PapC: A web tool for paper clustering associated with PDB files

Daniel Viana, Wesley Paulino Fernandes Maciel, Raquel Melo Minardi

UFMG

Abstract

The literature review is the basis and first step required for any scientific study. Usually, the researcher searches in physical and digital scientific article repositories with the purpose of identifying content related to the researched subject. A search involves from the terms representing entities of interest as well as qualified relationships among them. Manly by the large volume of information available on the internet, this search returns vast data sets which makes the analysis process hard and toilsome. An example of this type of search and what motivated this study is the search for information related to a specific PDB file in a clear and precise way. When a new protein structure is published, it is deposited in the Protein Data Bank, and its structure is made available through the PDB files. With these files, it is possible to identify data related to some a publication. With this information, we can relate, through an API provided by NCBI to work with PubMed, all the papers relevant to the study of the respective structure. Hence, it can recover the related papers and all the articles that cited it. Therefore, this tool contains a database of relevant papers to each structure available at the PDB. The tool was created using Python, PHP, HTML, javascript, and CSS. To create the data visualization, we used the d3.js library. Preliminary results show that through a data visualization in graph form, it is possible to identify clusters of paper that can be related and relevant for a detailed literature review on the subject. Finally, it is also possible to identify collections of papers associated with several PDB ids that may indicate their relevance to the study of structures.

Funding: CAPES