Shell command line assignment – PDB files

First, mkdir STRUCTURE will create a directory called STRUCTURE, and then cd STRUCTURE will change the current directory and access STRUCTURE.

```
haya@LAPTOP-8APDPGH1:~$ mkdir STRUCTURE
haya@LAPTOP-8APDPGH1:~$ cd STRUCTURE
```

Question 1:

To retrieve the pdb files from the website, we use the curl command with the identifier: 1W[0-9][A-Z], [0-9] to match any digit, and [A-Z] to match any upper-case letter.
This is the curl command I used: **curl -# -O** **http://files.rcsb.org/view/1W[0-9][A-Z].pdb**

-# will use the simplest progress bar while downloading the files.
-O saves the output of the curl command into separate files while naming them like their respective identifiers.

```
haya@LAPTOP-8APDPGH1:~/STRUCTURE$ curl -# -O http://files.rcsb.org/view/1W[0-9][A-Z].pdb

[1/260]: http://files.rcsb.org/view/1W0A.pdb --> 1W0A.pdb
--_curl_--http://files.rcsb.org/view/1W0A.pdb
##  #  #                                                                    -=O=-

[2/260]: http://files.rcsb.org/view/1W0B.pdb --> 1W0B.pdb
--_curl_--http://files.rcsb.org/view/1W0B.pdb
                     #    #      #     #              -=O=-
```

