INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Life of Extremophiles: Database and Knowledge Graph of Microbes Living in Extreme Conditions (Acidic)

M.Tech. Minor Project Project Report(MIP591)

Submitted by- Dheeraj Pandey MT23034 M.Tech CSE	Under theSupervision of Dr. N. Arul Murugan Sir

Table of Contents

1. Project Description

Objective: The project involves the stability analysis of proteomes of organisms thriving in extreme conditions such as Acidic Environment.

Tasks: In this project, I have conducted extensive literature searches using PubMed, Europe PMC, and Google Scholar to gather research papers on extremophiles thriving in acidic environments. Utilizing tools like PubMed Miner (an R package), I analyze these papers to extract relevant information. Additionally, I explore databases to identify crucial data related to genes, chemicals, diseases, and species associated with these extremophiles. This comprehensive approach integrates insights from scientific papers and structured datasets to achieve a thorough understanding of extremophiles, their habitats, and associated characteristics.

2. Extremophiles Overview

Extremophiles are organisms that thrive in environments with extreme conditions, such as extreme temperatures, high radiation, high salinity, or unusual pH levels. These organisms exhibit remarkable adaptability to conditions that are typically lethal to most life forms. The deep sea, in particular, is an extreme environment characterized by the absence of sunlight, frigid temperatures, and immense hydrostatic pressures. Specific habitats within the deep sea, such as hydrothermal vents and deep hypersaline anoxic basins (DHABs), pose even more challenging conditions for survival.

3. Research and Data Collection

The project involved extensive research, primarily using databases such as PubMed, Europe PMC, and Google Scholar. Keywords used for searches included 'Life of extremophiles in Acidic Environment.'

Analysis: Acidic Environment Extremophiles :

Overview

- The dataset comprises bibliographic details of research articles centered on extremophiles in acidic environments and associated studies. It includes various columns such as PMID, Title, Authors, Citation, First Author, Journal/Book, Publication Year, Create Date, PMCID, NIHMS ID, and DOI.

  - The dataset comprises a total of 182 records.

Time Span of Publications:

- The publications range from the year 2008 to 2021.

Journals and Books:

- The research articles are published in various reputable journals and books, such as Science, Prog Retin Eye Res, Proc Biol Sci, Nat Prod Rep, and Adv Mar Biol.

Top Authors:

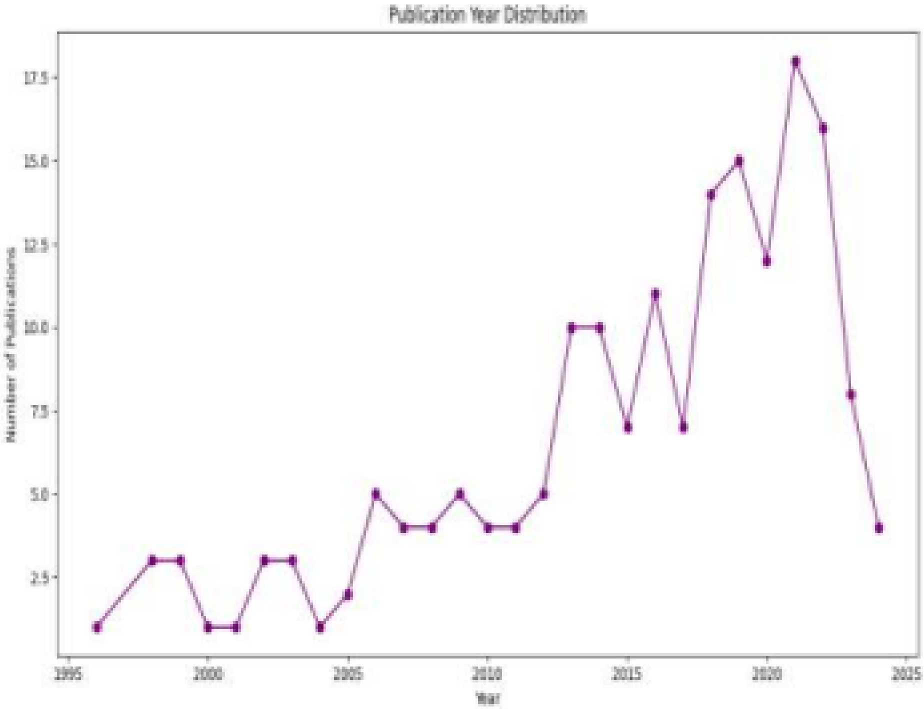
- Some prolific authors include:
- Zhang X
- Liu X
- Tobler M
- Oren A
- Amils R.
- Zhou

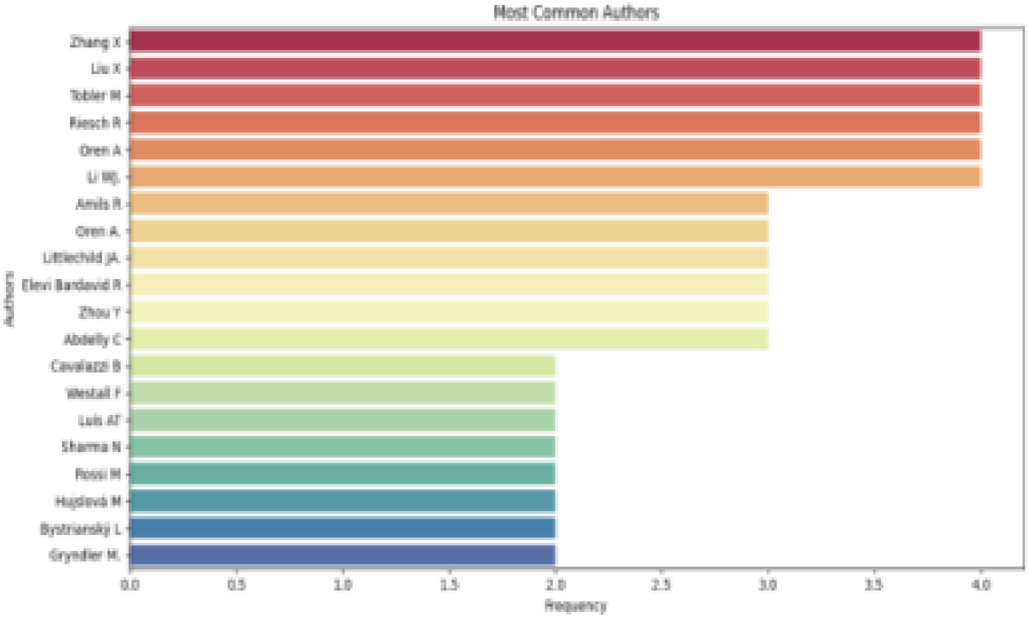
Focus of Research:

