Using a combination of sequence-, structure- and function-based arguments one can gain insights about the evolution of proteins (Doolittle, 1981; Murzin, 1998; Petsko and Ringe, 2004). The inferred evolutionary connections can further help in deducing structures and functions of homologs, and the forces that shaped the protein world and the biological pathways (Doolittle, 1981; Grishin, 2001a; Koonin and Galperin, 2013; Schwede and Peitsch, 2008).

We have worked towards achieving all the proposed objectives of this study.

Objective-1 was to identify and classify all ZFDs in the PDB which would reflect their structural, functional and evolutionary relationships. We have classified all structures in the PDB which bind zinc for structure-stability and where the overall fold of the zinc-bound protein domain could not be related to other known protein folds. We classified a total of 1953 ZFDs from 5823 PDB structures of 1479 proteins into 104 ZF families. These families could be grouped into eight folds types based on structure. The family level represents grouping based on sequence similarity and thus plausible homology, and the fold group level is based on overall 3D structural similarity. The eight fold groups are the C2H2-like, Gag knuckle, treble clef, zinc ribbon, Zn2/Cys6, TAZ2 domain-like, zinc-binding loops and metallothionein-like. The most highly populated fold groups are the C2H2-like, treble clef and zinc ribbons. We observe an increase of ~10 times in the number of structurally-characterized ZFDs as compared to their previous structural classification (Krishna et al., 2003). The overall framework of the classification and the major fold-types which are stabilized by zinc remain unchanged, although, we could identify several novel members with interesting variations.

Analysis of ZFDs reveals enormous structure, sequence, and functional diversity. The structure variability can be seen not only at the fold-group level, wherein zinc stabilizes different types of SSEs with different topological connections; but also within various fold groups wherein we observe diverse insertions and extensions to the common structural scaffold in different proteins/protein families. The aminoacids involved in chelating zinc are predominantly observed to be Cys and His, although the precise sequence motif and residue-

configuration which chelates zinc is not always absolutely conserved among different ZFD families and sometimes even among members of a given family. The distance among the zinc-binding half-sites and between the two ligands of a single half-site is also seen to vary considerably. The ZFDs are mostly seen to play a structural role in stabilizing the protein fold and mediating interactions with other biomolecules, such as DNA, RNA, proteins and small molecules including lipids, sugars, nucleosides and nucleotides. A small fraction of ZFDs are also seen to possess enzymatic functions.

We have grouped together all ZFDs in the PDB under a common classification schema. Many of these are not presently classified in Pfam and SCOP, and some ZFDs are just annotated by comments/remarks in larger domains in which the former might be inserted. Further, based on evolutionary analysis we have also grouped ZFDs which are presently classified as independent families (or clans) in Pfam, such as, the DNL ZF (PF05180) and CSL ZF zf-CSL (PF05207); the UBR-box (PF02207) and the RING-like ZFs; zf-CGNR (DUF1470, PF11706) and the TRASH-like treble clefs; DUF3222 (PF11519) and the IAP BIR domain; DUF1364 (PF07102) and His-Me family, etc.

Unlike the previous evolutionary classification of ZFs (Krishna et al., 2003), we include the zinc-binding domain of Ada DNA repair protein in our classification, based on our understanding that the zinc-binding site of zinc fingers can be involved in other functions apart from structure-stabilization. We are now aware of many examples where the zinc-chelating residues can have multiple roles, i.e. besides chelating metal for structural-stability, they can also help in mediating functional interactions and can also be reactive as in Ada and PKC and mediate catalysis as in MerB. Besides this example, we have moved the B-box domain from the zinc ribbon fold group to the treble clef fold group based on the currently available structural and sequence data.

Objective-2 was to discover plausible novel ZFDs/proteins which may have lost their zinc binding in the course of evolution and predict their functions through sequence and structure analysis. The novel ZFDs that were found through our analysis include the duplicated and fused treble clefs in the catalytic domain of MerB (Kaur and Subramanian,

2014), zinc ribbon in the ID1 of α_2 glyRS (Kaur and Subramanian, 2015a) and Cren7, the segment-swapped zinc ribbons in MarR proteins PA1607 and PA1374 (Kaur and Subramanian, 2015b) and the RING-like domain in α-COP. In all of these structurally characterized proteins, the zinc-chelating residues are absent and therefore, do not bind zinc. However, homologous sequences of these proteins have conserved metal-chelating residues at positions which recapitulate the pattern seen in bonafide ZFDs and would thus, likely chelate zinc. Based on a combination of structure- and sequence-based arguments we have proposed a likely ZF origin for these protein domains.

