How to mine x-ray and cryo-EM data from the RSC PDB database
2023.10.16

It has become apparent that even well into a given year, the deposits in the PDB from the previous year may not be complete and indeed this can continue to increase over several years. It is therefore recommended to do a full data mine using the below procedure, starting from the historically first deposit. Note also that in 2019, PDB had its first x-ray entry in 1972; in 2021, this entry seems to have disappeared and the first entry is at 1976. It is therefore important to keep an eye on these (somewhat inexplicable) changes. My approach is simply to take the most recent full data set.
This is the procedure I use to mine the data for the parsing program `xrayCryoDataParse.m`.

- Go to https://www.rcsb.org/search
- In the field "Structure Attributes", select methods -> experimental methods and enter either x-ray diffraction or electron microscopy
- Add a Subquery: Structure Details -> Deposition Date and enter ">" and the first date (e.g., 01.01.1976)
- Press SEARCH button
- Click on Tabular Report-> Create Custom Report
- Select deposit date, structure molecular weight, resolution IN THAT ORDER so columns are date, MW, resn in the CSV file then click on "run report"
- Click on download CSV file. If there are more entries than 2500 for the search you have done, they are downloadable as chunks of 2500 at a time. In this case, you need to concatenate them (at the time of writing 2023.10.16, the entire x-ray data set required me to download 90 files of 2500 entries each, which cost me almost half an hour!). Do this by:
    - Collect all the CSV files in a single folder
    - Open terminal in your mac
    - Change directory to this folder, e.g. `cd Desktop/temp/`
    - Type in: `cat *.csv >yourFilename.csv`
- Open this CSV file in Numbers (if you're using a mac, otherwise, you have to work out the next bit yourself. Soz)
- Go to the column header of the deposition date – this should be Column B
- In the dropdown menu of this column, choose "sort ascending"
- Click on first date entry (it should be the oldest) then scroll to bottom row and click on the last resolution entry while holding down shift key – this should highlight all three columns. Press cmd-C (copy)
- Open a new file in TexShop and press cmd-V (paste) – all the data should now be in the tex file.
- Now you need to remove all dashes in the dates (e.g. `2020-05-05` changes to `2020  05  05`, that is, with tabs in between. Spaces don't work, as matlab then thinks it is a string element). To do this simply press cmd-F (find) and enter "–" in the field "Find" and a tab in "Replace With", then click on "Replace All". To get the tab in the replace field, open a blank tex file, enter a tab, then copy-paste it into the field
- Save the file with the appropriate name e.g., allXrayData_20YYMMDD.dat, or allCryoData_20YYMMDD.dat

- There might be some rows with multiple resolutions separated by a comma. The way I resolve this is as follows: Open up TeXShop. In TeXShop preferences, select OgreKit Find panel (if you aren't already using this, you will need to restart TeXShop). Then press ctrl-f, click on 'regular expressions' in the options, and enter a comma followed by a space then (.+) in find [i.e. type in: `,  (.+)`], and in replace, type in nothing. Do this for all instances of a comma. Not very clean, but hey-ho
- The relevant columns are:
    - 1: The year
    - 4: The molecular weight in kDa
    - 5: The resolution in Angstroms