After all the files were downloaded, by using **ls** we can see all the files in the directory.

```
haya@LAPTOP-8APDPGH1:~/STRUCTURE$ ls
1W0A.pdb  1W0P.pdb  1W1E.pdb  1W1T.pdb  1W2I.pdb  1W2X.pdb  1W3M.pdb  1W4B.pdb  1W4Q.pdb  1W5F.pdb  1W5U.pdb  1W6J.pdb  1W6Y.pdb  1W7N.pdb  1W8C.pdb  1W8R.pdb  1W9G.pdb
1W9V.pdb
1W0B.pdb  1W0Q.pdb  1W1F.pdb  1W1U.pdb  1W2J.pdb  1W2Y.pdb  1W3N.pdb  1W4C.pdb  1W4R.pdb  1W5G.pdb  1W5V.pdb  1W6K.pdb  1W6Z.pdb  1W7O.pdb  1W8D.pdb  1W8S.pdb  1W9H.pdb
1W9W.pdb
1W0C.pdb  1W0R.pdb  1W1G.pdb  1W1V.pdb  1W2K.pdb  1W2Z.pdb  1W3O.pdb  1W4D.pdb  1W4S.pdb  1W5H.pdb  1W5W.pdb  1W6L.pdb  1W7A.pdb  1W7P.pdb  1W8E.pdb  1W8T.pdb  1W9I.pdb
1W9X.pdb
1W0D.pdb  1W0S.pdb  1W1H.pdb  1W1W.pdb  1W2L.pdb  1W3A.pdb  1W3P.pdb  1W4E.pdb  1W4T.pdb  1W5I.pdb  1W5X.pdb  1W6M.pdb  1W7B.pdb  1W7Q.pdb  1W8F.pdb  1W8U.pdb  1W9J.pdb
1W9Y.pdb
1W0E.pdb  1W0T.pdb  1W1I.pdb  1W1X.pdb  1W2M.pdb  1W3B.pdb  1W3Q.pdb  1W4F.pdb  1W4U.pdb  1W5J.pdb  1W5Y.pdb  1W6N.pdb  1W7C.pdb  1W7R.pdb  1W8G.pdb  1W8V.pdb  1W9K.pdb
1W9Z.pdb
1W0F.pdb  1W0U.pdb  1W1J.pdb  1W1Y.pdb  1W2N.pdb  1W3C.pdb  1W3R.pdb  1W4G.pdb  1W4V.pdb  1W5K.pdb  1W5Z.pdb  1W6O.pdb  1W7D.pdb  1W7S.pdb  1W8H.pdb  1W8W.pdb  1W9L.pdb
1W0G.pdb  1W0V.pdb  1W1K.pdb  1W1Z.pdb  1W2O.pdb  1W3D.pdb  1W3S.pdb  1W4H.pdb  1W4W.pdb  1W5L.pdb  1W6A.pdb  1W6P.pdb  1W7E.pdb  1W7T.pdb  1W8I.pdb  1W8X.pdb  1W9M.pdb
1W0H.pdb  1W0W.pdb  1W1L.pdb  1W2A.pdb  1W2P.pdb  1W3E.pdb  1W3T.pdb  1W4I.pdb  1W4X.pdb  1W5M.pdb  1W6B.pdb  1W6Q.pdb  1W7F.pdb  1W7U.pdb  1W8J.pdb  1W8Y.pdb  1W9N.pdb
1W0I.pdb  1W0X.pdb  1W1M.pdb  1W2B.pdb  1W2Q.pdb  1W3F.pdb  1W3U.pdb  1W4J.pdb  1W4Y.pdb  1W5N.pdb  1W6C.pdb  1W6R.pdb  1W7G.pdb  1W7V.pdb  1W8K.pdb  1W8Z.pdb  1W9O.pdb
1W0J.pdb  1W0Y.pdb  1W1N.pdb  1W2C.pdb  1W2R.pdb  1W3G.pdb  1W3V.pdb  1W4K.pdb  1W4Z.pdb  1W5O.pdb  1W6D.pdb  1W6S.pdb  1W7H.pdb  1W7W.pdb  1W8L.pdb  1W9A.pdb  1W9P.pdb
1W0K.pdb  1W0Z.pdb  1W1O.pdb  1W2D.pdb  1W2S.pdb  1W3H.pdb  1W3W.pdb  1W4L.pdb  1W5A.pdb  1W5P.pdb  1W6E.pdb  1W6T.pdb  1W7I.pdb  1W7X.pdb  1W8M.pdb  1W9B.pdb  1W9Q.pdb
1W0L.pdb  1W1A.pdb  1W1P.pdb  1W2E.pdb  1W2T.pdb  1W3I.pdb  1W3X.pdb  1W4M.pdb  1W5B.pdb  1W5Q.pdb  1W6F.pdb  1W6U.pdb  1W7J.pdb  1W7Y.pdb  1W8N.pdb  1W9C.pdb  1W9R.pdb
1W0M.pdb  1W1B.pdb  1W1Q.pdb  1W2F.pdb  1W2U.pdb  1W3J.pdb  1W3Y.pdb  1W4N.pdb  1W5C.pdb  1W5R.pdb  1W6G.pdb  1W6V.pdb  1W7K.pdb  1W7Z.pdb  1W8O.pdb  1W9D.pdb  1W9S.pdb
1W0N.pdb  1W1C.pdb  1W1R.pdb  1W2G.pdb  1W2V.pdb  1W3K.pdb  1W3Z.pdb  1W4O.pdb  1W5D.pdb  1W5S.pdb  1W6H.pdb  1W6W.pdb  1W7L.pdb  1W8A.pdb  1W8P.pdb  1W9E.pdb  1W9T.pdb
1W0O.pdb  1W1D.pdb  1W1S.pdb  1W2H.pdb  1W2W.pdb  1W3L.pdb  1W4A.pdb  1W4P.pdb  1W5E.pdb  1W5T.pdb  1W6I.pdb  1W6X.pdb  1W7M.pdb  1W8B.pdb  1W8Q.pdb  1W9F.pdb  1W9U.pdb
```

Some of these files do not exist in the Database. While downloading, the curl command will fill these files with an HTML code stating that the file is not found, with a size of 260. To find these

files we use the grep command “ **grep -il "not found" *.pdb** ”. 7 files with this pattern were found.

-i makes the search case insensitive.
-l displays only the name of the file that matches the pattern.
In this case, we are looking for the pattern “not found” in every .pdb file downloaded in this directory.

```
haya@LAPTOP-8APDPGH1:~/STRUCTURE$ grep -il "not found" *.pdb
1W0L.pdb
1W2J.pdb
1W4D.pdb
1W6A.pdb
1W6D.pdb
1W6E.pdb
1W7Y.pdb
```

To delete these files, the xargs with rm command is piped to the grep command: “**grep -il "not found" *.pdb | xargs rm**”. The names of the files are fed to xargs, and then xargs feeds them to rm which in turn deletes the files.

After running this command, ls shows the remaining files proving that the chosen files were deleted.

