Published in final edited form as:

Proteins. 2009 May 15; 75(3): 760–773. doi:10.1002/prot.22287.

Structural genomics reveals EVE as a new ASCH/PUA-related domain

Claudia Bertonati^{a,b,c,d,*‡}, Marco Punta^{a,b,c,‡}, Markus Fischer^{a,b,c}, Guy Yachdav^{a,b,c}, Farhad Forouhar^e, Weihong Zhou^e, Alexander P. Kuzin^e, Jayaraman Seetharaman^e, Mariam Abashidze^e, Theresa A. Ramelot^f, Michael A. Kennedy^f, John R. Cort^{g,h}, Adam Belachewⁱ, John F. Hunt^e, Liang Tong^e, Gaetano T. Montelione^{c,j}, and Burkhard Rost^{a,b,c}

^aDepartment of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, NY 10032, USA

^bColumbia University Center for Computational Biology and Bioinformatics (C2B2), 1130 St. Nicholas Ave., New York, NY 10032, USA

^cNortheast Structural Genomics Consortium (NESG), Columbia University, 1130 St. Nicholas Ave. Rm. 802, New York, NY 10032, USA

^dDepartment of Biochemistry 'A. Rossi Fanelli', University of Rome "La Sapienza", 00185 Rome, Italy

^eDepartment of Biological Sciences, Northeast Structural Genomics Consortium, Columbia University, New York, NY 10027

^fDepartment of Chemistry and Biochemistry and Northeast Structural Genomics Consortium (NESG), Miami University, Oxford, Ohio 45056

^gWashington State University Tri-Cities, Richland WA, 99354, USA

^hBiological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, 99352, USA and Northeast Structural Genomics Consortium

ⁱWeill medical College of Cornell, Pharmacology department, 1300 York Avenue, New York, NY 10021

^jCenter for Advanced Biotechnology, Dept of Molecular Biology and Biochemistry, Rutgers University, and Northeast Structural Genomics Consortium, 679 Hoes Lane, Piscataway, NJ 08854 and Dept. of Biochemistry, Robert Wood Johnson Medical School, University of Medicine and Dentistry of NJ, Piscataway, NJ 08854

Summary

We report on several proteins recently solved by structural genomics consortia, in particular by the Northeast Structural Genomics consortium (NESG). The proteins considered in this study differ substantially in their sequences but they share a similar structural core, characterized by a pseudobarrel five-stranded beta sheet. This core corresponds to the PUA domain-like architecture

*Corresponding author: claudia.bertonati@uniroma1.it, Tel: +39 06 49690276, fax: +39 06 4440062 / +39 06 4453933.

‡These authors contributed equally

in the SCOP database. By connecting sequence information with structural knowledge, we characterize a new subgroup of these proteins that we propose to be distinctly different from previously described PUA domain-like domains such as PUA proper or ASCH. We refer to these newly defined domains as *EVE*. Although *EVE* may have retained the ability of PUA domains to bind RNA, the available experimental and computational data suggests that both the details of its molecular function and its cellular function differ from those of other PUA domain-like domains. This study of *EVE* and its relatives illustrates how the combination of structure and genomics creates new insights by connecting a cornucopia of structures that map to the same evolutionary potential. Primary sequence information alone would have not been sufficient to reveal these evolutionary links.

Keywords

structural genomics; protein function prediction; PUA domain-like domains; X-ray crystallography; NMR

Introduction

Structural Genomics has high impact on sequence space

Structural Genomics (SG) is a worldwide effort that has many objectives. One goal is the increase of the structural coverage, or the percentage of proteins for which we have direct experimental or indirect computational information about their three-dimensional (3D) structure. One way of realizing this goal is to identify families for which we currently do not have 3D coverage and that are large, so that each experimental structure yields reliable models for many proteins. Experimentally determining the structure for at least one representative in this sequence-structure family yields high impact ¹. Over the last few years, SG consortia have deposited thousands of structures ^{2,3} into the Protein Data Bank (PDB) ⁴. If the aim is to increase structural coverage as much as possible given certain resources, then SG undoubtedly has become the most efficient way of reaching this aim ⁵. The systematic identification of large families with no previous structural characterization leads to a new scenario. While, traditionally, the details of experimental structures have confirmed hypotheses about protein function, now the 3D structures themselves generate functional hypotheses. In other words, the approach shifts from hypothesis-driven to hypothesis-generating science.

Large families are biologically important

One assumption of target selection in structural genomics is that the larger and more diverse a protein family, the more likely its proteins perform essential functions ⁶; this assumption has been proven in hundreds of cases ⁷. Solving the structure for one family-representative may unravel evolutionary relations with a different family, and may thereby provide new insights into the functional traits of the latter as well as into the function of the target protein. Indeed, SG is likely to impact biology more by unraveling relations that are undetectable from sequence analysis alone, than by discovering new exotic folds ⁸. In its targeting of large un-annotated families, SG is clearly orthogonal to traditional structural

biology, which focuses more on providing a better understanding of what is partially known than on exploring what is largely unknown.

Structural genomics data reveals a new PUA domain-like domain—Here, we investigated several proteins recently solved by the Northeast Structural Genomics consortium (NESG) ⁹ and other SG consortia. Many of these proteins have levels of sequence similarity that pairwise sequence comparisons would consider as below random. Still, they can be defined by a common structural framework/architecture, namely the PUA domain-like fold in SCOP ¹⁰. These proteins constitute a family with (1) still limited experimental exploration, (2) a great number of sequences, and (3) with a large number of diverse structures. This combination is exactly the target that structural genomics has successfully been addressing over the last eight years. The SCOP fold is named after the pseudouridine synthase and archaeosine transglycosylase (PUA) domain, a highly conserved RNA binding motif ¹¹. PUA is present, among other proteins, in RNA modification enzymes and ribonucleoproteins; in human, it has been related to disease, including cancer ¹². SCOP ¹⁰ classifies PUA domain-like domains (PUA-like domains hereafter) by their conserved 5-stranded core pseudo-barrel beta sheet (strands 1 to 5, Fig. 1A). Many family members additionally feature an alpha helix (helix A) between strands 1 and 2.

We revisited previous studies in light of the new structural information available and provided an in depth-analysis of PUA-like domains. Although we consider information from sequence including alignments and genomic profiles, our study is primarily based on a comparative structural analysis of the domains. We use structural alignments, cavity identification, surface residue conservation and surface electrostatic potential calculations. We show that seven of the proteins with a PUA-like architecture recently solved by SG consortia exhibit features clearly distinct from other PUA domain-like domains. These structural differences predict that they also carry different functional traits. We refer to these newly identified domains as *EVE* (following the PDB identifier of one of the available structures: 2eve ¹³). Overall, this study illustrates how structural knowledge can be essential not only for evolutionary linking proteins without recognizable sequence similarity but also for highlighting differences between structurally related domains that may reflect more subtle divergences in function and history. Sequence and structural data for the proteins that we analyze are available at <http://luna.bioc.columbia.edu/honiglab/PUAlike> (see also Methods).

Results

PUA domain-like structures from NESG suggest EVE as a new structural family

Structural comparisons using Skan ^{14,15} (Methods) revealed that at least ten of the proteins for which structures have been experimentally determined by the NESG consortium ⁹ included domains with a PUA-like architecture ^{16,11} (Table 1) (note that, in this paper, we will refer to individual proteins by preferentially using their PDB identifier). Interestingly, querying Pfam ¹⁷ with our NESG protein sequences produced no match with the PF01472 PUA proper family. Instead, three proteins were assigned to the PF04266 ASCH (Activating Signal Cointegrator 1 Homology) family, four to the PF01878-DUF55 family of unknown

function, one (1zbo¹⁸) to the PF02190 ATP-dependent protease La family, one (1nxz¹⁹) to the PF04452 RNA methyltransferase family, and one (1t5y²⁰) could not be assigned to any Pfam-A family. Aspects of function could be inferred by homology for only two out of the ten proteins (ASCH domains have been connected to RNA metabolism²¹, but more detailed characterizations are yet missing).

