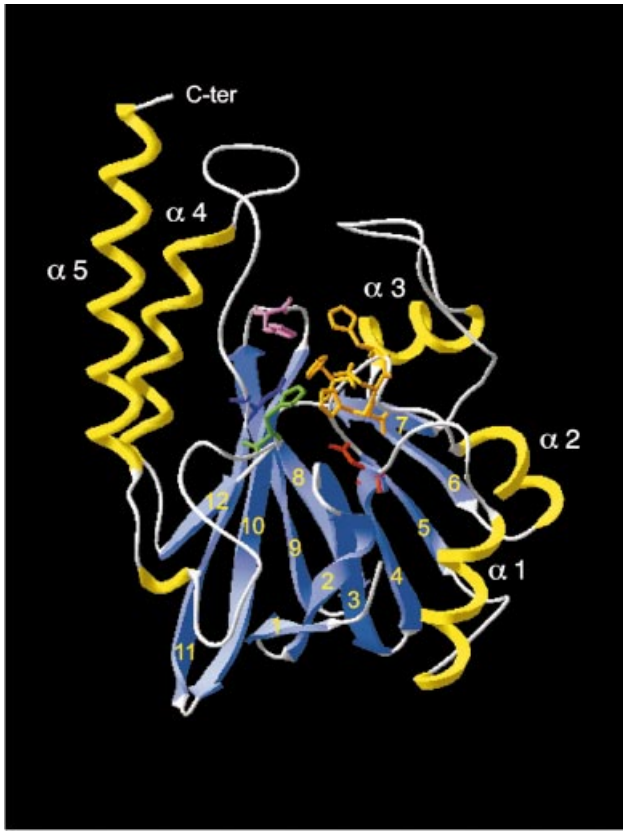


Figure 1. Three-dimensional representation of a metallo-β-lactamase domain. The structure of *Stenotrophomonas maltophilia* metallo-β-lactamase (26; PDB identifier 1SML) is used for illustration purposes. The two β-sheets are formed by strands β1–β7 and β8–β12, respectively. Side chains of amino acids participating in zinc binding are coloured red (motif 1 after strand β5), yellow (motif 2 after strand β6), pink (motif 3 after strand β9/), violet (motif 4 after strand β11) and green (motif 5 after strand β12). Clear sequences similarities are found with members of the metallo-β-lactamase superfamily up to motif 4, whereas the region linking motif 4 to motif 5, comprising here helix α4, is proposed to be extremely large for members of the β-CASP family discussed here. This figure was prepared using Swiss-PdbViewer (27).

region, the length of which should correspond to that including the metallo-β-lactamase motif 5, but share features of a distinct globular domain that we named the β-CASP motif, after metallo-β-lactamase-associated CPSF Artemis SNM1/PSO2. Using a combination of profile-based and bidimensional methods of sequence analysis, we highlighted all detectable extant sequences that make part of the 'β-CASP family' in the three primary kingdoms (eukaryotes, bacteria and archaea) and found several conserved motifs, some of which are yet undescribed. These highly conserved motifs, including two histidine and an acidic residue, are likely to play a key role in the structure and/or function of this family within the metallo-β-lactamase superfamily. Moreover, specific features of these motifs allow distinguishing between enzymes acting on DNA substrates from those involved in RNA metabolism. The comprehensive sequence analysis presented here is more useful for the characterisation of the β-CASP family functions as sequence divergence hampers its fully automatic description.

MATERIALS AND METHODS

Domain databases searches were performed using RPS-BLAST 2.2.1 running at the National Center for Biological Information (NCBI; http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) and a HMM search running at the Sanger Center (http://www.sanger.ac.uk/Software/Pfam/search.shtml).

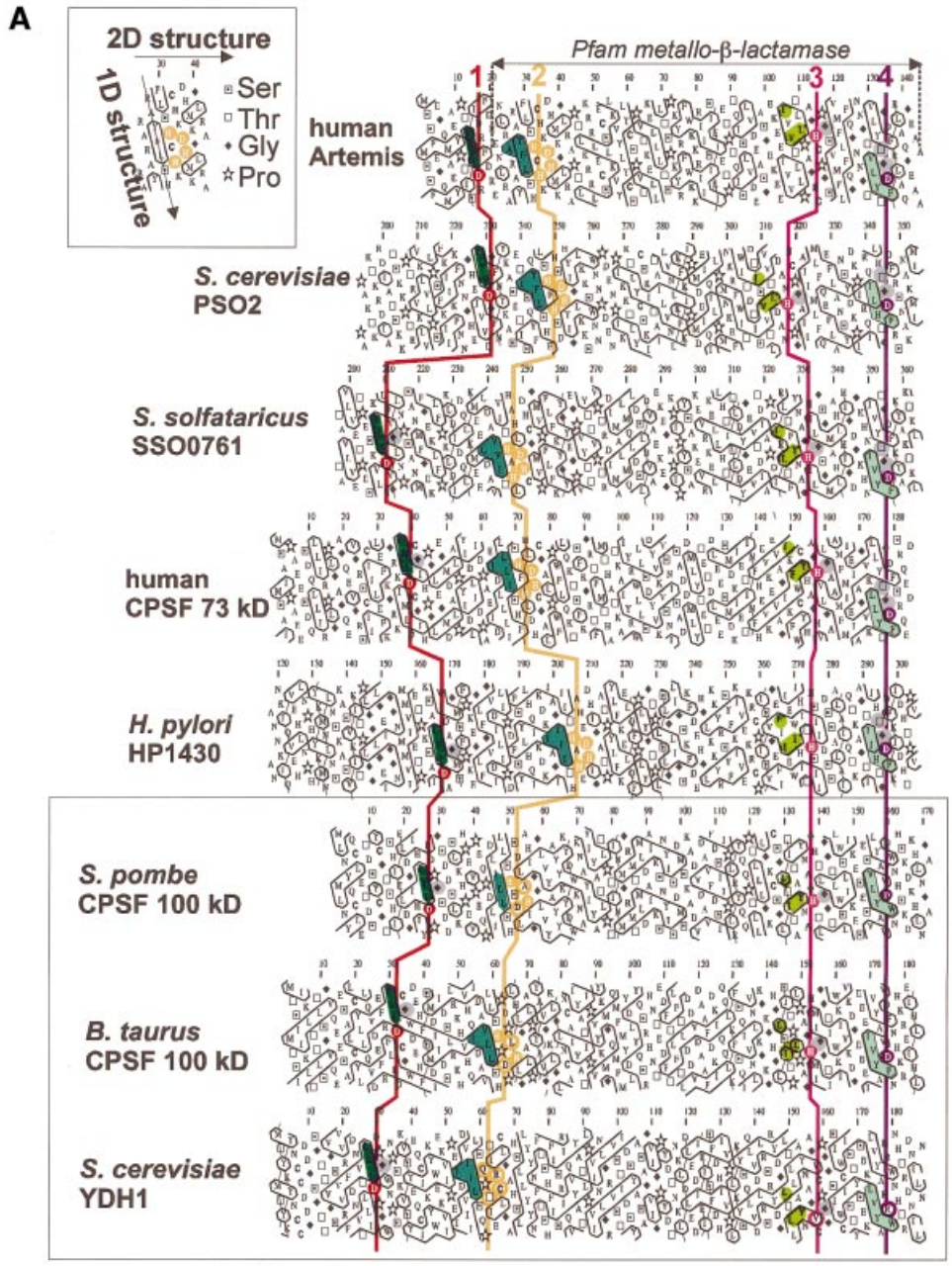
The non-redundant database (NRDB) at NCBI was searched using PSI-BLAST (7). We also used the bidimensional hydrophobic cluster analysis (HCA) (8,9), which offers the possibility to add information about secondary structures to the lexical analysis of the considered sequences. The sequence is handled on a duplicated α-helical net in which hydrophobic amino acids (V, I, L, F, M, Y, W) are contoured. The defined hydrophobic clusters [i.e. hydrophobic amino acids that are separated from each other by at least four non-hydrophobic residues (connectivity distance linked to the use of a α-helical support)] were shown to mainly correspond to the internal faces of regular secondary structures (α-helices or β-strands) (10). Conservation of hydrophobic cluster features which participate in the protein core, together with a similar texture of sequence similarities, are associated with the maintenance of a similar structure and often help and/or allow the alignment procedure for highly divergent sequences (typically in the 10–20% sequence identity range, below the so-called 'twilight zone' of 25–30% identity). The sensitivity of this approach, combined with profile-based lexical tools, has often been successfully used to identify new domains (11,12) and/or to link orphan sequences to particular structural and functional families (13–15).