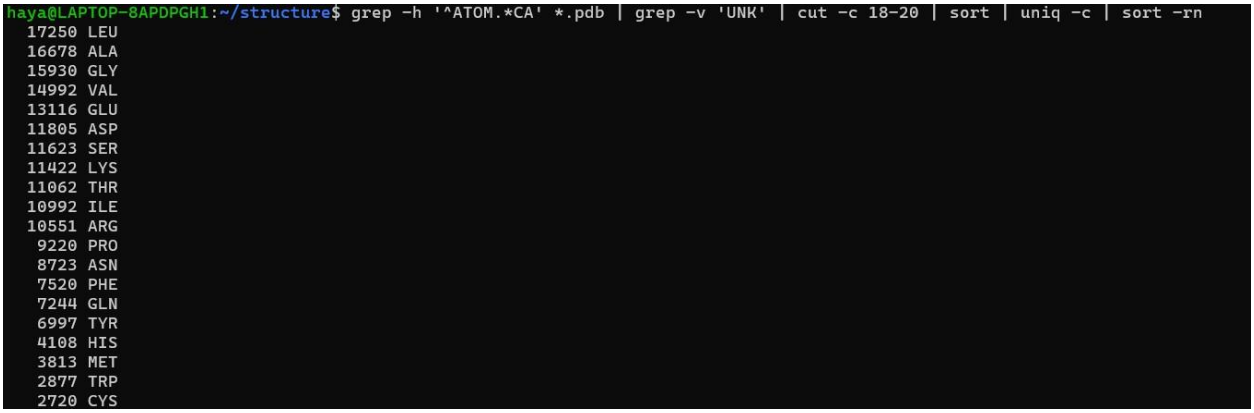
```
haya@LAPTOP-8APDPGH1:~/structure$ grep -il "not found" *.pdb | xargs rm
haya@LAPTOP-8APDPGH1:~/structure$ ls
1W0A.pdb  1W0Q.pdb  1W1F.pdb  1W1U.pdb  1W2K.pdb  1W2Z.pdb  1W3O.pdb  1W4E.pdb  1W4T.pdb  1W5I.pdb  1W5X.pdb  1W6P.pdb  1W7E.pdb  1W7T.pdb  1W8J.pdb  1W8Y.pdb  1W9N.pdb
1W0B.pdb  1W0R.pdb  1W1G.pdb  1W1V.pdb  1W2L.pdb  1W3A.pdb  1W3P.pdb  1W4F.pdb  1W4U.pdb  1W5J.pdb  1W5Y.pdb  1W6Q.pdb  1W7F.pdb  1W7U.pdb  1W8K.pdb  1W8Z.pdb  1W9O.pdb
1W0C.pdb  1W0S.pdb  1W1H.pdb  1W1W.pdb  1W2M.pdb  1W3B.pdb  1W3Q.pdb  1W4G.pdb  1W4V.pdb  1W5K.pdb  1W5Z.pdb  1W6R.pdb  1W7G.pdb  1W7V.pdb  1W8L.pdb  1W9A.pdb  1W9P.pdb
1W0D.pdb  1W0T.pdb  1W1I.pdb  1W1X.pdb  1W2N.pdb  1W3C.pdb  1W3R.pdb  1W4H.pdb  1W4W.pdb  1W5L.pdb  1W6B.pdb  1W6S.pdb  1W7H.pdb  1W7W.pdb  1W8M.pdb  1W9B.pdb  1W9Q.pdb
1W0E.pdb  1W0U.pdb  1W1J.pdb  1W1Y.pdb  1W2O.pdb  1W3D.pdb  1W3S.pdb  1W4I.pdb  1W4X.pdb  1W5M.pdb  1W6C.pdb  1W6T.pdb  1W7I.pdb  1W7X.pdb  1W8N.pdb  1W9C.pdb  1W9R.pdb
1W0F.pdb  1W0V.pdb  1W1K.pdb  1W1Z.pdb  1W2P.pdb  1W3E.pdb  1W3T.pdb  1W4J.pdb  1W4Y.pdb  1W5N.pdb  1W6F.pdb  1W6U.pdb  1W7J.pdb  1W7Z.pdb  1W8O.pdb  1W9D.pdb  1W9S.pdb
1W0G.pdb  1W0W.pdb  1W1L.pdb  1W2A.pdb  1W2Q.pdb  1W3F.pdb  1W3U.pdb  1W4K.pdb  1W4Z.pdb  1W5O.pdb  1W6G.pdb  1W6V.pdb  1W7K.pdb  1W8A.pdb  1W8P.pdb  1W9E.pdb  1W9T.pdb
1W0H.pdb  1W0X.pdb  1W1M.pdb  1W2B.pdb  1W2R.pdb  1W3G.pdb  1W3V.pdb  1W4L.pdb  1W5A.pdb  1W5P.pdb  1W6H.pdb  1W6W.pdb  1W7L.pdb  1W8B.pdb  1W8Q.pdb  1W9F.pdb  1W9U.pdb
1W0I.pdb  1W0Y.pdb  1W1N.pdb  1W2C.pdb  1W2S.pdb  1W3H.pdb  1W3W.pdb  1W4M.pdb  1W5B.pdb  1W5Q.pdb  1W6I.pdb  1W6X.pdb  1W7M.pdb  1W8C.pdb  1W8R.pdb  1W9G.pdb  1W9V.pdb
1W0J.pdb  1W0Z.pdb  1W1O.pdb  1W2D.pdb  1W2T.pdb  1W3I.pdb  1W3X.pdb  1W4N.pdb  1W5C.pdb  1W5R.pdb  1W6J.pdb  1W6Y.pdb  1W7N.pdb  1W8D.pdb  1W8S.pdb  1W9H.pdb  1W9W.pdb
1W0K.pdb  1W1A.pdb  1W1P.pdb  1W2E.pdb  1W2U.pdb  1W3J.pdb  1W3Y.pdb  1W4O.pdb  1W5D.pdb  1W5S.pdb  1W6K.pdb  1W6Z.pdb  1W7O.pdb  1W8E.pdb  1W8T.pdb  1W9I.pdb  1W9X.pdb
