

# SCIENTIFIC REPORTS



OPEN

## Evolutionary convergence and divergence in archaeal chromosomal proteins and Chromo-like domains from bacteria and eukaryotes

Gurmeet Kaur<sup>1</sup>, Lakshminarayan M. Iyer<sup>1</sup>, Srikrishna Subramanian<sup>2</sup> & L. Aravind<sup>1</sup>

SH3-fold- $\beta$ -barrel domains of the chromo-like superfamily recognize epigenetic marks in eukaryotic proteins. Their provenance has been placed either in archaea, based on apparent structural similarity to chromatin-compacting Sul7d and Cren7 proteins, or in bacteria based on the presence of sequence homologs. Using sequence and structural evidence we establish that the archaeal Cren7/Sul7 proteins emerged from a zinc ribbon (ZnR) ancestor. Further, we show that the ancestral eukaryotic chromo-like domains evolved from bacterial versions, likely acquired from early endosymbioses, which already possessed an aromatic cage for recognition of modified amino-groups. These bacterial versions are part of a radiation of secreted SH3-fold domains, which spawned both chromo-like domains and classical SH3 domains in the context of peptide-recognition in the peptidoglycan or the extracellular matrix. This establishes that Cren7/Sul7 converged to a "SH3"-like state from a ZnR precursor via the loss of metal-chelation and acquisition of stronger hydrophobic interactions; it is unlikely to have participated in the evolution of the chromo-like domains. We show that archaea possess several Cren7/Sul7-related proteins with intact Zn-chelating ligands, which we predict to play previously unstudied roles in chromosome segregation during cell-division comparable to the PRC barrel and CdvA domain proteins.

Three-dimensional structures or folds of proteins are evolutionarily less prone to change than their sequences<sup>1–3</sup>. In the absence of statistically-significant sequence similarity, the detection of structural equivalences can be used to assess evolutionary relatedness<sup>1,4,5</sup>. However, the evidence can, in some instances, be equivocal regarding structural convergence versus divergence: the moot point in these cases is whether the structural similarity in folds in question is a signal of a divergent origin from a common ancestor or independent convergence to a common scaffold<sup>1,6–9</sup>. Automated sequence- and structure-similarity search tools, though widely used for gauging relatedness among proteins, are often of limited utility in these cases. Tracing the correct relationships demands careful case-by-case analysis<sup>5,10</sup> and on multiple occasions, has helped untangle convergence from extreme divergence, which had otherwise eluded automated similarity search tools<sup>11–17</sup>. In this work, we present such a case regarding the SH3 fold and certain zinc ribbons (ZnRs), with bearing on the function and evolution of key domains involved in chromatin structure and chromosome segregation in archaea, recognition of epigenetic marks in eukaryotes, and bacterial cell-wall dynamics.

The Src homology 3 (SH3) is a small  $\beta$ -barrel domain, comprised of five or six  $\beta$ -strands that are tightly packed into two orthogonal  $\beta$ -sheets<sup>18</sup>. The eponymous SH3 domains are involved in eukaryotic signaling pathways where they mediate protein-protein interactions by binding proline-rich peptide sequences via a conserved cluster of aromatic residues<sup>18–20</sup>. The discovery of bacterial homologs of the SH3 domain presented an interesting contrast as they were primarily found as extracellular domains in periplasmic or cell-wall associated proteins<sup>21–23</sup>. Members of the larger SH3-like  $\beta$ -barrel fold (hereinafter SH3 fold) include a vast collection of superfamilies found in diverse

<sup>1</sup>National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA. <sup>2</sup>Bioinformatics Centre, CSIR-Institute of Microbial Technology, Sector 39A, Chandigarh, 160036, India. Correspondence and requests for materials should be addressed to S.S. (email: [krishna@imtech.res.in](mailto:krishna@imtech.res.in)) or L.A. (email: [aravind@mail.nih.gov](mailto:aravind@mail.nih.gov))

biological functional contexts. The best characterized of them are implicated in a variety of key protein-protein interactions via recognition of short peptide motifs<sup>24</sup>. SH3-fold  $\beta$ -barrels also mediate interactions with nucleic acids<sup>24,25</sup>. For example, PAZ (Piwi ArgonAUT and ZWille), a SH3-fold  $\beta$ -barrel domain, found in the Piwi and the Dicer proteins in the RNAi system interacts with RNA<sup>26–28</sup>. Likewise, certain representatives of other families with the SH3 fold such as the CarD<sup>29,30</sup>, chromo<sup>31</sup>, and TUDOR domains<sup>32,33</sup> have been shown to bind DNA.

In recent years, it has become clear that in eukaryotes a large superfamily of domains with the SH3 fold, the chromo-like superfamily, plays a key role in recognition of short peptide-motifs, especially those with covalently modified side-chains in chromatin (chiefly histones) and RNA-processing proteins. This superfamily includes the classical chromo (chromatin organization modifier) domains, BAM/BAH, BMB/PWWP and Tudor-like domains<sup>34–36</sup>. The Tudor-like domains further include within them the Tudor, MBT (malignant brain tumor), Agenet, DUF3590, DUF1325, RAD53BP, Tudor-knot, AuxRF(PF06507) and the MORC C-terminal domains<sup>37,38</sup> (PFAM clan CL0049). The conserved structural core of the chromo-like domains features a SH3-fold  $\beta$ -barrel with 5 strands that is often capped by a C-terminal helix<sup>34,35</sup>. They share a broadly conserved mode of interaction with peptides, specifically recognizing covalent modifications of positively charged side-chains via cation- $\pi$  interactions with conserved aromatic residues<sup>39</sup>. While most members bind peptides with methylated lysines, representatives of the Tudor-like domains specialize in binding peptides with methylated arginines.

These covalent modifications along with the chromo-like domains that bind them are defining features of all eukaryotes, which set them apart from prokaryotes<sup>37</sup>. Hence, understanding the provenance of eukaryotes depends on a proper explanation for their origins in the stem eukaryotic lineage. Sequence and structure comparison studies have proposed two distinct possibilities for their origins. In the first, based on structural similarity, an evolutionary relationship was proposed between the eukaryotic chromo-like domains and the SH3-like fold of archaeal Sul7d-like chromatin-compaction proteins<sup>40</sup>. These in turn are structurally and functionally related to the major pan-crenarchaeal Cren7 protein, also involved in DNA-compaction and supercoiling<sup>41,42</sup>. Thus, in this scenario, the eukaryotic chromo-like domains were derived from an archaeal chromatin protein in the context of chromatin function<sup>40</sup>. In the alternative scenario, the presence of unambiguous sequence homologs of the chromo-like domains in bacterial proteins suggest that the eukaryotic versions evolved from the bacterial precursors<sup>38</sup>.

To distinguish between these alternatives and better understand the function of the bacterial chromo-like domains we sought to utilize the wealth of new genomic and structural data. Using sequence and structure comparisons, we demonstrate that the Cren7/Sul7-like “SH3 fold” domains are members of the zinc ribbon fold with certain members, including Cren7, secondarily losing their ability to chelate a zinc ion. We further shown that the Sul7 proteins most likely arose in Sulfobacterales as a paralogous family of Cren7. We also show that the eukaryotic chromo-like domains instead evolved from bacterial precursors as part of the radiation of extracellular SH3 fold domains in the context of bacterial cell-wall dynamics. Based on these considerations, we hypothesize that the SH3-like  $\beta$ -barrel architecture convergently emerged from an ancestral ZnR fold in Cren7/Sul7-like proteins.