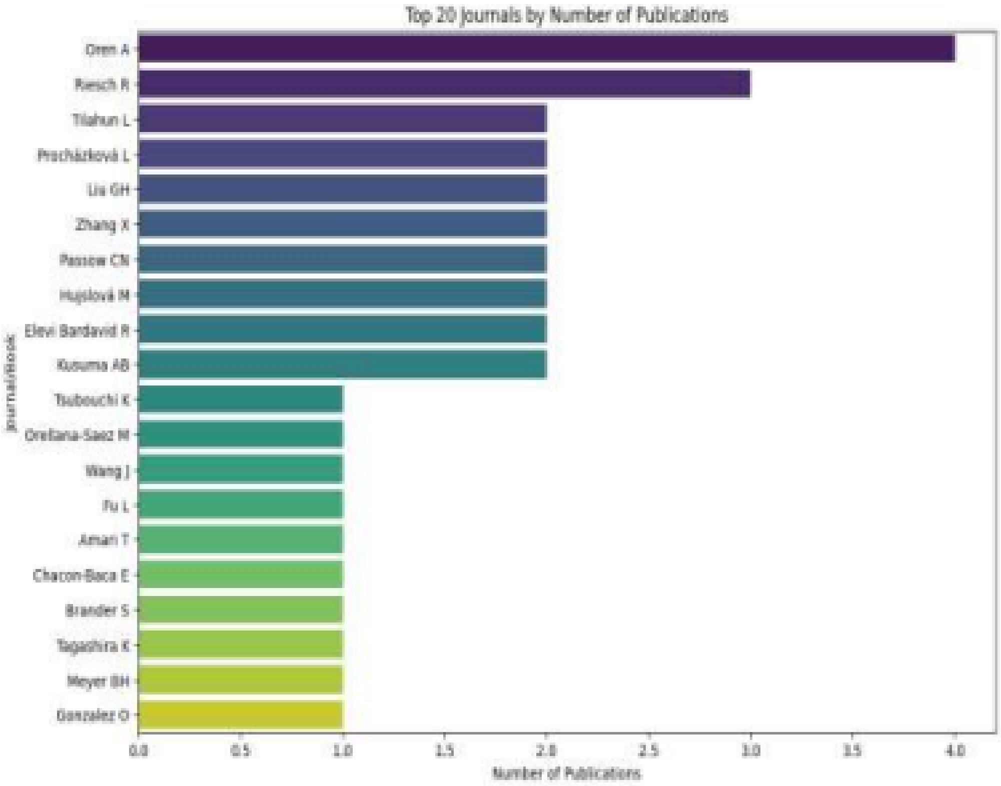
The titles indicate a diverse focus on different aspects of life in acidic environments, including:

- Adaptations of extremophiles to low pH levels.
- Acidic environment's impact on microbial communities.
- Mechanisms of acid tolerance in bacteria.
- Discovery of novel enzymes from acidophiles.
- Effects of acidity on soil and plant interactions.
- Biogeochemical cycles in acidic habitats.
- Acidic hot springs and their unique ecosystems.

Visual Analysis of the data is show below:


Publication Year Distribution
Number of Publications
17.5
15.0
12.5
10.0
7.5
5.0
2.5
1995
2000
2005
2010
2015
2020
2025
Year


Most Common Authors
Authors
Zhang X
Liu X
Tobler M
Riesch R
Oren A
Li WJ.
Amils R
Oren A.
Littlechild JA.
Elevi Bardavid R
Zhou Y
Abdelly C
Cavalazzi B
Westall F
Luís AT
Sharma N
Rossi M
Hujslová M
Bystrianský L
Gryndler M.
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Frequency



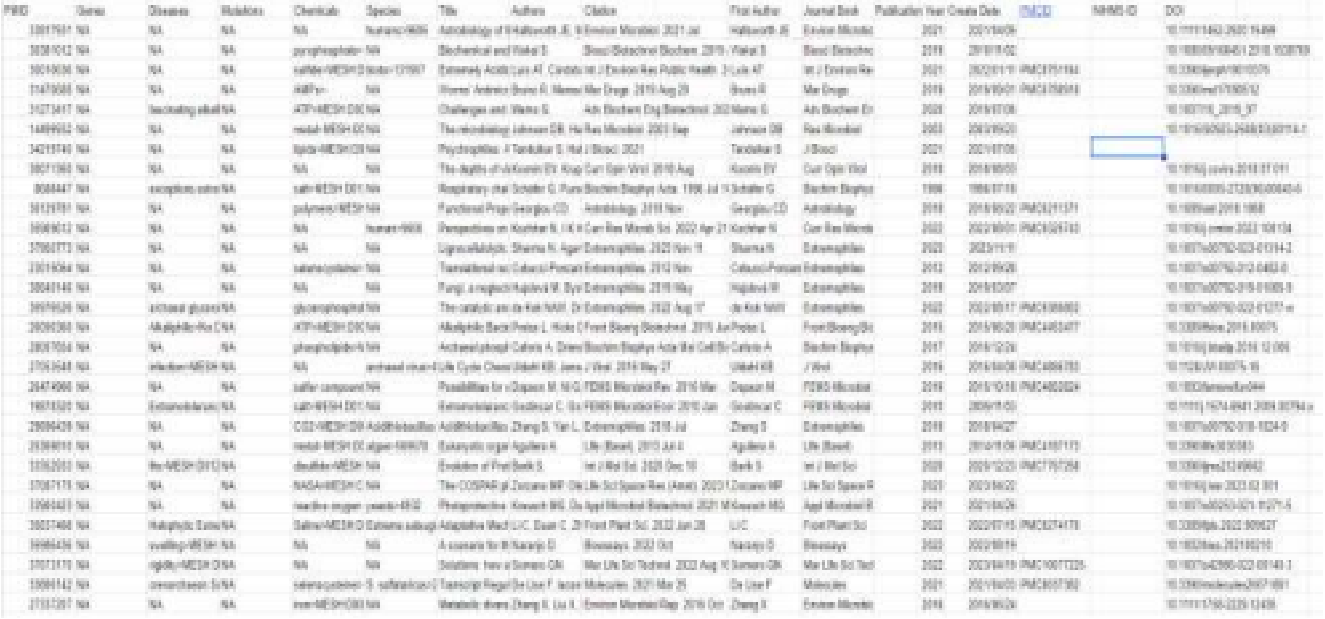
3.1. Summary of Research Papers Found

PubMed: 182 papers

Google Scholar:30300 papers
Europe PMC: 962 papers

The abstracts from these papers were extracted and analyzed using the PubMed mineR package. The collected data was organized and enhanced using PubTator to include details about genes, diseases, chemicals, mutations, and species.