Most PUA-like structures from NESG can be grouped into two topological themes of variation of the 5-stranded pseudo-barrel beta sheet core architecture of SCOP. The first theme (Fig. 1B) is a long insertion (INS1) of about 40 residues, featuring a number of short helices, between strands 4 and 5. Proteins carrying this insertion (1s04^{21,22}, 1xne²³, and 1te7²⁴; Table 1) belong to the ASCH family²¹ (details below). The second theme (Fig. 1C) is a shorter insertion (INS1: 10–20 residues) between strands 4 and 5 coupled with a long (about 40 residues) second insertion (INS2) at the C-terminus (1zce²⁵, 2eve¹³, 2gbs²⁶ and 2g2x²⁷; Table 1). INS2 contains several helices of variable lengths and an additional strand (strand 6 in Fig. 1C) that pairs in anti-parallel fashion with strand 1 of the core beta sheet. Following the PDB identifier of one of the representatives from this group (2eve¹³, Table 1), we dubbed this latter topological theme *EVE*. The addition of a strand that pairs with strand 1 is not unique to EVE: PUA domains (e.g. 1j2b²⁸, Table 1) also have an extra strand pairing to strand 1 (strand 2', Fig. 1D). However, topologically this strand is very different from strand 6 in EVE. In fact, since PUA domains lack INS2, strand 2' is located between strands 2 and 3. Also, although strands 6 (Fig 1C) and 2' (Fig 1D) seem to pair with strand 1 in a similar way, they have a very different effect on the solvent accessible surface of their respective domains (below).

Among the remaining NESG structures, 1t5y²⁰ is similar to PUA, while 1nxz¹⁹ resembles the SCOP-core architecture (Fig. 1A). Protein 1nxz is classified by Pfam as an RNA methyltransferase, due to an additional C-terminal domain carrying this annotation. Finally, 1zbo¹⁸ features a long (~100 residues) C-terminal insertion unrelated to INS2 and is classified by Pfam as part of the ATP-dependent protease La domain family (Table 1).

EVE domains share structural features and signature conserved residues

Four of the NESG structures are characterized by the topology of 2eve that we named *EVE* (Fig. 1C: 2eve, 2gbs, 2g2x, 1zce). Structural comparisons with Skan^{14,15} revealed three more proteins with similar topology in the PDB (solved by other SG consortia), namely 2ar1²⁹, 2p5d³⁰ and 2hd9/1wmm (1wmm³¹ and 2hd9³² are two crystal structures of the same protein; Table 1). Structural alignments highlighted a strong sequence/structure relation between these seven single-domain proteins (Table 2 and Fig. 2); a subset of these relations has been analyzed before²⁹. 2hd9³² and 2p5d³⁰ are relatively sequence distant from the others (<25% pairwise sequence identity in the structural alignment); all align at Evalues <10⁻³⁹ to the same Pfam family (PF01878-DUF55) (data obtained via InterProScan³³).

Sequence conservation analysis (Methods) identified several conserved aromatic residues at the N-terminus of EVE; these included Y3, W4 and W26 (Fig. 2; sequence and numbering as in the 2eve¹³ PDB file). Note that the sequence-based Pfam seed alignment (PF01878-DUF55) of the EVE domain N-terminus is incorrect (Supplementary Online Material).

While W4 is mostly conserved, Y3 is often substituted by other aromatic residues such as phenylalanine and histidine. This appears to reflect the biophysical features of Y3 as well as its characteristic interactions in EVE: in all known structures, Y3 seems to form a π - π bond with W142, another well-conserved aromatic residue located in the C-terminal insertion. W142 variants include phenylalanine or tyrosine. Highly conserved W4, on the other hand, interacts with highly conserved D45 at the beginning of strand 3. Hence, Y3 and W4, interacting with sequence-distant residues, seem to be important for stabilizing the domain structure. In contrast, the strong conservation of W26 may have functional implications. W26 extends its bulky side chain into the domain's main surface cavity, which is lined by helix A, strands 1, 2, and 6, and has an overall surface area of about 200 Å² (Fig. 3A). In three of the six available crystal structures this cavity is partially occupied by a ligand. Note that the structure of 2gbs is from NMR data, i.e. data from which ligand identification is difficult. The observed ligands include a molecule of 3[N-morpholino]propane sulfonic acid (MPO) in 2eve¹³ (Fig. 3A and 3B) and a molecule of glycerol (GOL) in both 2ar1²⁹ and 2hd9³². In all three instances, W26 appears to play a major role in shaping the binding pocket. Neither MPO nor GOL, however, are functional ligands. Their presence in the structures is related to their use as buffer or cryoprotectant²⁹ substances. Note that MPO should not be confused with morpholino oligonucleotides, which are nucleotide analogs that have become an important knockdown tool in developmental biology³⁴. Indeed, no nucleobase is present in the MPO molecule bound to 2eve.

Among the residues lining the cavity, V29 and F13 are also conserved (Fig. 3A); however, F13 is less conserved than W26 and V29 (in particular, it is mutated into non-aromatics in 2hd9³² and 2g2x²⁷). Overall, the region around the cavity appears to be well conserved in sequence (Fig. 3B). An analysis of the electrostatic potential of the cavity by GRASP2¹⁴ noted the shift from a slightly negatively charged surface potential in 2eve and 2ar1 to a slightly positive potential in 2hd9 (data not shown).

Another interesting pattern of conservation is constituted by the previously mentioned D45, along with G44 and Y50, which are all highly conserved in the EVE family (Fig. 2). While G44 and D45 are also conserved in other PUA-like domains (see below), the conservation of Y50 seems to be unique to EVE domains. Y50 is at the center of a crevice lined by strands 1, 3, 4, and INS2, which can be seen as a continuation of the cavity containing F13, W26 and V29 (Fig. 3A and 3B). In 2eve, this additional cavity hosts a small non-functional ligand, a molecule of tris-hydroxymethyl-methyl-ammonium that interacts with Y50 and with a series of positively charged residues (K7, K115, and R131; mostly non-conserved).

YTH domains: EVE-like structure of a putative splicing factor in human—

Structural alignments^{14,15} revealed two additional EVE-like structures in the PDB, namely the human proteins 2yud³⁵ and 2yu6³⁶ (Table 1) belonging to the putative splicing factor YT521-B homology (YTH) family (PF04146)³⁷. Both 2yud and 2yu6 feature the characteristic C-terminal insertion INS2 including strand 6, while they differ substantially in sequence from EVE domains with maximally 19% pairwise sequence identity (between 2yud and 2hd9; Table S1, Supplementary Online Material). Structural alignments to EVE (Fig. S1, Supplementary Online Material) and multiple sequence alignments of the family³⁷ showed that YTH domains conserve the N-terminal pair of aromatic residues Y3W4 (YF in

2yu6 and FF in 2yud) and the cavity-lining W26 typical of EVE. Y50 is instead mutated into a highly conserved serine. The main surface cavity of EVE is also present in YTH and it is of comparable size. On the other hand, EVE-conserved residues V29, G44, D45 and W142 (Fig. S1, Supplementary Online Material) are not found in YTH and YTH domains feature several well-conserved residues not present in EVE (including several aromatic residues³⁷). Also, while EVE domains are well represented in prokaryotes, YTH domains are only found in eukaryotes. Differences notwithstanding, YTH domains appear to be as the most similar to EVE among all other PUA-like domains currently in the PDB.

