Life of extremophiles : Database and Knowledge Graph of microbes living in extreme conditions(Alkaline)

Progress Report

Week 1:-

INTRODUCTION:-

The database and knowledge graph focusing on extremophiles, particularly those thriving in alkaline environments, provide a comprehensive understanding of these remarkable microbes. Extremophiles are organisms capable of surviving and even thriving in environments considered extreme by human standards, such as elevated temperatures, acidity, salinity, or alkalinity. Alkaline environments, characterized by high pH levels, pose significant challenges to most life forms due to their harsh conditions. However, extremophiles adapted to alkaline conditions have evolved unique biochemical and physiological mechanisms to withstand and use these environments to their advantage. The database and knowledge graph contain information on various aspects of alkaline extremophiles, including their taxonomy, habitat preferences, metabolic pathways, molecular adaptations, and potential applications. Researchers can use this resource to explore the diversity, evolution, and ecological roles of alkaline extremophiles, as well as to discover novel enzymes, biomolecules, and biotechnological applications associated with these organisms.

In summary, the database and knowledge graph serve as valuable tools for researchers interested in unravelling the secrets of extremophiles living in alkaline environments, offering insights that could have implications for fields ranging from biotechnology and bioengineering to astrobiology and environmental science.

So before dive into the next step in this weak I have gathered all the information regarding the Alkaline and also fetch the csv files from PubMed and EuroPMC. Specifically, I obtained 806 data entries from Europe PMC and 409 data entries from PubMed.

Snippets to demonstrate the data forms

	A	B	C	D	E	F	G	H	I	J	K
1	PMID	Title	Authors	Citation	First Auth	Journal/B	Publicatic	Create Dat	PMCID	NIHMS ID	DOI
2	30796503	Genomics	Lebre PH,	Adv Bioch	Lebre PH	Adv Bioch	2020	########			10.1007/10_2018_83
3	37118007	Chemical	Murray KA	Nat Rev Ch	Murray KA	Nat Rev Ch	2022	########			10.1038/s41570-022-00407-4
4	15046570	Alkalitheri	Wiegel J, K	Biochem S	Wiegel J	Biochem S	2004	########			10.1042/bst0320193
5	33763123	Comparat	Choure K,	Front Gen	Choure K	Front Gen	2021	########	PMC7982539		10.3389/fgene.2021.643423
6	28007654	Archaeal	Caforio A,	Biochim B	Caforio A	Biochim B	2017	########			10.1016/j.bbalip.2016.12.006
7	29266533	Recent Ad	Harding T,	J Eukaryot	Harding T	J Eukaryot	2018	########			10.1111/jeu.12495
8	26780356	Wide pH r	Dhakar K,	Appl Micr	Dhakar K	Appl Micr	2016	########			10.1007/s00253-016-7285-2
9	29290045	Abiostress	Wei Y, Zha	Extremopl	Wei Y	Extremopl	2018	########			10.1007/s00792-017-0986-3
10	37490857	Extremopl	Gostinčar	Curr Biol.	Gostinčar	Curr Biol	2023	########			10.1016/j.cub.2023.06.011
11	32323057	Nature an	Uma G, Ba	World J Mi	Uma G	World J Mi	2020	########			10.1007/s11274-020-02841-2
12	32533304	Microorga	Novak Bal	Appl Micr	Novak Bal	Appl Micr	2020	########	PMC7347518		10.1007/s00253-020-10719-4
13	26733008	High-Up:	Albarracín	Front Micr	Albarracín	Front Micr	2015	########	PMC4679917		10.3389/fmicb.2015.01404
14	26647770	Forged Un	Albarracín	Photocher	Albarracín	Photocher	2016	########			10.1111/php.12555
15	37367588	Alkaliphili	Fernández	J Fungi (Ba	Fernández	J Fungi (Ba	2023	########	PMC10301932		10.3390/jof9060652
16	30796504	Isolation	Kevbrin V	Adv Bioch	Kevbrin V	Adv Bioch	2020	########			10.1007/10_2018_84
17	8688447	Respirato	Schäfer G,	Biochim B	Schäfer G	Biochim B	1996	########			10.1016/0005-2728(96)00043-6
18	31541933	Metal and	Giovanella	J Hazard M	Giovanella	J Hazard M	2020	########			10.1016/j.jhazmat.2019.121024
19	35688350	Extremopl	Vivek K, Sa	Biotechno	Vivek K	Biotechno	2022	########			10.1016/j.biotechadv.2022.108002
20	11809961	Antarctic	Thomas D	Science. 2	Thomas D	Science	2002	########			10.1126/science.1063391
21	37317247	Investigati	Pham VHT	Microorga	Pham VHT	Microorga	2023	########	PMC10223213		10.3390/microorganisms11051273
22	33643258	Bioprospe	Sysoev M,	Front Micr	Sysoev M	Front Micr	2021	########	PMC7902512		10.3389/fmicb.2021.630013

So, this is the file that I obtain while downloading from the PubMed.

Week 2: -

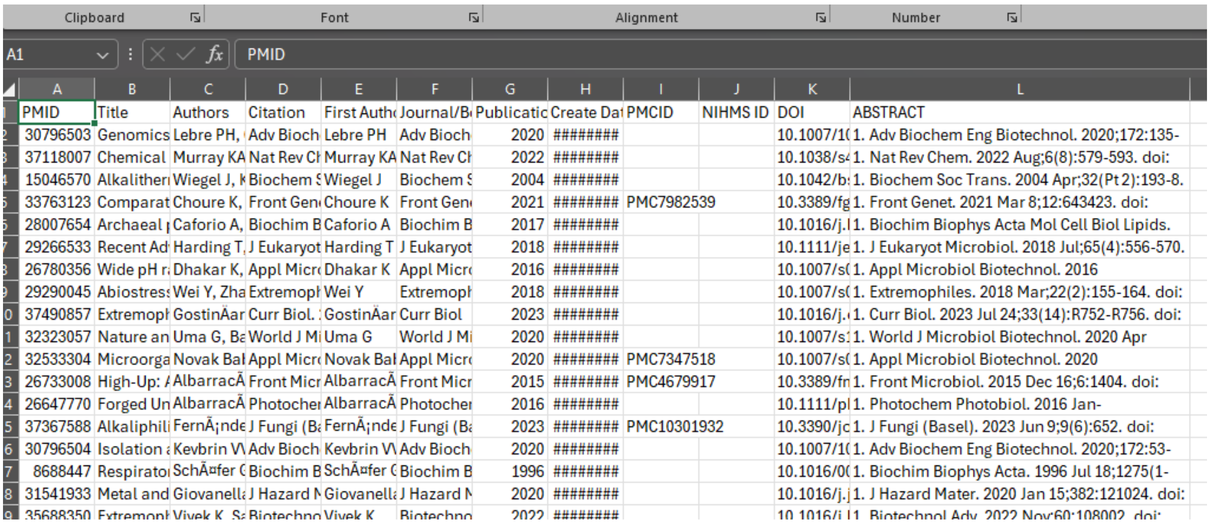
In the second week, I focused on PubMed .csv files to extract relevant data based on the PubMed ID (PMID). The process involves searching for the corresponding abstract using the PMID. If an abstract is found, I systematically create a new column and populate it with the abstract text. In cases where no abstract is found, the entry is marked as NULL.

```
Successfully installed Bio-1.6.2 biopython-1.83 biothings-client-0.3.1 gprofiler-official-1.0.0 mygene-3.2.2
```

```python
from Bio.Entrez import efetch
from Bio import Entrez
import pandas as pd
df = pd.read_csv("/content/extremophilesOrigina.csv") #this file is on zip folder
def print_abstract(pmid):
    try:
        handle = efetch(db='pubmed', id=pmid, retmode='text', rettype='abstract')
        abstract = handle.read()
        df.loc[df['PMID'] == pmid, 'ABSTRACT'] = abstract
        # print(abstract)
    except Exception as e:
        print(f"Error fetching abstract for PMID {pmid}: {str(e)}")


df.head()
```

PMID	Title	Authors	Citation	First Author	Journal/Book	Publication Year	Create Date	PMCID	NIHMS ID	DOI
30796503	Genomics of Alkaliphiles	Lebre PH, Cowan DA.	Adv Biochem Eng Biotechnol. 2020;172:135-	Lebre PH	Adv Biochem Eng Biotechnol	2020	2019/02/24	NaN	NaN	10.1007/10_2018_83

