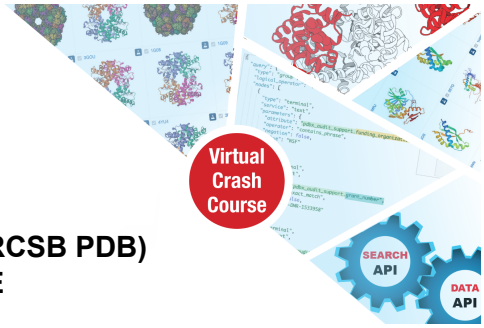


RCSB PROTEIN DATA BANK (RCSB PDB)
API CRASH COURSE

Leveraging RCSB PDB APIs for Bioinformatics Analyses and Machine Learning

Part 2: Hands-on Exercise

Thursday, October 19th, 2023, 16:00-18:30 UTC
and
Thursday, October 27th, 2023, 00:00-02:30 UTC

Announcement: https://www.rcsb.org/news/feature/64f9d026d78e004e766a9699

Agenda with quick links to sections in document (times in UTC):

16:00-16:10 - Introduction

Speaker: Dennis Piehl, Ph.D. - RCSB PDB, Rutgers University
Synopsis: Introduction to Part 2 of the RCSB PDB API virtual crash course.

Relevant material:

- Part 1 recordings:
    - YouTube: https://www.youtube.com/watch?v=ZyCQeentCtw
    - PDB-101: https://pdb101.rcsb.org/train/training-events/api
    - Slides and notes will be made available at both YouTube and PDB-101 soon as well

- Google Colab notebooks:
    - Accessing RCSB PDB APIs via Python: https://colab.research.google.com/drive/1ACwyJx3bROHvBuGrZOChjvczEK5ACDQe
    - Preparing a dataset for ML/AI-based prediction of heterodimer binding sites: https://colab.research.google.com/drive/12JLLB3UtxXNv-jGS89-JQbHbNGAEE34u

- Other useful materials/references:
    - RCSB.org API web services: https://www.rcsb.org/docs/programmatic-access/web-services-overview
    - Organization of PDB data:
        - https://data.rcsb.org/index.html#data-organization
        - https://www.rcsb.org/docs/general-help/organization-of-3d-structures-in-the-protein-data-bank
    - Identifiers in the PDB: https://www.rcsb.org/docs/general-help/identifiers-in-pdb
    - Search API query schema: https://search.rcsb.org/redoc/index.html#tag/Search-Service
        - After navigating to the page, click on “json” to open up the list of query parameters, and continue clicking on sub-items to expand further.
    - Search and Data API attributes: https://data.rcsb.org/data-attributes.html
        - Also see: https://rcsb.org/docs/search-and-browse/advanced-search/attribute-details
    - Data API schema: https://data.rcsb.org/index.html#data-schema
    - PDBx/mmCIF dictionary:
        - Documentation: http://mmcif.wwpdb.org
        - Virtual crash course: https://pdb101.rcsb.org/train/training-events/mmcif

Questions from the participants:

16:10-17:00 - Group-wide hands-on examples

Speaker: Sebastian Bittrich, Ph.D. and Joan Segura, Ph.D. - RCSB PDB, UCSD

Synopsis: Introduction to Google Colab notebooks, learn how to make simple Search & Data API requests using Python. Then, pivot into a practical example to compile a dataset for training protein-protein binding site prediction ML/AI methods.

Relevant material:

- Google Colab notebooks:
  - Accessing RCSB PDB APIs via Python: https://colab.research.google.com/drive/1ACwyJx3bROHvBuGrZOChjvczEK5ACDQe
  - Preparing a dataset for ML/AI-based prediction of heterodimer binding sites: https://colab.research.google.com/drive/12JLLB3UtxXNv-jGS89-JQbHbNGAEE34u
- RCSB PDB Search API Python package: https://rcsbsearchapi.readthedocs.io/en/latest/

Questions from the participants:

Q: where can we find a detailed list of all the possible different queries operators and values (i.e like the pair rcsb_entry_info.experimental_method - exact_match - X-ray)?