EVE differ from ASCH and PUA domains in important ways

Iyer et al.²¹ identified ten families (including ASC-1) within what they defined as the *ASCH superfamily* (Fig. 1B). Experimental structures are now known for five ASCH domains. These five single-domain proteins vary substantially in sequence; they map to four of the ten families described by Iyer et al.²¹ (Table 1). In Pfam, ASCH domains have recently been reclassified into a single family (ASCH-PF04266) of unknown function. We have already described the main differences between the EVE and ASCH structural themes. Sequence-wise, ASCH domains have an N-terminal sequence signature²¹ that is conserved across all ten families and includes a glycine and a lysine (G19-x-K21; numbers as in the 1xne²³ PDB file). In most ASCH domains, two conserved, adjacent residues (E24 and R26) contribute to the longer G-x-K-x-x-[ETS]-x-R motif. This motif is in the “loop” region connecting helix A and strand 2. Notably, EVE domains do not conserve any of these residues (Fig. 2 and Fig. 4). Similarly, most of the EVE signatures discussed previously are not conserved in ASCH domains, specifically: residues Y3 and W4 at the N-terminus, W26 and V29 in the main surface cavity, Y50 in the second EVE cavity. Finally, residue W142 in INS2 is also not conserved in ASCH, as INS2 is completely absent from these domains. The largest surface cavity in ASCH (>300 Å²) roughly corresponds to the main cavity in EVE, although the latter is ~30% smaller due to the extra strand 6. Residue conservation within the cavity is markedly different in ASCH and EVE. In ASCH, most of the conserved residues protruding into the cavity are polar, acidic or basic, in particular: E24, R26 and in at least some cases K21 (e.g. in 1xne). E24 and R26 structurally align to hydrophobic conserved residues W26 and V29 in EVE (Fig. 4), giving a completely different ‘flavor’ to the ASCH cavity. An exception is constituted by aromatic residue Y12, often mutated into a phenylalanine or a tryptophan (but a leucine in ASCH family 4). It corresponds to F13 in EVE. Note that this conserved position in ASCH domains was not identified by Iyer et al.²¹ (due to their use of sequence-based alignments) and that one of the tRNA binding residues of PUA domain 1j2b, F519, is found in a similar structural position (see below). Although the hypothesis awaits experimental confirmation, the largest surface cavity of ASCH and EVE, with its size, conservation and capability of accommodating small ligands (Fig. 3A and 3B), is likely to play an important functional role. The significant differences observed between the conserved residues lining the cavity in ASCH and EVE, along with the additional presence of a nearby cavity surrounding highly conserved Y50 in EVE, seem hence to point to important differences in the molecular function of the two domains.

Most of the conserved residues characterizing EVE domains and analyzed above (i.e. Y3, W4, W26, V29, Y50 and W142) are not conserved in the PUA domains. Although some

PUA domains feature a phenylalanine at the end of strand 2 (F527 in 1j2b, see Fig. 2 in ¹¹), structural alignments show that it does not correspond to W26 in EVE (Fig. S2, Supplementary Online Material). The PUA cavity corresponding to the main EVE (and ASCH) cavity is very small (surface area of $< 100 \text{ \AA}^2$). This is due to the presence of strand 2' (Fig. 2D), which contrary to the shorter strand 6 of EVE fully extends into the core of the cavity (Fig. S2, Supplementary Online Material). Indeed, a simple structural superposition by Skan ^{14,15} indicated that there is no space to dock the ligands bound to EVE (specifically, MPO in 2eve ¹³ and GOL in 2ar1 ²⁹ and 2hd9 ³²) into the PUA domain structures of 1j2b ²⁸, 2hvy ³⁸ and 1r3e ³⁹. These differences may lead to important functional consequences as we discuss in the next paragraph.

EVE structure compatible with RNA binding but RNA binding, if any, likely to be different than in PUA

PUA domains are known to interact with RNA as part of multi-domain RNA modifying proteins. Although no EVE domain appears to be fused to any known RNA modifying protein, EVE domains could in principle act in concert with other proteins to bind and modify RNA. To further investigate this hypothesis, we compared the regions involved in RNA binding in PUA with the corresponding regions in EVE.

Four PUA-RNA complexes are available, including archaeosine tRNA-guanine transglycosylase ArcTGT (PDB identifier 1j2b ²⁸) and pseudouridine synthases Cbf5 and TruB (PDB identifiers 2hvy ³⁸, 1r3e ³⁹ and 1k8w ⁴⁰). Although RNA binding in known PUA-RNA complexes presents important differences, some common elements have been identified ¹¹. In particular, three structural elements seem to always participate in RNA recognition: helix A, strand 2 and strand 5 (note, naming and numbering used in Perez-Arellano et al. ¹¹ differ from ours). While strand 5 and the loopy region connecting helix A and strand 2 bind to the double stranded RNA stems, a cleft between helix A and strand 2 provides recognition for single stranded RNA overhangs. As detailed above, in EVE and ASCH, this shallow cleft is part of a much larger cavity. Double stranded RNA binding residues include polar and charged residues found in the loop connecting helix A and strand 2 and on strand 5. Single stranded RNA overhangs are instead anchored to the PUA domain differently in the ArcTGT-RNA and in the Cbf5-RNA complex: via a cluster of phenylalanines in ArcTGT (F519, F527, F530) and via a C-terminal tryptophan (W337) and hydrogen bonds with the protein backbone in Cbf5. Note that no single stranded RNA binds the PUA domain in TruB structures 1r3e ³⁹ and 1k8w ⁴⁰.

Structural alignments of ArcTGT PUA with EVE revealed that EVE domains expose several positively charged residues at the end of helix A and on strand 4 and 5 that may be able to mediate recognition of double stranded RNAs (Fig. S3–S5, Supplementary Online Material). To better investigate the putative binding of single stranded RNA overhangs to EVE domains, we performed the following simple experiment. We first superimposed the structures of 2eve ¹³ and of the PUA domain of ArcTGT 1j2b ²⁸ (residues 507–582) (using GRASP2 ¹⁴). Then, we removed 1j2b while leaving its bound tRNA and 2eve in place (Fig. 5A and 5B). The resulting virtual complex provides us with two important indications. First, only one of the three phenylalanines that bind RNA in the PUA domain is conserved in EVE

(Fig. 5A and S2, Supplementary Online Material) (the only conserved residue in EVE is F13, sometimes mutated into non-aromatic hydrophobic residues; in ASCH it corresponds to conserved residue Y12). Second, although the 2eve region where the single stranded RNA overhang is found in the virtual complex is part of the previously characterized large EVE cavity, it does not correspond to the highly conserved pocket where MPO (and GOL) is found (Fig. 5A and 5B). Neither it corresponds to the small cavity into which conserved residue Y50 protrudes (Fig. 5A). So, hypothesizing that indeed EVE binds RNA, it is likely that its large cavity either hosts a cofactor (in place of MPO and GOL) or allows single stranded RNAs to bind more deeply into the domain (with W26 possibly playing a major role in binding). Another possibility is that single stranded RNAs anchor themselves to EVE by binding on the opposite side of the cavity with respect to what we see in PUA domains. In conclusion, our analysis points to important differences between EVE and PUA in a functionally relevant region of the PUA domain.

Another indirect evidence supporting an RNA binding role for EVE comes from the EVE-like eukaryotic YTH domains. Stoilov et al.³⁷ suggested that YTH (2yud³⁵ and 2yu6³⁶) could act as mRNA-binding domains, taking part in alternative splicing. In multi-domain proteins, YTH can be found fused to zinc fingers, RRM motifs, and HA2 domains, all of which are thought or known to be involved in nucleic acid binding. Although a role in alternative splicing for EVE domains can be excluded given their extensive presence in prokaryotes (see below), it is tempting to postulate that they could still perform a function similar to YTH at the molecular level.