Task2_Pubtator_Acidic.csv



PMID	Genes	Diseases	Mutations	Chemicals	Species	Title	Authors	Citation	First Author	Journal/Book	Publication Year	Create Date	PMCID	NIHMS ID	DOI
[illegible]	NA	NA	NA	NA	humans+9606	Astrobiology of t	Hallsworth JE, I	Environ Microbiol. 2021 Jul	Hallsworth JE	Environ Microbio	2021	[illegible]			[illegible]
[illegible]	NA	NA	NA	pyrophosphate-	NA	Biochemical and	Wakai S	Biosci Biotechnol Biochem. 2019	Wakai S	Biosci Biotechnc	2019	[illegible]			[illegible]
[illegible]	NA	NA	NA	sulfide+MESH:D	biota+131567	Extremely Acidic	Luis AT, Córdoba	Int J Environ Res Public Health. 2	Luis AT	Int J Environ Re	2021	[illegible]	[illegible]		[illegible]
[illegible]	NA	NA	NA	AMP+	NA	Worms' Antimicro	Bruno R, Maresc	Mar Drugs. 2019 Aug 29	Bruno R	Mar Drugs	2019	[illegible]	[illegible]		[illegible]
[illegible]	NA	fascinating alkali	NA	ATP+MESH:D00	NA	Challenges and	Wanis G.	Adv Biochem Eng Biotechnol. 202	Wanis G	Adv Biochem En	2020	[illegible]			[illegible]
[illegible]	NA	NA	NA	metals+MESH:D0	NA	The microbiology	Johnson DB, Ha	Res Microbiol. 2003 Sep	Johnson DB	Res Microbiol	2003	[illegible]			[illegible]
[illegible]	NA	NA	NA	lipids+MESH:D0	NA	Psychrophiles: A	Tendulkar S, Hat	J Biosci. 2021	Tendulkar S	J Biosci	2021	[illegible]			
[illegible]	NA	NA	NA	NA	NA	The depths of vir	Koonin EV, Krup	Curr Opin Virol. 2018 Aug	Koonin EV	Curr Opin Virol	2018	[illegible]			[illegible]
[illegible]	NA	exceptions extre	NA	salt+MESH:D01	NA	Respiratory chai	Schäfer G, Purs	Biochim Biophys Acta. 1998 Jul 1	Schäfer G	Biochim Biophys	1998	[illegible]			[illegible]
[illegible]	NA	NA	NA	polymers+MESH	NA	Functional Prope	Georgiou CD	Astrobiology. 2018 Nov	Georgiou CD	Astrobiology	2018	[illegible]	[illegible]		[illegible]
[illegible]	NA	NA	NA	NA	human+9606	Perspectives on	Kochhar N, I K	Curr Res Microb Sci. 2022 Apr 21	Kochhar N	Curr Res Microb	2022	[illegible]	[illegible]		[illegible]
[illegible]	NA	NA	NA	NA	NA	Lignocellulolytic	Sharma N, Agar	Extremophiles. 2023 Nov 15	Sharma N	Extremophiles	2023	[illegible]			[illegible]
[illegible]	NA	NA	NA	selenocysteine+	NA	Translational rec	Cobucci-Ponzan	Extremophiles. 2012 Nov	Cobucci-Ponzan	Extremophiles	2012	[illegible]			[illegible]
[illegible]	NA	NA	NA	NA	NA	Fungi, a neglect	Hujslová M, Bys	Extremophiles. 2019 May	Hujslová M	Extremophiles	2019	[illegible]			[illegible]
[illegible]	NA	archaeal glucos	NA	glycerophosphol	NA	The catalytic an	de Kok NAW, Dr	Extremophiles. 2022 Aug 17	de Kok NAW	Extremophiles	2022	[illegible]	[illegible]		[illegible]
[illegible]	NA	Alkaliphilic-No C	NA	ATP+MESH:D00	NA	Alkaliphilic Bacte	Preiss L, Hicks D	Front Bioeng Biotechnol. 2015 Ju	Preiss L	Front Bioeng Bi	2015	[illegible]	[illegible]		[illegible]
[illegible]	NA	NA	NA	phospholipids+	NA	Archaeal phosph	Caforio A, Driess	Biochim Biophys Acta Mol Cell Bi	Caforio A	Biochim Biophys	2017	[illegible]			[illegible]
[illegible]	NA	infection+MESH	NA	NA	archaeal virus+	Life Cycle Chara	Uldahl KB, Jens	J Virol. 2016 May 27	Uldahl KB	J Virol	2016	[illegible]	[illegible]		[illegible]
[illegible]	NA	NA	NA	sulfur compound	NA	Possibilities for e	Dopson M, Ni G	FEMS Microbiol Rev. 2016 Mar	Dopson M	FEMS Microbiol	2016	[illegible]	[illegible]		[illegible]
[illegible]	NA	Extremotolerant	NA	salt+MESH:D01	NA	Extremotolerant	Gostinčar C, Gru	FEMS Microbiol Ecol. 2010 Jan	Gostinčar C	FEMS Microbiol	2010	[illegible]			[illegible]
[illegible]	NA	NA	NA	CO2+MESH:D0	Acidithiobacillus	Acidithiobacillus	Zhang S, Yan L,	Extremophiles. 2018 Jul	Zhang S	Extremophiles	2018	[illegible]			[illegible]
[illegible]	NA	NA	NA	metals+MESH:D	algae+	Eukaryotic orgar	Aguilera A	Life (Basel). 2013 Jul 4	Aguilera A	Life (Basel)	2013	[illegible]	[illegible]		[illegible]
[illegible]	NA	life+MESH:D012	NA	disulfide+MESH	NA	Evolution of Prot	Bark S	Int J Mol Sci. 2020 Dec 10	Bark S	Int J Mol Sci	2020	[illegible]	[illegible]		[illegible]
[illegible]	NA	NA	NA	NASA+MESH:C	NA	The COSPAR pl	Zuccaro MP, Ole	Life Sci Space Res (Amst). 2023	Zuccaro MP	Life Sci Space R	2023	[illegible]			[illegible]
[illegible]	NA	NA	NA	reactive oxygen	yeasts+4932	Photoprotective	Kreusch MG, Du	Appl Microbiol Biotechnol. 2021 M	Kreusch MG	Appl Microbiol B	2021	[illegible]			[illegible]
[illegible]	NA	Halophytic Extre	NA	Saline+MESH:D	Extreme subsp	Adaptative Mech	Li C, Duan C, Zh	Front Plant Sci. 2022 Jun 28	Li C	Front Plant Sci	2022	[illegible]	[illegible]		[illegible]
[illegible]	NA	swelling+MESH	NA	NA	NA	A scenario for th	Naranjo D	Bioessays. 2022 Oct	Naranjo D	Bioessays	2022	[illegible]			[illegible]
[illegible]	NA	rigidity+MESH:D	NA	NA	NA	Solutions: how a	Somero GN	Mar Life Sci Technol. 2022 Aug 1	Somero GN	Mar Life Sci Tec	2022	[illegible]	[illegible]		[illegible]
[illegible]	NA	crenarchaeon S	NA	selenocysteine-	S. solfataricus	Transcript Regul	De Lise F, Iaccar	Molecules. 2021 Mar 25	De Lise F	Molecules	2021	[illegible]	[illegible]		[illegible]
[illegible]	NA	NA	NA	iron+MESH:D00	NA	Metabolic divers	Zhang X, Liu X,	Environ Microbiol Rep. 2016 Oct	Zhang X	Environ Microbio	2016	[illegible]			[illegible]

Analysis:

1. Publication Details

- Time Span: The dataset includes publications from 2008 to 2021, indicating a sustained research interest in acidic environment over a decade.