A: Here is a full list: https://search.rcsb.org/rcsbsearch/v2/metadata/schema. Also see: https://www.rcsb.org/docs/search-and-browse/advanced-search/attribute-details (click “Expand all”)

A: I like to use the rcsb.org frontend to discover available attributes and the values and operators that are compatible with it. (E.g., search query). Then you can use the Search API button in the top-right to see the corresponding Search API query.

A: See https://search.rcsb.org/structure-search-attributes.html and https://search.rcsb.org/chemical-search-attributes.html (doesn’t have values). Also see https://www.rcsb.org/docs/search-and-browse/advanced-search/attribute-details.

Q: How can we find the documents for rcsbsearchapi.search.Attr, such as for all the options?

A: https://www.rcsb.org/docs/search-and-browse/advanced-search/attribute-details is a helpful resource to discover the various attributes that are available.

Q: Hi, is the return type of API able to return attributes other than the PDB IDs?

A: Search API is designed to return only IDs, to get data for other attributes you need to use Data API.

A: In case you’re asking about more fine-grained results, you can also search for individual chains, entities, assemblies etc: https://search.rcsb.org/#return-type

Q: So we need the search API to filter the PDBs and then call Data API to extract the attributes?

A: Exactly.

Q: Can the Data API be called from Python as well?

A: Yes certainly! See the section "Interacting with RCSB Data API" in the first Colab notebook (https://colab.research.google.com/drive/1ACwyJx3bROHvBuGrZOChjvczEK5ACDQe), as well as all the examples in the second notebook (https://colab.research.google.com/drive/12JLLB3UtxXNv-jGS89-JQbHbNGAEE34u)

Q: I have another question about can do the sequence similarity search via search API? with taking a given sequence and returning the PDB IDs that contain similar sequence elements?

A: Yes! It can surely be done via Search API: https://search.rcsb.org/#search-example-3

17:00-18:00 - Independent/small-group work

Breakout room: General search questions
Tutor: Dennis Piehl and Brinda Vallat
Notes (e.g., links to queries, summary of findings, remaining questions):
Based on the first notebook, I have generic questions but not sure if should share these in the open chat.

(Q): Can the Search API be used to find macromolecular structures within a specific resolution range, e.g., filtering structures based on their quality?

(A): Yes, using the "rcsb_entry_info.resolution_combined" attribute

(A): You can also do the same for computed structure models (CSMs) by using the global quality score, e.g. this search example:

```
"parameters": {
      "attribute": "rcsb_ma_qa_metric_global.ma_qa_metric_global.value",
      "operator": "greater",
      "negation": false,
      "value": 50
    }
```

(Q): Is it possible to access NMR spectroscopy and explore structural and spectral information of biomolecules?

(A): You can search for structures that were determined by NMR spectroscopy and download the associated chemical shift and restraint data where available. Restraint files are available for download from RCSB.org but spectral data isn't available. Some data exists on https://bmrb.io/.

(A): You can also use the "rcsb_external_references.type" attribute and value "BMRB" to search for structures that have data from the BMRB associated with them.

(Q): How the Search API find enzyme structures and their associated functional annotations to learn about enzyme mechanisms and active sites?

(A): The Annotations tabs provides some functional annotations like GO terms, see e.g. https://www.rcsb.org/annotations/4CHA#go. You can also find the EC number of individual entities on the Structure Summary Page: https://www.rcsb.org/structure/4CHA#macromoleculespanel. Both properties can be searched for by clicking on the link.

(Q): Can electron density maps for specific PDB entries be accessed? What tools or software can be used to visualise and analyse the 3D density distribution of macromolecules?