Our analysis has further allowed us to draw insights about the plausible functions and/or the evolution of functions in these proteins. For example, we propose a plausible origin of the organometal-binding active site in MerB from the structure-stabilizing zinc-binding site of treble clefs (Kaur and Subramanian, 2014). In the case of zinc ribbon domain of ID1 from α_2 glyRS, our analysis and literature review suggests that Asp146, an essential residue for synthesis of the neurotransmitter Ap4A, is found to be mutated in some patients with hereditary motor neuropathies; thus, plausibly linking the progression of the diseased state to the decrease in the Ap4A on mutation of Asp146 (Kaur and Subramanian, 2015a).

Our analysis helped fetch out a paralogous family of Cren7 which have all zinc chelating residues intact in their sequences and likewise, function as DNA binding proteins (Annexure I). In case of α-COP, we proposed that the RING-like ZFD may be involved in oligomerization of the coat or help tether the coated vesicle on the target membranes, similar to functions being performed by its plausible homologs in other cellular transport pathways. We also proposed that the RING-like domain seen in α-COP and in proteins of other membrane associated complexes, all of which share similar domain architectures, are distantly related in evolution, although they do not retrieve each other in sequence-based searches.

Objective-3 was to understand the evolution of new folds and functions in zinc finger families. Our analysis revealed that the origin of many protein folds proposed to be novel could actually be traced back to ZFs. Of all the examples which we studied, three

phenomenon, *viz.*, duplication and fusion, circular permutation, and loss of zinc binding accompanied by the gain of substitutional stabilizing forces, were seen to be responsible for emergence of novel 3D protein structures. The duplication and fusion of treble clef and zinc ribbons accompanied by other structural modifications had plausibly led to the emergence of the novel catalytic domain in MerB and the fold seen in MAL13P1.257, respectively. Likewise, CHORD has possibly emerged by the duplication and fusion of a Btk-like ZF ancestor.

Our analysis revealed that the UBR-box domain could be related to the RING-like binuclear treble clefs by a circular permutation resulting in a split zinc knuckle, which is likely stabilized by structural extensions that chelate an additional zinc ion. These modifications to the core of the RING-like scaffold possibly aided the emergence of a novel peptide binding site and a novel fold (Kaur and Subramanian, 2015c).

Further, in proteins like Cren7 and MarR proteins PA1607, the loss of zinc binding has led to emergence of a barrel-like topology resembling that of SH3-like folds. Indeed in the case of archaeal chromatin-associated proteins we could suggest a plausible path for the emergence of chromo SH3-like folds from a zinc ribbon-like ancestor with intact metal-chelating configuration.

Our analysis revealed that several ZFs possess enzymatic functions, which was previously thought to be a rare phenomenon. During the first structural classification of ZFDs the only known examples of catalytic ZFs were those of the His-Me family (Krishna et al., 2003). Our analysis has helped compile a list of several other catalytic ZFs. In most of the examples, the catalytic residues are outside the zinc-binding core of the domain. Thus, there is no overlap between the structure-stabilizing center and the functional center. However, in a unique and rare example, i.e. of MerB, we show that the structure-stabilizing site of an ancestral treble clef may evolve into a functional site (Kaur and Subramanian, 2014).

During our study we encountered numerous peculiar examples where there were structural elements corresponding to two folds overlapping in a single protein domain. We referred to these as overlapping domains and were fascinated and puzzled by the evolutionary

phenomenon which might have led to their emergence. These protein domains may be viewed as examples of continuity of protein fold space or we could at best attribute their existence to the *de novo* emergence of zinc-binding sites and/or domain atrophy.

In conclusion, this evolutionary analysis of ZFDs was a timely revisit to the previous structural classification of ZFs (Krishna et al., 2003). It helped gain novel insights about the sequence, structure and function of ZFDs. Many new ZFDs were discovered and the underlying importance of a combination of sequence, structure and function in understanding protein evolution was drawn. ZFDs appear to be one of the ancient protein domains as suggested previously (Burroughs et al., 2011; Grishin, 2001d) and have provided a robust scaffold for the emergence of novel folds and functions.