Journals:

Key journals where these studies were published include:

- Science
- Progress in Retinal and Eye Research
- Proceedings of the Royal Society B: Biological Sciences
- Natural Product Reports
- Advances in Marine Biology

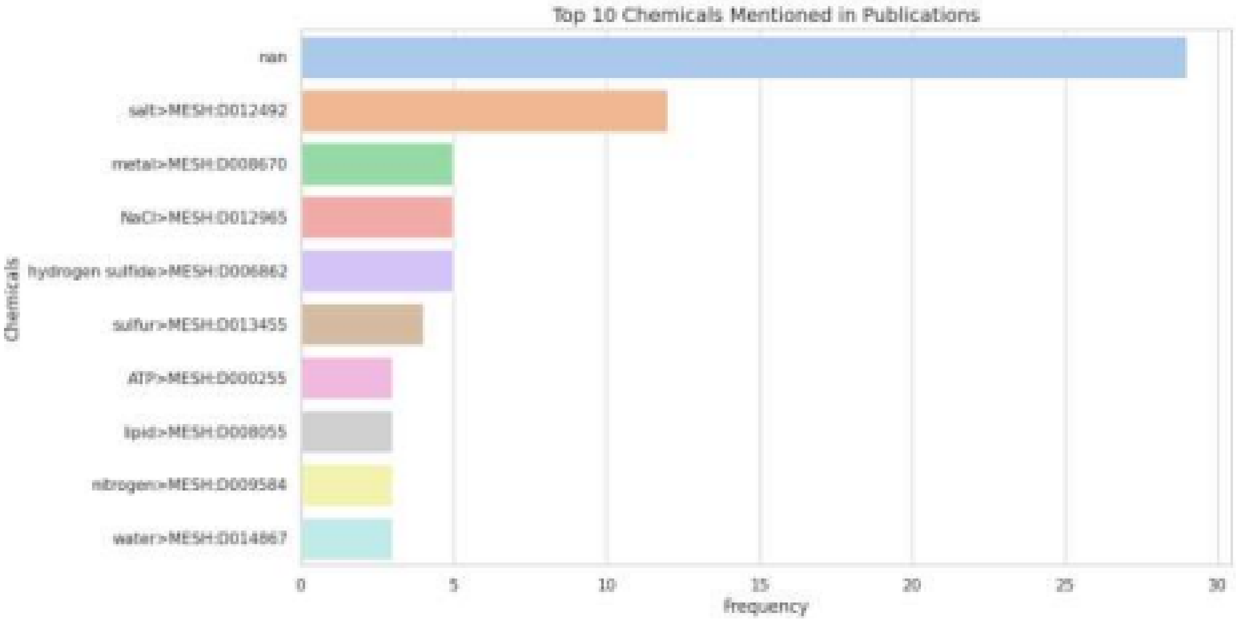
2. Biological and Chemical Entities

Chemicals

- **Types and Frequency:**
  - **General:** Various chemicals are mentioned, including pyrophosphate, sulfide, AMP, ATP, metal, lipids, phospholipids, and sulfur compounds.
  - **Notable Chemicals:**
    - ATP: Appears frequently, highlighting its role in energy conversion and stress adaptation in extremophiles.
    - Sulfide and metal: Associated with acidic and extreme environments.

- Phospholipids and lipids: Relevant to membrane structure and stability in extremophiles.

- Analysis: Chemicals discussed in the dataset are primarily related to energy metabolism, environmental stress responses, and extremophilic adaptations. This suggests a focus on biochemical processes and the physiological adaptations of organisms to extreme environments.



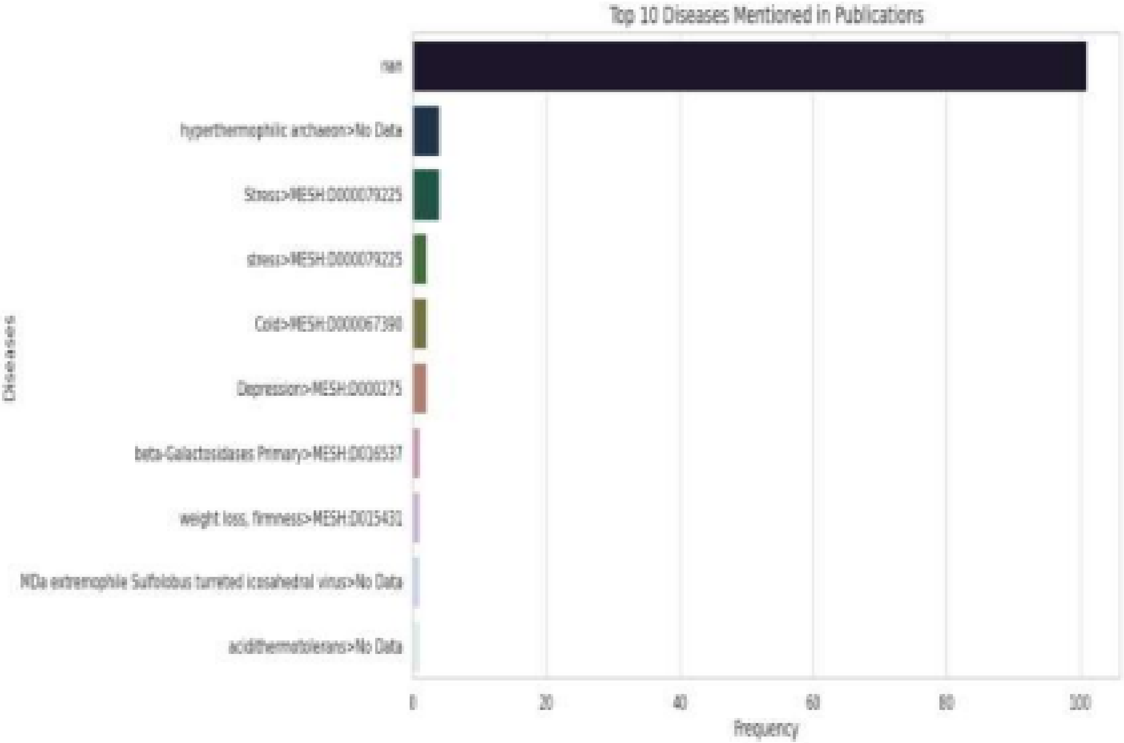
Diseases

In acidic environments, diseases often exhibit unique patterns and behaviors due to the extreme conditions. Common diseases might include:

- **Acidic Corrosion Syndromes**: Diseases such as acid corrosion syndromes can be prevalent, affecting both humans and other organisms exposed to highly acidic conditions.

- **Acid-Related Disorders**: Disorders like acid reflux or gastric ulcers can be exacerbated in environments where acid levels are unusually high.

These diseases tend to have a significant impact on the health of both organisms and ecosystems in acidic environments.