Output file snippets are:-



PMID	Title	Authors	Citation	First Auth	Journal/B	Publicatio	Create Dat	PMCID	NIHMS ID	DOI	ABSTRACT
30796503	Genomics	Lebre PH,	Adv Bioch	Lebre PH	Adv Bioch	2020	########			10.1007/1(	1. Adv Biochem Eng Biotechnol. 2020;172:135-
37118007	Chemical	Murray KA	Nat Rev Ch	Murray KA	Nat Rev Ch	2022	########			10.1038/s4	1. Nat Rev Chem. 2022 Aug;6(8):579-593. doi:
15046570	Alkalitheri	Wiegel J, K	Biochem S	Wiegel J	Biochem S	2004	########			10.1042/b:	1. Biochem Soc Trans. 2004 Apr;32(Pt 2):193-8.
33763123	Comparat	Choure K,	Front Gen	Choure K	Front Gen	2021	########	PMC7982539		10.3389/fg	1. Front Genet. 2021 Mar 8;12:643423. doi:
28007654	Archaeal p	Caforio A,	Biochim B	Caforio A	Biochim B	2017	########			10.1016/j.l	1. Biochim Biophys Acta Mol Cell Biol Lipids.
29266533	Recent Ad	Harding T,	J Eukaryot	Harding T	J Eukaryot	2018	########			10.1111/je	1. J Eukaryot Microbiol. 2018 Jul;65(4):556-570.
26780356	Wide pH r	Dhakar K,	Appl Micr	Dhakar K	Appl Micr	2016	########			10.1007/s(	1. Appl Microbiol Biotechnol. 2016
29290045	Abiostress	Wei Y, Zha	Extremoph	Wei Y	Extremoph	2018	########			10.1007/s(	1. Extremophiles. 2018 Mar;22(2):155-164. doi:
37490857	Extremoph	GostinÄar	Curr Biol.	GostinÄar	Curr Biol	2023	########			10.1016/j.	1. Curr Biol. 2023 Jul 24;33(14):R752-R756. doi:
32323057	Nature an	Uma G, Ba	World J Mi	Uma G	World J Mi	2020	########			10.1007/s1	1. World J Microbiol Biotechnol. 2020 Apr
32533304	Microorga	Novak Bal	Appl Micr	Novak Bal	Appl Micr	2020	########	PMC7347518		10.1007/s(	1. Appl Microbiol Biotechnol. 2020
26733008	High-Up: A	AlbarracÃ	Front Micr	AlbarracÃ	Front Micr	2015	########	PMC4679917		10.3389/fn	1. Front Microbiol. 2015 Dec 16;6:1404. doi:
26647770	Forged Un	AlbarracÃ	Photocher	AlbarracÃ	Photocher	2016	########			10.1111/pl	1. Photochem Photobiol. 2016 Jan-
37367588	Alkaliphili	FernÃ¡nde	J Fungi (Ba	FernÃ¡nde	J Fungi (Ba	2023	########	PMC10301932		10.3390/jc	1. J Fungi (Basel). 2023 Jun 9;9(6):652. doi:
30796504	Isolation a	Kevbrin V\	Adv Bioch	Kevbrin V\	Adv Bioch	2020	########			10.1007/1(	1. Adv Biochem Eng Biotechnol. 2020;172:53-
8688447	Respirator	SchÃ¤fer (	Biochim B	SchÃ¤fer (	Biochim B	1996	########			10.1016/0(	1. Biochim Biophys Acta. 1996 Jul 18;1275(1-
31541933	Metal and	Giovanella	J Hazard N	Giovanella	J Hazard N	2020	########			10.1016/j.j	1. J Hazard Mater. 2020 Jan 15;382:121024. doi:
35688350	Extremoph	Vivek K, Sa	Biotechno	Vivek K	Biotechno	2022	########			10.1016/j.l	1. Biotechnol Adv. 2022 Nov;60:108002. doi:

Week 3:-

In this weak I was working on the task 2 where I am using Pubtator function to generate Genes Disease, Mutations, Chemicals And Species

1. Genes: PubTator finds mentions of genes within the text and provides annotations linking them to specific gene identifiers or symbols, allowing researchers to quickly find genes associated with topics or diseases.

2. Diseases: PubTator annotates mentions of diseases or medical conditions in the text, providing links to standardized disease names or identifiers from biomedical ontologies or databases.

3. Mutations: PubTator can detect references to genetic mutations or variations within the text, providing annotations that link these mutations to specific genes or diseases when applicable.

4. Chemicals: PubTator finds references to chemicals, drugs, or other chemical compounds mentioned in the text, providing annotations that link them to standardized chemical identifiers or names.

5. Species: PubTator recognizes mentions of species within the text, providing annotations that specify the species mentioned, which is particularly useful in biomedical research where species-specific information is important.

The PubTator tool typically provides a web-based interface or an API (Application Programming Interface) that allows users to programmatically access its functionalities. With the API, users can send text documents or PubMed article identifiers and retrieve annotations for genes, diseases, mutations, chemicals, and species mentioned in those documents.

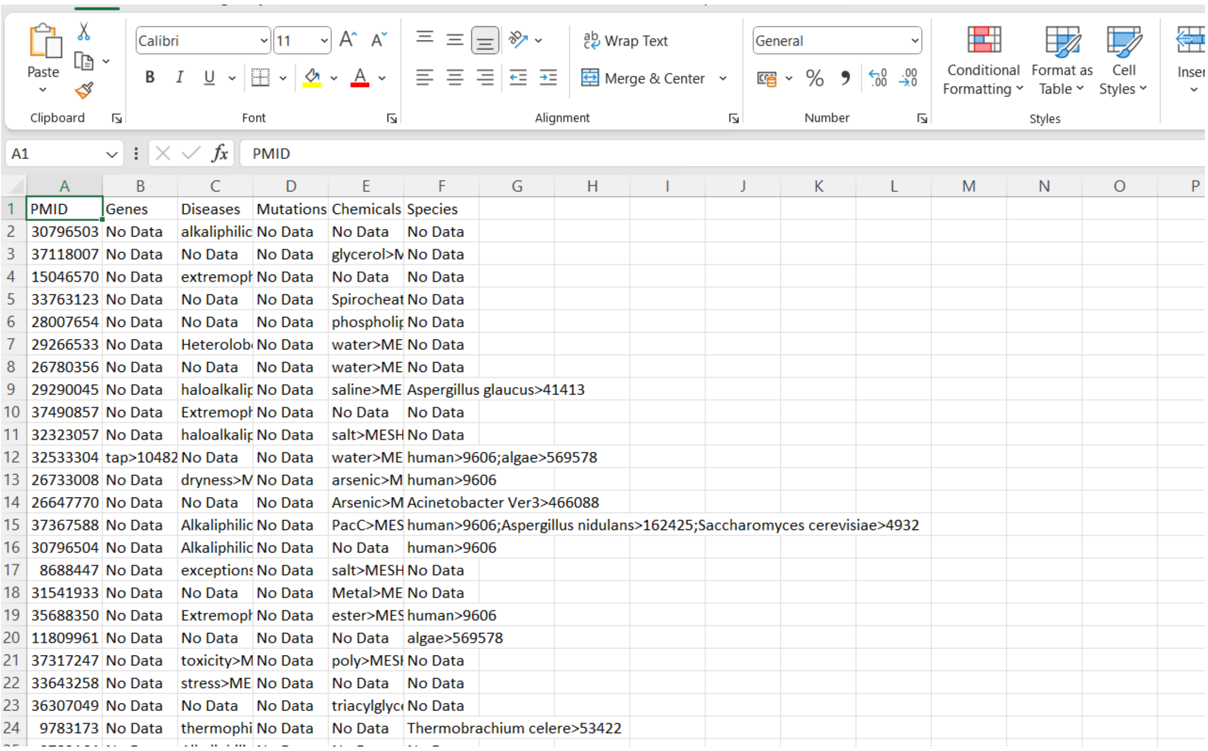
So, how does I proceeds with the Input and the output files

Input Data:

we would provide PubMed IDs or text from biomedical literature as input to PubTator. PubMed IDs uniquely find articles in the PubMed database.

Output Entities:

PubTator outputs the recognized entities (genes, diseases, mutations, chemicals, and species) along with their respective annotations in the provided text.



	A	B	C	D	E	F
1	PMID	Genes	Diseases	Mutations	Chemicals	Species
2	30796503	No Data	alkaliphilic	No Data	No Data	No Data
3	37118007	No Data	No Data	No Data	glycerol>M	No Data
4	15046570	No Data	extremoph	No Data	No Data	No Data
5	33763123	No Data	No Data	No Data	Spirocheat	No Data
6	28007654	No Data	No Data	No Data	phospholip	No Data
7	29266533	No Data	Heterolob	No Data	water>ME	No Data
8	26780356	No Data	No Data	No Data	water>ME	No Data
9	29290045	No Data	haloalkalip	No Data	saline>ME	Aspergillus glaucus>41413
10	37490857	No Data	Extremoph	No Data	No Data	No Data
11	32323057	No Data	haloalkalip	No Data	salt>MESH	No Data
12	32533304	tap>10482	No Data	No Data	water>ME	human>9606;algae>569578
13	26733008	No Data	dryness>M	No Data	arsenic>M	human>9606
14	26647770	No Data	No Data	No Data	Arsenic>M	Acinetobacter Ver3>466088
15	37367588	No Data	Alkaliphilic	No Data	PacC>MES	human>9606;Aspergillus nidulans>162425;Saccharomyces cerevisiae>4932
16	30796504	No Data	Alkaliphilic	No Data	No Data	human>9606
17	8688447	No Data	exceptions	No Data	salt>MESH	No Data
18	31541933	No Data	No Data	No Data	Metal>ME	No Data
19	35688350	No Data	Extremoph	No Data	ester>MES	human>9606
20	11809961	No Data	No Data	No Data	No Data	algae>569578
21	37317247	No Data	toxicity>M	No Data	poly>MESH	No Data
22	33643258	No Data	stress>ME	No Data	No Data	No Data
23	36307049	No Data	No Data	No Data	triacylglyc	No Data
24	9783173	No Data	thermophi	No Data	No Data	Thermobrachium celere>53422

Week 4:-

Steps to demonstrate the functionality

1. Obtaining Input Files: i obtained files from PubTator holding annotations for genes, diseases, mutations, chemicals, and species mentioned in PubMed csv or other

2. Parsing Input Files: i parsed the input files to extract the relevant information, focusing on the columns for genes, diseases, mutations, chemicals, and species.

