Data processing and analysis

CD-HIT clustering:

Protein fragments were clustered at 100% identity threshold and representative fragments from each cluster was further considered for the analysis. Depleted sets have more redundant fragments than enriched sets and were reduced to half their size after clustering (Suppl. Fig 1)

	Total.No.Of. fragments	Total.No.of fragment representatives
Pp depleted	141357	76048
Pp enriched	10404	8803
Sc depleted	136531	63657
Sc enriched	11625	6239

Table 1: Total number of fragments in the original data as given and the number of representative fragments after clustering at 100% identity from each sets.

PDB hits:

The representative fragments from each datasets were blasted against PDB database using standalone blast (ncbi-blast-2.6.0+). Three best hits were retained for each fragments. ~ 50% of the fragment representatives have structures with ≥ 30% sequence identities and in that ~ 30% of them are with identities and query coverage ≥ 70% (Suppl. Fig 2A-2D & Suppl. File 1A-1D)

	No.of fragments with hits (≥ 30% ident)	No.of fragments with hits (≥ 70% ident and ≥ 70% Q.cov)
Pp depleted	42294	15688
Pp enriched	4437	1831
Sc depleted	32434	11529
Sc enriched	3001	1148

Table 2: No.of fragments with PDB hits with greater than 30% sequence identities and fragments with more than 70% sequence identities and query coverage are shown.

Secondary structure assignment:

The percentage of secondary structural elements for each fragment with a PDB hit was calculated from its corresponding DSSP coordinates obtained from PDB. Depleted fragments tend to be more rigid with high helical propensity whereas the enriched fragments tend to be more flexible with coil/ unstructured propensity (Suppl. Fig 3 & 4). No significant difference for the beta sheets.

Domain Enrichment:

Representative fragments were scanned for Pfam domains using interproscan-5.24-63.0. Frequency of a particular domain in a dataset was obtained by removing duplicate entries (if a particular domain is present more than once for a particular fragment) in the dataset. Those unique domains were used for further analysis (Suppl. Fig 5).

Some statistics for domain analysis: domain length distribution (Suppl. Fig. 6) fragments with domain repeats (Suppl. Fig. 7), fragments having more than one domains (Suppl. Fig 8) and fragments with at least one domains (Suppl. Fig 9)

Domains which are found in common between the enriched and depleted datasets(for a specific species) were retained for comparison. (Suppl. File. 2 & 3)

Domains which are found in enriched sets but not in depleted (Suppl. File. 4 & 5) and vice versa (Suppl. File. 6 & 7)