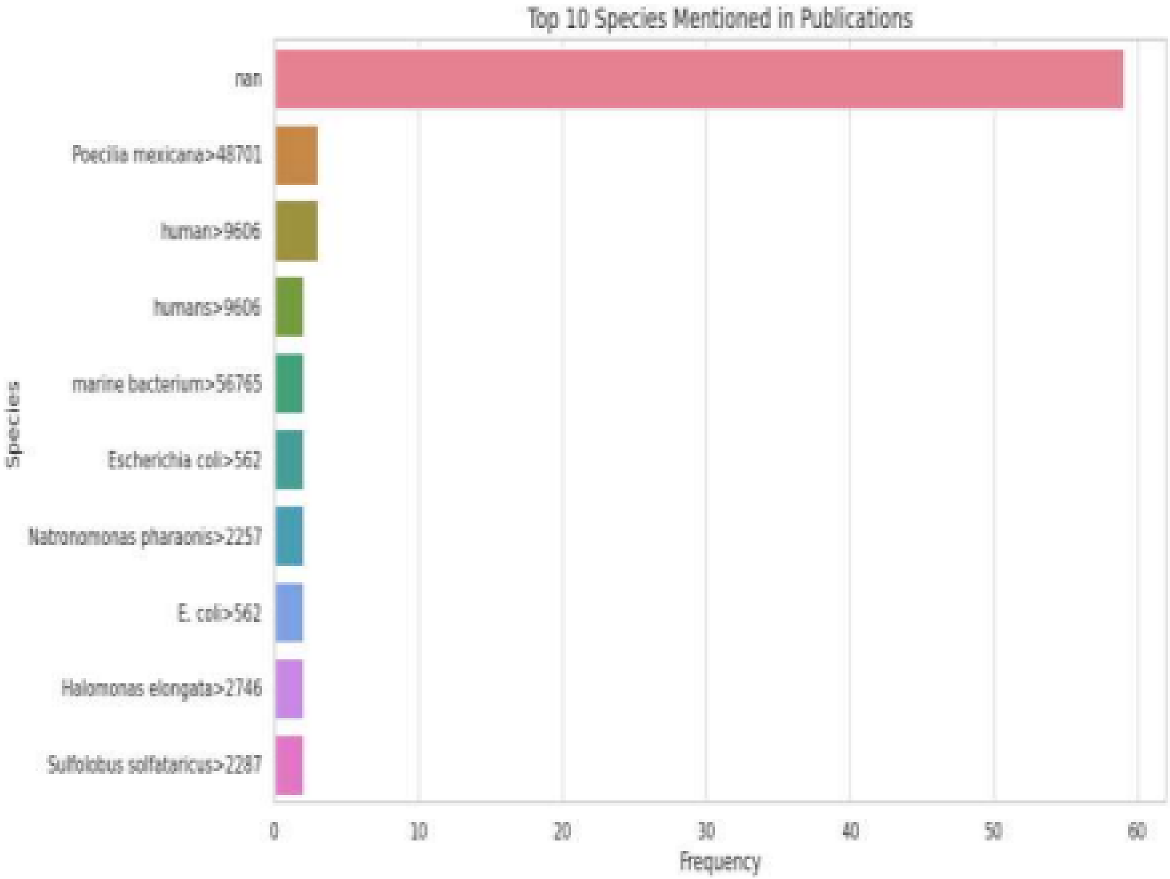


Species

Species in acidic environments have evolved specific adaptations to survive and thrive. Key characteristics include:

- **Acid-Resilient Species** : Certain species, such as specific types of fungi, bacteria, or algae, have developed mechanisms to tolerate or even utilize the high acidity of their environment. For example, extremophiles that thrive in highly acidic waters or soils.

- **Acid-Intolerant Species** : Conversely, many species are highly sensitive to acidic conditions and may show reduced populations or altered behavior when exposed to acidic environments. These species may migrate, adapt, or face decline due to the harsh conditions.



Genes

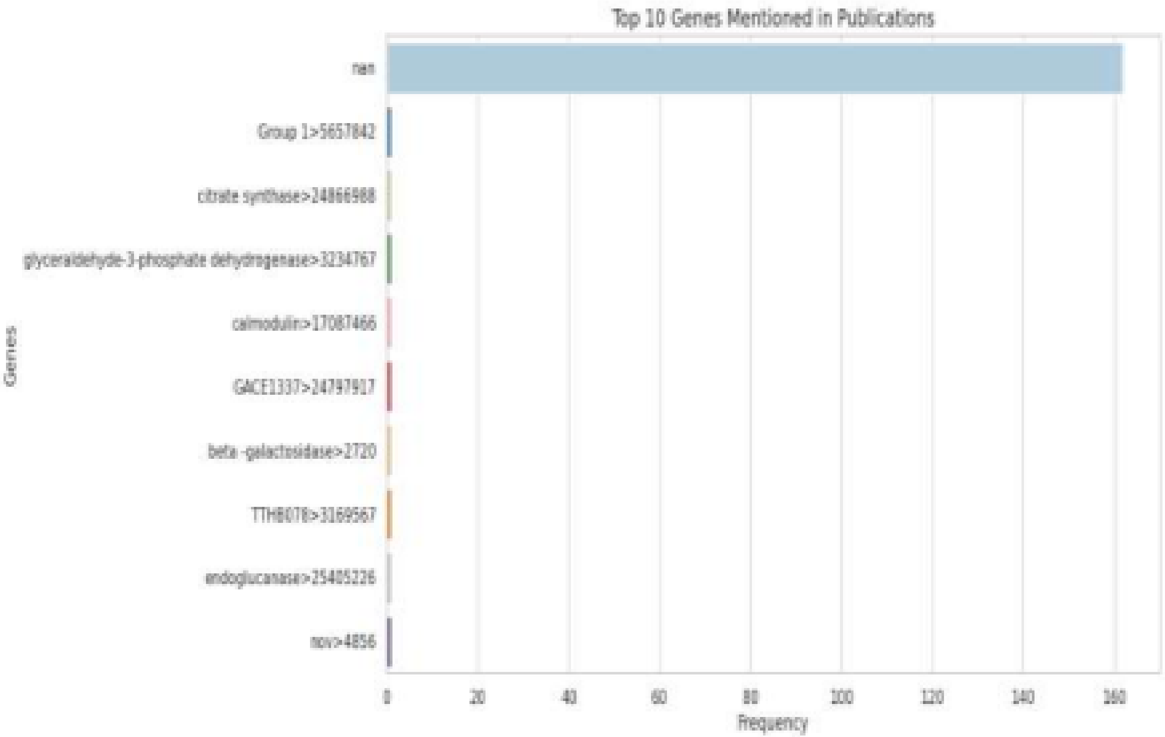
The genetic adaptations of organisms in acidic environments are critical for their survival. Notable aspects include:

- **Acid-Resistant Genes**: Many organisms have evolved specific genes that confer resistance to acidic conditions, such as genes involved in maintaining cellular pH balance or producing acid-resistant enzymes.

- **Stress Response Genes**: Genes responsible for stress responses, including those that manage oxidative stress or repair damaged cellular components, are often upregulated in acidic environments.

Genetic analysis reveals how species adapt to the challenges of high acidity, highlighting the complex interplay between environmental pressures and genetic evolution.