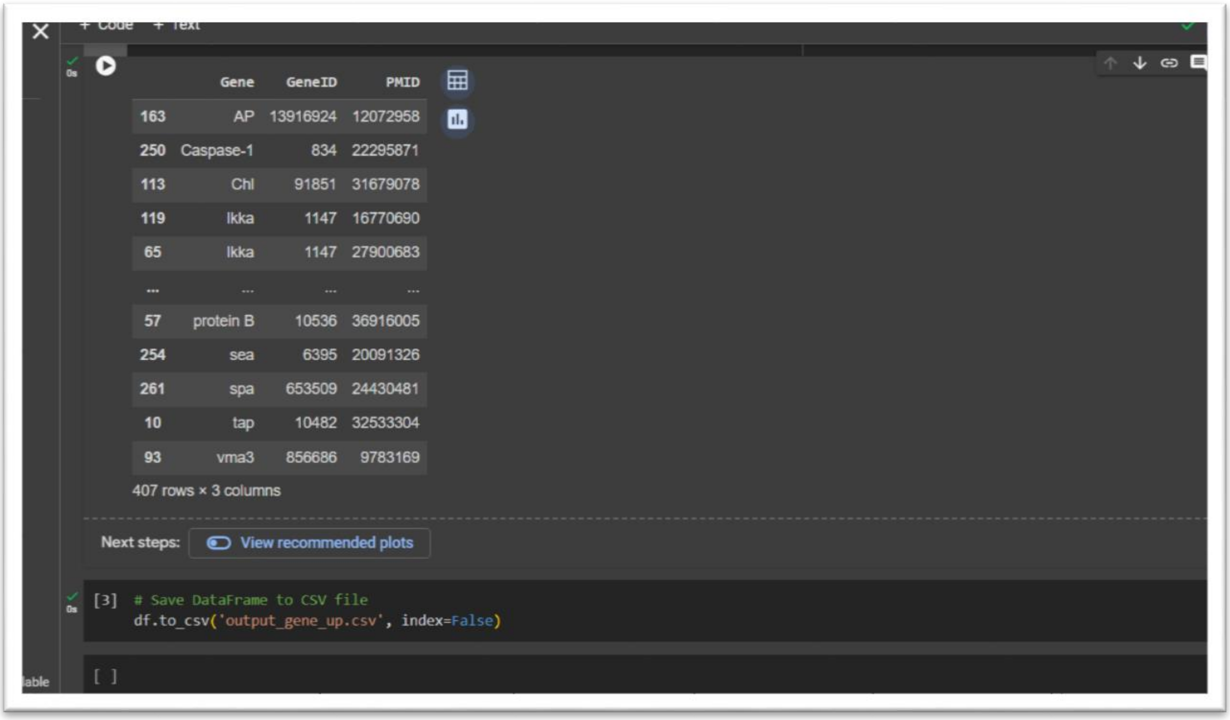
3. Creating Separate Files: i created four separate output files, each holding information related to one of the following categories:

- Genes
- Diseases
- Species
- Chemicals

4. Saving Output Files: Finally, i saved each of the four output files separately, making them available for further processing or analysis as needed.

By following these steps, I was able to extract and organize the information obtained from PubTator into four distinct files, each focusing on a specific aspect of the annotations: genes, diseases, mutations, and chemicals/species. This process facilitated further analysis or research tasks related to the identified biomedical entities.

Outputs are:-For Gene



	Gene	GeneID	PMID
163	AP	13916924	12072958
250	Caspase-1	834	22295871
113	Chl	91851	31679078
119	Ikka	1147	16770690
65	Ikka	1147	27900683
...	...	...	...
57	protein B	10536	36916005
254	sea	6395	20091326
261	spa	653509	24430481
10	tap	10482	32533304
93	vma3	856686	9783169

407 rows × 3 columns


```
# Save DataFrame to CSV file
df.to_csv('output_gene_up.csv', index=False)
```

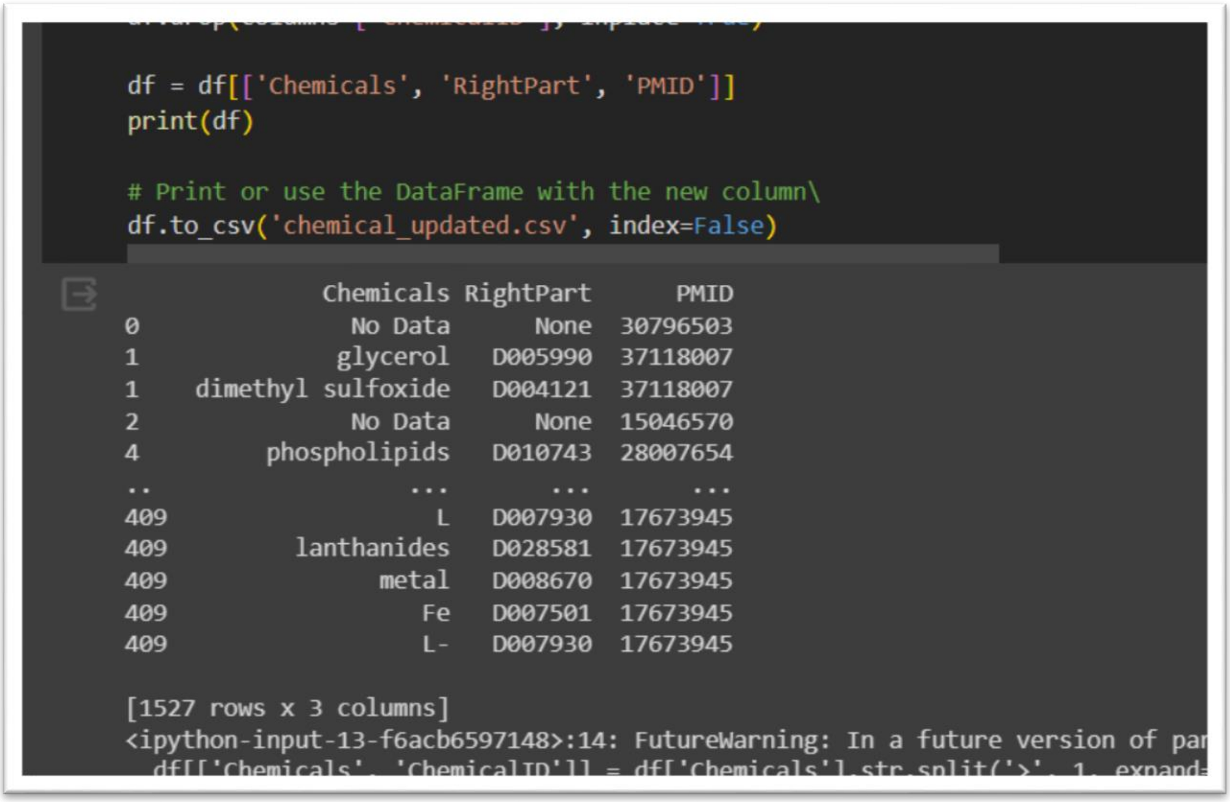
Outputs are:-For Disease

```
df
```

	Diseases	RightPart	PMID
0	alkaliphilic	No Data	30796503
1	No Data	None	37118007
2	extremophilic organisms	D019965	15046570
2	alkaliphilic	No Data	15046570
3	No Data	None	33763123
...	...	...	...
405	alkaliphilic	No Data	11778838
406	extremophilic proteins	D018455	7908011
407	alkaliphilic Bacillus YN-2000	D000881	9672682
408	archaeon Sulfolobus solfataricus	No Data	10741831
409	toxicity	D064420	17673945