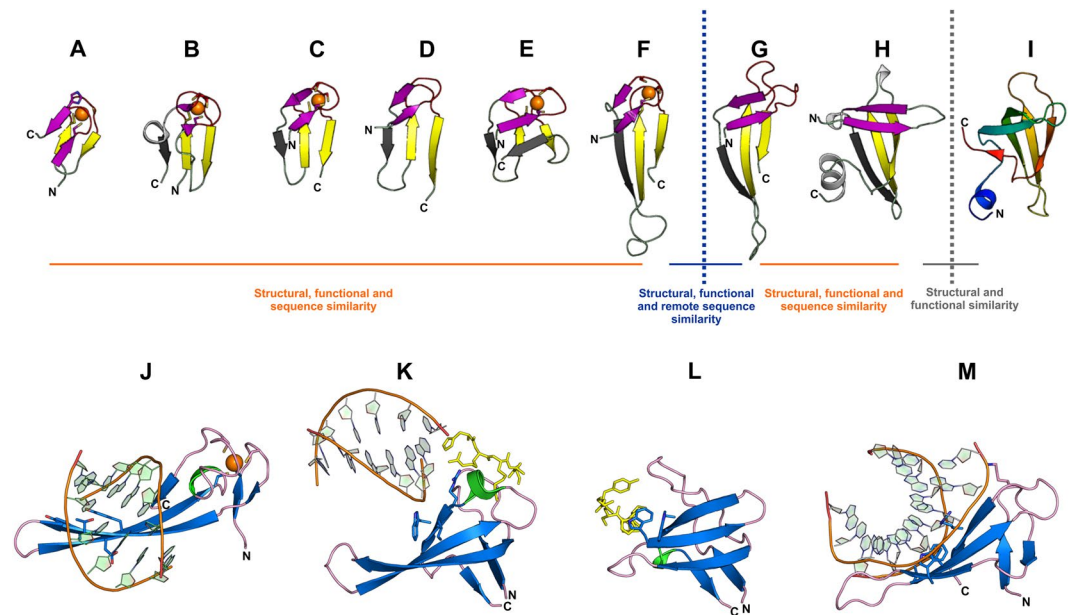
## Results and Discussion

### Structural diversity of ZnR domains and the relationship of certain versions to the SH3-like $\beta$ -barrel fold.

ZnRs are small domains that lack an extensive hydrophobic core beyond the presence of two  $\beta$ -hairpins, being primarily stabilized by chelation of a metal ion<sup>43–46</sup>. The metal ion is chelated by two cysteine ligands from each of the “knuckles” with the consensus motif ‘CPxCG’ situated in the turn of each  $\beta$ -hairpin<sup>45–47</sup>. A comparative structural analysis revealed two major types of the ZnR fold, type-1 and type-2. The type-1 ZnRs have a distinct separation between the N-terminal and C-terminal  $\beta$ -hairpins such that no contiguous  $\beta$ -sheets incorporating the two knuckles or parts thereof are observed. Type-1 ZnRs can be further distinguished into two sub-types based on the presence of an additional  $\beta$ -strand at the C-terminus. Type-1A ZnRs (Fig. 1A) have a structural core made up of only two  $\beta$ -hairpins. Examples of these ZnRs include the Ran-binding family (PDB: 3CH5\_B), the pre-SET domain associated with certain SET domain protein methylases (PDB: 2L9Z\_A) and cytochrome-c oxidase polypeptide Vb (PDB: 1OCC\_F). In Type-1B ZnRs, (Fig. 1B), the C-terminal  $\beta$ -hairpin extends into an additional  $\beta$ -strand that forms a three-stranded  $\beta$ -sheet with the N-terminal  $\beta$ -hairpin. The type-1B ZnRs are typified by the rubredoxins (PDB: 1DX8), the ZnRs of methionyl tRNA synthetases (PDB: 1A8H), class IIA glycyl tRNA synthetases (PDB: 2ZT5)<sup>48</sup> and threonine synthase (PDB: 1VB3)<sup>49</sup>, and the ubiquitin-binding ZnRs of deubiquitinating peptidases of the UBCH family (PDB: 2GFO).

In type-2 ZnRs the N-terminal  $\beta$ -hairpin is extended into a  $\beta$ -strand that forms hydrogen bonds with the C-terminal  $\beta$ -hairpin resulting in a sheet formed by a three-stranded  $\beta$ -meander (Fig. 1C–F). Type-2 ZnRs tend to be prevalent in nucleic acid-binding proteins (PDB: 1TFI, 1L1O) and in a strand-swapped version in the DNA-binding Ku and MarR family ZnR domains (PDB: 1JFY, 2F2E)<sup>46,50,51</sup>. Despite these structural differences the two classes of the ZnRs are likely to be related because they possess a common four-stranded core, conserved geometry of the Zn-chelating residues and can be structurally superimposed.

Interestingly, while developing this classification of ZnRs, we noticed a consistent structural similarity between type-2 ZnRs and SH3 domains (Fig. 1G–I). For example, a DALI<sup>52</sup> search initiated with the ZnRs from the RNA polymerase subunit RBP9 (PDB: 1QYP\_A), in addition to retrieving various ZnRs (e.g. Ribonuclease P protein component 4, PDB: 1X0T\_A, Z-score = 3.9, RMSD = 1.2 Å, lali = 38; zinc finger protein ZPRI, PDB: 2QKD\_A, Z-score = 3.3, RMSD = 3.3 Å, lali = 40), recovered several distinct SH3-like  $\beta$ -barrel domains albeit with lower Z-scores such as the classic SH3 domains (e.g. Myosin VI, PDB: 2VB6\_A, Z-score = 2.2, RMSD = 1.8 Å, lali = 33; Rho guanine exchange factor 16, PDB: 1X6B, Z-score = 2.2, RMSD = 5.1 Å, lali = 43), the Tudor domain (e.g. Survival motor neuron protein, PDB: 4A4E\_A, Z-score = 2.5, RMSD = 2.9 Å, lali = 41) and the chromo domain (e.g. Mortality factor 4-like protein 1, PDB: 2F5K\_F, Z-score = 2.0, RMSD = 2.6 Å, lali = 39). Several SH3-like  $\beta$ -barrel folds were also retrieved in structural searches initiated with the ZnRs of peptide:N-glycanase (PDB: 1X3W\_A), ORF131 of *Pyrobaculum* spherical virus (PDB: 2X5C\_A), anaerobic ribonucleotide-triphosphate reductase (PDB: 1HK8\_A) and lysyl-tRNA synthetase (PDB: 1IRX\_A). Further, visual inspection confirmed this