RESULTS

The Artemis sequence was searched against domain databases using two different programs: HMM search and RPS-BLAST. A metallo-β-lactamase domain (Pfam00753; 16) was highlighted between amino acids 5 and 173 after the HMM search (*E*-value 0.062) or amino acids 21 and 145 with RPS-BLAST (*E*-value 8×10^{-5}). The alignment performed using RPS-BLAST ends after the conserved Gly–Asp sequence (motif 4) whereas the alignment obtained using the HMM search is much longer. However, in this last case, the conserved histidine residue of the metallo-β-lactamase family (motif 5) is missing in Artemis, being aligned with a phenylalanine, which cannot substitute histidine for zinc binding. Moreover, the C-terminal region of the HMM alignment (amino acids 142–173), after the conserved Gly–Asp sequence (motif 4), shares much less sequence identity/similarity with the Pfam metallo-β-lactamase profile, indicating that this part of the alignment is probably fortuitous. This observation prompted us to analyse further sequences after motif 4, with the particular aim of identifying the missing conserved histidine (motif 5) of the metallo-β-lactamase signature.

Identification of the members of the β-CASP family

Thus, we searched for similarities of the Artemis sequence with other proteins after motif 4 and up to amino acid 370 (ending a globular domain, as indicated after HCA, star in Fig. 2B). Using this query (amino acids 142–370), PSI-BLAST searches against the NRDB (891 607 sequences) at NCBI revealed by iteration 1 significant similarities with



several proteins including mouse SNM1. Further iterations highlighted similarities with other proteins, including yeast PSO2 [17 significant matches (*E*-value <0.002) at convergence by iteration 3]. Marginal but interesting similarities were also observed just above the threshold *E*-value (*E* = 1.5) with a hypothetical protein from *Sulfolobus solfataricus*, described as a putative mRNA 3′-end polyadenylation factor (SSO0761, see Table 3). This similarity was supported at the structural level as a true relationship using HCA (see Materials and Methods for details; Fig. 2B). Moreover, a metallo-β-lactamase signature was also detected just before the similarity region using RPS-BLAST (Fig. 2A). Thus, the *S.solfataricus* sequence was included in a new position specific score matrix (PSSM), which was then used for additional iterations. This led to the identification at convergence by iteration 16 of numerous proteins, including various hypothetical prokaryotic proteins, the 73 kDa subunit of CPSF from mammals (iteration 4) and yeast (Ysh1, iteration 5) as well as the 100 kDa subunit of CPSF from mammals and from fission yeast (iteration 8) (Tables 1–3).

The similarities of the detected sequences were confirmed by reciprocal iterative strategies, sometimes extending the detected family to other members, as in the case of the *Saccharomyces cerevisiae* Cft2/Ydh1 sequence, orthologous to the mammalian CPSF 100 kDa subunits. Most of the matching proteins possess, before the detected regions,