563 rows × 3 columns

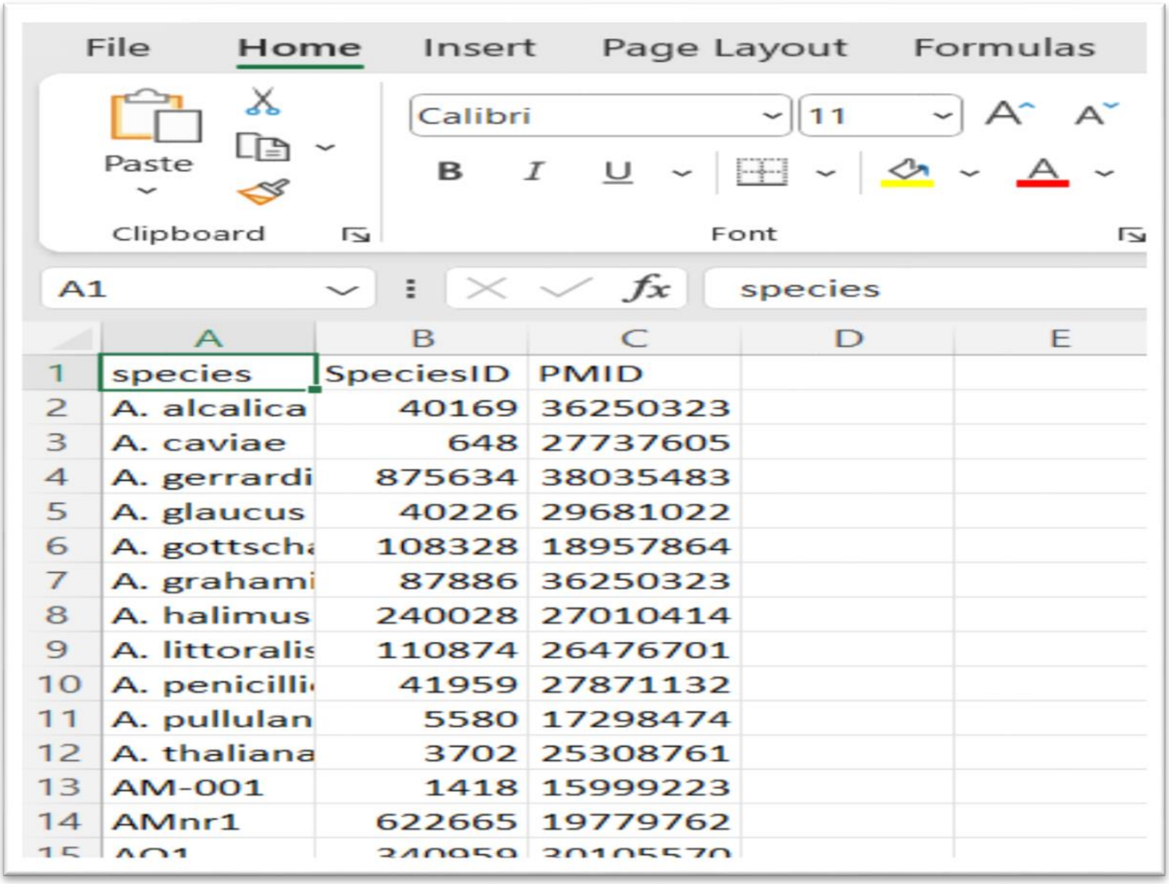
Outputs are:-For Chemical



```
df.drop(columns=['ChemicalID'], inplace=True)

df = df[['Chemicals', 'RightPart', 'PMID']]
print(df)

# Print or use the DataFrame with the new column\
df.to_csv('chemical_updated.csv', index=False)
```

```
               Chemicals RightPart      PMID
0                No Data      None  30796503
1               glycerol   D005990  37118007
1     dimethyl sulfoxide   D004121  37118007
2                No Data      None  15046570
4          phospholipids   D010743  28007654
..                   ...       ...       ...
409                    L   D007930  17673945
409          lanthanides   D028581  17673945
409                metal   D008670  17673945
409                   Fe   D007501  17673945
409                   L-   D007930  17673945

[1527 rows x 3 columns]
<ipython-input-13-f6acb6597148>:14: FutureWarning: In a future version of par
  df[['Chemicals', 'ChemicalID']] = df['Chemicals'].str.split('>', 1, expand=
```

Outputs are:-For Species



	A	B	C	D	E
1	species	SpeciesID	PMID		
2	A. alcalica	40169	36250323		
3	A. caviae	648	27737605		
4	A. gerrardi	875634	38035483		
5	A. glaucus	40226	29681022		
6	A. gottscha	108328	18957864		
7	A. grahami	87886	36250323		
8	A. halimus	240028	27010414		
9	A. littoralis	110874	26476701		
10	A. penicilli	41959	27871132		
11	A. pullulan	5580	17298474		
12	A. thaliana	3702	25308761		
13	AM-001	1418	15999223		
14	AMnr1	622665	19779762		
15	AO1	340959	30105570		

Week 5:-

Genes	GeneIDs	PMID
AP	13916924	12072958
Caspase-1	834	22295871
Chl	91851	31679078
Ikka	1147	16770690
Ikka	1147	27900683
L-1	16728	28045976
NapB	63908	19050822
Ndh-2	1660	37577439
ORF-1	1115973	11057908
ORF-4	1115988	11057908
PEP	828706	31338597
PH	5053	27276261
PIP1	856754	34256694
PRB	10536	36916005
Rhbg	57127	36250323
TAK	1025	34093984
acid 1	81857	16808526

Chemical	ChemicalI	PMID
glycerol	D005990	37118007
dimethyl s	D004121	37118007
phospholi	D010743	28007654
Phospholi	D010743	28007654
lipids	D008055	28007654
isoprenoi	D013729	28007654
glycerol-1	C029620	28007654
fatty acids	D005227	28007654
ester	D004952	28007654
glycerol-	D005990	28007654
lipid	D008055	28007654
Lipid	D008055	28007654
Lipids	D008055	28007654
water	D014867	29266533
salt	D012492	29266533
lipid	D008055	29266533
water	D014867	26780356
saline	D012965	29290045

species	SpeciesID	PMID
A. alcalica	40169	36250323
A. caviae	648	27737605
A. gerrard	875634	38035483
A. glaucus	40226	29681022
A. gottsch	108328	18957864
A. graham	87886	36250323
A. halimu	240028	27010414
A. littorali	110874	26476701
A. penicill	41959	27871132
A. pullulai	5580	17298474
A. thalian	3702	25308761
AM-001	1418	15999223
AMnr1	622665	19779762
AO1	340959	30105570
ATCC 4309	13769	22559199
ATCC BAA-	159292	12728359
Acacia ge	875634	38035483
Acinetoba	466088	26647770

Diseases	DiseasesI	PMID
extremopl	D019965	15046570
Extremopl	D019965	37490857
Stress	D0000792	37490857
haloalkali	C537702	32323057
dryness	D014987	26733008
Alkaliphili	D006934	37367588
Alkaliphili	D006934	37367588
alkaliphili	D015163	30796504
exception	C537702	8688447
Extremopl	614025	35688350
cold activ	D0000673	35688350
alkaliphili	614025	35688350
extremopl	614025	35688350
toxicity	D064420	37317247
stress	D0000792	33643258
alkalitheri	C537702	9783173
alkaliphili	D000881	9783164
cancer	D009369	31734456

So bsaically what I did in this week is that I just updated the files like there are few entreis in the both file which has no PMID and there were few entries like there were no specific domain name like Diseases and Species and Gene and chemical there were just PMID so I just remove them from the entries.

Week 6:-

So in this week I tranform the data in the forms of group the reason there were some chemical which hade similar PMID and there were few Diseases and species also which hade similar PMID and the difrent name so I basically merege them based on the names and applied group by on it and

Chemical:	ChemicalI	Length	PMID
S)-(+)-1-|	C033198	1	[32418069]
S)-amine	D000588	1	[32418069]
1-10-phen	C025205	1	[12910392]
1-aminoc	D002264	1	[37474779]
1-butanol	D020001	1	[12382117]
1-butyl-3-i	C502841	1	[27142029]
1-butyl-3-i	C532403	1	[27142029]
1-ethyl-3-r	C556629	1	[27142029]
1-hydroxyl	C011852	1	[32617733]
12-methyl	C069642	1	[10972188]
2-2'-azino	C002502	1	[12892493]
2-2'-bipyri	D015082	1	[12910392]
2-3-butan	C026978	1	[28425950]
2-4-D	D015084	1	[11778838]
2-4-diami	C005959	1	[28737704]
2-4-dichlc	D015084	1	[11778838]
2-methyl-4	D008456	1	[11778838]
2-propanc	D019840	1	[32617734]
20-60 C	C069837	1	[37847305]
3-amino-1	D000640	1	[12892493]
3-chlorob	C036427	1	[11778838]
30-60 C	C069837	1	[38010865]
5'-GMP	D006157	1	[28764042]
5'-IMP	D007291	1	[28764042]
5-HT	D012701	1	[12932132]

Diseases	DiseasesI	Length	PMID
Alcolapia	D011507	1	[36250323]
Alcolapia	D018457	1	[36250323]
Alkaliphili	D000881	1	[9783168]
Alkaliphili	D006934	1	[37367588]
Alkaliphili	C537702	1	[26090360]
Alkaliphili	D019965	1	[30457468]
Alkaliphili	D006934	1	[37367588]
Antarctic I	D003424	1	[34228196]
Antarctic I	C537702	1	[33255932]
Antarctic I	D018459	1	[30282060]
C3 extrem	C565169	1	[37667571]
CVDs	D002318	1	[33208066]
Cancer	D009369	2	[22295871, 16808526]
Cardiovas	D002318	1	[33208066]
Chromobl	D002862	1	[29538737]
Cold	D0000673	3	[27900683, 32833498, 27209
CotB anch	C537277	1	[26026992]
Death	D003643	1	[26543264]
Extremoph	D0000792	1	[28418707]
Extremoph	D000193	1	[33977442]
Extremoph	D054882	1	[34458243]
Extremoph	D002181	1	[26859958]
Extremoph	D000193	1	[33977442]
Extremoph	614025	1	[35688350]

Genes	GeneIDs	Length	PMID
AP	13916924	1	[12072958]
Caspase-1	834	1	[22295871]
Chl	91851	1	[31679078]
Ikka	1147	2	[16770690, 27900683]
L-1	16728	1	[28045976]
NapB	63908	1	[19050822]
Ndh-2	1660	1	[37577439]
ORF-1	1115973	1	[11057908]
ORF-4	1115988	1	[11057908]
PEP	828706	1	[31338597]
PH	5053	1	[27276261]
PIP1	856754	1	[34256694]
PRB	10536	1	[36916005]
Rhbg	57127	1	[36250323]
TAK	1025	1	[34093984]
acid 1	81857	1	[16808526]
alkaline p	13916924	2	[22212656, 12072958]
caspase-1	834	1	[22295871]
fog	161882	1	[24927538]
hMDH	4191	1	[12382117]
interleukin	3553	1	[22295871]
ml-1	16728	1	[28045976]

species	SpeciesID	Length	PMID
A. alcalica	40169	1	[36250323]
A. caviae	648	1	[27737605]
A. gerrard	875634	1	[38035483]
A. glaucus	40226	1	[29681022]
A. gottsch	108328	1	[18957864]
A. graham	87886	1	[36250323]
A. halimus	240028	1	[27010414]
A. littorali	110874	1	[26476701]
A. penicill	41959	1	[27871132]
A. pullulai	5580	1	[17298474]
A. thaliana	3702	1	[25308761]
AM-001	1418	1	[15999223]
AMnr1	622665	1	[19779762]
AO1	340959	1	[30105570]
ATCC 4309	13769	1	[22559199]
ATCC BAA-	159292	1	[12728359]
Acacia gei	875634	1	[38035483]
Acinetoba	466088	1	[26647770]
Acinetoba	470	1	[36094301]
Acinetoba	472	1	[33645540]
Acinetoba	466088	2	[33645540, 30485446]
Aeluropus	110874	1	[26476701]
Aeromona	648	1	[27737605]
Agarivorai	1872412	1	[19002649]

Week 7:-

Task Summary: Extracting and Mapping Gene and Species Information with Corresponding Sentences from PubTator Data

Objective:

The primary objective this week was to enhance the extracted information from the PubTator output by fetching specific sentences related to genes and species from the corresponding PubMed articles. This involved mapping PubMed IDs (PMIDs) to PubMed Central IDs (PMCIDs) and extracting relevant sentences from the articles.