(A): Yes. FIrst, to find all structures that have an electron density map associated with them, you need to first search for all structures determined by EM and have experimental data, e.g., using this

query. Then, you'll need to extract the EMD ID from each entry by pulling the attribute "rcsb_entry_container_identifiers.emdb_ids" using the data API, like this query. Then, you can download the electron density maps for structures via the following URL format (changing the emdb ID as needed): https://files.rcsb.org/pub/emdb/structures/EMD-1015/map/emd_1015.map.gz. The electron density maps can be viewed directly in the Mol* interface (e.g., https://www.rcsb.org/3d-view/1DYL?preset=electronDensityMaps)

(Q): Can structures related to a specific organism or species be searched?

(A): Yes, you can use the organism specific attributes such as "rcsb_entity_source_organism.ncbi_scientific_name". For a full list of available attributes, check out this page: https://www.rcsb.org/docs/search-and-browse/advanced-search/attribute-details, click "Expand all" in the upper right, then search for "organism". Also, I encourage you to use the "contains" operator instead of "exact match", because sometimes an organism may have many different variations (e.g., E.coli has tons of different options due to the variety of genetic strains).

(Q): I have a list of UniProt IDs and I want to know which of them have experimental structures in the database. How do you recommend performing a query from a list like this, where not everything is expected to return a result?

(A): You can use the uniprot search attribute, "rcsb_polymer_entity_container_identifiers.reference_sequence_identifiers.database_accession" set to the uniprot value. And then use boolean operators (in this case, "or") to perform it for multiple uniprot ids at the same time.

(Q): How can I use the results from Search API with library like OpenMM for Molecular Dynamics Simulation?

(A): You can retrieve a list of relevant identifiers from Search API. Then, you can download structure data of each file using e.g. this link: https://files.rcsb.org/download/4CHA.cif.gz. Other formats are available as well, see the Download Files dropdown on a Structure Summary Page. These files can be used as input of MDS.

(Q): How can I "format" the API access to interactively integrate it to another webservice? i.e. without directly making the queries with fixed parameters, but defining them "on the fly" by the user inputs and then sending them to the API

(A): You can build your query however you want depending on the attributes or search types you want to perform, then you can add specific options (like if the user wants only "experimental" structures) by simply modifying the JSON object (e.g., you would set "results_content_type": ["experimental"]).

(Q): Is there any RCSB GQL client (python or other language) that can programmatically (not interactive) build a query as the rcsbsearchapi does? Because I have been building a very basic one (https://gitlab.com/martingonzalezbuitron/rcsb_services) to get all the data (json files) related to Entries, Entities and Assemblies and now I have some issues to get other related data, like annotations.

(A): That's awesome! We don't have an RCSB-maintained one yet, it was something on our plans to eventually do once we got the rcsbsearchapi ready. So instead we've been recommending the Python graphql module. But I'm super interested to learn about yours!

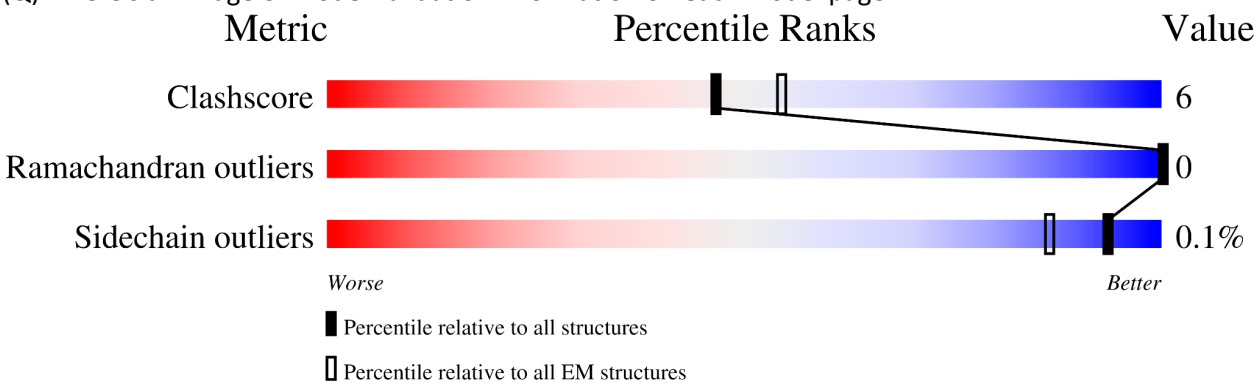
(Q): I don't know if this is a part of API or not, but the information I want is as follows. I'm taking this as an example https://www.rcsb.org/3d-view/8DNF; if I enable the assembly symmetry option, I get to see the symmetry axis. What I want to know is if any information (coordinates/direction cosines/..etc.) about this axis is accessible from data api? Also, how is this axis calculated? I tried to search about it online, I just found that PDB uses Quatsymm? and weekly updates the symmetries of newly added assemblies.

