Advanced Searches of the Protein Data Bank using Python in Jupyter Notebooks

Victoria Greever, Anna Rack, Murphy Pick, Lee Schoneman, Paul A. Craig
Rochester Institute of Technology, Rochester NY, USA

Abstract

The Protein Data Bank (PDB) holds an extensive amount of information, and can be a vital tool when performing background research for biochemical work. In an attempt to make the information in the PDB more accessible, the RCSB Search API was employed within Jupyter Notebooks to create more customizable and user-friendly tools with Python code. Areas of focus include searches targeting ligands with specific characteristics, searches for FDA Approved Drugs, as well as sequence searches, used to search for entries based on different sequence characteristics. This code has been built into Jupyter Notebook templates that include examples of these searches as well as annotated code that users can customize to more efficiently run advanced searches on the PDB and download structure and small molecule files returned by the search. These notebooks also walk users through different ways to organize or utilize the returns from advanced searches. Future plans include increasing the amount and type of information available from a search, improved ease of access for visualizing and downloading search results, and expanding the scope of our notebooks to cover more types of searches. This research was supported by NSF-IUSE award number 2142033.

Methods

The Jupyter Notebooks were developed utilizing the rcsbsearchapi library and online PDB documentation to replicate sample searches provided by the PDB. Code was written to replicate example searches and explain the different inputs and outputs for the advanced searches in Jupyter Notebooks. In addition, markdown cells were developed to explain the code, the libraries imported (Table 1)and the changes users can make to customize the output to their own searches rather than the examples searches provided by the PDB. Once the code was created to perform the example searches, further code was developed to demonstrate different data analysis approaches with the search outputs.

Note - the rcsbsearchapi library is being phased into the rcsb-api library, and there are plans to update accordingly.

Library	Abbreviation	Contents	Source
json	N/A	library for working with JavaScript Object Notation for data interchange	json — JSON encoder and decoder
rcsbsearchapi	N/A	library for automated searching of the RCSB Protein Data Bank	py-rcsbsearchapi on GitHub
rcsbattributes	attrs	sublibrary from rcsbsearchapi to specifically make *insert here* searches from the PDB	
AttributeQuery	Attr	sublibrary from rcsbsearchapi to specifically make *insert here* seraches from the PDB	
os	N/A	library for manipulating files within various operating systems	https://docs.python.org/3/library/os.html
sys	N/A	library to deal with system specific parameters and functions	https://docs.python.org/3/library/sys.html
glob	N/A	library to handle specific file manipulations	https://docs.python.org/3/library/glob.html
Bio (Biopython)	N/A	many libraries that are useful for dealing with fasta files and multiple sequence alignments	https://biopython.org/
subprocess	N/A	library that allows you to call executable files from your computer	https://docs.python.org/3/library/subprocess.html
requests	N/A	used to grab information from websites	https://pypi.org/project/requests/
numpy	np	library for a variety of data manipulation	https://numpy.org/
panel	pn	library for graphs and visualizing information	https://panel.holoviz.org/
panel.widgets	pnw	library to build off of panel and allow you to adjust your field of view in Jupyter notebooks	https://panel.holoviz.org/api/panel.widgets.html
bokeh	N/A	library to help with the visualization of everything	https://bokeh.org/

Table 1 - Example documentation explaining the library usage and installlation from a Jupyter Notebook markdown cell