Process:

1. Input Files Preparation:

The previous week's output consisted of a CSV file with columns: Gene, GeneID, Length, and PMID.

This file needed to be processed further to include the corresponding sentences from PubMed articles that mention the specific genes and species.

2. Mapping PMIDs to PMCIDs:

Using the initial CSV file, each PMID was mapped to its corresponding PMCID. This mapping is crucial as PMCIDs are required to fetch full-text articles from PubMed Central, which contain the sentences of interest.

3. Fetching Sentences:

A custom function, `give_sentence`, was employed to extract sentences from full-text articles that mention the particular genes and species. This function likely utilized the mapped PMCIDs to access the articles and retrieve the relevant text.

The sentences were then associated with the specific Gene and Species entries.

4. Creating the Output CSV:

The final output was structured into a new CSV file that included the Gene, GeneID, Length, PMID, PMCID, and the extracted sentences.

This new CSV file provided a comprehensive dataset linking gene and species mentions to specific sentences in the corresponding PubMed articles, facilitating further analysis and research.

Outcome:

The resultant CSV file now contains detailed information, including the original Gene, GeneID, Length, PMID, and the newly added PMCID and relevant sentences.

Genes	GeneIDs	PMID	PMCID
Caspase-1	834	22295871	PMC3330824
caspase-1	834	22295871	PMC3330824
interleukir	3553	22295871	PMC3330824
L-1	16728	28045976	PMC5207672
mL-1	16728	28045976	PMC5207672
Ndh-2	1660	37577439	PMC1041664
PH	5053	27276261	PMC4906265
PIP1	856754	34256694	PMC8278772
PRB	10536	36916005	PMC1011134
protein B	10536	36916005	PMC1011134
Rhbg	57127	36250323	PMC9672858
TAK	1025	34093984	PMC8148631
fog	161882	24927538	PMC4156692
osteocalc	12097	30400922	PMC6220464
spa	653509	24430481	PMC4030231

species	SpeciesID	PMID	PMCID
A. alcalica	40169	36250323	PMC9672858
A. graham	87886	36250323	PMC9672858
Alcolapia	87886	36250323	PMC9672858
AMnr1	622665	19779762	PMC2797408
Bacillus a	85682	19779762	PMC2797408
Bacillus s	622665	19779762	PMC2797408
enrichmei	1566338	19779762	PMC2797408
enrichmei	1566338	19779762	PMC2797408
ATCC 4309	13769	22559199	PMC3403918
Natrialba	13769	22559199	PMC3403918
Natrialba	547559	22559199	PMC3403918
Acinetoba	470	36094301	PMC9602519
Clostridiu	1294142	36094301	PMC9602519
Pseudomo	208964	36094301	PMC9602519
mammali	9606	36094301	PMC9602519
Alkalibaci	1193119	22887673	PMC3415526
Anditalea	1048983	26171779	PMC4501810

Diseases	DiseasesI	PMID	PMCID
dryness	D014987	26733008	4679917
Alkaliphili	D006934	37367588	10301932
Alkaliphili	D006934	37367588	10301932
toxicity	D064420	37317247	10223213
stress	D0000792;	33643258	7902512
Extremopl	D015163	37839067	10577106
Toxicity	D064420	28737704	5532675
paralytic	D007418	28737704	5532675
Extremopl	D054882	34458243	8387880
low toxici	D009800	34458243	8387880
extremop	D054882	34458243	8387880
halo-alka	D055882	37982082	10651602
extremop	D00007106	37982082	10651602
dryness	D014987	36414646	9681764
Alkaliphil	C537702	26090360	4453477
alkaliphil	C537702	26090360	4453477
new alkal	C0006572	36581887	9798632

A1 Chemicals

Chemical	ChemicalI	PMID	PMCID
water	D014867	32533304	PMC7347518
arsenic	D001151	26733008	PMC4679917
salts	D012492	26733008	PMC4679917
O2	D010100	26733008	PMC4679917
ozone	D010126	26733008	PMC4679917
PacC	C406277	37367588	PMC10301932
poly	C017937	37317247	PMC10223213
ammoniu	D000645	31388592	PMC6667821
Ba2+	C080430	31388592	PMC6667821
Ca2+	D00006928	31388592	PMC6667821
EDTA	D004492	31388592	PMC6667821
N-(2-amir	C028791	28737704	PMC5532675
2-4-diami	C005959	28737704	PMC5532675
#NAME?	C089595	28737704	PMC5532675
microcyst	D052998	28737704	PMC5532675
nodularin	C063998	28737704	PMC5532675
Polymer	D011108	31087781	PMC6828557
poly(vinyl	D011142	31087781	PMC6828557

Week 8:-

Week 8 Summary: Extracting Sentences for Genes and Species Using PMCID Mapping

Objective:

The goal for this week was to enhance the dataset by including specific sentences from PubMed articles that mention particular genes and species. This required mapping PMIDs to PMCIDs and using these IDs to extract relevant sentences.

Process:

1. Input File Preparation:

The input file contained columns for the name, PMID, and corresponding PMCID.

This file served as the basis for retrieving specific sentences related to the genes and species of interest.

2. Mapping PMIDs to PMCIDs:

Each PMID in the input file was mapped to its corresponding PMCID, enabling access to the full-text articles available in PubMed Central.

This mapping was essential to facilitate the extraction of detailed textual information from the articles.

3. Sentence Extraction Using give_sentences Function:

The give_sentences function was utilized to extract sentences from the full-text articles based on the PMCID.

This function scanned the articles for mentions of the specific genes and species, extracting and compiling relevant sentences.

4. Creating the Enhanced Output File:

The output file was generated to include the original columns (name, PMID, PMCID) along with the extracted sentences.

This new CSV file provided a comprehensive dataset with contextual information, linking each gene and species to specific sentences in the articles.

Outcome:

The final output file now includes detailed sentences from PubMed articles, providing valuable context for each gene and species mention.

A	B	C	D	E	F	G
Chemical:	Chemicals	PMID	PMCID	Sentences		
water	D014867	32533304	7347518	While many studies have in		
arsenic	D001151	26733008	4679917	HAAL proved to be a rich sou		
salts	D012492	26733008	4679917	HAAL proved to be a rich sou		
O2	D010100	26733008	4679917	The modern stromatolites a		
ozone	D010126	26733008	4679917	The modern stromatolites a		
PacC	C406277	37367588	10301932	In both biological models, t		
poly	C017937	37317247	10223213	Therefore, extremophilic mi		
ammoniu	D000645	31388592	6667821	After precipitation using am		
Ba2+	C080430	31388592	6667821			
Ca2+	D000069288	31388592	6667821			
EDTA	D004492	31388592	6667821	The activity completely dimi		
N-(2-amir	C028791	28737704	5532675			
2-4-diami	C005959	28737704	5532675			
#NAME?	C089595	28737704	5532675			
microcyst	D052998	28737704	5532675	Cyanotoxins detected in the		
nodularin	C063998	28737704	5532675	Cyanotoxins detected in the		
Polymer	D011108	31087781	6828557	This also explains why man		
poly(vinyl	D011142	31087781	6828557			

Diseases

A	B	C	D	E
iseases	DiseasesI	PMID	PMCID	Sentences
ryness	D014987	26733008	4679917	HAAL proved t
lkaliphili	D006934	37367588	10301932	
lkaliphili	D006934	37367588	10301932	Alkaliphilic ar
oxicity	D064420	37317247	10223213	This review pr
tress	D0000792	33643258	7902512	These organis
xtremopl	D015163	37839067	10577106	
oxicity	D064420	28737704	5532675	
aralytic	D007418	28737704	5532675	Cyanotoxins d
xtremopl	D054882	34458243	8387880	
ow toxici	D009800	34458243	8387880	Compared to
xtremop	D054882	34458243	8387880	In this article
alo-alka	D055882	37982082	10651602	
xtremop	D0000710	37982082	10651602	
ryness	D014987	36414646	9681764	Salt-in strateg
lkaliphil	C537702	26090360	4453477	Alkaliphilic ba
lkaliphil	C537702	26090360	4453477	Each of these
ew alkal	C00065724	36581887	9798632	A set of new al