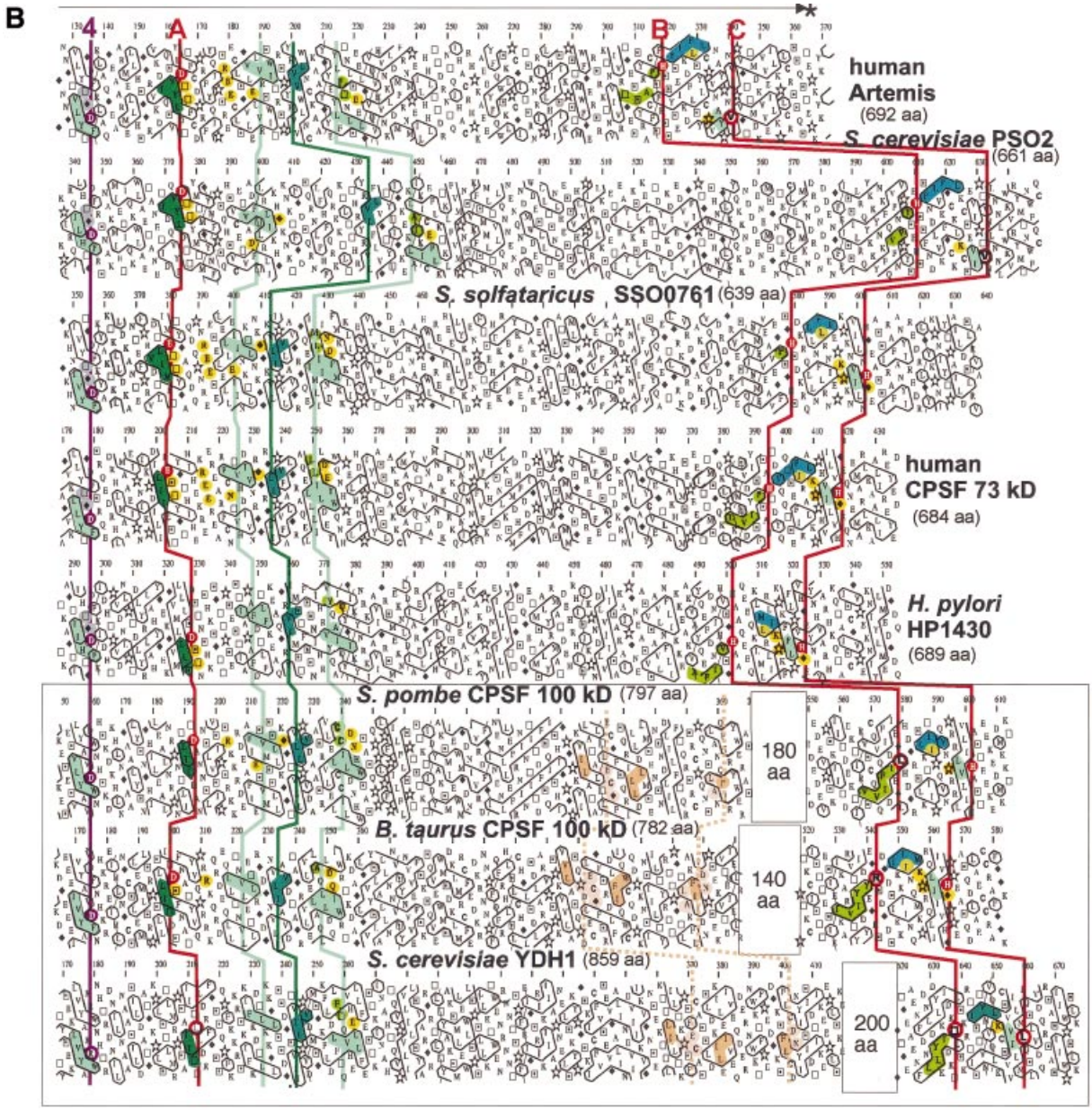


Figure 2. (Opposite and above) Comparison of the HCA plots of several members of the β-CASP family. (**A**) The first four metallo-β-lactamase conserved motifs; (**B**) the β-CASP region. The interest of using HCA for sequence comparison, especially at low levels of identity, lies in the possibility that such a representation offers to combine lexical analysis (1D structure) and secondary structure localisation. Briefly, the sequence is written on a duplicated α-helical net in which strong hydrophobic amino acids (V, I, L, F, M, Y, W) are contoured. These are not randomly distributed but instead form clusters that were shown to mainly correspond to the internal faces of regular secondary structures (α-helices and β-strands). A cluster thus includes hydrophobic but also non-hydrophobic amino acids lying between its first and last hydrophobic amino acids. Non-hydrophobic sequences separating clusters mainly correspond to loop regions. The observed good correspondence between hydrophobic clusters and regular secondary structures is linked to the use of an appropriate connectivity distance (i.e. the minimum distance separating two hydrophobic amino acids belonging to two different clusters) related to the α-helical pitch. The way to read the sequence and special symbols are indicated in the inset. As secondary structures are often much more stable than sequences, a good conservation of key hydrophobic clusters, whose hydrophobic amino acids participate in the protein core, is generally observed. Hence, it is therefore possible to appreciate the required fold conservation for remote sequence relationships. Conservation of key hydrophobic clusters is indicated in green, with vertical lines highlighting correspondences. Highly conserved amino acids of the four metallo-β-lactamase domain motifs are coloured red, yellow, pink and violet (A), according to Figure 1. The β-CASP ones are coloured red (B, motifs A–C). Amino acids that substitute these highly conserved amino acids in some sequences (e.g. CPSF 100 kDa) are contoured accordingly. Other conserved, non-hydrophobic amino acids are coloured grey and yellow. Local similarities between the CPSF 100 kDa proteins are indicated in orange. The C-terminal limit of the globular domain following the four metallo-β-lactamase motifs (β-CASP motif) is indicated with a star and the total number of amino acids within each protein is given within brackets. The three sequences belonging to the CPSF 100 kDa group are boxed. These have lost part or most of the highly conserved amino acids. Moreover, long intervening sequences separate motifs B and C from the rest of the domain.

Table 1. Members of the βCASP family: eukaryotes

Species	Protein/gene name	gi number	Total	Motifs A/B/C	Specific features
Homo sapiens	Artemis	gi|13872809|	692	D165/H319/V341	
Homo sapiens	FLJ12810	gi|12383082|	532	D145/H276/V298	SNM1C (1)
Homo sapiens	KIAA0086	gi|577303|	1040	D838/H994/V1016	SNM1/PSO2 group
Homo sapiens	CPSF 73 kD	gi|7706427|	684	E204/H396/H418	CPSF 73 kD
Homo sapiens	CPSF 100 kD	gi|17477312|	782	D200/R543/H565	CPSF 100 kD
Homo sapiens	DKFZp434A1923	gi|12053137|	600	E203/H392/H414	CPSF 73 kD-related (2)
Macaca fascicularis	AB045994.1	gi|9280039|	328	E139 /nd/nd(3)	CPSF 73 kD-related (2)
Mus musculus	BC011094.1	gi|15029751|	452	D56/H187/V209	SNM1C (1)
Mus musculus	SNM1 protein	gi|9055350|	1023	D821/H977/V999	SNM1/PSO2 group
Mus musculus	CPSF73 kD	gi|9055194|	684	E204/H396/H418	CPSF 73 kD
Mus musculus	CPSF100 kD	gi|8393762|	782	D201/R543/H565	CPSF 100 kD
Mus musculus	AK010425.1	gi|12845859|	600	E203/H392/H414	CPSF 73 kD-related (2)
Bos taurus	CPSF73 kD	gi|1707412|	684	E204/H396/H418	CPSF 73 kD
Bos taurus	CPSF100 kD	gi|1706103|	782	D201/R543/H565	CPSF 100 kD
Xenopus laevis	CPSF 100 kD	gi|4927240|	783	D201/R543/H565	CPSF 100 kD
Oikopleura dioica	BAC0001.26	gi|18029276|	765	D200/R560/R582	CPSF 100 kD
Drosophila melanogaster	CG10018	gi|7296732|	763	D403/H553/V575	SNM1/PSO2 group
Drosophila melanogaster	CG7698	gi|7300421|	705	E210/H402/H424	CPSF 73 kD
Drosophila melanogaster	LD14168	gi|18488729|	756	D201/R554/H576	CPSF 100 kD
Drosophila melanogaster	CG1972	gi|18488805|	597	E203/H392/H414	CPSF 73 kD-related (2)
Caenorhabditis elegans	F39H2.1	gi|7503021|	587	D408/H539/C560	SNM1C (1) (4)
Caenorhabditis elegans	Y67H2A.1	gi|17543978|	1252	E205/H397/H419	CPSF 73 kD, (5)
Caenorhabditis elegans	F09G2.4	gi|17559452|	843	G200/I577/H599	CPSF 100 kD
Caenorhabditis elegans	F10B5.8	gi|7498794|	474	E208/H396/H418	CPSF 73 kD-related (2)
Arabidopsis thaliana	F14D16.17	gi|8778278|	612	D219/H385/V407	SNM1C (1)
Arabidopsis thaliana	F17L21.20	gi|15223519|	422	D146/H313/V335	SNM1C (1)
Arabidopsis thaliana	At1g66730	gi|15219728|	1417	D198/H355/V378	SNM1C, domain homologous to eukaryotic DNA ligase I (1011-1383)
Arabidopsis thaliana	orf12	gi|1495267|	484	D284/H441/V463	SNM1/PSO2 group
Arabidopsis thaliana	F17K2.23	gi|7485634|	723	D506/H684/V706	SNM1/PSO2 group
Arabidopsis thaliana	At1g61010	gi|15219848|	693	E214/H406/H428	CPSF 73 kD
Arabidopsis thaliana	CPSF 100 kDa	gi|9082326|	739	D200/R547/H569	CPSF 100 kD
Arabidopsis thaliana	At2g01730	gi|15226342|	837	E202/H392/H414	CPSF 73 kD-related (2)
Arabidopsis thaliana	At5g63420	gi|15242755|	528	D299/H472/H492	UPF0036 (6), 38 % identity over 444 aa with *C. glutamicum* ORF4
Saccharomyces cerevisiae	PSO2	gi|6323786|	661	D376/H611/V633	SNM1/PSO2 group
Saccharomyces cerevisiae	Ysh1p	gi|6323307|	779	E209/H408/H430	CPSF 73 kD
Saccharomyces cerevisiae	Ydh1p	gi|6323144|	859	R179/T638/L660	CPSF 100 kD
Schizosaccharomyces pombe	C56F8.17C	gi|6176585|	539	D305/H482/V504	SNM1/PSO2 group
Schizosaccharomyces pombe	SPAC17G6.16c	gi|19115175|	775	E229/H421H443	CPSF 73 kD
Schizosaccharomyces pombe	SPBC1709.15c	gi|19112240|	797	D193/L580/H602	CPSF 100 kD
Neurospora crassa	B8L21.110	gi|18376069|	850	E220/H442/H464	CPSF 73 kD
Plasmodium falciparum	MAL3P6.24	gi|16805270|	1017	E339/H527/H549	CPSF 73 kD-related (2)
Encephalitozoon cuniculi	ECU10_0900	gi|19074699|	730	E286/H478/H500	CPSF 73 kD-related (2)
Encephalitozoon cuniculi	ECU10_1350	gi|19074744|	496	E198/H386/H408	CPSF 73 kD-related (2)

Members of βCASP family, sorted by kingdom. These were identified by running to convergence the PSI-BLAST program against the NRDB at NCBI (BLASTP 2.2.2, 891 607 sequences, default values) and using the Artemis sequence as the query (amino acids 142–370). Proteins were then manually checked and clustered by sequence similarity, as deduced from PSI-BLAST searches using as the query each protein sequence found in the significant results of the initial PSI-BLAST. These 'secondary' searches also allowed highlighting of the most divergent members, such as the yeast Ydh1p that was only found using the other CPSF 100 kDa sequences as queries.