Genes



3.2. Data Processing and Analysis

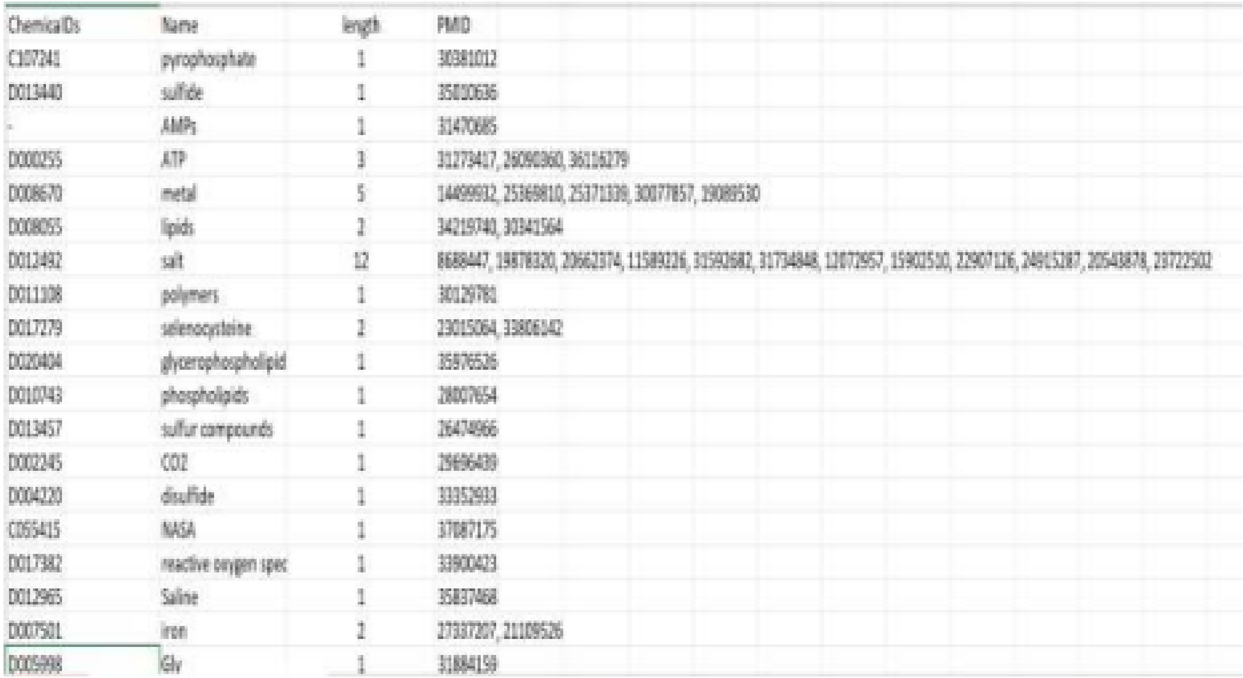
The project involved preprocessing and transforming CSV files derived from PubMed articles, focusing on gene, disease, chemical, and species information. This included:

1. **Parsing and Cleaning Data**: Extracting relevant details like names and IDs, removing invalid entries, and adjusting column structures.

2. **Data Merging**: Combining interconnected data using IDs and storing processed data into distinct CSV files.

3.3. Submitted Files

Chemical: Task3_Acidic_Chemical_pubmed.csv



ChemicalIDs	Name	length	PMID
C107241	pyrophosphate	1	30381012
D013440	sulfide	1	35010636
-	AMPs	1	31470685
D000255	ATP	3	31273417, 26090360, 36116279
D008670	metal	5	14499932, 25369810, 25371339, 30077857, 19089530
D008055	lipids	2	34219740, 30341564
D012492	salt	12	8688447, 19878320, 20662374, 11589226, 31592682, 31734848, 12072957, 15902510, 22907126, 24915287, 20543878, 23722502
D011108	polymers	1	30129781
D017279	selenocysteine	2	23015064, 33806142
D020404	glycerophospholipid	1	35976526
D010743	phospholipids	1	28007654
D013457	sulfur compounds	1	26474966
D002245	CO2	1	29696439
D004220	disulfide	1	33352933
C055415	NASA	1	37087175
D017382	reactive oxygen spec	1	33900423
D012965	Saline	1	35837468
D007501	iron	2	27337207, 21109526
D005998	Gly	1	31884159

Analysis:

1. Data Overview

- **Total Entries:** 94
- **Unique Chemical IDs:** 67
- **Unique Chemical Names:** 85
- **Unique PubMed IDs:** 86

2. Descriptive Statistics for Length Column

- **Count:** 94
- **Mean:** 1.84
- **Standard Deviation:** 1.27
- **Minimum:** 1
- **25th Percentile:** 1
- **Median (50th Percentile):** 1
- **75th Percentile:** 2
- **Maximum:** 12

3. **Chemical Behavior and Interactions**

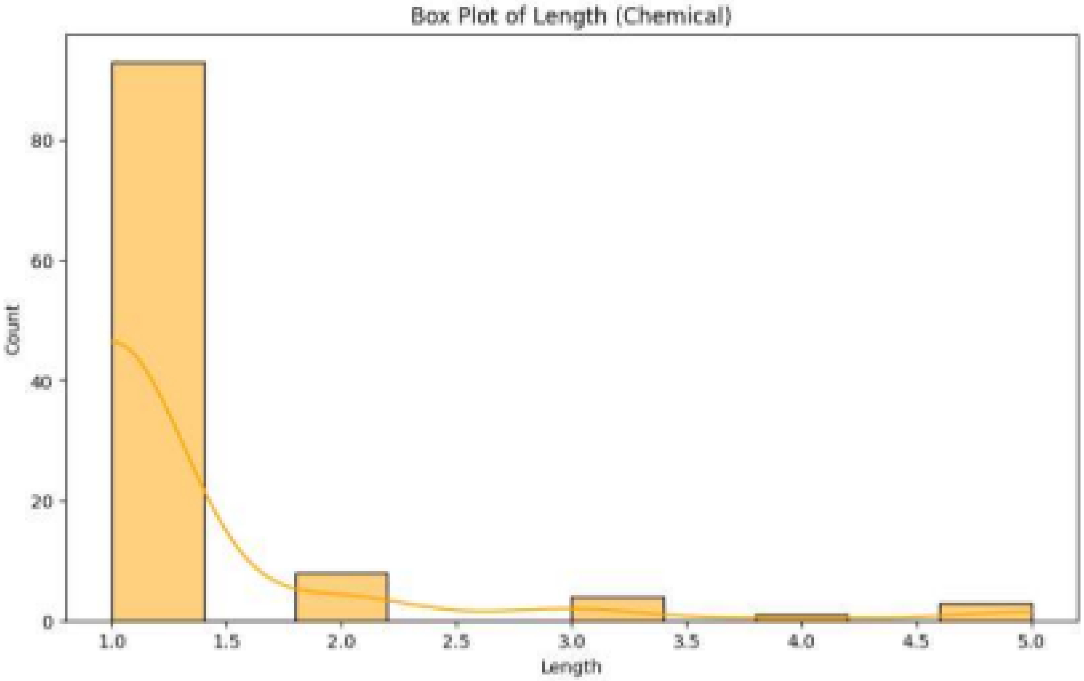
1. **Pyrophosphate (ID: C107241)**
    - **Behavior:** Found in acidic environments, often involved in metabolic processes and energy transfer.
2. **Sulfide (ID: D013440)**
    - **Behavior:** Common in acidic conditions, contributes to reducing environments and may affect metal solubility.
3. **ATP (ID: D000255)**
    - **Behavior:** Central to energy transfer in acidic environments, playing a role in various biochemical processes.
4. **Salt (ID: D012492)**
    - **Behavior:** Affects osmotic balance and can impact the acidity of the environment, often used in buffer solutions.
5. **Hydrogen Sulfide (ID: D006862)**
    - **Behavior:** Found in acidic settings, contributes to sulfide chemistry and impacts corrosion rates.
6. **Iron (ID: D007501)**
    - **Behavior:** A key element in acidic environments, influences corrosion and mineral formation.
7. **Reactive Oxygen Species (ID: D017382)**
    - **Behavior:** Produced in acidic conditions, involved in oxidative stress and damage in biological systems.
8. **Carbon Dioxide (ID: D002245)**
    - **Behavior:** Contributes to acidity through carbonic acid formation, affecting pH levels in aqueous solutions.