Genes	GeneIDs	PMID	PMCID	Sentences
Caspase-1	834	22295871	3330824	Caspase-1
caspase-1	834	22295871	3330824	The comp
interleukii	3553	22295871	3330824	
L-1	16728	28045976	5207672	The lake
mL-1	16728	28045976	5207672	
Ndh-2	1660	37577439	10416648	Within this
PH	5053	27276261	4906265	This work
PIP1	856754	34256694	8278772	Furthermo
PRB	10536	36916005	10111349	
protein B	10536	36916005	10111349	
Rhbg	57127	36250323	9672858	Neverthele
TAK	1025	34093984	8148631	The
fog	161882	24927538	4156692	Studies of
osteocalc	12097	30400922	6220464	Furthermo
spa	653509	24430481	4030231	Studies ha
tan	10482	32533304	7347518	While mar

species	SpeciesID	PMID	PMCID	Sentences
A. alcalica	40169	36250323	9672858	Using in situ hyb
A. graham	87886	36250323	9672858	In contrast, the
Alcolapia	87886	36250323	9672858	Comparing amn
AMnr1	622665	19779762	2797408	Bacillus
Bacillus a	85682	19779762	2797408	A pure culture o
Bacillus s	622665	19779762	2797408	Bacillus sp. stra
enrichmer	1566338	19779762	2797408	A single, stable
enrichmer	1566338	19779762	2797408	A single, stable
ATCC 4309	13769	22559199	3403918	The genome seq
Natrialba	13769	22559199	3403918	Natrialba magac
Natrialba	547559	22559199	3403918	B. Synteny plot o
Acinetoba	470	36094301	9602519	The peptide sho
Clostridiu	1294142	36094301	9602519	This work report
Pseudomo	208964	36094301	9602519	The peptide sho
mammali	9606	36094301	9602519	The activity of In
Alkalibaci	1193119	22887673	3415526	Alkalibacillus ha
Anditalea	1048983	26171779	4501810	Here, we report

Week 9:-

Week 9 Summary: Research on Alkaline-Adapted Extremophiles

Objective:

The task for this week involved conducting research to gather detailed information about alkaline-adapted extremophiles. The goal was to compile a comprehensive dataset that includes the names, descriptions, publication links, and year of publication of relevant studies.

Process:

1. Literature Search:

Conducted an extensive search of scientific databases to identify research articles and publications related to alkaline-adapted extremophiles.

Utilized keywords such as "alkaline extremophiles," "alkaline environments," and "alkaliphilic organisms" to find relevant studies.

2. Data Compilation:

Collected the names of identified alkaline-adapted extremophiles.

Summarized the descriptions of these extremophiles, focusing on their unique adaptations and ecological significance.

Recorded the publication links for each study, ensuring easy access to the full text of the articles.

Noted the year of publication for each study to provide a temporal context for the research.

3.Creating the Output File:

Compiled the gathered information into a structured format, creating a comprehensive dataset.

The output file included the following columns: Name, Description, Publication Link, and Year of Publication.

Ensured accuracy and completeness of the data, providing a reliable resource for further research.

Outcome:

The resulting dataset offers a detailed overview of alkaline-adapted extremophiles, including their names, descriptions, publication links, and years of publication

A	B	C	D	E
Name	Description	Link	PublicationLink	yearOfPublication
Alkaliphile Gen	This database specifically f	https://www.ncbi.nlm.nih.gov/pmc/a	https://www.ncbi.nlm.nih.g	2014 Nov
Alkaline Enzyme	This database catalogs alka	https://www.ncbi.nlm.nih.gov/pmc/a	https://www.ncbi.nlm.nih.g	**2019 Dec**
Alkali-Resistant	This database compiles info	https://www.ncbi.nlm.nih.gov/pmc/a	https://www.ncbi.nlm.nih.g	2007
ALKATLAS	ALKATLAS is a database that	https://www.ncbi.nlm.nih.gov/pmc/a	https://www.ncbi.nlm.nih.g	2002 Apr
ALKGENDB	ALKGENDB is a repository of	https://www.ncbi.nlm.nih.gov/pmc/a	https://www.ncbi.nlm.nih.g	2003 Feb

Week 10:-

Week 10 Summary: Examination of UniProt Data for Insights into Genes and Proteins**

Objective:

The task for this week focused on examining UniProt data to gain crucial insights into the genes and proteins relevant to the study of alkaline-adapted extremophiles. The goal was to understand their potential roles in viral assembly, replication, and the infection process.

Process:

1. Data Retrieval from UniProt:

Accessed the UniProt database to retrieve detailed information about genes and proteins associated with alkaline-adapted extremophiles.

Utilized specific search criteria and filters to identify relevant entries within the UniProt database.

2. Data Analysis:

Analyzed the retrieved UniProt data to identify genes and proteins with potential roles in viral assembly, replication, and the infection process.

- Examined protein functions, domains, and interactions to understand their involvement in these processes.

3. Compilation of Insights:

Compiled comprehensive insights regarding the roles of identified genes and proteins.

Highlighted key findings about their contributions to molecular mechanisms and pathogenicity associated with alkaline environments.

4. Documentation and Reporting:

Documented the findings in a structured report, detailing the potential roles of each gene and protein.

Provided a clear and concise summary of the insights gained from the UniProt data, emphasizing their relevance to further research.

Outcome:

The examination of UniProt data has furnished crucial insights into the genes and proteins pertinent to the study of alkaline-adapted extremophiles. This data offers valuable understanding regarding their possible roles in viral assembly, replication, and the infection process.

Entry	Entry Name	Gene Nam	GeneID	Length	PubMed ID	Protein names
O28523	APGM1_A	apgM1 AF_	24795495;	408	9389475; 1	2,3-bisphosphoglycerate-independent phosphoglycerate mutase 1 (BPG-independent PGAM 1) (Phosphoglyceromutase 1) (aPGAM 1) (EC 5.4.2.12)
Q5WD76	Q5WD76_S	prs ABC31		317	7632397;	Ribose-phosphate pyrophosphokinase (RPPK) (EC 2.7.6.1) (5-phospho-D-ribosyl alpha-1-diphosphate synthase) (Phosphoribosyl diphosphate synthase) (Phosp
Q5WFG5	Q5WFG5_S	murE ABC		485	7632397;	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--2,6-diaminopimelate ligase (EC 6.3.2.13) (Meso-A2pm-adding enzyme) (Meso-diaminopimelate-adding enzyme
Q5WFJ2	Q5WFJ2_S	pyrAB carE		1062	7632397;	Carbamoyl phosphate synthase large chain (EC 6.3.4.16) (EC 6.3.5.5) (Carbamoyl phosphate synthetase ammonia chain)
Q5WFK5	Q5WFK5_S	coaBC ABC		401	7632397;	Coenzyme A biosynthesis bifunctional protein CoaBC (DNA/pantothenate metabolism flavoprotein) (Phosphopantothenoylcysteine synthetase/decarboxylase) (
Q5WFV4	Q5WFV4_S	asd ABC22		350	7632397;	Aspartate-semialdehyde dehydrogenase (ASA dehydrogenase) (ASADH) (EC 1.2.1.11) (Aspartate-beta-semialdehyde dehydrogenase)
Q5WL63	Q5WL63_S	hmp ABC0		411	7632397;	Flavohemoprotein (Flavohemoglobin) (Hemoglobin-like protein) (Nitric oxide dioxygenase) (NO oxygenase) (NOD) (EC 1.14.12.17)
Q5WLC1	Q5WLC1_S	nnrE nnrD		511	7632397;	Bifunctional NAD(P)H-hydrate repair enzyme (Nicotinamide nucleotide repair protein) [Includes: ADP-dependent (S)-NAD(P)H-hydrate dehydratase (EC 4.2.1.1
Q5WLV6	Q5WLV6_S	ftsH ABC0		662	7632397;	ATP-dependent zinc metalloprotease FtsH (EC 3.4.24.-)
A0A430QV	A0A430QV	deoB CSW		380	30536130	Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)
A0A430R0	A0A430R0	deoB CSW		380	30536130	Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)
A0A430RH	A0A430RH	deoB CSW		380	30536130	Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)
A0A430RK	A0A430RK	deoB CSW		380	30536130	Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)
A0A430S5	A0A430S5	deoB CSW		380	30536130	Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)
A0A430UE	A0A430UE	deoB CSW		380	30536130	Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)
A0A430UK	A0A430UK	deoB CSW		380	30536130	Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)
A0A430UT	A0A430UT	deoB CSW		380	30536130	Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)
A0A430UV	A0A430UV	deoB CSW		380	30536130	Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)
A0A430V9	A0A430V9	deoB CSW		380	30536130	Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)
A0A7R6PX	A0A7R6PX	gpmI TTHT		505	22212657	2,3-bisphosphoglycerate-independent phosphoglycerate mutase (BPG-independent PGAM) (Phosphoglyceromutase) (iPGM) (EC 5.4.2.12)
A0A7R6SX	A0A7R6SX	deoB TTHT		391	22212657	Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)
A0A7Z0LS	A0A7Z0LS	deoB HZS8		414	12579380	Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)
A0A7Z0LT	A0A7Z0LT	gpmI HZS8		525	12579380	2,3-bisphosphoglycerate-independent phosphoglycerate mutase (BPG-independent PGAM) (Phosphoglyceromutase) (iPGM) (EC 5.4.2.12)
A0A838CN	A0A838CN	deoB H026		392	15064986	Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)
A0A838CU	A0A838CU	gpmI H026		512	15064986	2,3-bisphosphoglycerate-independent phosphoglycerate mutase (BPG-independent PGAM) (Phosphoglyceromutase) (iPGM) (EC 5.4.2.12)
B1YLD9	B1YLD9_E	gpmI Exig_		515	16489412;	2,3-bisphosphoglycerate-independent phosphoglycerate mutase (BPG-independent PGAM) (Phosphoglyceromutase) (iPGM) (EC 5.4.2.12)
C6X5T7	C6X5T7_Fl	gpmI FIC_0		509	18622572	2,3-bisphosphoglycerate-independent phosphoglycerate mutase (BPG-independent PGAM) (Phosphoglyceromutase) (iPGM) (EC 5.4.2.12)
F5L4L9	F5L4L9_CA	gpmI Cath		512	21685297;	2,3-bisphosphoglycerate-independent phosphoglycerate mutase (BPG-independent PGAM) (Phosphoglyceromutase) (iPGM) (EC 5.4.2.12)
F5L874	F5L874_CA	deoB Cath		400	21685297;	Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)
F7PLH2	F7PLH2_9	gpmI HLRT	23799984;	505	21705593;	2,3-bisphosphoglycerate-independent phosphoglycerate mutase (BPG-independent PGAM) (Phosphoglyceromutase) (iPGM) (EC 5.4.2.12)