Genomics and phylogeny of EVE domains—PUA domains almost always occur in multi-domain proteins, often in fusion with known RNA modifying enzymes. ASCH domains, although mostly found as single-domain proteins, were originally identified at the C-terminus of ASC-1 proteins that in vertebrates are known to physically interact with proteins involved in RNA processing²¹. Different from PUA and ASCH, EVE domains seem to exist almost exclusively as single-domain proteins. Exceptions are proteins from *Staphylococcus aureus* and *Staphylococcus saprophyticus*, in which EVE is flanked by N- and C-terminal regions that have not yet been functionally characterized. In Eukarya, most proteins feature an insertion N-terminal to EVE that may or may not constitute a separate domain. In fact, the only examples of annotated domains associated to EVE are found in *Aspergillus fumigatus*, *Aspergillus oryzae* and *Aspergillus nidulans*. In these genomes, proteins that contain the EVE domain carry three copies of the AT-hook motif at the N-terminus. This domain is typically found in proteins that bind preferentially to the minor groove of AT-rich regions in double-stranded DNA^{41,42}. Finally, several eukaryotic EVE domains (including one in *Homo sapiens* protein Q9HC20_HUMAN) are annotated as *thymocyte proteins* (e.g. in UniProt⁴³), where thymocytes are the T-cell precursors that mature in the thymus.

EVE, ASCH and PUA domains occur in all three super-kingdoms of life (Eukaryota, Bacteria, and Archaea). Although the analysis of EVE sequences included in the PF01878-DUF55 Pfam family reveals a predominant presence of EVE domains in *Bacteria*, EVE are further present in both the *Crenarchaeota* and *Euryarchaeota* phyla of *Archaea* and in *Eukaryota*, including representatives in *Euglenozoa*, *Metazoa*, *Fungi* and *Viridiplantae*. In

Eukaryota, however, EVE domains are relatively rare. This may suggest that the acquisition of the EVE domain by eukaryotic organisms has occurred via lateral gene transfers from *Bacteria*. The phyletic distribution of YTH (PF04146), exclusively found in *Eukaryota*, is very similar to that one of EVE (PF01878-DUF55). This observation, along with the discussed similarities in sequence and structure, may indicate that YTH originated from EVE domains following gene duplication in *Eukaryota*. ASCH domains (PF04266) also have a phyletic pattern similar to EVE and ASCH although, contrary to the other two, they seem to be practically absent in *Fungi*.

Discussion

EVE: a novel PUA-like domain

We characterized a new PUA-like domain, that we named EVE (Fig. 1C). Structure comparison revealed problems in alignments based on sequence alone (Supplementary Online Material), and showed that EVE domains are related but importantly distinct from ASCH and PUA domains (Table 3). The evidence includes the following observations: (i) differences in topology (Fig. 1B, 1C and 1D), in particular the characteristic EVE C-terminal insertion (INS2); (ii) residue conservation, with Y3, W4, W26, V29, Y50 and W142 that appear to be conserved only in EVE and not in other PUA-like domains; (iii) surface cavity properties, including cavities' size and residue conservation within them; in EVE the main surface cavity is lined by three conserved hydrophobic residues, two of which (W26 and V29) structurally align to conserved E24 and R26 in ASCH. In PUA domains the same cavity is partially occluded by the presence of an additional strand (strand 2', Fig. 1D).

While PUA and ASCH domains are sometimes found as part of multi-domain proteins in which one domain is an RNA modifying enzyme, we were not able to identify any such case for EVE. Analysis of PUA-RNA complexes revealed that EVE sequence and structural features are compatible with RNA binding. The comparison, however, also points to important differences between the two domains. Conserved EVE cavities, in particular, much reduced or absent in PUA, are likely to be relevant for function. This suggests that RNA binding in EVE, if any, may be very different from the one observed in PUA. Another hint for a possible RNA binding role of EVE comes from eukaryotic YTH domains, which have basically the same topology as EVE (albeit some interestingly different conservation patterns). In fact, YTH domains have been suggested to bind mRNA. Overall, the exact functional role of EVE domains remains unclear and awaits experimental investigation.

EVE and other PUA-like domains are evolutionary related

Although our evidence suggested important functional differences between EVE and other PUA-like domains, the strong conservation of a structural core between all PUA-like domains suggests an evolutionary connection. The GD sequence motif found in several families (including EVE, ASCH and PUA but not YTH) may be a surviving legacy of this common ancestry. The glycine (G at the inception of strand 3, Fig. 2) may provide the protein backbone with the flexibility needed for accommodating the insertions found immediately before this strand (e.g. strand 2', Fig. 1D). Thereby, it may help preventing the disruption of the core structure. The aspartate (D) often interacts with residues on strand 1,

likely increasing the stability of the pairing between strands 1 and 3 and hence the stability of the whole domain.

Considering that EVE, PUA and ASCH domains all span the three super-kingdoms, it is possible that PUA-like domains all originated from an early common ancestor. Later they may have diverged functionally, giving rise to several functional families.

Suggested alterations to Pfam

Pfam groups families that are likely to have arisen from a single evolutionary origin⁴⁴ into clans. Several PUA-like domains are included into the PUA (CL0178) clan, in particular: PUA (PF01472), ASCH (PF04266), PF02594-DUF167 and PF03657-UPF0113 (Table 1). Our results seem to clearly indicate that EVE (PF01878-DUF55) and YTH (PF04146) domains should be added to this clan. On the other hand, structure analysis of family PF02594 (dubbed Uncharacterized ACR, YggU family and including PDB proteins 1jrm, 1n9l and 1yh5) indicates that the PUA-like core architecture (Fig. 1A) is not conserved in these domains. This suggests that PF02594 may not have the same evolutionary origin as the other families in the PUA clan, and that it should therefore be removed from the clan. Other Pfam families that could possibly be included in the clan are PF09157 (Pseudouridine synthase II TruB, C-terminal, see 1k8w⁴⁰), PF09142 (tRNA Pseudouridine synthase II, C terminal, see 1sgv⁴⁵) and PF02190 (ATP-dependent protease La (LON) domain, see 1zbo¹⁸).

More PUA-like targets for structural genomics

What are the most interesting PUA-like domains still lacking a structural representative in the PDB? Among EVE domains, it would be interesting to target proteins in the PF01878-DUF55 Pfam family seed alignment such as YOUNG_BACLI and Y3206_RHILO. These proteins feature, at the N-terminus, two conserved histidines and in some instances one cysteine. Given their position along the sequence, these residues could protrude into the main surface cavity of the domain, thus potentially affecting electrostatic potential and putative ligand binding properties within the cavity. Among ASCH domains, the ones that stand out are those found in the ASC-1 family. Functionally, they are part of transcriptional coactivator proteins while, structurally, they feature an about 30 residue-long insertion between strand 3 and strand 4, not found in any of the available structures that we analyzed.

Conclusions

Structural genomics consortia continuously deposit high-resolution structural data for a great number of proteins into the public domain. Most of these proteins are outstandingly important as revealed by their evolutionary conservation. Surprisingly often we have no functional annotations from sequence alone for many of these important discoveries. Adding so many new structures is an important achievement *per se*, irrespective of what is known today about the function of these proteins. Sequences from genome sequencing projects are an obvious analogy. The simultaneous availability of both a wealth of sequence and structural data for evolutionary related proteins allows for detailed investigations of

functional similarities and differences. Sequence analysis is in this a first step, but often it requires structures to generate hypotheses that will then drive further experimental studies.

Here, we applied comparative sequence-structure analysis to characterize a novel domain that we called EVE. Taking advantage of several structures within the PUA domain-like SCOP-fold solved by different Structural Genomics consortia, we showed that differences in topology, residue conservation and surface cavities clearly separate EVE from other PUA-like domains (e.g. PUA and ASCH). While the molecular function of EVE is most likely RNA-binding, we predict that EVE will have functional roles not shared by other PUA-like domains. Further functional characterization of EVE domains will have to await systematic experimental studies.