The positions of the motifs A, B and C described here were checked and sometimes revealed using HCA. Motifs A, B and C most often correspond to D or E, H and H, respectively, but this consensus is not respected in some cases (underlined amino acids). Motif C histidine is indeed substituted by other amino acids (V or C) in proteins involved, or predicted to be involved, in DNA metabolism, whereas one or several of the highly conserved amino acids of the three motifs A, B and/or C are not present in proteins thought to be devoid of catalytic activity (e.g. CPSF 100 kDa). These predicted inactive proteins also miss some conserved amino acids of motifs 1–4, in contrast to proteins involved in DNA metabolism which always conserve them.

The presence of defined domains was searched for using the NCBI CD-search and completed by standard similarity searches which reveal the position of conserved but more divergent domains in orthologous sequences (e.g. the KH domain was not significantly detected using CD searches in all of the orthologous sequences of *Halobacterium* Epf2, although it is clearly detected in the same sequences by standard similarity searches).

(1) SNM1C: these proteins are clearly related to the SNM1 and Artemis sequences (the sequence of the human FLJ12810 β-CASP domain shares 35 and 22% of identity with human SNM1 and Artemis, respectively, as well as highly conserved islets of amino acids around motifs A, B and C).

(2) CPSF 73 kDa-related: the proteins of this group share a high level of sequence similarity with the CPSF 73 kDa proteins, although being clearly distinct (e.g. the human DKFZp434A1923 β-CASP sequence shares 40% identity with human CPSF 73 kDa).

(3) Nd, not determined (sequence not included in the predicted protein).

(4) The N-terminal region of this predicted protein shares high sequence similarities with telomere-binding proteins (*C.elegans* F57C2.3.p and 3R5.1.p).

(5) Predicted end of the sequence at amino acid 567, the C-terminal of the predicted sequence being identical to synaptojanin UNC-26A.

(6) UPF0036: the β-CASP sequence of these proteins matches the uncharacterised Pfam family UPF0036.

(7) α conserved: proteins which are highly conserved in alpha subdivision of Proteobacteria.

sequences belonging to the metallo-β-lactamase superfamily, as detected against the Pfam database (presence of motifs 1–4). Those sequences that do not match the Pfam metallo-β-lactamase profile, however, possess a metallo-β-lactamase fold, as shown by global similarity searches, but have lost some or all of the conserved amino acids of the four consecutive motifs (Tables 1–3). This observation suggests that the considered region, ranging from amino acids 142 to 370 in Artemis and specific to the defined β-CASP family (and for this reason, named the 'β-CASP' motif), does not form an independent domain, but rather should be associated with, or even integrated to the metallo-β-lactamase domain for playing a particular role, probably relative to nucleic acids as all of the characterised detected members appear specific for this kind of substrate (see below). The addition of N-terminal sequences to the β-CASP signature of Artemis in a new database search

Table 2. Members of the βCASP family: bacteria[a]

Species	Protein/gene name	gi number	Total	Motifs A/B/C	Specific features
Agrobacterium tumefaciens str. C58	Atu0839	gi|17934747|	331	E133/H278/H301	α conserved (7), =gi|15888180|(352 aa)
Agrobacterium tumefaciens str. C58	Atu1285	gi|17935185|	555	D198/H370/H392	UPF0036 (6) , =gi|15888614|(579 aa)
Bacillus halodurans	BH2398	gi|15614961|	555	D196/H368/H390	UPF0036
Bacillus halodurans	BH2662	gi|15615225|	555	D196/H368/H390	UPF0036, 53 % identity with BH2398
Bacillus subtilis	ykqc	gi|16078517|	555	D185/H368/H390	UPF0036
Bacillus subtilis	ymfA	gi|16078741|	515	D124/H328/H350	UPF0036, 49 % identity with ykqc
Brucella melitensis	BMEI1143	gi|17987426|	558	D200/H372/H394	UPF0036
Campylobacter jejuni	Cj1710c	gi|15793013|	664	D304/H477/H504	UPF0036
Carboxydothermus hydrogenoformans	AF244644.1	gi|10802743|	108	nd/H21/H55	
Caulobacter crescentus	CC0811	gi|16125064|	530	E213/H407/H431	
Caulobacter crescentus	CC1934	gi|16126177|	559	D200/H373/H400	UPF0036
Caulobacter crescentus	CC3644	gi|16127874|	342	E144/H289/H311	α conserved
Clostridium acetobutylicum	CAC1683	gi|15894960|	555	D195/H368/H390	UPF0036
Clostridium perfringens	CPE1775	gi|18310757|	581	D221/H394/H416	UPF0036
Corynebacterium glutamicum	ORF4	gi|18266923|	718	D341/H514/H536	UPF0036
Deinococcus radiodurans R1	DRA0069	gi|15807737|	499	E222/H418/H440	
Deinococcus radiodurans R1	DR2417m	gi|15807662|	579	D209/nd/nd	UPF0036
Helicobacter pylori 26695	HP1430	gi|15646039|	689	D329/H502/H524	UPF0036
Helicobacter pylori J99	JHP1323	gi|15612388|	692	D332/H505/H527	UPF0036
Lactococcus lactis subsp. lactis	yciH	gi|15672272|	560	D198/H371/H393	UPF0036
Lactococcus lactis subsp. lactis	yqgA	gi|15673600|	570	S192/H365/Q387	UPF0036, 35 % identity with yciH
Listeria innocua	lin1026	gi|16800095|	555	D195/H368/H390	UPF0036
Listeria innocua	lin1473	gi|16800541|	555	D196/H368/H390	UPF0036, 45 % identity with lin1026
Listeria monocytogenes EGD-e	lmo1027	gi|16803067|	555	D195/H368/H390	UPF0036
Listeria monocytogenes EGD-e	lmo1434	gi|16803474|	555	D196/H368/H390	UPF0036, 45 % identity with lmo1027
Mesorhizobium loti	mll1350	gi|13471391|	556	D200/H372/H394	UPF0036
Mesorhizobium loti	mll5484	gi|13474573|	335	E138/H283/H305	α conserved
Mesorhizobium loti	mlr6574	gi|13475489|	535	E219/H412/H436	
Mycobacterium leprae	ML1512	gi|15827796|	558	D202/H375/H397	UPF0036
Mycobacterium tuberculosis H37Rv	Rv2752c	gi|15609889|	558	D202/H365/H397	UPF0036
Mycoplasma fermentans	orf550	gi|4587468|	550	D192/R363/Q385	UPF0036
Mycoplasma genitalium	MG139	gi|12044991|	569	E203/H377/H399	UPF0036
Mycoplasma genitalium	MG423	gi|12045283|	561	D163/Q365/G387	UPF0036, 22 % identity with MG139
Mycoplasma pneumoniae	MPN280	gi|13508019|	569	E203/H377/H399	UPF0036
Mycoplasma pneumoniae	MPN621	gi|2496428|	561	D163/N365/S387	UPF0036, 23 % identity with MPN280
Mycoplasma pulmonis	MYPU_7040	gi|15829175|	631	D191/H365/H387	UPF0036
Mycoplasma pulmonis	MYPU_7050	gi|15829176|	546	D190/P362/Q383	UPF0036, 25 % identity with MYPU_7040
Nostoc sp PCC 7120	all3220	gi|17230712|	555	E205/H377/H394	
Nostoc sp PCC 7120	all3678	gi|17231170|	589	D199/H372/H394	UPF0036
Pseudomonas aeruginosa	PA3614	gi|15598810|	467	E213/H412/H435	
Rickestsia prowazekii	RP441	gi|15604306|	560	D200/H372/H394	UPF0036
Ralstonia solanacearum	RSc0201	gi|17544920|	452	E206/H395/H419	
Rickettsia conorii	RC0613	gi15892536|	560	D200/H372/H394	UPF0036
Sinorhizobium meliloti	SMa1131	gi|16263064|	531	E215/H408/H432	
Sinorhizobium meliloti	SMc01929	gi|15965034|	564	D207/H379/H401	UPF0036
Sinorhizobium meliloti	SMc03176	gi|15966660|	336	E138/H284/H306	α conserved
Staphylococcus aureus subsp aureaus Mu50	SAV1089	gi|15924079|	565	D195/H368/H390	UPF0036,=subsp.aureus N315 (gi|15926675|)
Staphylococcus aureus subsp aureaus Mu50	SAV1275	gi|15924265|	557	D198/H370/Q392	UPF0036,=subsp.aureus N315(gi|15926858|), 40 % identity with SAV1089
Streptococcus pneumoniae TIGR4	SP0121	gi|15900063|	559	D196/H369/H391	UPF0036, 99 % identity *S. pneumoniae R6* Spr0125(gi|15902169|)
Streptococcus pneumoniae TIGR4	SP0613	gi|15900521|	553	D192/H365/Q387	UPF0036 , identical to *S. pneumoniae R6* Spr0538 (gi|15902582|), 36 % identity with SP0121
Streptococcus pyogenes M1 GAS	Spy1876	gi|15675695|	560	D197/H370/H392	UPF0036
Streptococcus pyogenes M1 GAS	Spy1020	gi|15675021|	553	D182/H365/Q387	UPF0036, 36 % identity with Spy1876
Streptomyces coelicolor	SC9A10.09	gi|7479794|	561	D205/H367/H389	UPF0036
Streptomyces toyocaensis	AF039028.1	gi|4104709|	528	D172/H344/H366	UPF0036
Synechocystis sp. PCC 6803	sll0514	gi|16332017|	554	E188/H368/H390	
Synechocystis sp. PCC 6803	slr0551	gi|16332072|	640	D189/H362/H394	UPF0036
Thermotoga maritima	Ta0613	gi|15643428|	522	E373/H477/H499	
Ureaplasma urealyticum	UU570	gi|13358135|	596	E196/H370/H392	UPF0036
Ureaplasma urealyticum	UU509	gi|13358072|	556	G195/T364/N386	UPF0036
Vibrio cholerae	wbfZ	gi|3724326|	446	E200/H401/H425	

