Project Description

Group 12: Deeptha (s210230), Eric (s212514), Jonathan (s212697), Laura (s212775)

2022-04-06

We had the idea that we could recreate and analyze the Protein Data Bank statistics. We discovered that the protein database (RCSB) uses pie charts for data visualization.

We would like to recreate these plots, so they are more useful and only contain the subsets of data that we are interested in. The RCSB statistics contains nucleic acid structures and small molecule structures in addition to protein structures. We want to filter out the small molecules and nucleic acids and perform analyses only on the proteins. Therefore we need to use multiple meta data files found on https://www.wwpdb.org/ftp/pdb-ftp-sites, join the data and tidy it afterwards. On first glance, we are primarily going to work with the *entries.idx*, *source.idx* and *pdb_entry_type.txt*.