Methods

General approach to sequence and structural analysis of PUA-like domains

On our main web page <http://luna.bioc.columbia.edu/honiglab/PUALike>, we show the PDB identifiers of PUA-like domains separated into 5 groups: EVE, EVE-like, ASCH, PUA, and others. For EVE (1zce²⁵, 2eve¹³, 2gbs²⁶, 2g2x²⁷, 2ar1²⁹, 2p5d³⁰, 2hd9³²/1wmm³¹), EVE-like (2yud³⁵, 2yu6³⁶) and ASCH (1s04^{21,22}, 1xne²³, 1te7²⁴, 1wk2⁴⁶, 1t62⁴⁷) we included all such domains that we could find in the PDB. For the other two groups, we only included a few representative domains (i.e., those that are mentioned throughout the paper; for PUA: 1j2b²⁸, 2hvy³⁸ and 1r3e³⁹) (Table 1). PDB identifiers for all these proteins are linked to the pages containing the sequence and structure analysis results. Our database is accessible to any JavaScript capable browser. For protein structure visualization Java has to be enabled. Protein structures are visualized using the AstexViewer 2.0⁴⁸, which allows for interactive manipulation.

We submitted the sequence and structure of each of the proteins found on our main web page to several servers. Structure comparison methods (Skan^{14,15} and Dali⁴⁹) identified proteins with similar structures. The protein surface was scanned for cavities using SCREEN⁵⁰. In parallel, the target sequence was PSI-blasted⁵¹ against UniProt⁴³ to identify related sequences and submitted to InterProScan³³ to provide links to databases such as Pfam^{17,44}. Proteins with related sequences were then aligned using ClustalW⁵²; the resulting multiple sequence alignment was submitted to ConSurf⁵³ assigning a conservation score to each residue in the target. All these data were then used for the manual analysis of function presented in this paper. All details on how the servers and programs were run and on how data were collected and processed follow.

Sequence analysis

Sequence conservation was assessed by three different means: (1) structure-based sequence alignments, i.e. from 3D superimpositions, for all available 3D structures (Skan^{14,15}; Dali⁴⁹; details below), (2) ClustalW⁵⁴ alignments with two versions of generating the input: (i) sequences retrieved by PSI-BLAST⁵¹ using as query the sequence each protein of known structure in our dataset, and (ii) sequences from the Pfam-A family associated with each protein of known structure in our dataset, and (3) the original Pfam family

alignments¹⁷. For the domains within the ASCH group, we also considered the alignments of ten distinct ASCH families reported by Iyer et al.²¹ (in that paper, these correspond to families one to nine plus ASC-1).

For each protein in our dataset (Table 1), we ran three iterations of PSI-BLAST⁵¹ against UniProt⁴³ (release 9.0) and using the E-value cutoffs 10^{-1} , 10^{-3} , and 10^{-10} for the final iteration. Redundant sequences were removed by filtering with CD-HIT⁵⁵ at 80% (E-values 10^{-1} and 10^{-3}) and at 90% (E-value 10^{-10}) pairwise sequence identity. The remaining sequences were aligned by ClustalW 1.83⁵² using default parameters (gap open: 10, gap extension: 0.2, matrix: Gonnet⁵⁶). In addition to the PSI-BLAST based sequence alignment that we just described, we produced two alternative multiple sequence alignments: one by aligning the target to the Pfam^{17,44} seeds, the other by aligning the target to the full Pfam family if available (i.e. the seed list expanded by using HMMer⁵⁷), using the profile alignment method implemented in CLUSTALW 1.83⁵².

We assigned each protein to a particular Pfam^{17,44} family, by submitting its sequence to InterPro⁵⁸ using InterProScan³³. Interpro runs the protein sequence against every Pfam family HMM model (using HMMer⁵⁷), returning a Pfam family assignment for that protein if a match is found. All ASCH and all EVE domains here analyzed (Results) were assigned with high reliability to a Pfam family (E-values $<10^{-10}$). We used Pfam to obtain several different types of information: 1) to extract sequences later aligned with CLUSTALW⁵²; 2) to extract the original Pfam alignments, which might differ from the automatic CLUSTALW alignments that use the same sequences since they are manually adjusted; 3) to analyze the distribution of the members of a Pfam family in Eukarya, Bacteria, and Archaea; and, finally, 4) to study gene fusion events between members of a given Pfam family and other domains. All this information is directly available from the Pfam website.

Structure analysis

We used the two structural alignment programs Skan^{14,15} and Dali⁴⁹ to identify proteins with structures similar to those of the targets. All hits (proposed structural relations) were ranked by a variety of scores: 1) the internal scores of the 3D alignment method (PSD and Z-score for Skan and Dali, respectively); 2) the RMSD; 3) the alignment length; 4) the percentage of pairwise sequence identity; and (5) the structure alignment score (SAS)⁵⁹, which combines alignment length and RMSD. We used VISTAL⁶⁰ for pairwise structure superimposition according to Skan alignments and to visualize aligned secondary structure elements. To identify residues that may play a key role in protein function, each multiple sequence alignment was input to ConSurf⁵³. ConSurf's scores were then mapped onto the targets for visual inspection. Solvent accessible cavities on the protein surface were identified by SCREEN⁵⁰. Cavities were visualized by displaying their molecular surface and with the option of mapping the ConSurf scores⁵³ onto them.

The electrostatic surface potentials were calculated using Delphi⁶¹ as implemented in Grasp2¹⁴ (with internal and external dielectric constants set to 1.0 and 80.0, respectively; constant salt concentration at 0.2 M and the probe radius equal to 1.4 Å). We used charmm22 parameters⁶² for atomic charges and atomic radii.

Genomic analysis

The genomic analysis of the PUA-like domains was performed using all sequences found in the corresponding Pfam families as a reference (e.g. family PF01878-DUF55 for EVE). Hence, hypotheses about the evolutionary history of the domains are based on the assumption that the sequences found in Pfam are representative of the domains' actual genomic distribution.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grant U54-GM074958-01 to the Northeast Structural Genomics consortium from the Protein Structure Initiative (PSI) of the National Institute of General Medical Science (NIGMS), National Institutes of Health (NIH). Guy Yachdav and Burkhard Rost were additionally supported by the grants R01-GM079767 and R01-LM07329 from the NIH; Claudia Bertonati was supported by Istituto Pasteur - Fondazione Cenci Bolognetti Università di Roma "La Sapienza". Marco Punta wishes to thank Marco DeVivo (Rib-X Pharmaceuticals) for helpful discussions. Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases. A portion of the research described in this paper was performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory.

Abbreviations used

ASCH (domain)	Activating Signal Cointegrator 1 Homology
DUF	Domains of Unknown Function
EVE (domain)	a newly characterized PUA-like domain
GOL	glycerol
MPO	3[N-morpholino]propane sulfonic acid
NESG	Northeast Structural Genomics
PDB	Proteins Data Bank
PUA (domain)	<u>P</u> seudouridine synthase and <u>A</u> rchaeosine transglycosylase
PUA-like domains	PUA domain-like domains
RMSD	Root Mean Square Displacement
SAS	Structure Alignment Score
SG	Structural Genomics
YTH (domain)	YT521-B Homology

References

1. Blundell TL, Mizuguchi K. Structural genomics: an overview. *Prog Biophys Mol Biol.* 2000; 73(5): 289–295. [PubMed: 11063776]
2. Bhattacharya A, Tejero R, Montelione GT. Evaluating protein structures determined by structural genomics consortia. *Proteins.* 2006