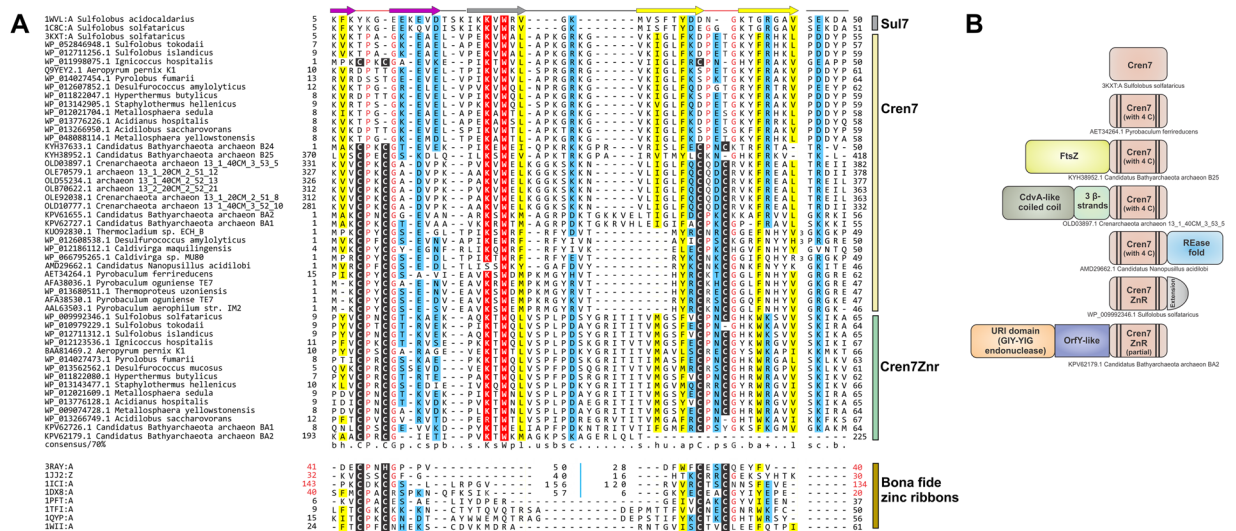


**Figure 1.** Comparative structural view of the major secondary elements (A–I) and binding interfaces (J–M) in ZnRs, SH3s and Cren7. (A) Type-1A ZnR (PR domain-containing protein 11, PDB: 3RAY\_A) (B) Type-1B zinc ribbon (Rubrerythrin, PDB: 1NNQ\_A) (C) Type-2 zinc ribbon (Archaeal exosome RNA binding protein CSL4, PDB: 2BA1\_A) (D) Type-2 zinc ribbon (Uncharacterized BCR *cg1592*, PDB: 2JNY\_A) (E) ZnRs with three strands on each side (Anaerobic ribonucleoside-triphosphate reductase, PDB: 1H8K\_A) (F) Type-2 zinc finger (Transcriptional Elongation Factor SII, PDB: 1TFI\_A) (G) Cren7 (PDB: 3KXT\_A) (H) Sso7 (PDB: 1C8C\_A) (I) SH3-like fold (Tudor domain-containing protein 3, PDB: 3PNW\_O). (J) DNA bound zinc ribbon domain (putative integrase [Bacteriophage A118], PDB: 4KIS\_A) (K) DNA and peptide bound chromodomain (Male specific Lethal-3, PDB: 3OA6\_A) (L) peptide bound SH3 domain (Tyrosine-protein kinase ABL1, PDB: 1BBZ\_A) (M) DNA bound Cren7 (PDB: 3KXT\_A). In all the panels, the Zn ion is represented by an orange sphere, side chains of zinc-chelating and other functionally-important amino acids are represented as sticks, bound peptides are yellow colored stick, the sugar-phosphate backbone of DNA is orange with nucleotides in green. Color scheme for (A–I): For type-2 zinc ribbons, the N-terminal  $\beta$ -hairpin is colored purple, the third  $\beta$ -strand is colored gray, the zinc knuckles are colored red and the C-terminal  $\beta$ -hairpin is colored yellow. The equivalent  $\beta$ -strands in type-2, -1A and -1B are colored alike. The Tudor domain is colored in a gradient of blue to red from the N- to the C-terminal. In panels (J–M): All  $\beta$ -strands are colored in blue, all loops in pink and  $\alpha$ -helices in green.

relationship, indicating a topological similarity in arrangement of the  $\beta$ -strands in type-2 ZnRs and the  $\beta$ -barrel of certain domains with the SH3 fold (Fig. 1E–I).

**Type-2 ZnRs and SH3-fold domains show comparable ligand-binding interfaces.** The detection of this relationship between type-2 ZnRs and SH3 fold domains led us to further compare the ligand-binding interfaces of members of the two folds. In type-2 ZnRs the ligand-binding surface is formed by the three-stranded  $\beta$ -sheet typical of these domains<sup>45</sup>. The interacting residues often emanate from strands  $\beta 3$  and  $\beta 4$  and frequently have aromatic or charged side chains (Fig. 1J). For example, such an interface is used by ZnRs in TFIIS, TBP N-terminal domain, Ku, RBP9 and the RNaseP Rpp21 to contact nucleic acids<sup>53–55</sup>. Interestingly, while chromo-like domains bind methylated peptides mainly via the open region of the barrel, representatives bind nucleic-acids using an interface that is spatially comparable to the equivalent interface of the Type-2 ZnRs (Fig. 1J,K). Classical SH3 domains use the same interface as the nucleic-acid-binding chromo-like domains, with a conserved tryptophan residue, to bind their proline-rich peptide ligand<sup>19,20</sup> (Fig. 1L). The position of this tryptophan corresponds to that of the nucleic acid-interacting residues in ZnRs. These observations suggest that in addition to the structural similarities, at least one binding interface of the SH3 fold domains is similar that of the type-2 ZnRs. This analysis also led us to the archaeal chromosomal proteins Cren7 (PDB: 3KXT; Fig. 1G) and Sac7d/Sul7 (PDB: 1WD0; Fig. 1H) proteins, which have been classified with the chromo-like SH3 fold domains<sup>3,56</sup>, and have even been proposed as their precursors<sup>40</sup>. In these proteins too, the residues responsible for DNA-binding are mainly contributed by the region of the triple-stranded  $\beta$ -sheet ( $\beta 3$ - $\beta 4$ - $\beta 5$ ) (Fig. 1M)<sup>41,57,58</sup> which presents a clear parallel to the type-2 ZnRs.

**The origin of Cren7 and Sul7 proteins from ZnRs and their diversification in archaea.** The above observations hinted that Cren7 with structural features resembling both the SH3 fold and ZnR domains might help better understand the connections between the two folds. Consistent with earlier findings<sup>41</sup>, DALI searches initiated with Cren7 protein (PDB: 3KXT\_A), recovered multiple SH3-like fold domains such as chromo domains, Tudor domains and myosin S1 fragment SH3 domains<sup>41</sup>. These searches also recovered the



**Figure 2.** Multiple sequence alignment (MSA) and domain architectures of Cren7 proteins. (A) Structure-based MSA of representative sequences of Sul7, Cren7, Cren7Znr and bona fide ZnRs. Accession number or the PDB identifier, and the alignment range are indicated for each sequence. For Sul7, Cren7 and Cren7Znr proteins, the organism names are also mentioned after the accession number/PDB. Zn-chelating residues are boxed in black and small amino acids (Gly, Pro) just after the Zn-binding ligands are colored in red. At positions with a 70% consensus, charged or polar residues are highlighted in blue, and hydrophobic, apolar and aliphatic residues are in yellow. Some insertions are not displayed for clarity, and the number of omitted residues is indicated. Regions of circular permutation in ZnRs are separated by a blue ‘|’ mark and the sequence numbers of the permuted ZnRs are shown in red. The consensus secondary structures are depicted above the alignment. (B) Domain architectures of Cren7 and Cren7Znr proteins. Representative sequence identifier with organism name are mention below individual architectures.

Sul7 protein<sup>41,42</sup> (PDB: 1WVL\_A), which is believed to be structurally and functionally related to Cren7 as one of the hits (PDB: 1WVL\_A, Z-score: 3.3, RMSD = 3.1 Å lali = 46). Concurrently, the search also retrieved hits to ZnRs, such as those in the sarcosine oxidase delta subunit (PDB: 1VRQ\_D, Z-score = 3.0, RMSD = 2.8 Å, lali = 44), PNGase (PDB: 3ESW\_A, Z-score = 2.8, RMSD = 2.6 Å, lali = 42), ZPRI (PDB: 2QKD\_A, Z-score = 2.4, RMSD = 3.4 Å, lali = 44), peptide:N-glycanase (PDB: 1 × 3W\_A, Z-score = 2.2, RMSD = 3.1 Å, lali = 43) and 50S ribosomal protein L44E (PDB: 1Q81\_4, Z-score = 2.0, RMSD = 3.2 Å, lali = 45). Manual structural superimposition of Cren7 (PDB: 3KXT\_A) and ZnRs (eg. sarcosine oxidase delta subunit, PDB: 1VRQ\_D) perfectly aligned all secondary structure elements of the Cren7 β-barrel onto the core of the ZnR fold (RMSD = 1.5 Å over 35 pairs of backbone C<sub>α</sub> atoms), where the turns of β1/β2 and of β4/β5 of Cren7 recapitulate the position of knuckles in ZnRs. Similar results were obtained by automated pairwise structural alignment of Cren7 (PDB: 3KXT\_A) and ZnRs (eg. sarcosine oxidase delta subunit, PDB: 1VRQ\_D) using TM-align<sup>59</sup> and Fr-TM-align<sup>60</sup> which gave a TM score of 0.52 (normalised by the length of 3KXT\_A), indicating a ‘fold level’ similarity between the two.