[a]Footnotes as in Table 1.

did not highlight new members of this family that we could have missed using the β-CASP motif *stricto sensu*.

Identification of conserved amino acids within the β-CASP motif

An in-depth analysis of all the identified members of the β-CASP family was then undertaken, in particular with the aim of refining the alignments proposed by PSI-BLAST and of identifying conserved residues that could play a critical role in their function.

HCA was used to manually localise on the bidimensional level (secondary structure context) conserved motifs, as proposed in the PSI-BLAST results (Figs 2 and 3 and Tables 1–3). Several major anchor points of the alignment were highlighted; each consisting of conserved hydrophobic residues gathered into a cluster, which represents the internal face of a conserved regular secondary structure, often accompanied (upstream, within and downstream) by identical or highly conserved non-hydrophobic residues. Motif A is characterised by an acidic residue (D or E) after a stretch of hydrophobic residues typical of a β-strand structure [ϕ-ϕ-ϕ-

Table 3. Members of the βCASP family: archaea[a]

Species	Protein/gene name	gi number	Total	Motifs A/B/C	Specific features
Acidianus ambivalens	Orf2	gi|267504|	140	D64/nd/nd	
Aeropyrum pernix	APE0181	gi|14600511|	420	E187/H364/H387	
Aeropyrum pernix	APE0522	gi|14600777|	676	E399/H599/H623	KH domain
Aeropyrum pernix	APE2295	gi|14601975|	286	D98/H239/A261	
Archaeoglobus fulgidus	AF0482	gi|11498093|	632	E373/H575/H599	KH domain
Archaeoglobus fulgidus	AF0532	gi|11498143|	407	E176/H352/H37	
Archaeoglobus fulgidus	AF2361	gi|11499938|	291	E137/H241/H263	
Halobacterium sp. NRC-1	Vng1149c	gi|10580687|	399	D158/H340/H362	UPF0036
Halobacterium sp. NRC-1	Epf1	gi|15791270|	410	E181/H359/H381	
Halobacterium sp. NRC-1	Epf2	gi|15789650|	641	E379/H584/H608	KH domain
Methanobacterium thermoautotrophicum	MTH49	gi|15678078|	450	E205/H388/H410	UPF0036
Methanobacterium thermoautotrophicum	MTH912	gi|15678932|	521	E271/H477/H499	UPF0036
Methanobacterium thermoautotrophicum	MTH1203	gi|15679214|	636	E377/H579/H603	KH domain
Methanococcus jannaschii	MJ0047	gi|15668217|	428	E192/H375/H397	= *M. jannaschii* YLR277c (gi|2129074|)
Methanococcus jannaschii	MJ0162	gi|15668334|	421	E178/H366/H389	
Methanococcus jannaschii	MJ0861	gi|15669052|	448	E205/H387/H409	UPF0036
Methanococcus jannaschii	MJ1236	gi|15669421|	634	E375/H577/H601	KH domain
Pyrobaculum aerophilum	PAE0820	gi|18312205|	634	E367/H577/H601	KH domain
Pyrobaculum aerophilum	PAE3309	gi|18313978|	430	E189/H369/H392	
Pyrobaculum aerophilum	PAE3418	gi|18314051|	314	E141/H262/H284	
Pyrococcus abyssi	PAB1035	gi|14521766|	516	E272/H468/H490	
Pyrococcus abyssi	PAB1751	gi|14521123|	451	E209/H388/H410	UPF0036
Pyrococcus abyssi	PAB1868	gi|14520957|	651	E392/H594/H618	KH domain
Pyrococcus furiosus DSM 3638	PF0596	gi|18976968|	314	E155/H262/H286	
Pyrococcus furiosus DSM 3638	PF1101	gi|18977473|	439	E203/H382/H404	UPF0036
Pyrococcus furiosus DSM 3638	PF1405	gi|18977777|	651	E392/H594/H618	KH domain
Pyrococcus horikoshii	PH0466	gi|14590378|	514	E270/H466/H488	UPF0036
Pyrococcus horikoshii	PH0724	gi|14590601|	274	E116/H223/H247	
Pyrococcus horikoshii	PH1071	gi|14590907|	450	E213/H392/H414	UPF0036
Pyrococcus horikoshii	PH1404	gi|7519131|	651	E392/H594/H618	KH domain
Sulfolobus tokodaii	ST0222	gi|15920404|	328	D135/H278/D299	
Sulfolobus tokodaii	ST0343	gi|15920543|	637	E376/H574/H598	KH domain
Sulfolobus tokodaii	ST2188	gi|15922516|	422	E189/H366/H389	
Sulfolobus solfataricus	SSO0188	gi|15897139|	328	E138/H281/G303	
Sulfolobus solfataricus	SSO0386	gi|15897318|	492	E260/H437/H460	
Sulfolobus solfataricus	SSO0761	gi|15897661|	639	E381/H579/H603	KH domain
Thermoplasma acidophilum	Ta0333	gi|16081465|	407	E182/H358/H380	
Thermoplasma acidophilum	Ta0613m	gi|16082538|	639	E380/H582/H606	KH domain
Thermoplasma volcanium	TVN0437	gi|13541268|	407	E182/H358/H380	
Thermoplasma volcanium	TVN0664	gi|13541495|	639	E380/H582/H606	KH domain

[a]Footnotes as in Table 1.