3. Levitt M. Growth of novel protein structural data. *Proc Natl Acad Sci U S A*. 2007; 104(9):3183–3188. [PubMed: 17360626]
4. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr*. 2002; 58(Pt 61):899–907. [PubMed: 12037327]
5. Liu J, Montelione GT, Rost B. Novel leverage of structural genomics. *Nature Biotechnology*. 2007 in press.
6. Galperin MY, Koonin EV. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res*. 2004; 32(18):5452–5463. [PubMed: 15479782]
7. Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I, Addou S, Lise S, Johnston C, Sillero A, Thornton J, Orengo C. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res*. 2005; 33(Database issue):D247–251. [PubMed: 15608188]
8. Montelione GT, Anderson S. Structural genomics: keystone for a Human Proteome Project. *Nat Struct Biol*. 1999; 6(1):11–12. [PubMed: 9886282]
9. Wunderlich Z, Acton TB, Liu J, Kornhaber G, Everett J, Carter P, Lan N, Echols N, Gerstein M, Rost B, Montelione GT. The protein target list of the Northeast Structural Genomics Consortium. *Proteins*. 2004; 56(2):181–187. [PubMed: 15211503]
10. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 1995; 247(4):536–540. [PubMed: 7723011]
11. Perez-Arellano I, Gallego J, Cervera J. The PUA domain - a structural and functional overview. *Febs J*. 2007; 274(19):4972–4984. [PubMed: 17803682]
12. Marrone A, Walne A, Dokal I. Dyskeratosis congenita: telomerase, telomeres and anticipation. *Curr Opin Genet Dev*. 2005; 15(3):249–257. [PubMed: 15917199]
13. PDB ID: 2EVE Forouhar F, Zhou W, Belachew A, Jayaraman S, Cio M, Xiao R, Acton TB, Montelione GT, Hunt JF, Tong L. Northeast Structural Genomics Consortium. Crystal Structure of the Conserved Hypothetical Protein from *Pseudomonas syringae*, Northeast Structural Genomics Target PsR62.
14. Petrey D, Honig B. GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol*. 2003; 374:492–509. [PubMed: 14696386]
15. Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J Mol Biol*. 2000; 301(3):665–678. [PubMed: 10966776]
16. Aravind L, Koonin EV. Novel predicted RNA-binding domains associated with the translation machinery. *J Mol Evol*. 1999; 48(3):291–302. [PubMed: 10093218]
17. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. *Nucleic Acids Res*. 2004; 32(Database issue):D138–141. [PubMed: 14681378]
18. PDB ID: 1ZBO Forouhar FYW, Conover K, Acton TB, Montelione GT, Tong L, Hunt JF. Northeast Structural Genomics Consortium. Crystal Structure of the Hypothetical Protein BPP1347 from *Bordetella parapertussis*, Northeast Structural Genomics Target BoR27.
19. Forouhar F, Shen J, Xiao R, Acton TB, Montelione GT, Tong L. Functional assignment based on structural analysis: crystal structure of the yggJ protein (HI0303) of *Haemophilus influenzae* reveals an RNA methyltransferase with a deep trefoil knot. *Proteins*. 2003; 53(2):329–332. [PubMed: 14517985]
20. PDB ID: 1T5Y Kuzin AP, Chen Y, Forouhar F, Acton TB, Shastry R, Ma L-C, Cooper B, Xiao R, Montelione G, Tong L, Hunt JF. Northeast Structurally Genomics Consortium. Crystal Structure of Northeast Structural Genomics Consortium Target HR2118: A Human Homolog of *Saccharomyces cerevisiae* Nip7p.

21. Iyer LM, Burroughs AM, Aravind L. The ASCH superfamily: novel domains with a fold related to the PUA domain and a potential role in RNA metabolism. *Bioinformatics*. 2006; 22(3):257–263. [PubMed: 16322048]
22. PDB ID: 1S04 Liu G, Xiao R, Sukumaran DK, Acton T, Montelione GT, Szyperski T. Northeast Structurally Genomics Consortium. Solution Structure Of The Hypothetical Protein PF0455 From *Pyrococcus furiosus*: Northeast Structural Genomics Consortium Target Pfr13.
23. PDB ID: 1XNE Liu G, Xiao R, Parish D, Ma L, Sukumaran D, Acton T, Montelione GT, Szyperski T. Northeast Structural Genomics Consortium. Solution Structure of *Pyrococcus furiosus* Protein PF0470: The Northeast Structural Genomics Consortium Target Pfr14.
24. Shen Y, Atreya HS, Liu G, Szyperski T. G-matrix Fourier transform NOESY-based protocol for high-quality protein structure determination. *J Am Chem Soc*. 2005; 127(25):9085–9099. [PubMed: 15969587]
25. PDB ID: 1ZCE Forouhar F, Chen Y, Conover K, Acton TB, Montelione GT, Hunt JF, Tong L. Northeast Structural Genomics Consortium. Crystal Structure of the Hypothetical Protein Atu2648 from *Agrobacterium tumefaciens*, Northeast Structural Genomics Target AtR33.
26. PDB ID: 2GBS Ramelot TA, Cort JR, Xiao R, Montelione GT, Kennedy MA. Northeast Structural Genomics Consortium. NMR structure of Rpa0253 from *Rhodopseudomonas palustris*.
27. PDB ID: 2G2X Kuzin AP, Chen Y, Abashidze M, Acton T, Conover K, Janjua H, Ma L-C, Ho CK, Cunningham K, Montelione G, Hunt JF, Tong L. Northeast Structural Genomics Consortium. Crystal structure of the conserved hypothetical protein from *Pseudomonas putida* Q88CH6. Northeast Structural Genomics target PpR72.
28. Ishitani R, Nureki O, Nameki N, Okada N, Nishimura S, Yokoyama S. Alternative tertiary structure of tRNA for recognition by a posttranscriptional modification enzyme. *Cell*. 2003; 113(3):383–394. [PubMed: 12732145]
29. Arakaki T, Le Trong I, Phizicky E, Quartley E, DeTitta G, Luft J, Lauricella A, Anderson L, Kalyuzhnyi O, Worthey E, Myler PJ, Kim D, Baker D, Hol WG, Merritt EA. Structure of Lmaj006129AAA, a hypothetical protein from *Leishmania major*. *Acta Crystallograph Sect F Struct Biol Cryst Commun*. 2006; 62(Pt 3):175–179.
30. PDB ID: 2P5D Sugahara M, Kunishima N. RIKEN Structural Genomics/Proteomics Initiative (RSGI) Crystal structure of MJECL36 from *Methanocaldococcus jannaschii* DSM 2661.
31. PDB ID: 1WMM Sugahara M, Kunishima N. RIKEN Structural Genomics/Proteomics Initiative (RSGI). Crystal structure of PH1033 from *Pyrococcus horikoshii* OT3.
32. PDB ID: 2HD9 Sugahara M, Kunishima N. RIKEN Structural Genomics/Proteomics Initiative (RSGI). Crystal structure of PH1033 from *Pyrococcus horikoshii* OT3.
33. Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. 2001; 17(9):847–848. [PubMed: 11590104]
34. Karkare S, Bhatnagar D. Promising nucleic acid analogs and mimics: characteristic features and applications of PNA, LNA, and morpholino. *Appl Microbiol Biotechnol*. 2006; 71(5):575–586. [PubMed: 16683135]
35. PDB ID: 2YUD He F, Muto Y, Inoue M, Kigawa T, Shirouzu M, Tarada T, Yokoyama S. RIKEN Structural Genomics/Proteomics Initiative (RSGI) Solution structure of the YTH domain in YTH domain-containing protein 1 (Putative splicing factor YT521).
36. PDB ID: 2YU6 Endo R, Muto Y, Inoue M, Kigawa T, Shirouzu M, Tarada T, Yokoyama S. RIKEN Structural Genomics/Proteomics Initiative (RSGI) Solution structure of the YTH domain in YTH domain-containing protein 2.
37. Stoilov P, Rafalska I, Stamm S. YTH: a new domain in nuclear proteins. *Trends Biochem Sci*. 2002; 27(10):495–497. [PubMed: 12368078]
38. Li L, Ye K. Crystal structure of an H/ACA box ribonucleoprotein particle. *Nature*. 2006; 443(7109):302–307. [PubMed: 16943774]
39. Pan H, Agarwalla S, Moustakas DT, Finer-Moore J, Stroud RM. Structure of tRNA pseudouridine synthase TruB and its RNA complex: RNA recognition through a combination of rigid docking and induced fit. *Proc Natl Acad Sci U S A*. 2003; 100(22):12648–12653. [PubMed: 14566049]