(A): Transformation operations are provided in the mmCIF file, in which the axis is implicitly determinable. We have a variety of attributes in our schema that let you get information about symmetry (e.g., take a look at https://www.rcsb.org/docs/search-and-browse/advanced-search/attribute-details, click "Expand all" in the upper right, then search for "symmetry".

(Q): What is the easiest way to contact the team in regards specifically to usage questions/problems?

(A): https://www.rcsb.org/pages/contactus or email us at info@rcsb.org

(Q): There's an image of model validation information on each model page:



Is this information available through the APIs? All I can find is a steric clash count.

(A): Not yet, but it will be soon. Keep an eye out for a news bulletin.

Breakout room: Structure similarity search and structure motif ("strucmotif") search
Tutor: Sebastian Bittrich
Notes (e.g., links to queries, summary of findings, remaining questions):

Q: How to search for a specific sequence and find similar/related structures?

A: Either use the Structure Summary Page or the Advanced Query Builder to define your query. The result looks like this.

Q: What's the best way to retrieve all identifiers as a simple list?

A: Combine return_all_hits and results_verbosity. E.g., like this.

Q: How to search for a structural motif and a particular Pfam annotation at the same time?

A: Use the Mol* 3D viewer to define your motif, use the Annotations tab to discover the relevant Pfam annotation. Both criteria can be combined into this query. AND and OR are supported when composing predicates.

Q: How to find proteins with or without mutations?

A: "entity_poly.rcsb_mutation_count" helps with that. A query for entries with mutations could look like this.

Breakout room: Protein complexes and interfaces
Tutor: Joan Segura
Notes (e.g., links to queries, summary of findings, remaining questions): …

(Q): How can I identify and analyse protein-protein interfaces? What details can be obtained, including chain IDs and binding sites?

A: This query requests the information for assembly interfaces. It includes accessible surface areas of bound and unbound conformations. The example in the colab provides code examples of how to parse interface data and find the protein interacting residues.

Breakout room: Annotations and positional features
Tutor: Yana Rose
Notes (e.g., links to queries, summary of findings, remaining questions):

Breakout room: Chemical search and ligand annotations
Tutor: Jose Duarte and Chenghua Shao
Notes (e.g., links to queries, summary of findings, remaining questions): …
Q: how to find structures of a specific protein with covalently bound ligands? Furthermore, how to find structures with a ligand bound to specific amino acids, say, Lys. (Answered by Jose and Chenghua)

A: This query (inspired by one of the links in the ligand summary page) should search for all structures with covalently linked ligands. Finding what's the amino acid that is linked to is possible with Data API: this query. However the coverage of this data is limited to Ligands of Interest.

Q: Can we combine a search using sequence searches (finding all sequences similar to a reference one) and then retrieve a list of the identifiers of the ligands bound to each of the similar structures?

A: you can surely find ligands that are present in the same complex with the target sequences using search API but filtering the cases where the ligand is bound to the target sequence might need more post processing

18:00-18:30 - Report out and closing remarks