(D,E)-(T,S)-T, where ϕ is a hydrophobic amino acid]. Motif B includes a histidine ending an amphiphilic β-strand structure and followed by a α-helical structure. C-terminal to this last α-helix, and at the end of another predicted β-strand, a conserved histidine (motif C) can also be found in all the sequences of the β-CASP family, with the exception of a few sequences including those of the Artemis/SNM1/PSO2 group in which this histidine is most often substituted by a valine (red circle in Fig. 2B). Other conserved motifs were detected along the compared sequences, but none include highly conserved polar residues.

The three conserved polar amino acids characterising motifs A, B and C are all located at the end of predicted β-strands, like all of the zinc-binding residues of canonical metallo-β-lactamases (Fig. 1). This particular position together with their conservation suggests that they could be located in the vicinity of the metallo-β-lactamase active site and that they could play a specific role in the probable enzymatic function of the defined β-CASP family. Accordingly, mutation of the Asp165 residue (motif A) in Artemis, as well as that of His319 (motif B), strongly compromise both the V(D)J recombinase activity and the *in vitro* endonuclease activity (5; our unpublished observations).

Motif B histidine is conserved in all members of the β-CASP family, in contrast to motif C histidine which is substituted by a valine in the Artemis/SNM1/PSO2 group. Therefore, it is possible that motif C histidine, which is otherwise conserved for CPSF 73 kDa and almost all of the CPSF 100 kDa (with the exception of yeast Ydh1p), may play an important role in the specificity of members of the β-CASP family towards RNA targets. It is worth noting that it is also conserved for most of the bacterial β-CASP members, suggesting that these as yet uncharacterised proteins could act specifically on RNA as well (Tables 2 and 3).

Interestingly, our database mining also revealed an as yet uncharacterised group of proteins with a non-histidine motif C (Table 1), which shares similarities but is distinct from Artemis and SNM1 (the human β-CASP sequence of this protein shares 22 and 35% of sequence identity with Artemis and human SNM1, respectively). This protein has been previously named SNM1C (17–19). According to our observations, the SNM1C group can thus be predicted to play a role in DNA processing, which remains to be experimentally investigated.

Refinement of alignments using HCA—the particular case of predicted inactive members of the β-CASP family

In some cases, the pairwise similarities proposed in the PSI-BLAST results, although ranging beyond motif 4 of the metallo-β-lactamase signature, were patchy and do not always encompass conserved motifs of the emerging multiple alignment (e.g. in the PSI-BLAST results relative to the

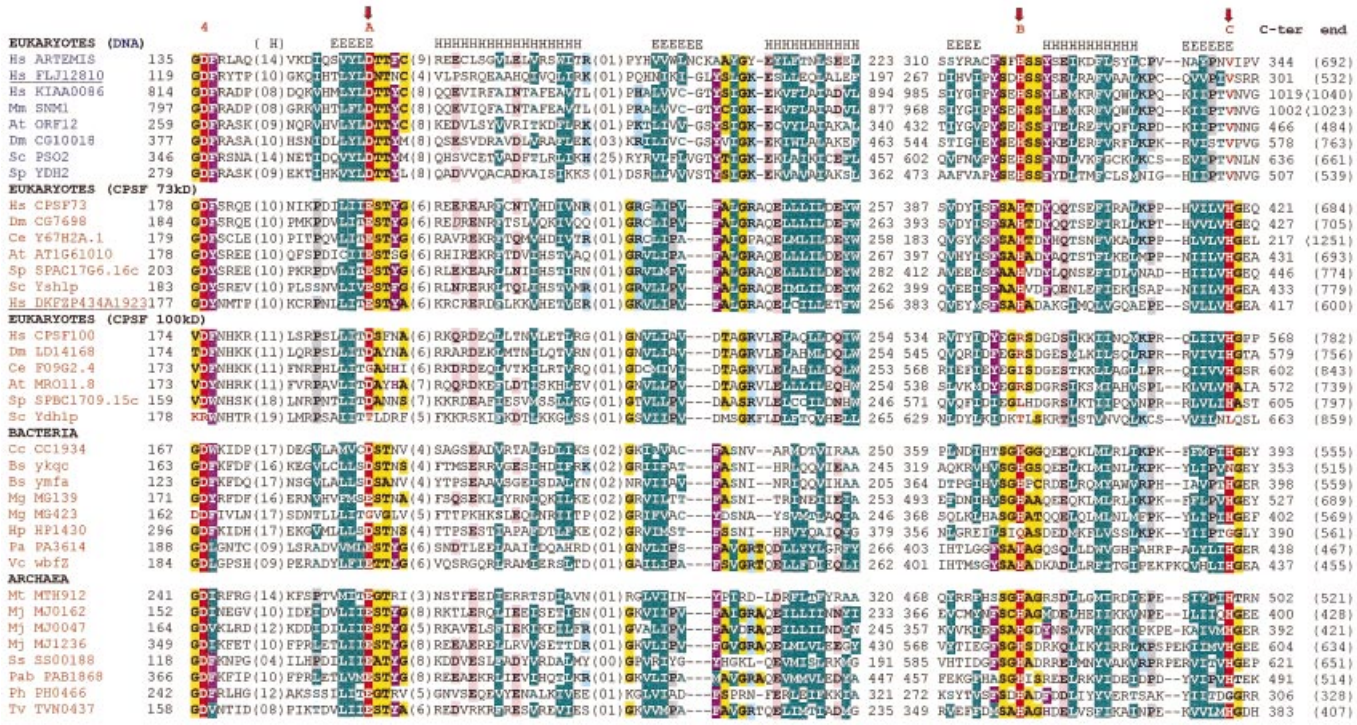


Figure 3. Multiple alignment of conserved motifs of representative members of the β-CASP family. The alignment is divided into two main blocks, the first one including metallo-β-lactamase motif 4 (with the conserved aspartic acid in red) and the β-CASP motif A (with the conserved aspartic or glutamic acid in red), the second one centred on the two β-CASP motifs B and C (conserved histidine residues in red). The identifiers of proteins that are involved, or are predicted to be involved, in DNA and RNA metabolism are coloured blue and red, respectively. Positions of the alignment N- and C-termini are indicated by the number of residues. Distances between conserved blocks, as well as the C-terminal end position of sequences (right), are indicated in brackets. Predicted secondary structures are shown up to the alignment [H and E stand for helix and strand (extended), respectively]. Conserved hydrophobic amino acids (V, I, L, F, M, Y, W) are boxed in green, conserved aromatic amino acids (F, Y, W) in violet, acidic and basic amino acids in pink and blue, respectively, small amino acids (G, A, T, S, C) in yellow. Amino acids that can substitute these residues in some circumstances are shown in bold (e.g. A, T, C and S can substitute hydrophobic positions). Underlined sequences correspond to human SMN1C and CPSF 73 kDa-related proteins.