To investigate if this structural relationship to ZnRs also extends to sequence similarity, we initiated iterative sequence similarity searches with *S. solfataricus* Cren7 (PDB: 3KXT\_A) using the PSI-BLAST<sup>61</sup> and the JACKHMMER<sup>62</sup> programs. Interestingly, these searches recovered multiple orthologous sequences from crenarchaeota and bathyarchaeota that contained two to four cysteine residues at positions corresponding to turns between strands β1/β2 and β4/β5 (eg. *Ignicoccus hospitalis*, WP\_011998075.1) (Fig. 2A). The positions of these cysteines suggest that when four of them are present they are likely to chelate a Zn ion. Moreover, we found another group of Cren7 homologs from several crenarchaea which contained a Cren7 domain as part of a larger multidomain protein (see below). These Cren7 domains were characterized by the presence of all four expected Zn-chelating cysteines. Additionally, our searches also retrieved a distinct paralogous family of crenarchaeal proteins (Cren7Znr family; e.g. APE\_2454.1 from *Aeropyrum pernix*; BAA81469.2) typified by an extended C-terminal region, most of whose members possess intact Zn-chelating cysteines (Fig. 2A). Profile-profile sequence similarity searches<sup>63</sup> initiated with the Cren7 and Cren7Znr domains consistently retrieved ZnRs. For example, HHPRED searches with *I. hospitalis* (WP\_011998075.1) found matches to many ZnR proteins such as the lysine biosynthesis protein LysX/ArgZ (PDB: 5K2M\_E, E-value = 0.0005), ZitP pilus assembly/motility regulator (PDB: 2NB9\_A, E-value = 0.0054) and transcriptional regulator MqsA (PDB: 3O9X\_A, E-value = 0.091). Further, even among the Cren7Znr proteins some representatives show loss of one of the Zn-chelating cysteines exactly paralleling the situation observed among the classic Cren7 proteins (Fig. 2A).

Thus, multiple lines of evidence support Cren7-like proteins (PDB: 3KXT\_A) being ZnRs, with secondary loss of Zn-chelating residues in some versions such as the *S. solfataricus* Cren7. Although the Sul7 proteins were not retrieved in these sequence searches, they show several features that suggests their derivation from Cren7 proteins. (1) They are so far only found in a limited number of crenarchaeal lineages (*Sulfolobus*, *Acidanus* and *Metallosphaera*) unlike Cren7, which is conserved across crenarchaeota and the bathyarchaeota<sup>64</sup>. (2) They

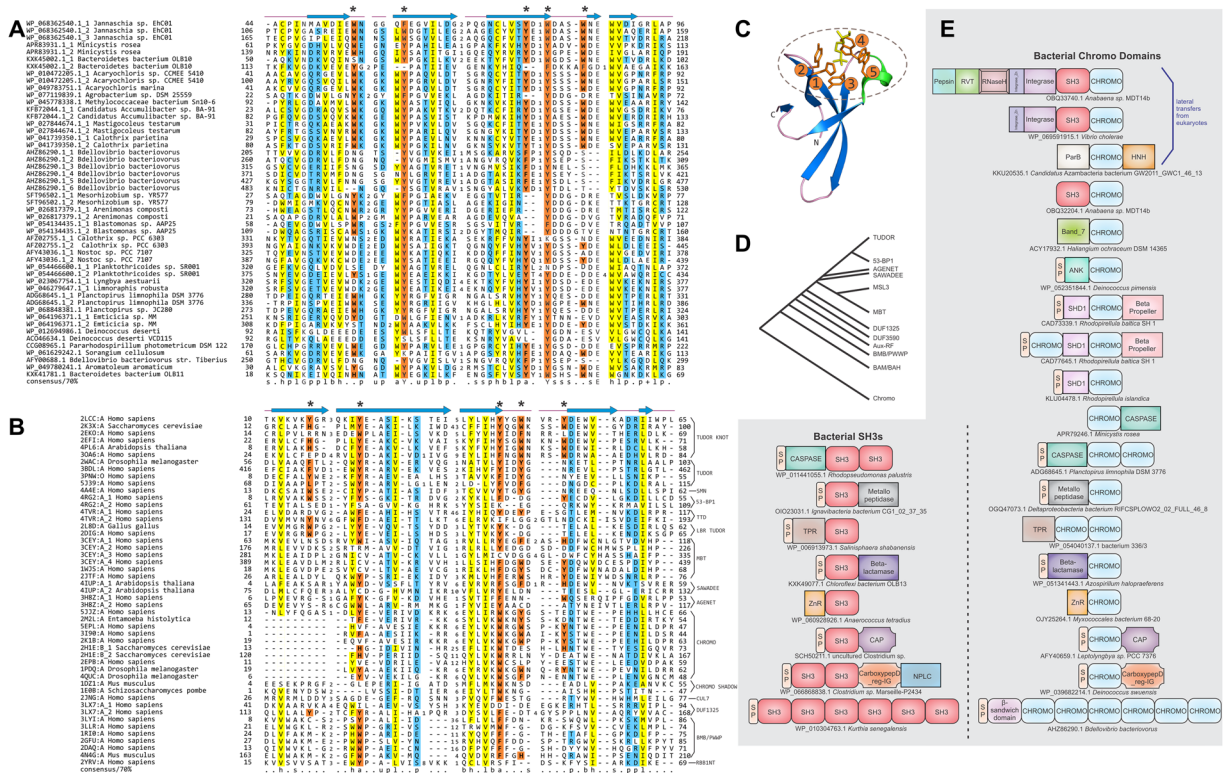
share a common DNA-compactation function in crenarchaeal chromatin<sup>41,42,58</sup>, which is mediated via a similar protein-DNA interface, with positionally- and chemically- equivalent residues involved in DNA contact. (3) Experiments to make the Cren7 protein more “Sul7d-like” by mutating the loop between strands  $\beta 3/\beta 4$  results in an exact superposition of the DNA-binding interface<sup>41,42,58</sup>.

Presence of Cren7/Sul7 proteins only in crenarchaeota and bathyarchaeota, together with their provenance from ZnRs, which we establish above supports the following evolutionary scenario: the ancestral Cren7 proteins likely arose from a DNA-binding ZnR of which several are found in the pan-archaeo-eukaryotic transcription apparatus<sup>46</sup>. Consistent with this, we have found versions of Cren7, which still retain the ancestral Zn-chelating residues. The strengthening of the hydrophobic core due to interactions in the strand regions appear to have facilitated the loss of Zn-chelation in more than one version of the family. This appears to have concomitantly supported the emergence of a more  $\beta$ -barrel-like geometry that converged to a SH3-like state. This was followed by the emergence of Sul7 only in Sulfolobaceae as a specialized DNA-packaging protein from a Cren7-like protein that had already lost its Zn-chelating residues.

Our recovery of novel members of the Cren7-like family suggests that they underwent functional diversification in the archaeal lineages that possess them. Notably, in one group of these proteins, Cren7 is the C-terminal domain of much larger protein (e.g. OLD03897.1) where it is combined with an N-terminal CdvA-like coiled coil domain and a central small 3-stranded domain. In crenarchaea, the CdvA-like coiled-coil domain proteins are components of the cell-division system and form filamentous double-helical complexes with DNA<sup>65</sup>. In bathyarchaea, we found a fusion of the Cren7 domain to an N-terminal FtsZ domain (KYH38952.1), a key component of the cell-division apparatus related to the tubulin-like cytoskeletal proteins. These architectures suggest that representatives of the Cren7 family play a specific role in cell-division, probably in anchoring the DNA. All Cren7Znr proteins contain a conserved C-terminal motif with an absolutely conserved acidic residue beyond the core Cren7 domain, which is likely to adopt an extended conformation. It is possible that this region also helps in specific interactions with cell-division components. We also found an instance of bathyarchaeal Cren7 (KPV62179.1) fused to the URI domain (GIY-YIG endonuclease) and an uncharacterized enzymatic domain of the OrfY-like superfamily<sup>66</sup>, and a nanoarchaeal Cren7 (AMD29662.1) with a C-terminal REase fold nuclease domain (Fig. 2B; Supplementary Material S1, S2).

