Analysis of Primary Citations (References) of PDB Deposits

Joanna Lenkiewicz[1], Michal Gucwa[2], David Cooper[3], Wladek Minor[4]

[1]University of Virginia [2]University of Virginia, [3]University of Virginia, [4]University of Virginia

joanna_len@iwonka.med.virginia.edu

Researchers worldwide from almost every biomedical discipline perform basic searches of the PDB, so the essential information in a PDB deposit must be as informative as possible. On a larger scale, inaccurate or misleading metadata can skew data mining efforts. The title and keywords of PDB deposits may play an essential role in the data mining of the PDB. The primary citation (reference) title may help in such a search, yet many deposits have notable discrepancies between the structure title and the primary reference title. Moreover, we have observed that the fraction of deposits with the status "To be published" has grown in recent years. We also analyze the similarity of titles, the number of citations for various classes of structures, and the primary reference keywords. Finally, the information about crystallization conditions is compared between PDB and the methods section from the primary citation. Several noteworthy examples are presented.