Schizosaccharomyces pombe CPSF 100 kDa: motifs B and C are missing, Fig. 2B). This observation suggests that (i) the lexical procedure used by PSI-BLAST for alignment was not sufficient in itself to align highly divergent sequences, (ii) large insertions or deletions could interfere with the recognition of conserved motifs, and/or (iii) motifs could really be absent in some proteins. Thus, HCA was also used to refine the proposed relationships, as illustrated in Figure 2B with the particular case of proteins of the CPSF 100 kDa family. On the one hand, these alignments do indeed tolerate very large insertions just before motif B, which hampers alignment beyond, at least in the case of the *S.pombe* CPSF 100 kDa sequence, although clusters typical of motifs B and C can be easily detected farther on. On the other hand, these sequences have lost in part or totally the highly conserved residues of the different motifs, although the global fold of the domain was conserved. The histidine of motif B is indeed not present in any of the three CPSF 100 kDa sequences whereas histidine of motif C is only absent in the yeast sequence (see also Table 1). The absence of critical residues is also found in motifs 1–4 of the metallo-β-lactamase domain preceding the aligned region (Fig. 2A), reinforcing the hypothesis that conserved residues of motifs A–C, together with motifs 1–4, play a key role in the probable enzymatic function of the β-CASP family. This function should be lost in some proteins of the family, including the CPSF 100 kDa subunits. The missing motifs, highlighted using HCA, were confirmed as true relationships using them as queries in PSI-BLAST searches in a 'reverse' strategy.

Some predicted 'non-catalytic' members of the β-CASP family, in particular missing motifs A and B, were also highlighted in bacterial genomes (Tables 2 and 3).

Conserved motifs specific to the β-CASP family relative to the metallo-β-lactamase active site

Our initial aim was to identify within the β-CASP family a highly conserved amino acid, which could be located at a similar position relative to the active site than that of the motif 5 histidine of canonical metallo-β-lactamases (Fig. 1). As stated above, we identified not one but three highly conserved amino acids, suggesting that the active site of members of the β-CASP family could accommodate more critical residues than those of canonical metallo-β-lactamases.

Histidine of motif B appears to be the best candidate to correspond to motif 5 histidine of canonical metallo-β-lactamases, as already noticed by Aravind (1), since it is conserved in all of the members of the β-CASP family, in contrast to motif C histidine. If the sought-after motif 5 histidine actually corresponds to the highlighted motif B histidine, sequences intervening between motifs 4 and 5 should correspond to a large insertion within the metallo-β-lactamase domain. Another possibility is that no large

insertion should occur between metallo-β-lactamase motifs 4 and 5 and that motif 5 should actually correspond to the highlighted motif A. Both hypotheses are in fact supported by the loss of Artemis activity in Asp165 and His319 mutants (5; our unpublished observations). Like motif 5, corresponding to strand β12 and the subsequent loop of the β-lactamase structure shown in Figure 1, motif A also corresponds to a β-strand ended by a conserved polar residue. Moreover, like motif 5 in canonical metallo-β-lactamase structures, motif A is separated from motif 4 by a short peptide including a helix (Figs 2B and 3; helix α4 in Fig. 1). Regarding this hypothesis, the canonical motif 5 histidine should, however, be substituted by an acidic residue and sequences located after motif A should correspond to a distinct additional domain accompanying the metallo-β-lactamase domain. This hypothesis is further supported by a comparison of the three-dimensional structure of canonical metallo-β-lactamases with that of glyoxalase II (20), in which the strand ended by motif 5 histidine is followed by a distinct globular domain, differing from the C-terminal helix of metallo-β-lactamases. Regarding this possibility, it could be hypothesised that substrates of the β-CASP family bind at the domain interface, as substrates of glyoxalase II do.

DISCUSSION

The β-CASP family described here is a very large family including proteins of the three primary kingdoms (eukaryotes, bacteria and archaea) and appears to be specialised towards nucleic acids, as suggested by functional data gained for some members or by domains accompanying the metallo-β-lactamase domain. Some archaeal ORFs indeed possess N-terminal KH domains (1), known as RNA-binding motifs (21), whereas an *Arabidopsis* ORF of the Artemis/SNM1/PSO2 group has a module homologous to the eukaryotic DNA ligase I downstream of the metallo-β-lactamase/β-CASP domain (Tables 1 and 3).

On the basis of an in-depth sequence analysis, we showed that members of the β-CASP family that specifically interact with DNA targets can be distinguished from those involved in RNA metabolism regarding the nature of a particular amino acid included in a conserved sequence motif (motif C), which is always a histidine in RNA-specific proteins, whereas it is substituted by a hydrophobic amino acid in proteins acting on DNA. This distinctive sequence feature can usefully be considered for functional characterisation in wide-scale genome analyses.

Proteins of the β-CASP family involved in DNA metabolism

The recently identified Artemis protein makes part of the DNA double-strand break (DSB) repair machinery, as inferred from the phenotype of patients with RS-SCID who possess defects in V(D)J recombination leading to an early arrest of B- and T-cell maturation (4). Given the potential enzymatic function of its metallo-β-lactamase/β-CASP domain, it has been hypothesised that Artemis could be involved in the opening of hairpin-sealed coding ends, as generated by the RAG1/RAG2 complex (4). Hence, Artemis could be integrated in the DNA non-homologous end joining cascade, in addition to the Ku70/Ku80 complex, DNA-PKcs subunit and XRCC4/DNA ligase 4. Recently, Ma *et al.* (5) demonstrated that Artemis do indeed possess an hydrolase catalytic activity. Moreover, when complexed to, and phosphorylated by, the DNA-PKcs, Artemis is capable of opening and processing hairpin structures generated by Rag1 and Rag2. This activity is strictly dependent on the continuous association of Artemis with DNA-PKcs. The metallo-β-lactamase/β-CASP domain of Artemis is located N-terminus and is followed by a large, essentially non-globular domain which does not share any obvious similarity with other proteins. In contrast, SNM1 and PSO2 have a C-terminal metallo-β-lactamase/β-CASP domain that is preceded by a large N-terminal sequence including a zinc-finger domain. SNM1 and PSO2 are also involved in DNA repair, but they are specialised in DNA interstrand crosslink repair and are not sensitive to ionising radiations (17,18). Here again, their precise role is not yet known, but it could be hypothesised that their possible enzymatic activity could imply DNA cleavage to help remove the crosslink.

As mentioned in the Results, we found evidence in the human genome of a third sequence [named SNM1C by Dronkert *et al.* (17,18) and Wood (19)] which, like Artemis and SNM1, possesses the hallmark of an enzyme with a DNA substrate (a valine at the motif C position) (Table 1). The corresponding uncharacterised protein may thus play an important role in DNA repair, which has to be yet uncovered. This SNM1C protein is conserved in human, mouse, *Caenorhabditis elegans* and *Arabidopsis thaliana* but is apparently absent from yeast, suggesting a metazoan-specific activity.

Members of the β-CASP family involved in RNA metabolism

Predicted active members of the β-CASP family also include the 73 kDa subunit of the mammalian CPSF and its yeast orthologue, Ysh1/Brr5. Mammalian CPSF, composed of four subunits (30, 73, 100 and 160 kDa), plays a central role in the endonucleolytic cleavage and the polyadenylation of 3′ ends of most eukaryotic messenger RNAs (mRNAs) (22,23). It recognises AAUAAA hexanucleotides found upstream of the polyadenylation site via the 160 kDa subunit. The exact role of CPSF 73 kDa/Ysh1 in the mRNA processing remains largely unknown. Consistent with its predicted 'active' metallo-β-lactamase/β-CASP domain, it could also be directly involved in an enzymatic function. This hypothesis is supported by a Brr5/Ysh1 mutant identified in a screen for cold-sensitive pre-mRNA splicing mutants (24). Moreover, depletion of Bbr5/Ysh1 resulted in inhibition of both cleavage and polyadenylation (24). The two CPSF subunits, CPSF 73 kDa/Ysh1 and CPSF 100 kDa/Ydh1, may have evolved from a common ancestor, as suggested by the similarities that these proteins share within their metallo-β-lactamase/β-CASP domain. However, CPSF 100 kDa/Ydh1 are predicted to be inactive as they lack part of all the conserved amino acids that should be involved in the enzymatic function. This 'loss of function' is particularly marked for Ydh1. Accordingly, CPSF 100 kDa/Ydh1 could be confined to a modulatory function helping in regulating enzymatic activity, as already suggested by Aravind (1). Acquisition of new functions beyond the ancestral enzymatic one is also possible (1).