1W0M.pdb  1W1B.pdb  1W1Q.pdb  1W2F.pdb  1W2V.pdb  1W3K.pdb  1W3Z.pdb  1W4P.pdb  1W5E.pdb  1W5T.pdb  1W6L.pdb  1W7A.pdb  1W7P.pdb  1W8F.pdb  1W8U.pdb  1W9J.pdb  1W9Y.pdb
1W0N.pdb  1W1C.pdb  1W1R.pdb  1W2G.pdb  1W2W.pdb  1W3L.pdb  1W4A.pdb  1W4Q.pdb  1W5F.pdb  1W5U.pdb  1W6M.pdb  1W7B.pdb  1W7Q.pdb  1W8G.pdb  1W8V.pdb  1W9K.pdb  1W9Z.pdb
1W0O.pdb  1W1D.pdb  1W1S.pdb  1W2H.pdb  1W2X.pdb  1W3M.pdb  1W4B.pdb  1W4R.pdb  1W5G.pdb  1W5V.pdb  1W6N.pdb  1W7C.pdb  1W7R.pdb  1W8H.pdb  1W8W.pdb  1W9L.pdb
1W0P.pdb  1W1E.pdb  1W1T.pdb  1W2I.pdb  1W2Y.pdb  1W3N.pdb  1W4C.pdb  1W4S.pdb  1W5H.pdb  1W5W.pdb  1W6O.pdb  1W7D.pdb  1W7S.pdb  1W8I.pdb  1W8X.pdb  1W9M.pdb
```

Question 2:

In order to extract the amino acids and count their frequencies, we search in each pdb file for the lines starting with “ATOM” and containing CA using the grep command “ grep -h '^ATOM.*CA' *.pdb ”. **-h** will hide the file's name in the output which will be useful for later. We then pipe this output into another grep command to exclude the lines containing ‘UNK’ by using -v:
“grep -v 'UNK' ”. Now to get the amino acid residues, which are between the characters 18 to 20, we pipe our last output into a cut command “cut -c 18-20”. -c will return the specified characters.

By using -h in the first grep and removing the file's name, the character count will be correct, and we will get the necessary output. Our next step is to get the frequencies of the residues by using the uniq command. However, uniq works by looking at consecutive identical lines and leaving one unique representative, so we need to sort them first alphabetically by using the command sort. We now pipe the output of sort to "uniq -c". -c will count the number of occurrences of each amino acid. Finally, to reverse numerical order and place the most abundant amino acid at the top we sort them again using "sort -rn": -r to sort in reverse order, and -n to sort by numeric value not alphabetical.

The final command will be:
grep -h '^ATOM.*CA' *.pdb | grep -v 'UNK' | cut -c 18-20 | sort | uniq -c | sort -rn



Haya Mazyad
202207736