Week 11:-

Week 11 Summary: Research on the Prevalence and Geographic Spread of Extremophiles

Objective:

The task for this week focused on gathering and analyzing data on the prevalence and geographic distribution of extremophiles across various regions and historical epochs. The aim was to uncover the narrative of their adaptation to challenging environments, particularly in relation to varying radiation exposures.

Process:

1. Data Collection:Collected information on the geographic locations and time periods where extremophiles have been identified.

2. Analysis of Adaptation:

Analyzed the collected data to understand how extremophiles have adapted to different environmental challenges, with a focus on varying levels of radiation exposure.

Investigated the mechanisms that enable these organisms to survive and thrive in high-radiation environments.

3. Geographic and Temporal Mapping:

Mapped the prevalence and distribution of extremophiles across diverse regions and historical epochs.

Identified patterns and trends in their geographic spread and adaptation over time.

4. Narrative Construction:

Highlighted key findings related to their resilience and survival strategies in the face of varying radiation exposures.

Outcome:

The gathered data on the prevalence and geographic spread of extremophiles across diverse regions and historical epochs reveals a captivating narrative of adaptation to challenging environments shaped by varying radiation exposures. This research provides valuable insights into the evolutionary processes and survival mechanisms of extremophiles, contributing to a deeper understanding of their ecological and biological significance.

Location	Epidemic/Panc	Description	Link	Publication Link	Year Of Publication
Calumet region in southe		Mark Twain or	https://se	https://serc.carlet	1999
Alkaline Environments (b		Alkaline enviro	https://wv	https://www.ncbi.	2023 Jun 9
Extreme environments wi		This overview	https://aq	https://aquaticbio	Sep-09
Egypt		This minireviev	https://wv	https://www.ncbi.	2012 june
USA San Francisco		The passage de	https://sp	https://spaceref.c	May/June 2002

Week 12:- Task8

Summary of Findings: Proteins and Genes Vital for Alkaline Extremophiles

Objective:

To identify proteins and genes critical for the survival of extremophiles in alkaline environments.

Process:

Collected and analyzed data on proteins and genes that enable extremophiles to thrive in highly alkaline conditions.

Focused on adaptations that facilitate survival in these extreme environments

Outcome:

The gathered information highlights key proteins and genes that are pivotal for survival in alkaline environments, detailing specific adaptations that allow extremophiles to endure and thrive under such demanding conditions.

B5 Alkaline shock protein A

Gene	Protein	Description	Link	Publication Link
phrB	Alkaline phosphatase	Alkaline phospha	https://pu	https://www.ncbi.nlm.nih.gov/gene/?term=phrB
cbbM	RuBisCO (Ribulose-1,5-	RuBisCO is a key	https://wv	https://pubmed.ncbi.nlm.nih.gov/?term=RuBisCO+extremophiles
lipA	Alkaline lipase	Alkaline lipases a	https://wv	https://pubmed.ncbi.nlm.nih.gov/?term=Alkaline+lipase+extremophiles
hspA	Alkaline shock protein A		https://wv	https://pubmed.ncbi.nlm.nih.gov/?term=Alkaline+shock+protein+extremophiles

Week 13:-

Summary of Findings: Chemicals and Pharmaceutical Compounds

Objective:

To compile a comprehensive dataset of chemicals and pharmaceutical compounds, including their names, IDs, references, stages of clinical trials, and other pertinent information.

Process:

Gathered data on various pharmaceutical compounds from relevant sources.

Collected detailed information including names, unique IDs, references, and clinical trial stages.

Outcome:

The dataset encompasses a variety of pharmaceutical compounds, providing comprehensive details such as names, IDs, references, and stages of clinical trials. This dataset serves as a valuable resource for further research and analysis in the pharmaceutical field.

Chemical Name	Chemical ID	Refrence	Phase Of Trial
Strontium	DB13987	https://go.drugbank	Approved
Magnesium cation	DB01378	https://go.drugbank	Approved, Nutraceutical
Barium sulfate	DB11150	https://go.drugbank	Approved
Magnesium hydroxide	DB09104	https://go.drugbank	Approved, Investigational
Magnesium trisilicate	DB09281	https://go.drugbank	Approved
Potassium citrate	DB09125	https://go.drugbank	Approved, Investigational, Vet approved
Potassium bicarbonate	DB11098	https://go.drugbank	Approved
Methenamine	DB06799	https://go.drugbank	Approved, Vet approved
Morphine	DB00295	https://go.drugbank	Approved, Investigational
Sulfametopyrazine	DB00664	https://go.drugbank	Approved, Withdrawn
Alginic acid	DB13518	https://go.drugbank	Approved, Investigational

Week 14:-

Week 14 Summary: Synonym Clubbing for Gene, Species, Disease, and Chemical IDs

Objective:

To create a script that consolidates synonyms for Gene, Species, Disease, and Chemical IDs, organizing them into lists for each respective ID.

Process:

Developed a script to identify and compile synonyms associated with each Gene, Species, Disease, and Chemical ID.

Processed the data to ensure that all synonyms are accurately listed in front of their corresponding IDs.

Submitted Files:

1. Task10_Clubbed_Species.csv: Contains Species IDs and their respective synonyms.

2. Task10_Clubbed_Gene.csv: Contains Gene IDs and their respective synonyms.

3. Task10_Clubbed_Disease.csv: Contains Disease IDs and their respective synonyms.

4. Task10_Clubbed_Chemicals.csv: Contains Chemical IDs and their respective synonyms.

Outcome:

The script successfully created comprehensive lists of synonyms for each Gene, Species, Disease, and Chemical ID. The resultant files provide a consolidated view of all relevant synonyms, facilitating easier reference and analysis.

species	SpeciesID	Length	PMID
A. alcalica	40169	2	36250323, 26547282
A. caviae, /	648	2	27737605, 27737605
A. gerrardi	875634	2	38035483, 38035483
A. glaucus	40226	1	29681022
A. gottscha	108328	2	18957864, 18957864
A. grahami	87886	2	36250323, 36250323
A. halimus	240028	3	27010414, 27010414, 27010414
A. littoralis	110874	2	26476701, 26476701
A. penicilli	41959	2	27871132, 27871132
A. pullulan	5580	3	17298474, 17298474, 30400922
A. thaliana	3702	9	25308761, 31338597, 24214268, 34256694, 25496221, 31781937, 33030592, 25308761, 313
AM-001, B	1418	3	15999223, 15999223, 17429572
AMnr1	622665	1	19779762
AO1, Bacil	340959	2	30105570, 30105570
ATCC 4309	13769	4	22559199, 21894491, 21894491, 22559199
ATCC BAA-	159292	6	12728359, 12728359, 16932842, 12728359, 12728359, 16932842
Acinetobac	466088	3	26647770, 33645540, 30485446
Acinetobac	470	1	36094301
Acinetobac	472	1	33645540
Agarivoran	1872412	1	19002649
Agarivoran	507618	1	19002649
Agromyces	758919	1	24817611
Alicycloba	61169	2	30656425, 30656425
Alkalibacil	1193119	1	22887673
Alkalibacte	235931	1	15127306
Alkalibacte	1581170	1	27362528
[illegible]	[illegible]	2	[illegible]

