**Construction of the dataset**

The dataset was constructed by repeating a similar process as suggested in Holland et al.’s[1] work in several steps as follows:

1. All chains which had the same number of domains as assigned by CATH and SCOP were selected resulting in 88,986 chains.
2. These chains were clubbed into topology groups where each group contains chains having same Class, Architecture & Topology as defined by CATH. A total of 1313 such groups were obtained.
3. Now for each topology group, a **N** domain protein was picked. Where **N** varies from 1 to 4. For single domain proteins, selection is simple and direct. For multi domain proteins, selection was done in such a way so that at least one domain is a unique representative of a topology group.
4. This resulted in a total of 1391 chains with the distribution as shown in table 1.

Table 1: Distribution of chains in the constructed dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **Type of Chain** | **Contiguous** | **Non Contiguous** | **Total** |
| 1-domain | 767 | 0 | 767 |
| 2-domain | 236 | 145 | 381 |
| 3-domain | 86 | 87 | 173 |
| 4-domain | 27 | 43 | 70 |
| Overall | 1117 | 275 | 1391 |

1: Holland TA, Veretnik S, Shindyalov IN, Bourne PE: Partitioning protein structures into domains: why is it so difficult? J Mol Biol 2006, 361:562-590.