**Construction of the dataset**

The dataset used for testing the performance of our algorithm is constructed by following the procedure proposed by Holland *et al* [1] and is described below. First, all PDB chains having the same annotation for the number of domains by CATH and SCOP are considered. This resulted in 88,986 chains. These chains are next clubbed into topology groups such that each group contains chains having same Class, Architecture & Topology as defined by CATH. A total of 1313 such groups are obtained. Now from each topology group, one chain representative of each topology group is selected for grouped in four different classes based on the number of domains, 1, 2, 3 and 4 domains respectively. For single domain proteins, the selection is straightforward, while for multi-domain proteins, care is taken to see that at least one domain is a unique representative of each topology group. 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

|  |  |  |  |
| --- | --- | --- | --- |
| **No. of Domains** | **Contiguous** | **Non-contiguous** | **Total** |
| 1 | 767 | 0 | 767 |
| 2 | 236 | 145 | 381 |
| 3 | 86 | 87 | 173 |
| 4 | 27 | 43 | 70 |
| Total | 1117 | 275 | 1391 |

Reference:

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.