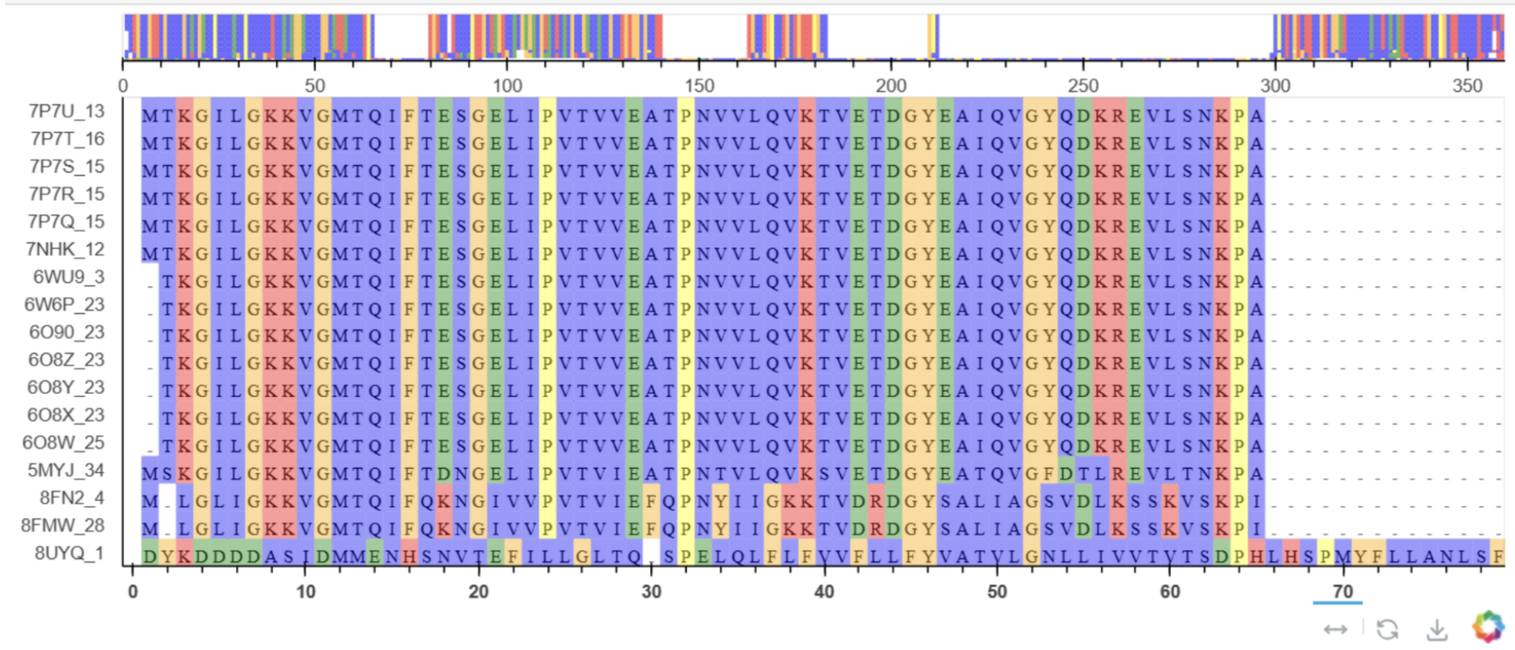

Figure 1 - Typical output from the notebook for a multiple sequence alignment using a Bokeh plot from the example sequence motif search "MTQIF". This example output is utilizing the Zappo color scheme which colors amino acids by their physical and chemical properties. This allows for the comparison of the resulting sequences and their properties.

Results

Code from the sequence alignment Notebook is shown in Figure 2 and can be found in our GitHub repository. They are being compiled into a Jupyter Book. At this point, there are Jupyter Notebooks for nearly every example search provided within the PDB documentation, alongside clear markdown to explain how to perform these searches with Python. There is also code developed that uses the requests library in Python to download the entries from the PDB and create a data folder on a users computer to easily analyze the entires in different file types (both FASTA and mmcif).

For the sequence searches notebook, code was created to allow users to manipulate the FASTA file entires within the PDB for different sequence analysis. Furthermore, code was developed that utilizes MUSCLE (MUltiple Sequence Comparison by Log-Expectation) to perform a multiple sequence alignment on the outputted entries from a specific query. The multiple sequence alignment is then viewed with bokeh panels (Figure 1) and allows for a user to interactively view the alignment under different color schemes. This demonstrates different ways for a user to take their PDB search results and analyze them compare them with computational work.

Future Work

Future work includes the development of a Jupyter Book containing the notebooks for different example searches and search types. For the sequence seach notebook, future work may expand on the available color schemes for the Bokeh panels or add the ability to perform further sequence analysis. In addition, further example searches may be developed to demonstrate different use cases for the Advanced Search function in the PDB.

References

The Protein Data Bank H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne (2000) Nucleic Acids Research, 28: 235-242. https://doi.org/10.1093/nar/28.1.235
Yana Rose, Jose M. Duarte, Robert Lowe, Joan Segura, Chunxiao Bi, Charmi Bhikadiya, Li Chen, Alexander S. Rose, Sebastian Bittrich, Stephen K. Burley, John D. Westbrook. RCSB Protein Data Bank: Architectural Advances Towards Integrated Searching and Efficient Access to Macromolecular Structure Data from the PDB Archive, Journal of Molecular Biology, 2020. DOI: 10.1016/j.jmb.2020.11.003
https://rcsbsearchapi.readthedocs.io
Edgar RC., Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nature Communications 13.1 (2022): 6968.
https://www.nature.com/articles/s41467-022-34630-w.pdf
Edgar RC. and Tolstoy I., Muscle-3D: scalable multiple protein structure alignment (2024) BioRxiv.
https://dmnfarrell.github.io/bioinformatics/bokeh-sequence-aligner
https://github.com/paulcraig/DiscoverBMB2024/blob/main/Sequence_analysis.ipynb