Other pairs of active/inactive metallo-β-lactamase/β-CASP domains are encountered in some bacterial genomes, such as those of *Mycoplasma genitalium* (MG139 and MG423) and *M.pneumoniae* (MPN280 and MPN261), *Staphylococcus aureus* (SA0940 and SA1118), *Lactobacillus lactis* (yciH and yqga), *Deinococcus radiodurans* (DRA0069 and DR2417m) and *Streptococcus pyogenes* (Spy1876 and Spy1020) (Table 2). These pairs of paralogous sequences, one active and the other inactive, can thus be good candidates to constitute the bacterial CPSF 73 kDa/100 kDa subunits. It is worth noting that all the bacterial and archaeal sequences described here seem to be specific to RNA targets, as they are more related to the CPSF 73 kDa proteins than to any other eukaryotic β-CASP proteins, and as they possess (at least those that are predicted to be active) a motif C histidine. According to Anantharaman *et al.* (25), the last universal common ancestor (LUCA) had probably a polyadenylation system that includes at least a CPSF 73 kDa-like enzyme that cleaves transcripts. However, the involvement of the described bacterial metallo-β-lactamase/β-CASP proteins in mRNA processing remains to be investigated.

Interestingly, similar to the way in which we highlighted a novel group of proteins distinct from Artemis and SNM1 which may act on DNA substrates, we also identified CPSF 73 kDa-related sequences (Table 1), which are related to, but distinct from CPSF 73 kDa. Thus, these proteins could play a pivotal role in mRNA processing, possibly within or in concert with the CPSF. Their exact role also remains to be unravelled.

In conclusion, the ubiquitous distribution of metallo-β-lactamases of the β-CASP family underlines its importance and the analysis presented here should help to elucidate their exact functions relative to nucleic acids, as well as their specificities.

ACKNOWLEDGEMENTS

This work was supported by institutional grants and grants from Association de Recherche sur le Cancer (ARC) and APEX (INSERM) to J.-P.V. and from the CNRS-INRA-INRIA-INSERM 'Action Bioinformatique' to I.C.

REFERENCES

1. Aravind,L. (1999) An evolutionary classification of the metallo-β-lactamase fold proteins. *In Silico Biol.*, **1**, 69–91.
2. Melino,S., Capo,C., Dragani,B., Aceto,A. and Petruzzelli,R. (1998) A zinc-binding motif conserved in glyoxalase II, β-lactamase and arylsulfatases. *Trends Biochem. Sci.*, **23**, 381–382.
3. Daiyasu,H., Osaka,K., Ishino,Y. and Toh,H. (2001) Expansion of the zinc metallo-hydrolase family of the β-lactamase fold. *FEBS Lett.*, **503**, 1–6.
4. Moshous,D., Callebaut,I., de Chasseval,R., Corneo,B., Cavazzana-Calvo,M., Le Deist,F., Tezcan,I., Sanal,O., Bertrand,Y., Philippe,N. *et al.* (2001) ARTEMIS, a novel DNA double-strand break repair/V(D)J recombination protein, is mutated in human severe combined immune deficiency. *Cell*, **105**, 177–186.
5. Ma,Y., Pannicke,U., Schwarz,K. and Lieber,M.R. (2002) Hairpin opening and overhang processing by an Artemis/DNA-dependent complex in non-homologous end joining and V(D)J recombination. *Cell*, **108**, 781–794.
6. Wang,Z., Fast,W., Valentine,A.M. and Benkovic,S.J. (1999) Metallo-β-lactamase: structure and mechanism. *Curr. Opin. Chem. Biol.*, **3**, 614–622.
7. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
8. Gaboriaud,C., Bissery,V., Benchetrit,T. and Mornon,J.-P. (1987) Hydrophobic cluster analysis. An efficient new way to compare and analyse amino-acid sequences. *FEBS Lett.*, **224**, 149–155.
9. Callebaut,I., Labesse,G., Durand,P., Poupon,A., Canard,L., Chomilier,J., Henrissat,B. and Mornon,J.-P. (1997) Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell. Mol. Life Sci.*, **53**, 621–645.
10. Woodcock,S., Mornon,J.-P. and Henrissat,B. (1992) Detection of secondary structure elements in proteins by hydrophobic cluster analysis. *Protein Eng.*, **5**, 629–635.
11. Callebaut,I. and Mornon,J.-P. (1997) From BRCA1 to RAP1: a widespread BRCT module closely associated to DNA repair. *FEBS Lett.*, **400**, 25–30.
12. Callebaut,I., de Gunzburg,J., Goud,B. and Mornon,J.-P. (2001) RUN domains: a new family of domains involved in Ras-like GTPase signaling. *Trends Biochem. Sci.*, **26**, 79–83.
13. Girault,J.-A., Labesse,G., Mornon,J.-P. and Callebaut,I. (1999) The N-termini of FAK and JAKs contains divergent band 4.1 domains. *Trends Biochem. Sci.*, **24**, 54–57.
14. Callebaut,I. and Mornon,J.-P. (1998) The V(D)J recombination activating protein RAG2 consists of a six-bladed propeller and a PHD finger-like domain, as revealed by sequence analysis. *Cell. Mol. Life Sci.*, **54**, 880–891.
15. Cornéo,B., Moshous,D., Callebaut,I., de Chasseval,R., Fischer,A. and de Villartay,J.P. (2000) Three-dimensional clustering of human RAG2 gene mutations in severe combined immune deficiency. *J. Biol. Chem.*, **275**, 12672–12675.
16. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S., Griffiths-Jones,S., Howe,K., Marshall,M. and Sonnhammer,E. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
17. Dronkert,M.L.G., De Witt,J., Boeve,M., Vasconcelos,M.L., Van Steeg,H., Tan,T.N.R., Hoeijmakers,J.H.J. and Kanaar,R. (2000) Disruption of mouse SNM1 causes increased sensitivity to the DNA interstrand cross-linking agent mitomycin C. *Mol. Cell. Biol.*, **20**, 4553–4561.
18. Dronkert,M.L.G. and Kanaar,R. (2001) Repair of DNA interstrand cross-links. *Mutat. Res.*, **486**, 217–247.
19. Wood,R.D., Mitchell,M., Sgouros,J. and Lindahl,T. (2001) Human DNA repair genes. *Science*, **291**, 1284–1289.
20. Cameron,A.D., Ridderström,M., Olin,B. and Mannervik,B. (1999) Crystal structure of human glyoxalase II and its complex with a glutathione thiolester substrate analogue. *Struct. Fold Des.*, **7**, 1067–1078.
21. Nagai,K. (1996) RNA–protein complexes. *Curr. Opin. Struct. Biol.*, **6**, 53–61.
22. Manley,J.L. and Takagaki,Y. (1996) The end of the message—another link between yeast and mammals. *Science*, **274**, 1481–1482.
23. Shatkin,A.J. and Manley,J.L. (2000) The ends of the affair: capping and polyadenylation. *Nature Struct. Biol.*, **7**, 838–842.
24. Chanfreau,G., Noble,S.M. and Guthrie,C. (1996) Essential yeast protein within unexpected similarity to subunits of mammalian cleavage and polyadenylation factor (CPSF). *Science*, **274**, 1511–1514.
25. 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.
26. Ullah,J.H., Walsh,T.R, Taylor,I.A., Emery,D.C., Verma,C.S., Gamblin,S.J. and Spencer,J. (1998) The crystal structure of the L1 metallo-beta-lactamase from *Stenotrophomonas maltophilia* at 1.7 Å resolution. *J. Mol. Biol.*, **284**, 125–136.
27. Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.