A database of high-quality protein residues for reference data, library construction, and motif analysis

Christopher J Williams, David C Richardson, and Jane S Richardson

Biochemistry Dept, Duke University
132 Nanaline Duke Bldg, 3711 DUMC, Durham NC 27710
christopher.j.williams@duke.edu
dcrjsr@kinemage.biochem.duke.edu

The Richardson Lab provides macromolecular structure validation through the MolProbity website, the Phenix crystallography package, and the open-source cctbx_project. Many of our validations rest on statistical expectations of protein behavior derived from datasets of the highest-quality available protein structures.

Here we present our latest dataset of high-quality protein structures - comprising over 16,000 diverse chains from crystallographic structures - and its development and capabilities. Residue-level filtering is an important and underappreciated step in the curation of a reliable dataset. Therefore, we present this dataset not just as a list of high-quality chains, but as the most reliably well-modeled residues within those chains. The prevalence of alternate conformations in high-resolution structures presented a particular challenge for the development of residue filters.

We take advantage of graph-based database software as a natural structure for storing protein information. This allows for fast and intuitive database queries. A graph database is also uniquely suited to modeling sequence relationships, and steric and hydrogen bonding contacts between residues, allowing rapid development of searches for structural motifs of interest, whether for scientific interest or library construction.