**Bacterial diversification of chromo-like SH3 fold domains.** The above rooting of the provenance of Cren7/Sul7-like proteins within the archaeal radiation of ZnRs and evidence for convergent acquisition of the SH3-like barrel morphology questions the evolutionary relationship of these archaeal chromosomal proteins and the chromo-like superfamily of SH3 fold, given the previous identification of bacterial versions of the latter<sup>38</sup>. To better understand the origins of the bacterial chromo-like domains, we ran iterative profile searches from the versions we had previously detected in bacteria<sup>38</sup>. Our current searches, greatly extended the phyletic spread of bacterial chromo-like domains, retrieving them from several different lineages such as proteobacteria (mainly  $\alpha$  and  $\delta$ ), cyanobacteria, Thermus/Deinococcus, bacteroidetes, planctomycetes, and spirochaetes and in rare cases in euryarchaea. Firmicutes and actinobacteria showed a strong under-representation of this domain (see Supplementary Material S3). A multiple sequence alignment of the bacterial chromo-like domains with various eukaryotic versions revealed that the bacterial homologs strongly conserve at least a subset of the aromatic residues corresponding to the aromatic cage involved in ligand-binding (Fig. 3A,B)<sup>38,67,68</sup>. Given that these residues are central to the recognition of methylated  $\epsilon$ -amino groups of lysine in the bound peptide in eukaryotic chromo-like domains (Fig. 3C), we posit a similar binding capacity for the bacterial versions. Across eukaryotic chromo-like domains, the Tudor assemblage strongly conserves the residues forming the aromatic cage, whereas the classical chromo domains show some variability in these residues (Fig. 4). This suggests that the bacterial versions are closer to the Tudor-like chromo domains and that the Tudor-like versions are likely to be closer to the ancestral mode of binding peptides. The plausible evolutionary relationship shared by the various chromo-like domains is depicted in Fig. 3D.

We observed that the bacterial chromo-like domains, in contrast to their eukaryotic counterparts, are consistently present in secreted proteins across the diverse lineages containing them. Some proteins with bacterial chromo-like domains also display cysteines likely to form disulphide bonds typical of extracellular proteins. Further, systematic analysis of their domain architectures revealed contexts unlike any seen in eukaryotes (Fig. 3E): in addition to being found in tandem repeats (2–6 domains per protein), one of the most common linkages of the secreted bacterial chromo-like domains is with the caspase-like peptidase domain. This architecture is found in planctomycetes, cyanobacteria, bacteroidetes and chloroflexi. Notably, less-frequent but parallel fusions are also observed with other extracellular peptide-bond hydrolase domains namely a zincin-like metallopeptidase and a  $\beta$ -lactamase domain (Fig. 3E). Additionally, these chromo-like domains are also combined in extracellular proteins with several other non-catalytic domains, such as another SH3-fold domain the Slap homology domain 1 (SHD1), WD40-like  $\beta$ -propeller domains, the EF-hand, the Ig-fold carboxypeptidase regulatory domain and TPR repeats. Interestingly, parallel domain architectures with multiple tandem domains and fusions to the caspase, zincin-like metallopeptidase,  $\beta$ -lactamase, WD40  $\beta$ -propellers, the carboxypeptidase regulatory and TPR repeat domains are seen for bacterial representatives of classical SH3 domains (Fig. 3B)<sup>21,23</sup>, again in contrast to their strictly intracellular location of the eukaryotic counterparts<sup>19</sup>. These bacterial SH3 domain proteins are distributed across a much wider phyletic range of taxa compared to the chromo-like domains. Recent studies have shown that secreted bacterial SH3 domains are likely involved in binding peptides in the peptidoglycan cell wall<sup>21–23</sup>. These parallels suggest that the bacterial chromo-like domains might function similarly to the bacterial SH3 in binding-specific peptides. However, they are likely to possess specificity for those containing methylated lysine-like moieties, in the murein or extracellular matrix peptides in bacteria with Gram-negative cell walls (given their near absence in firmicutes and actinobacteria).

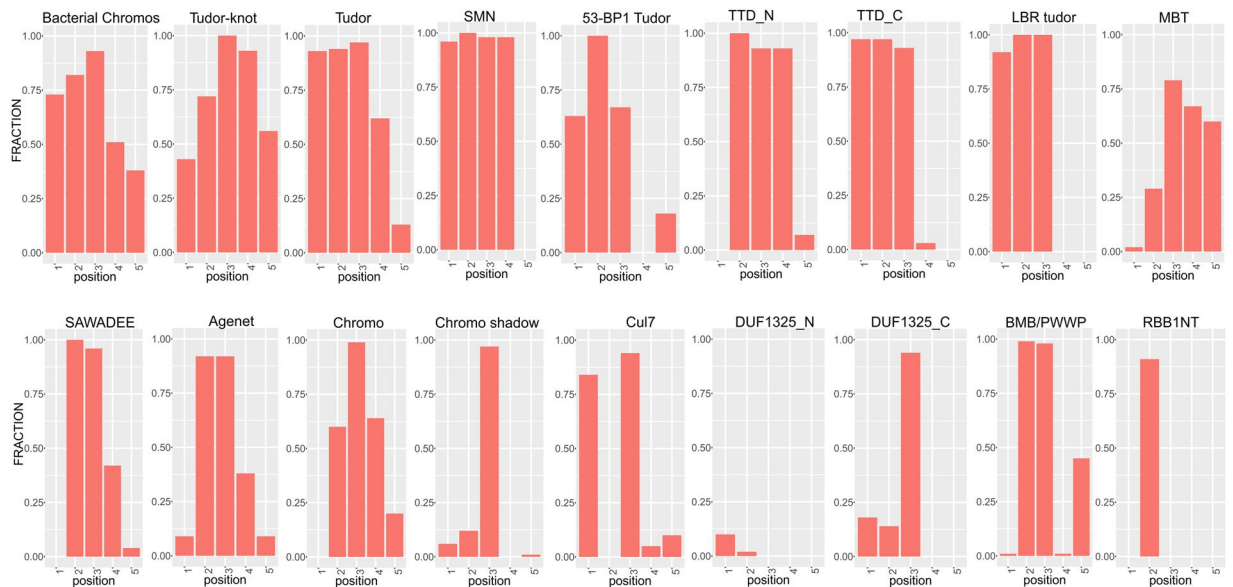


**Figure 3.** Multiple sequence alignment (MSA), probable phylogeny and domain architectures of chromo-like domains. **(A)** MSA of representative bacterial chromo-like domains. **(B)** Structure-based MSA of representative eukaryotic chromo-like domains. Coloring scheme for panel (A) and (B) follows Fig. 2(A). The amino acids at positions that are likely to be involved in forming the peptide-binding aromatic cage are highlighted in orange and marked with the asterisk above the MSA. **(C)** View of the peptide-binding aromatic cage of *Arabidopsis* Morf Related Gene (MRG) protein MRG2 chromo domain (PDB: 4PL6). The aromatic cage is marked by a dashed ellipse and the aromatic residue positions from the N- to C- terminus are marked by numbers 1–5. **(D)** Probable phylogenetic relationships between the various chromo-like domains. **(E)** Domain architectures of bacterial chromo domains. Parallel architectures among bacterial chromo and bacterial SH3 domains are grouped at the bottom. SP indicates signal peptide. Besides the architectures shown in the figure, there are various other bacteria proteins with 2 to 14 tandem SH3 domains and 2 to 6 tandem chromo domains. (see Supplementary Material S3 for details).

In addition to the above architectures, we also observed two unusual lateral transfers of the eukaryotic chromo domains to bacteria: (1) a Ty-3 family retrotransposon, which is commonly found in fungi is also found in multiple copies in *Anabaena* (for example, Accession no. OBQ33740.1, from *Anabaena* sp. MDT14b). Here the chromo domain is fused C-terminal to the polyprotein containing pepsin, reverse transcriptase, RNase H, integrase and SH3 domains. Given the association of certain fungi with cyanobacteria (e.g. example, in cyanolichens) this probably represents a transfer facilitated by such an association. (2) The other case is found in a single species, where a eukaryotic chromo domain is inserted into a mobile element of a bacterium with a ParB and HNH domains (accession: KKU20535.1, *Azambacteria bacterium* GW2011\_GWC1\_46\_13) (see Supplementary Material S3).