40. Hoang C, Ferre-D'Amare AR. Cocystal structure of a tRNA Psi55 pseudouridine synthase: nucleotide flipping by an RNA-modifying enzyme. *Cell*. 2001; 107(7):929–939. [PubMed: 11779468]
41. Friedmann M, Holth LT, Zoghbi HY, Reeves R. Organization, inducible-expression and chromosome localization of the human HMG-I(Y) nonhistone protein gene. *Nucleic Acids Res*. 1993; 21(18):4259–4267. [PubMed: 8414980]
42. Reeves R, Nissen MS. The A.T-DNA-binding domain of mammalian high mobility group I chromosomal proteins. A novel peptide motif for recognizing DNA structure. *J Biol Chem*. 1990; 265(15):8573–8582. [PubMed: 1692833]
43. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*. 2004; 32(Database issue):D115–119. [PubMed: 14681372]
44. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A. Pfam: clans, web tools and services. *Nucleic Acids Res*. 2006; 34(Database issue):D247–251. [PubMed: 16381856]
45. Chaudhuri BN, Chan S, Perry LJ, Yeates TO. Crystal structure of the apo forms of psi 55 tRNA pseudouridine synthase from *Mycobacterium tuberculosis*: a hinge at the base of the catalytic cleft. *J Biol Chem*. 2004; 279(23):24585–24591. [PubMed: 15028724]
46. PDB ID: 1WK2 Agari Y, Yokoyama S, Kuramitsu S. Northeast Structurally Genomics Consortium. Crystal structure of a hypothetical protein from *thermus thermophilus* HB8.
47. PDB ID: 1T62 Fedorov AA, Fedorov EV, Almo SC. Northeast Structurally Genomics Consortium. Crystal structure of conserved hypothetical protein [gi:29377587] from *Enterococcus faecalis* v583.
48. Hartshorn MJ. AstexViewer: a visualisation aid for structure-based drug design. *J Comput Aided Mol Des*. 2002; 16(12):871–881. [PubMed: 12825620]
49. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J Mol Biol*. 1993; 233(1):123–138. [PubMed: 8377180]
50. Nayal M, Honig B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*. 2006; 63(4):892–906. [PubMed: 16477622]
51. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997; 25(17):3389–3402. [PubMed: 9254694]
52. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 1994; 22(22):4673–4680. [PubMed: 7984417]
53. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res*. 2005; 33(Web Server issue):W299–302. [PubMed: 15980475]
54. Higgins DG. CLUSTAL V: multiple alignment of DNA and protein sequences. *Methods Mol Biol*. 1994; 25:307–318. [PubMed: 8004173]
55. Li W, Jaroszewski L, Godzik A. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*. 2002; 18(1):77–82. [PubMed: 11836214]
56. Benner SA, Cohen MA, Gonnet GH. Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng*. 1994; 7(11):1323–1332. [PubMed: 7700864]
57. Eddy SR. Profile hidden Markov models. *Bioinformatics*. 1998; 14(9):755–763. [PubMed: 9918945]
58. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R, Wu CH. InterPro, progress and status in 2005. *Nucleic Acids Res*. 2005; 33(Database issue):D201–205. [PubMed: 15608177]

59. Subbiah S, Laurents DV, Levitt M. Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr Biol.* 1993; 3(3):141–148. [PubMed: 15335781]
60. Kolodny R, Honig B. VISTAL--a new 2D visualization tool of protein 3D structural alignments. *Bioinformatics.* 2006; 22(17):2166–2167. [PubMed: 16837525]
61. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J Comput Chem.* 2002; 23(1):128–137. [PubMed: 11913378]
62. MacKerell JDA, Bashford D, Bellott M, Dunbrack JRL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE III, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J Phys Chem B.* 1998; 102:3586–3616. [PubMed: 24889800]
63. Ullrich TC, Blaesle M, Huber R. Crystal structure of ATP sulfurylase from *Saccharomyces cerevisiae*, a key enzyme in sulfate activation. *Embo J.* 2001; 20(3):316–329. [PubMed: 11157739]

Fig. 1. PUA-like domain topologies

In blue and orange we denote strands and helices respectively. Insertions (INS0, INS1, INS2) characteristic of specific domains are marked in red and additional strands (i.e. strands within insertions, namely 6 and 2') in lavender. The four panels represent: (A) the basic elements (core) of the PUA-like domains topology as reported in SCOP¹⁰, (B) the ASCH domain architecture²¹, (C) the EVE domain architecture, (D) the architecture of the PUA domains described in the text.

Fig. 2. Structure-based sequence alignment of 2eve¹³, 2ar1²⁹, 2g2x²⁷, 1zce²⁵, 2gbs²⁶, 2hd9/1wmm³²⁻³¹ and 2p5d³⁰ (EVE domains)

The sequence alignments are obtained automatically from pairwise structural alignments between 2eve and the other proteins (Table 2). Each protein is represented by two lines, one giving the sequence, the other the secondary structure assignment (according to Skan^{14,15}). Numbers at each end indicate the first and last residue found in the structure, following the PDB⁴ sequence numbering for each protein. Below the alignment we highlight the topology of the EVE domains (with strands 1–6, helix A and the two insertions INS1 and INS2). Boxes mark residues that are generally not only conserved in the structural alignment but also in sequence-based alignments from PSI-BLAST⁵¹ and CLUSTALW⁵² or Pfam¹⁷. Conserved residues and numbers reported above the boxes are as in the 2eve PDB file.

Fig. 3. Ligand binding and surface conservation in the main surface cavity of EVE domains
 (A) 2eve¹³ structure in cartoon representation (green) with co-crystallized ligands 3[N-morpholino]propane sulfonic acid (MPO) and tris-hydroxymethyl-methyl-ammonium (TRIS) in ball and stick representation (CPK colors). In red (licorice representation) we show conserved residues (Y3, W4, F13, W26, V29, D45, Y50, W142). (B) 2eve molecular surface colored by residue conservation score (as calculated by ConSurf⁵³), conserved residues are in purple, variable residues are in cyan, the orientation of 2eve is as in (A). MPO and TRIS are in ball and stick representation (CPK colors).

Fig. 4. Structure-based sequence alignment between ASCH and EVE domains

Notations are as in Fig. 2. The alignment is obtained from pairwise structural alignments between 1s04²¹²² and the other proteins, except for 2g2x²⁷, whose sequence was manually added taking its pairwise structural alignment to 2ar1²⁹ as a template (since Skan failed to identify 2g2x as a structural homolog of 1s04). For simplicity, alignments are reported only up to INS1, i.e. the insertion between strand 4 and 5. ASCH sequences are on top and EVE at the bottom. Numbers at each end of a sequence indicate the first and last residues following the PDB⁴ sequence numbering for each protein. As in Fig. 3, boxes highlight conserved residues generally indicating not only conservation in the structural alignment but also in sequence-based alignments obtained using PSI-BLAST⁵¹ and CLUSTALW⁵² or Pfam¹⁷. Type and numbers for these residues are as in the PDB sequence of 2eve¹³. Note that in this alignment W26 is slightly misaligned in the EVE domains (compare to Fig. 2) and so is the ASCH motif²¹ GxKxxE/T/SxR in the ASCH domains.