Genes	GeneIDs	Length	PMID
AP, alkalin	13916924	3	12072958, 22212656, 12072958
Caspase-1	834	2	22295871, 22295871
Chl	91851	1	31679078
Ikka	1147	2	16770690, 27900683
L-1, mL-1	16728	2	28045976, 28045976
NapB	63908	1	19050822
Ndh-2	1660	1	37577439
ORF-1	1115973	1	11057908
ORF-4	1115988	1	11057908
PEP	828706	1	31338597
PH, pH	5053	2	27276261, 26025020
PIP1	856754	1	34256694
PRB, prote	10536	2	36916005, 36916005
Rhbg	57127	1	36250323
TAK	1025	1	34093984
acid 1	81857	1	16808526
fog	161882	1	24927538
hMDH	4191	1	12382117
interleukin	3553	1	22295871
neuraminic	4758	1	33977442
osteocalci	12097	1	30400922
sea	6395	1	20091326
spa	653509	1	24430481
tap	10482	1	32533304
vma3	856686	1	9783169

iseases	DiseasesI[	Length	PMID
lcolapia a	D011507	1	36250323
lcolapia r	D018457	1	36250323
lkaliphilic	D000881	9	9783168, 9783164, 9680303, 10805564, 19411423, 11878564, 10972188, 9672682, 11057913
lkaliphilic	D006934	3	37367588, 35661272, 27737605
lkaliphilic	C537702	4	26090360, 33255932, 28478604, 37847305
lkaliphilic	D019965	3	30457468, 37490857, 15046570
lkaliphilic	D006934	3	37367588, 20703955, 33538376
ntarctic F	D003424	1	34228196
ntarctic li	D018459	1	30282060
3 extrem	C565169	1	37667571
VDs, Car	D002318	2	33208066, 33208066
ancer, ca	D009369	3	22295871, 16808526, 31734456
hromobla	D002862	1	29538737
old, cold,	D000067 3	10	27900683, 32833498, 27209523, 27900683, 33925342, 22297696, 20373120, 35688350, 16770690, 16642262
otB anch	C537277	1	26026992
eath	D003643	1	26543264
xtremoph	D0000792	22	28418707, 37490857, 28087527, 26909467, 34681129, 34256694, 31722266, 35190857, 38252174, 25308761, 35742848,
xtremoph	D000193	3	33977442, 22327111, 33404954
xtremoph	D054882	1	34458243
xtremoph	D002181	1	26859958
xtremoph	D000193	3	33977442, 30863668, 22327111
xtremoph	614025	3	35688350, 35688350, 35688350
xtremoph	D015163	2	37839067, 30796504
(1)F	102510	1	21600188
PD	D003922	1	36094301
lebsiella,	D007710	2	15064989, 19002649

	A	B	C	D
1	Chemical	ChemicalI	length	PMID
2	(S)-(+)-1-p	C033198	1	32418069
3	(S)-amine,	D000588	2	32418069, 32418069
4	1-10-phen	C025205	1	12910392
5	1-aminocy	D002264	1	37474779
6	1-butanol	D020001	1	12382117
7	1-butyl-3-r	C502841	2	27142029, 27142029
8	1-butyl-3-r	C532403	1	27142029
9	1-ethyl-3-r	C556629	1	27142029
10	1-hydroxyb	C011852	1	32617733
11	12-methyl	C069642	1	10972188
12	2-2'-azino-	C002502	2	12892493, 24915287
13	2-2'-bipyric	D015082	1	12910392
14	2-3-butane	C026978	1	28425950
15	2-4-D, 2-4-	D015084	2	11778838, 11778838
16	2-4-diamir	C005959	1	28737704
17	2-methyl-4	D008456	1	11778838
18	2-propano	D019840	1	32617734
19	20-60 C, 3	C069837	2	37847305, 38010865
20	3-amino-1	D000640	1	12892493
21	3-chlorobe	C036427	1	11778838
22	5'-GMP	D006157	1	28764042
23	5'-IMP	D007291	1	28764042
24	5-HT	D012701	1	12932132
25	5-hydroxyd	C052853	1	28425950
26	6-carboxyf	C024098	1	9783169
27	6-methoxy	C080190	1	38035483

Week 15:-

Week 15 Summary: Interaction Analysis Between Genes and Chemicals

Objective:

To analyze and identify interactions between genes and chemicals, and determine their regulation and interaction linked to specific PMIDs.

Process:

Combined two CSV files containing data on genes and chemicals.

Matched interactions between genes and chemicals using the data.

Extracted information on the regulation and interaction associated with specific PubMed IDs (PMIDs).

Outcome:

The analysis successfully identified and documented interactions between genes and chemicals, including details on their regulation. The results were linked to specific PMIDs, providing a clear reference for further research and validation.

PMID	PMCID	Sentence	Genes	Chemicals	Interaction	Regulation
32418069		(S)-(+)-1-p	AP, alkalin	(S)-(+)-1-p	Activation	Up
32418069		(S)-(+)-1-p	Caspase-1	(S)-(+)-1-p	Activation	Down
32418069		(S)-(+)-1-p	Chl	(S)-(+)-1-p	Activation	Down
32418069		(S)-(+)-1-p	Ikka	(S)-(+)-1-p	Activation	Up
32418069		(S)-(+)-1-p	L-1, mL-1	(S)-(+)-1-p	Activation	Down
32418069		(S)-(+)-1-p	NapB	(S)-(+)-1-p	Activation	Down
32418069		(S)-(+)-1-p	Ndh-2	(S)-(+)-1-p	Activation	Up
32418069		(S)-(+)-1-p	ORF-1	(S)-(+)-1-p	Activation	Down
32418069		(S)-(+)-1-p	ORF-4	(S)-(+)-1-p	Activation	Down
32418069		(S)-(+)-1-p	PEP	(S)-(+)-1-p	Activation	Up
32418069		(S)-(+)-1-p	PH, pH	(S)-(+)-1-p	Activation	Up
32418069		(S)-(+)-1-p	PIP1	(S)-(+)-1-p	Activation	Up
32418069		(S)-(+)-1-p	PRB, prote	(S)-(+)-1-p	Activation	Down
32418069		(S)-(+)-1-p	Rhbg	(S)-(+)-1-p	Activation	Down
32418069		(S)-(+)-1-p	TAK	(S)-(+)-1-p	Activation	Up
32418069		(S)-(+)-1-p	acid 1	(S)-(+)-1-p	Activation	Down
32418069		(S)-(+)-1-p	fog	(S)-(+)-1-p	Activation	Up

Week 16:-

Week 16 Summary: 3D Graph Representation of Gene and Chemical Interactions

Objective:

To visually represent the interactions between genes and chemicals using a 3D graph, indicating the nature of each interaction (inhibition, exhibition, or other).

Process:

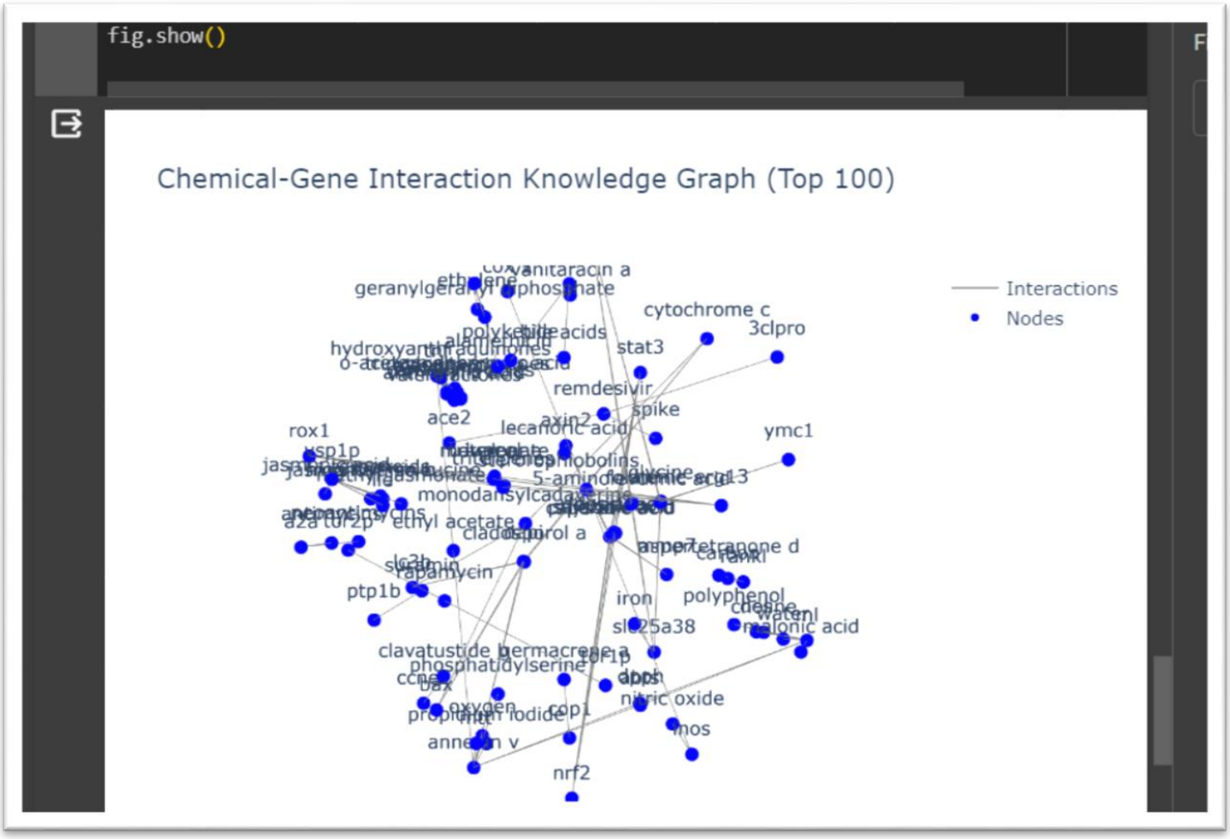
Mapped the interactions between genes and chemicals from the dataset.

Developed a 3D graph to illustrate these interactions.

Used edges in the graph to indicate the type of interaction (inhibition, exhibition, or other).

Outcome:

The 3D graph effectively represents the interactions between genes and chemicals. Each edge in the graph clearly indicates whether the interaction is inhibitory, exhibitory, or of another type, providing an intuitive visual tool for analyzing these relationships.



Week 17:-

 Updating in a graph show with the different colour like gene with the particular colour and then chemical with the particular colour