**Conclusions**

Small metal-chelating domains, such as ZnRs, likely originated with a relatively simple stabilizing core in the form of the Zn-chelating center. The ancestral ZnRs themselves could have emerged from a pair of small metal-stabilized, bi-cysteine, knuckle-like motifs that existed independently (similar to the minimal versions seen, for example, in Rad50 zinc-hook, PDB: 1L8D and the RNase E zinc-link domain, PDB: 2VMK). More structured ZnRs with  $\beta$ -hairpins developing around these Zn-knuckles at their turns (Fig. 1A–F) likely evolved from such versions and acquired the ability to exist as independent domains. These simplest versions of ZnRs might have resembled the type-1A scaffold as they have separate  $\beta$ -hairpins with the two metal chelating residue pairs (Fig. 1A). This provided a platform for considerable evolutionary variability and structural innovation with augmentation and/or supplanting of the original Zn-center via the emergence of further stabilizing hydrogen-bonding networks and hydrophobic cores in the form of new, more ordered secondary structure elements<sup>3,69–71</sup>. Such developments are seen in the form of the type-1B ZnRs, which developed an additional  $\beta$ -strand and finally the type-2 ZnRs, which appear to be related to the type-1B versions via a circular permutation. Here, the additional  $\beta$ -strand appears to have been further incorporated into the core of the fold to form a contiguous  $\beta$ -sheet (Fig. 1C–F). Indeed, loss of metal-ion chelation in such domains could accompany the development of alternative stabilizing hydrophobic cores, which might then be the progenitor of a distinct protein fold<sup>150,72,73</sup>. This could then be subject to further modifications via duplications and/or circular permutations<sup>45,74–77</sup>.



**Figure 4.** Comparative graphical view of the conservation of aromatic residues at the five positions that make up the peptide-binding aromatic cage in chromo-like domains. Individual graphs are labelled on the top by the family they represent. The X-axis represents positions corresponding to those shown in Fig. 3(C). The Y-axis represents the fraction of aromatic amino acids (Y, W, F, H) at those positions in the respective families.

In this study, we capture the evolutionary stages in one such transformation: the Cren7 domain, previously considered a SH3 fold  $\beta$ -barrel, actually emerged from a metal-binding ZnR domain. We find evidence that such convergence between ZnRs and SH3 fold domains might have happened independently more than once. For example, the segment-swapped Ku-bridge domain and its homolog: the C-terminal all- $\beta$  domain of the MarR-like transcription factors resemble segment-swapped SH3-like folds (labelled as SH3-like in SCOP and SCOP2; SCOP identifier 140307); however, sequence analysis clearly indicated the provenance of these domains from type-2 ZnRs<sup>50,51</sup>. We also observed that the type-2 ZnR at C-terminus of the nicotinate phosphoribosyltransferase converged to a SH3-like fold with the Zn-chelating sites lost alongside the evolution of compensatory hydrophobic interactions (Supplementary Figure S1). Further, this convergence can also extend to the substrate-binding mode of certain ZnRs and the SH3 fold domains (Fig. 1J–M). This raises the possibility that similar transitions from a ZnR might have spawned SH3-like fold domains on other occasions too but cannot be currently confirmed using sequence-based methods. Notably, both ZnRs and SH3 fold domains are found in proteins highly conserved across life, such as the ribosomal subunits<sup>78</sup>. Hence, it is possible that ancient representatives of the two folds also share an ultimately divergent relationship in a very early period of the evolution of protein universe prior to the last universal common ancestor (LUCA), with the SH3 fold emerging via loss of Zn-chelation from a ZnR via acquisition of stronger hydrophobic interactions.

Our findings help clarify the origins and new functions of key domains in chromosomal proteins. Compaction of genomic DNA into the limited intracellular space is a universal challenge for which multiple solutions have been selected across cellular life. In asgardarchaea (believed by many to be the closest sister group of eukaryotes), eukaryotes and euryarchaea this function is carried out by the  $\alpha$ -helical histone fold proteins and in bacteria by the HU/IHF superfamily of proteins<sup>79–82</sup>. However, in crenarchaeaota and at least certain bathyarchaeaota, the ancestral histones appear to have been displaced by a heterogeneity of chromatin-compacting proteins, namely Cren7, Sul7d and CC1<sup>83</sup>. Based on structural considerations it was earlier suggested that these archaeal chromosomal proteins might have an evolutionary relationship to the SH3-fold domains of the chromo-like superfamily, which are a hallmark of eukaryotic chromatin proteins<sup>40</sup>. Here, we present a clear evolutionary scenario for the convergent origin of a SH3-like morphology for Cren7/Sul7d-like proteins from ZnRs. We also present evidence for a functional diversification of these ZnRs in crenarchaeaota and bathyarchaeaota with potential roles in cell-division comparable to the CdvA-like proteins with the PRC-barrel domain<sup>65,84</sup>.

On the other hand, we present evidence that eukaryotic chromo-like domains have no close relationship to Cren7-like archaeal chromosomal proteins. While there are several SH3 fold  $\beta$ -barrels, eukaryotic chromo-like domains specifically share a peptide-binding function with the classical SH3 domains. We show that both these superfamilies are present in bacterial extracellular proteins and based on the evidence from the bacterial SH3 domains, we posit that the bacterial chromo-like domains too bind peptides in the peptidoglycan or in the periplasm. Further, sequence analysis suggests the bacterial chromo-like domains likely acquired specificity for methylated side-chains of basic amino acids even in these bacterial versions. Further, they might have even interacted with nucleic acids found in bacterial extracellular matrices. These observations suggest that the two SH3 fold superfamilies initially diverged and radiated primarily in the context of binding different peptides in the bacterial peptidoglycan. Notably, both these show parallel domain architectures (Fig. 3E) and are often coupled with peptidase domains, which might play a role in the degradation of proteins or peptides found in bacterial extracellular matrices. Thus, we predict that both domains might help anchor enzymes regulating the dynamics of the cell-wall or the extracellular matrices as part of cell-cell interactions in course of colony formation or conflicts with other bacteria.

These observations have important implications for the origin of eukaryotes. Bacterial chromo-like domains are found only in certain bacterial lineages, including alphaproteobacteria, unlike the bacterial SH3 domain which is more widely distributed. However, chromo-like domains are rare or absent in archaea (including asgard-archaeota). In contrast, we can trace at least 3 paralogous families of chromo-like domains in the LUCA<sup>38</sup>. Thus, the chromo-like superfamily was present in the stem eukaryote diversified prior to the common ancestor of all extant eukaryotes. This suggests that the eukaryotes possibly acquired their chromo-like and classical SH3 domains from the  $\alpha$ -proteobacterial mitochondrial progenitor. Their presence in the extracellular matrix might have facilitated interactions with “cytoplasmic” proteins and nucleic acids of archaeal component of the ancestral eukaryote upon endosymbiosis. Hence, we posit that this was the likely scenario that favoured their recruitment as intracellular peptide-binding domains in the ancestral eukaryote. In eukaryotes, the chromo-like and classical SH3 domains radiated extensively due to the “opening” of entirely new niches in the form of peptide-substrates from histone tails, positively charged RNA-binding complexes and cytoskeletal proteins respectively. This appears to have gone hand-in-hand with the radiation of a diverse suite of enzymes covalently-modifying histones and other proteins<sup>38</sup>. We show that the bacterial chromo-like domains already possessed the capacity to bind peptides through the aromatic cage in the open mouth of the  $\beta$ -barrel; thus, they were pre-adapted to binding methylated peptides in the eukaryotic chromatin niche.

We believe the Cren7-like and chromo-like domains identified in this study might help further experimental characterization of the biochemical and biological diversity of these domains.

## Methods

The PSI-BLAST and JACKHMMER programs were used for iterative sequence profile searches against the National Center for Biotechnology Information (NCBI) non-redundant (NR) and locally generated databases (e.g. nr.50: NR sequences clustered at 50% sequence identity)<sup>61,85</sup>. Additional sequence similarity searches were performed using the HHpred program<sup>63</sup> (against: PDB70\_12Feb17, SCOP95\_v1.75B and PfamA\_31.0, using MSA generation method HHblits run for 5 iterations, E-value threshold of 0.001). Multiple sequence alignments were constructed using Kalign<sup>86</sup> followed by manual corrections based on structural alignments. Sequence similarity-based clustering was performed using the BLASTCLUST program (<http://ftp.ncbi.nih.gov/blast/documents/blastclust.html>), by adjusting the length (L) and score (S) parameters based on need. Automated structure similarity searches were performed using the DALI server<sup>52</sup>. Structures were compared and superimposed in the molecular visualization program PyMOL by manually defining equivalent regions using the pair fitting wizard. Automated pairwise structure superimposition was performed using TM-align and Fr-TM-align tools<sup>59,60</sup>.