Fig. 5. Virtual complex of 2eve and tRNA from 1j2b

The figure was obtained according to the following procedure: first, 2eve¹⁹ was superimposed on the PUA domain of ArcTGT 1j2b complexed with tRNA²⁸ (residues 507–582) (using GRASP2¹⁴); then, the 1j2b PUA domain structure was removed while its bound tRNA and 2eve and its ligand (MPO) was kept in place. (A) 2eve structure is cartoon representation (blue), tRNA (CPK colors), MPO (green) are in ball and stick representation. Relevant residues (see text) are in licorice representation (blue). (B) 2eve molecular surface colored by residue conservation score (as calculated by ConSurf⁵³), conserved residues are in purple, variable residues are in cyan, the orientation of 2eve is as in (A). tRNA (CPK colors), MPO (green) are in ball and stick representation.

Table 1

Proteins analyzed in this paper *

PDBid [NESGid]	Our 3D classification	SCOP family	Pfam family	Experimental method	References
1I2B	PUA	PUA domain	PF01472	X-ray	28
2HVV	PUA	PUA domain	PF01472	X-ray	38
1R3E	PUA	PUA domain	PF01472	X-ray	39
1T5Y (95-170) [#] [HR2118]	PUA	PUA domain	PB010048 (PFAM-B)	X-ray	20
1NXZ (2-73) [#] [IR73]	Other (RNA methyltransferase)	YggJ N-terminal domain-like	PF04452	X-ray	19
1G8F(2-168)	Other (ATP sulfurylase N-terminal domain)	ATP sulfurylase N-terminal domain	PF01747	X-ray	63
1XNE [#] [PJR14]	ASCH (family 3 ²¹)	ProFAR isomerase associated	PF04266 (DUF)	NMR	23
1S04 [#] [PJR13]	ASCH (family 3 ²¹)	ProFAR isomerase associated	PF04266 (DUF)	NMR	2122
1TE7 [#] [ET99]	ASCH (family 6 ²¹)	yqfB-like	PF04266 (DUF)	NMR	24
1T62	ASCH (family 4 ²¹)	EF3133	PF04266 (DUF)	X-ray	2147
2DP9/1WK2	ASCH (family 1 ²¹)	TTHA0113	PF04266 (DUF)	X-ray	2146
1ZCE [#] [AUR33]	EVE	Atu2648/PHI1033-like	PF01878 (DUF)	X-ray	25
2EVE [#] [P-R62]	EVE	Atu2648/PHI1033-like	PF01878 (DUF)	X-ray	13
2AR1	EVE	Atu2648/PHI1033-like	PF01878 (DUF)	X-ray	29
2G2X [#] [PpR72]	EVE	Atu2648/PHI1033-like	PF01878 (DUF)	X-ray	27
2GBS [#] [RprR3]	EVE	Atu2648/PHI1033-like	PF01878 (DUF)	NMR	26
2HD9/1WMM	EVE	Atu2648/PHI1033-like	PF01878 (DUF)	X-ray	3132
2P5D	EVE	not in SCOP	PF01878 (DUF)	X-ray	30
2YUD	EVE-like	not in SCOP	PF04146	NMR	35
2YU6	EVE-like	not in SCOP	PF04146	NMR	36
1ZBO [#] [BoR27]	Other (ATP-dependent protease La)	LON domain like	PF02190	X-ray	18

* **PDBid:** PDB identifier ⁴; number signs mark structures determined by NESG (with NESG ID in squared parentheses); **Our 3D classification:** structural classification of the target based on the analysis performed by the authors (see text for more details); family numbers for ASCH domains follow Iyer et al. ²¹; **SCOP family:** all SCOP ¹⁰ families listed belong to the PUA domain-like superfamily and fold in SCOP; **Pfam family:** families annotated by Pfam ¹⁷ as belonging to the ASCH/PUA clan in bold; **Experimental method:** method used for structure determination; **References:** PDB reference or publications reporting on the experimental structures when available (note: the link to reference ²¹ does not indicate the structural paper but simply that the structures were discussed in ²¹).

Table 2

Conservation of structure and sequence among EVE domains *

RMSD/AL (%ID)	1ZCE	2EVE	2AR1	2G2X	2GBS	2HD9/ 1WMM	2P5D
1ZCE	0.0/146						
2EVE	1.5/132 (40%)	0.0/149					
2AR1	1.4/135 (45%)	0.8/147 (50%)	0.0/157				
2G2X	1.4/132 (42%)	0.6/147 (78%)	1.0/147 (50%)	0.0/149			
2GBS	1.3/139 (53%)	1.6/132 (40%)	1.9/136 (40%)	1.5/132 (38%)	0.0/145		
2HD9/ 1WMM	2.7/117 (22%)	2.8/121 (20%)	2.6/123 (19%)	2.8/120 (20%)	2.9/115 (22%)	0.0/143	
2P5D	3.0/116 (21%)	3.1/122 (19%)	2.6/123 (23%)	3.1/119 (18%)	2.9/115 (22%)	1.4/138 (59%)	0.0/145

* First row and first column: PDB identifiers of EVE domains. Other rows and columns: the first number represents the RMSD in (in Angstrom) between EVE domain pairs as obtained by means of the 3D alignment program Skan 14,15; the second number is the length of the alignment; the third number in parentheses is the percent sequence identity between the two domains as derived from the structural alignment (note: the diagonals show the length of the domains that varies from 143 to 157 residues). For instance, 1zce and 2eve have 1.5 Angstrom RMSD over 132 structurally aligned residues and 40% sequence identity. Cell colors reflect RMSD values, ranging from low=similar (dark) to high=dissimilar (light), and according to the following intervals (in Angstrom): 0.0-1.0 (dark), 1.1-1.5, 1.6-2.0, more than 2.0 (light).

Table 3

Summary of differences between EVE, ASCH and PUA domains *

	EVE	ASCH	PUA (1j2b, 2vhy, 1re3)
Topology	Add* strand 6 paired to strand 1; add INS1+INS2 (Fig. 1C)	Add INS1 (Fig 1B)	Add strand 2' paired to strand 1; add INS1+INS0 (Fig 1D)
Conserved residues	Y3, W4, F13, W26, V29, G44, D45, W142 (numbers as in 2eve)	G19-x-K21-x-x-E24-x-R26 (numbers as in 1xne) Y12 (aligned to F13 of EVE), G38 (aligned to G44), D39 (aligned to D45)	Depend on PUA domain
Cavity size	Lined by helix A, strand 1,2 and 6. Surface area ~ 200 Å ²	Lined by helix A, strand 1 and strand 2. Surface area ~ 300 Å ²	Lined by helix A and strand 2. Surface area < 100 Å ²
Residues lining the cavity	Mainly hydrophobic (F13, W26, V29)	Mainly charged and polar residues (Y12, E24 and R26)	Presence of a positive patch although not completely conserved throughout the family
Sequence identity (alignment length)	EVE/ASCH: 18%, (77) EVE/PUA: 9%, (66)	ASCH/EVE: 18%, (77) ASCH/PUA:10%, (68)	PUA/EVE:9%, (66) PUA/ASCH:10%, (68)
RMSD (alignment length)	EVE/ASCH: 3.4 Å, (84) EVE/PUA: 3.3 Å, (66)	ASCH/EVE:3.4 Å, (84) ASCH/PUA:3.7 Å, (68)	PUA/EVE:3.3 Å, (66) PUA/ASCH:3.7 Å, (68)

* 'Add' refers to structural changes with respect to the SCOP-fold core topology (Fig. 1A). Sequence identities in row 7 are the highest among any pair of domains belonging to the two structural themes indicated. For example, EVE/ASCH 18% sequence identity means that 18% is the maximum sequence identity (according to Skan structure-based sequence alignments) between any two EVE and ASCH domains. Similarly, 3.4 EVE/ASCH RMSD is the minimum RMSD between any two EVE and ASCH domains. Numbers in parentheses represent alignment lengths.