Domain architectures and other contextual information about the protein sequences were generated using both Pfam and a custom set of profiles. This analysis was automated using scripts in PERL. Graphs in Fig. 4 were generated using multiple sequence alignments (MSAs) of representative sequences for each family extracted from a database with NR sequences clustered down to 90% identity. The amino acid conservation at the five positions involved in the formation of aromatic cage was extracted from the text version of the multiple sequence alignments generated by us (bacterial chromo domains) or extracted from the Pfam database (previously known clades). Graphs were built using the ggplot2 package in R<sup>87</sup>.

The tree of relationships between different clades of chromo-like domains was generated thus: given that these domains are small in size and show rapid divergence they are not amenable to analysis by conventional phylogenetic methods. Hence, we established the relationships between them using two methods, namely the e-values reported by the profile-profile comparisons with the HHpred program<sup>63</sup> and the Z-scores using the DALI program. These searches were respectively run using representative sequences or structures for each of the clades against a library of HMM profiles developed from alignments in the Pfam database or structures in PDB. The e-values and Z-scores were recorded for the hits to all other clades. The clades were then clustered using single linkage clustering based on these values and the consensus clustering is rendered as a tree in Fig. 3D.

## References

1. Murzin, A. G. How far divergent evolution goes in proteins. *Current opinion in structural biology* **8**, 380–387 (1998).
2. Schwede, T. & Peitsch, M. C. *Computational Structural Biology: Methods and Applications*. (World Scientific Publishing Company, 2008).
3. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology* **247**, 536–540 (1995).
4. Swindells, M. B., Orengo, C. A., Jones, D. T., Hutchinson, E. G. & Thornton, J. M. Contemporary approaches to protein structure classification. *BioEssays: news and reviews in molecular, cellular and developmental biology* **20**, 884–891 (1998).
5. Orengo, C. A., Sillitoe, I., Reeves, G. & Pearl, F. M. Review: what can structural classifications reveal about protein evolution? *Journal of structural biology* **134**, 145–165 (2001).
6. Doolittle, R. F. Convergent evolution: the need to be explicit. *Trends Biochem Sci* **19**, 15–18 (1994).
7. Krishna, S. S. & Grishin, N. V. Structurally analogous proteins do exist! *Structure* **12**, 1125–1127 (2004).
8. Lupas, A. N., Ponting, C. P. & Russell, R. B. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *Journal of structural biology* **134**, 191–203 (2001).
9. Zhang, D., Iyer, L. M., Burroughs, A. M. & Aravind, L. Resilience of biochemical activity in protein domains in the face of structural divergence. *Current opinion in structural biology* **26**, 92–103 (2014).
10. Grishin, N. V. C-terminal domains of *Escherichia coli* topoisomerase I belong to the zinc-ribbon superfamily. *Journal of molecular biology* **299**, 1165–1177 (2000).
11. Grishin, N. V. Fold change in evolution of protein structures. *Journal of structural biology* **134**, 167–185 (2001).
12. Murzin, A. G. Biochemistry. Metamorphic proteins. *Science* **320**, 1725–1726 (2008).
13. Andreeva, A. Classification of proteins: available structural space for molecular modeling. *Methods Mol Biol* **857**, 1–31 (2012).
14. Grishin, N. V. KH domain: one motif, two folds. *Nucleic acids research* **29**, 638–643 (2001).
15. Alva, V., Koretke, K. K., Coles, M. & Lupas, A. N. Cradle-loop barrels and the concept of meta-folds in protein classification by natural descent. *Current opinion in structural biology* **18**, 358–365 (2008).



16. Roessler, C. G. *et al.* Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity but different folds. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 2343–2348 (2008).
17. Anantharaman, V. & Aravind, L. The SHS2 module is a common structural theme in functionally diverse protein groups, like Rpb7p, FtsA, GyrI, and MTH1598/TM1083 superfamilies. *Proteins* **56**, 795–807 (2004).
18. Kuriyan, J. & Cowburn, D. Structures of SH2 and SH3 domains: Current opinion in structural biology 1993, 3:828–837. *Current opinion in structural biology* **3**, 828–837 (1993).
19. Pawson, T. Protein modules and signalling networks. *Nature* **373**, 573–580 (1995).
20. Cohen, G. B., Ren, R. & Baltimore, D. Modular binding domains in signal transduction proteins. *Cell* **80**, 237–248 (1995).
21. Anantharaman, V. & Aravind, L. Evolutionary history, structural features and biochemical diversity of the NlpC/P60 superfamily of enzymes. *Genome biology* **4**, R11 (2003).
22. Xu, Q. *et al.* Insights into Substrate Specificity of NlpC/P60 Cell Wall Hydrolases Containing Bacterial SH3 Domains. *mBio* **6**, e02327–02314 (2015).
23. Ponting, C. P., Aravind, L., Schultz, J., Bork, P. & Koonin, E. V. Eukaryotic signalling domain homologues in archaea and bacteria. *Ancient ancestry and horizontal gene transfer. Journal of molecular biology* **289**, 729–745 (1999).
24. Kishan, K. V. & Agrawal, V. SH3-like fold proteins are structurally conserved and functionally divergent. *Current protein & peptide science* **6**, 143–150 (2005).
25. Dalgarno, D. C., Botfield, M. C. & Rickles, R. J. SH3 domains and drug design: ligands, structure, and biological function. *Biopolymers* **43**, 383–400 (1997).
26. Lingel, A., Simon, B., Izaurralde, E. & Sattler, M. Structure and nucleic-acid binding of the Drosophila Argonaute 2 PAZ domain. *Nature* **426**, 465–469 (2003).
27. Yan, K. S. *et al.* Structure and conserved RNA binding of the PAZ domain. *Nature* **426**, 468–474 (2003).
28. Burroughs, A. M., Ando, Y. & Aravind, L. New perspectives on the diversification of the RNA interference system: insights from comparative genomics and small RNA sequencing. *Wiley interdisciplinary reviews. RNA* **5**, 141–181 (2014).
29. Subramanian, G., Koonin, E. V. & Aravind, L. Comparative genome analysis of the pathogenic spirochetes *Borrelia burgdorferi* and *Treponema pallidum*. *Infection and immunity* **68**, 1633–1648 (2000).
30. Nicolas, F. J., Cayuela, M. L., Martinez-Argudo, I. M., Ruiz-Vazquez, R. M. & Murillo, F. J. High mobility group I(Y)-like DNA-binding domains on a bacterial transcription factor. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 6881–6885 (1996).
31. Bouazoune, K. *et al.* The dMi-2 chromodomains are DNA binding modules important for ATP-dependent nucleosome mobilization. *The EMBO journal* **21**, 2430–2440 (2002).
32. Gong, W., Wang, J., Perrett, S. & Feng, Y. Retinoblastoma-binding protein 1 has an interdigitated double Tudor domain with DNA binding activity. *The Journal of biological chemistry* **289**, 4882–4895 (2014).
33. Charier, G. *et al.* The Tudor tandem of 53BP1: a new structural motif involved in DNA and RG-rich peptide binding. *Structure* **12**, 1551–1562 (2004).
34. Koonin, E. V., Zhou, S. & Lucchesi, J. C. The chromo superfamily: new members, duplication of the chromo domain and possible role in delivering transcription regulators to chromatin. *Nucleic acids research* **23**, 4229–4233 (1995).
35. Jones, D. O., Cowell, I. G. & Singh, P. B. Mammalian chromodomain proteins: their role in genome organisation and expression. *BioEssays: news and reviews in molecular, cellular and developmental biology* **22**, 124–137 (2000).
36. Iyer, L. M., Anantharaman, V., Wolf, M. Y. & Aravind, L. Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. *International journal for parasitology* **38**, 1–31 (2008).
37. Maurer-Stroh, S. *et al.* The Tudor domain ‘Royal Family’: Tudor, plant Agenet, Chromo, PWWP and MBT domains. *Trends Biochem Sci* **28**, 69–74 (2003).
38. Aravind, L., Abhiman, S. & Iyer, L. M. Natural history of the eukaryotic chromatin protein methylation system. *Progress in molecular biology and translational science* **101**, 105–176 (2011).
39. Xu, C., Cui, G., Botuyan, M. V. & Mer, G. In Histone Recognition. (ed. M.-M. Zhou) 49–82 (Springer International Publishing, Cham; 2015).
40. Ball, L. J. *et al.* Structure of the chromatin binding (chromo) domain from mouse modifier protein 1. *The EMBO journal* **16**, 2473–2481 (1997).
41. Guo, L. *et al.* Biochemical and structural characterization of Cren7, a novel chromatin protein conserved among Crenarchaea. *Nucleic acids research* **36**, 1129–1137 (2008).
42. Zhang, Z., Gong, Y., Guo, L., Jiang, T. & Huang, L. Structural insights into the interaction of the crenarchaeal chromatin protein Cren7 with DNA. *Molecular microbiology* **76**, 749–759 (2010).
43. Klug, A. The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annual review of biochemistry* **79**, 213–231 (2010).
44. Grishin, N. V. Treble clef finger—a functionally diverse zinc-binding structural motif. *Nucleic acids research* **29**, 1703–1714 (2001).
45. Krishna, S. S., Majumdar, I. & Grishin, N. V. Structural classification of zinc fingers: survey and summary. *Nucleic acids research* **31**, 532–550 (2003).
46. Aravind, L. & Koonin, E. V. DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic acids research* **27**, 4658–4670 (1999).
47. Chen, H. T., Legault, P., Glushka, J., Omichinski, J. G. & Scott, R. A. Structure of a (Cys3His) zinc ribbon, a ubiquitous motif in archaeal and eucaryal transcription. *Protein science: a publication of the Protein Society* **9**, 1743–1752 (2000).
48. Kaur, G. & Subramanian, S. The insertion domain 1 of class IIA dimeric glycyl-tRNA synthetase is a rubredoxin-like zinc ribbon. *Journal of structural biology* **190**, 38–46 (2015).
49. Kaur, G. & Subramanian, S. Evolutionary analysis of a novel zinc ribbon in the N-terminal region of threonine synthase. *Cell Cycle* **16**, 1918–1926 (2017).
50. Krishna, S. S. & Aravind, L. The bridge-region of the Ku superfamily is an atypical zinc ribbon domain. *Journal of structural biology* **172**, 294–299 (2010).
51. Kaur, G. & Subramanian, S. The Ku–Mar zinc finger: A segment-swapped zinc ribbon in MarR-like transcription regulators related to the Ku bridge. *Journal of structural biology* (2015).
52. Holm, L. & Sander, C. Dali: a network tool for protein structure comparison. *Trends Biochem Sci* **20**, 478–480 (1995).
53. Awrey, D. E. *et al.* Yeast transcript elongation factor (TFIIS), structure and function. II: RNA polymerase binding, transcript cleavage, and read-through. *The Journal of biological chemistry* **273**, 22595–22605 (1998).
54. Olmsted, V. K. *et al.* Yeast transcript elongation factor (TFIIS), structure and function. I: NMR structural analysis of the minimal transcriptionally active region. *The Journal of biological chemistry* **273**, 22589–22594 (1998).
55. Amero, C. D., Boomershine, W. P., Xu, Y. & Foster, M. Solution structure of *Pyrococcus furiosus* RPP21, a component of the archaeal RNase P holoenzyme, and interactions with its RPP29 protein partner. *Biochemistry* **47**, 11704–11710 (2008).
56. Cheng, H., Liao, Y., Schaeffer, R. D. & Grishin, N. V. Manual classification strategies in the ECOD database. *Proteins* **83**, 1238–1251 (2015).
57. Robinson, H. *et al.* The hyperthermophile chromosomal protein Sac7d sharply kinks DNA. *Nature* **392**, 202–205 (1998).
58. Zhang, Z., Gong, Y., Chen, Y., Li, H. & Huang, L. Insights into the interaction between Cren7 and DNA: the role of loop beta3-beta4. *Extremophiles: life under extreme conditions* (2015).

59. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **33**, 2302–2309 (2005).
60. Pandit, S. B. & Skolnick, J. Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. *BMC bioinformatics* **9**, 531 (2008).
61. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389–3402 (1997).
62. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic acids research* **39**, 18 (2011).
63. Soding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research* **33**, W244–248 (2005).
64. Evans, P. N. *et al.* Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* **350**, 434–438 (2015).
65. Moriscot, C. *et al.* Crenarchaeal CdvA forms double-helical filaments containing DNA and interacts with ESCRT-III-like CdvB. *PLoS one* **6**, e21921 (2011).
66. Kryshchuk, A. *et al.* Challenging the state of the art in protein structure prediction: Highlights of experimental target structures for the 10th Critical Assessment of Techniques for Protein Structure Prediction Experiment CASP10. *Proteins* **82**(Suppl 2), 26–42 (2014).
67. Nielsen, P. R. *et al.* Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. *Nature* **416**, 103–107 (2002).
68. Kim, D. *et al.* Corecognition of DNA and a methylated histone tail by the MSL3 chromodomain. *Nature structural & molecular biology* **17**, 1027–1029 (2010).
69. Salgado, E. N., Radford, R. J. & Tezcan, F. A. Metal-directed protein self-assembly. *Accounts of chemical research* **43**, 661–672 (2010).
70. Arnold, F. H. & Zhang, J. H. Metal-mediated protein stabilization. *Trends in biotechnology* **12**, 189–192 (1994).
71. Aravind, L., Iyer, L. M. & Koonin, E. V. Comparative genomics and structural biology of the molecular innovations of eukaryotes. *Current opinion in structural biology* **16**, 409–419 (2006).
72. Aravind, L. & Koonin, E. V. The U box is a modified RING finger - a common domain in ubiquitination. *Current biology: CB* **10**, R132–134 (2000).
73. Kaur, G. & Subramanian, S. Repurposing TRASH: emergence of the enzyme organomercurial lyase from a non-catalytic zinc finger scaffold. *Journal of structural biology* **188**, 16–21 (2014).
74. Burroughs, A. M., Iyer, L. M. & Aravind, L. Functional diversification of the RING finger and other binuclear treble clef domains in prokaryotes and the early evolution of the ubiquitin system. *Molecular bioSystems* **7**, 2261–2277 (2011).
75. Kaur, G. & Subramanian, S. Evolutionary relationship between the cysteine and histidine rich domains (CHORDs) and Btk-type zinc fingers. *Bioinformatics*, bty041–bty041 (2018).
76. Kaur, G. & Subramanian, S. The UBR-box and its relationship to binuclear RING-like treble clef zinc fingers. *Biology direct* **10**, 36 (2015).
77. Kaur, G. & Subramanian, S. Classification of the treble clef zinc finger: noteworthy lessons for structure and function evolution. *Scientific reports* **6**, 32070 (2016).
78. Anantharaman, V., Koonin, E. V. & Aravind, L. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic acids research* **30**, 1427–1464 (2002).
79. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
80. Dillon, S. C. & Dorman, C. J. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nature reviews. Microbiology* **8**, 185–195 (2010).
81. Reeve, J. N. *et al.* Archaeal histones: structures, stability and DNA binding. *Biochemical Society transactions* **32**, 227–230 (2004).
82. Burroughs, A. M., Kaur, G., Zhang, D. & Aravind, L. Novel clades of the HU/IHF superfamily point to unexpected roles in the eukaryotic centrosome, chromosome partitioning, and biologic conflicts. *Cell Cycle* **16**, 1093–1103 (2017).
83. Zhang, Z., Guo, L. & Huang, L. Archaeal chromatin proteins. *Science China Life Sciences* **55**, 377–385 (2012).
84. Anantharaman, V. & Aravind, L. The PRC-barrel: a widespread, conserved domain shared by photosynthetic reaction center subunits and proteins of RNA metabolism. *Genome biology* **3**, Research0061 (2002).
85. Finn, R. D. *et al.* HMMER web server: 2015 update. *Nucleic acids research* **43**, W30–38 (2015).
86. Lassmann, T., Frings, O. & Sonnhammer, E. L. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic acids research* **37**, 858–865 (2009).
87. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (Springer-Verlag New York, 2009).

## Acknowledgements

This work was supported by the funds of the Intramural Research Program of the National Library of Medicine, USA (LA, LMI and GK) and the department of Biotechnology, India (BTISNET, GAP001: SS). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

## Author Contributions

S.S. and L.A. conceived the idea. G.K., L.M.I., S.S. and L.A. performed the research and wrote the paper.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-24467-z>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018