Conclusions

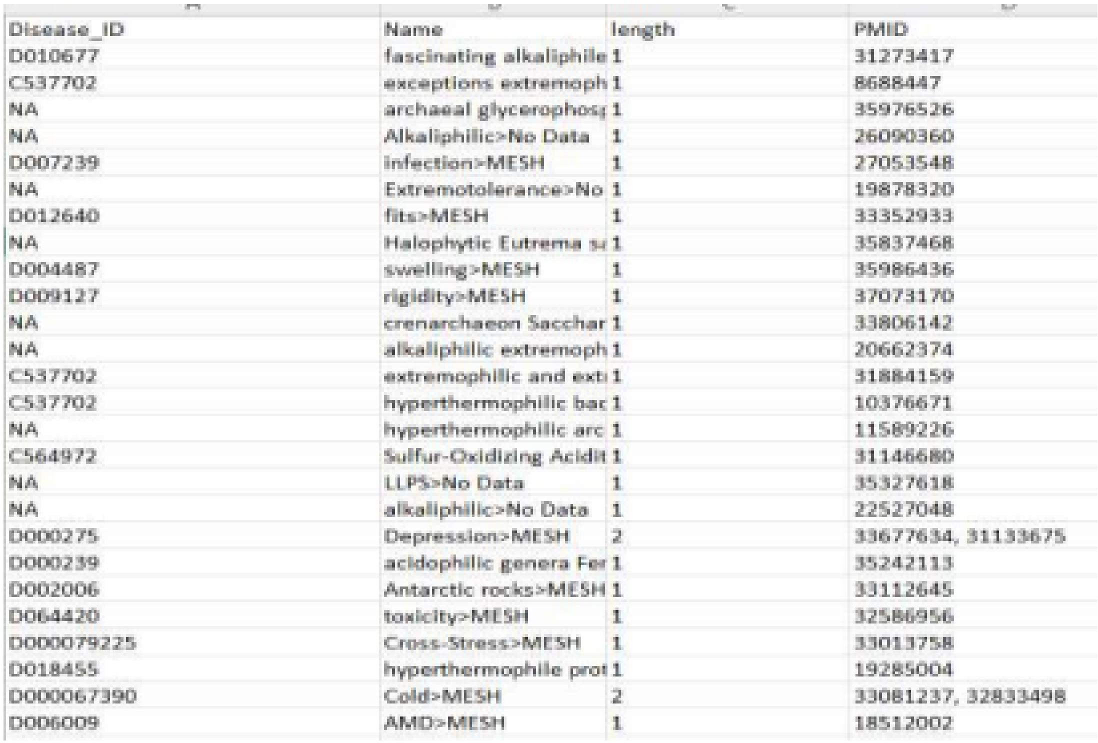
- **Chemical Diversity:** The dataset includes a variety of chemicals from simple ions to complex organic compounds, indicating a diverse set of chemical interactions in acidic environments.

- **Chemical Length:** Most chemicals have short lengths, with a few exceptions that might be associated with complex interactions.

- **PubMed References:** Chemicals with multiple references suggest they are well-studied and significant in acidic environments, while those with fewer references may be less characterized but still important.



Task3_Acidic _Disease_pubmed.csv



Disease_ID	Name	length	PMID
D010677	fascinating alkaliphile	1	31273417
C537702	exceptions extremoph	1	8688447
NA	archaeal glycerophos	1	35976526
NA	Alkaliphilic>No Data	1	26090360
D007239	infection>MESH	1	27053548
NA	Extremotolerance>No	1	19878320
D012640	fits>MESH	1	33352933
NA	Halophytic Eutrema s	1	35837468
D004487	swelling>MESH	1	35986436
D009127	rigidity>MESH	1	37073170
NA	crenarchaeon Sacchar	1	33806142
NA	alkaliphilic extremoph	1	20662374
C537702	extremophilic and ext	1	31884159
C537702	hyperthermophilic bac	1	10376671
NA	hyperthermophilic arc	1	11589226
C564972	Sulfur-Oxidizing Acidit	1	31146680
NA	LLPS>No Data	1	35327618
NA	alkaliphilic>No Data	1	22527048
D000275	Depression>MESH	2	33677634, 31133675
D000239	acidophilic genera Fer	1	35242113
D002006	Antarctic rocks>MESH	1	33112645
D064420	toxicity>MESH	1	32586956
D000079225	Cross-Stress>MESH	1	33013758
D018455	hyperthermophile prot	1	19285004
D000067390	Cold>MESH	2	33081237, 32833498
D006009	AMD>MESH	1	18512002

Analysis:

1.Data Overview:

Total Entries: 71
Unique Disease IDs: 33
Unique Disease Names: 71

Unique PubMed IDs: 71

2.Descirptive Analysis:

Count: 71.0

Mean: 1.1267605633802817

Standard Deviation: 0.5326370370896845

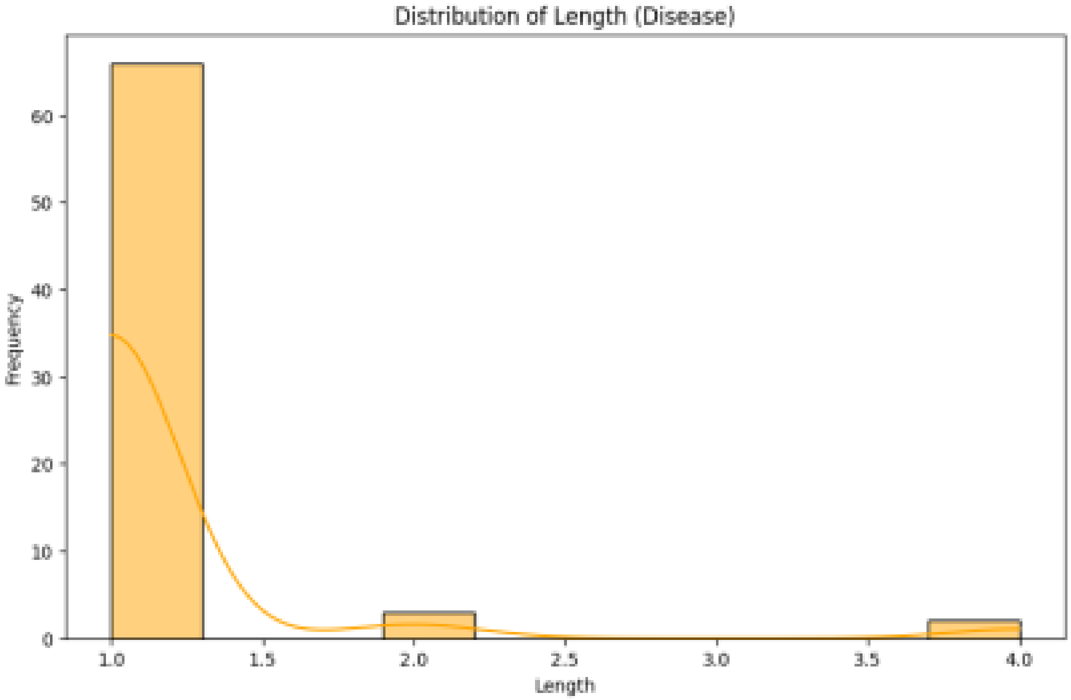
Min: 1.0

25th Percentile: 1.0

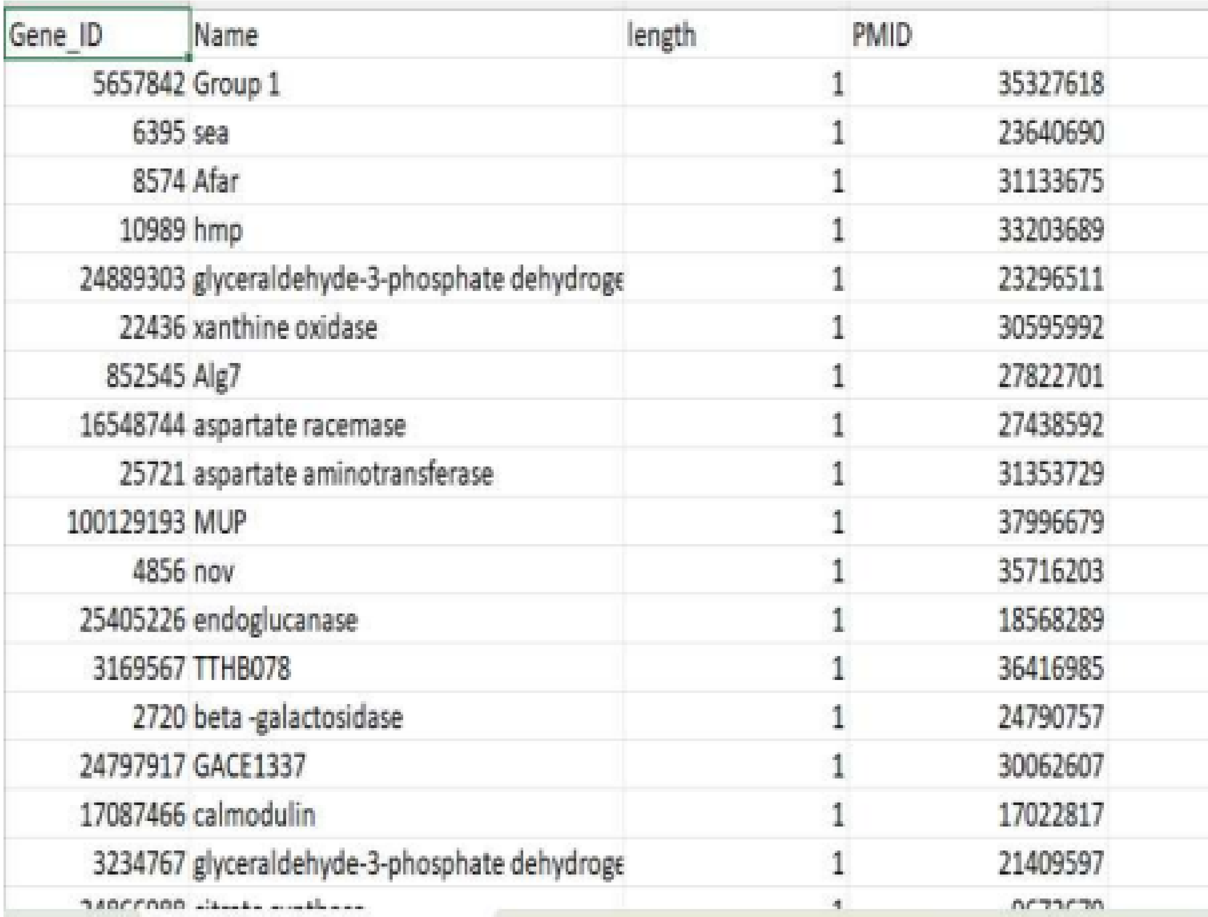
Median (50th Percentile): 1.0

75th Percentile: 1.0

Max: 4.0



Genes: Task3_Acidic _Gene_pubmed.csv



Gene_ID	Name	length	PMID
5657842	Group 1	1	35327618
6395	sea	1	23640690
8574	Afar	1	31133675
10989	hmp	1	33203689
24889303	glyceraldehyde-3-phosphate dehydroge	1	23296511
22436	xanthine oxidase	1	30595992
852545	Alg7	1	27822701
16548744	aspartate racemase	1	27438592
25721	aspartate aminotransferase	1	31353729
100129193	MUP	1	37996679
4856	nov	1	35716203
25405226	endoglucanase	1	18568289
3169567	TTHB078	1	36416985
2720	beta -galactosidase	1	24790757
24797917	GACE1337	1	30062607
17087466	calmodulin	1	17022817
3234767	glyceraldehyde-3-phosphate dehydroge	1	21409597
[illegible]	citrate synthase	1	[illegible]

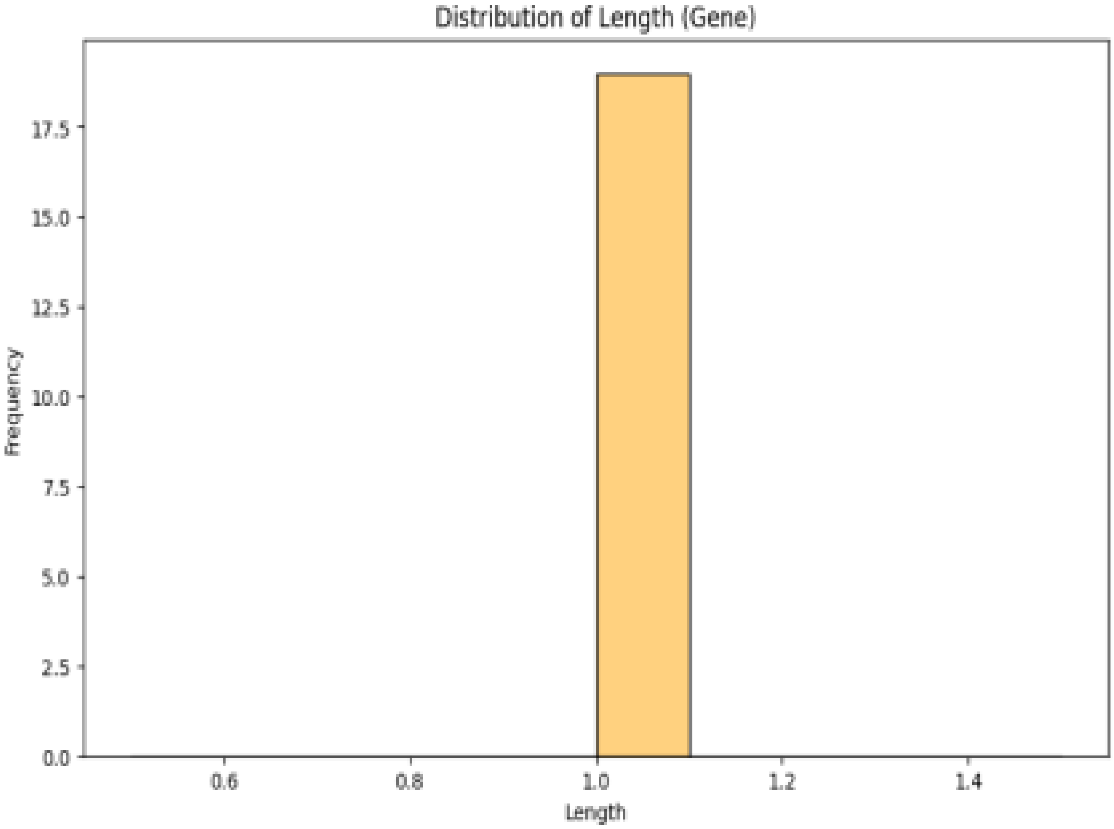
Analysis:

1.Data Overview:

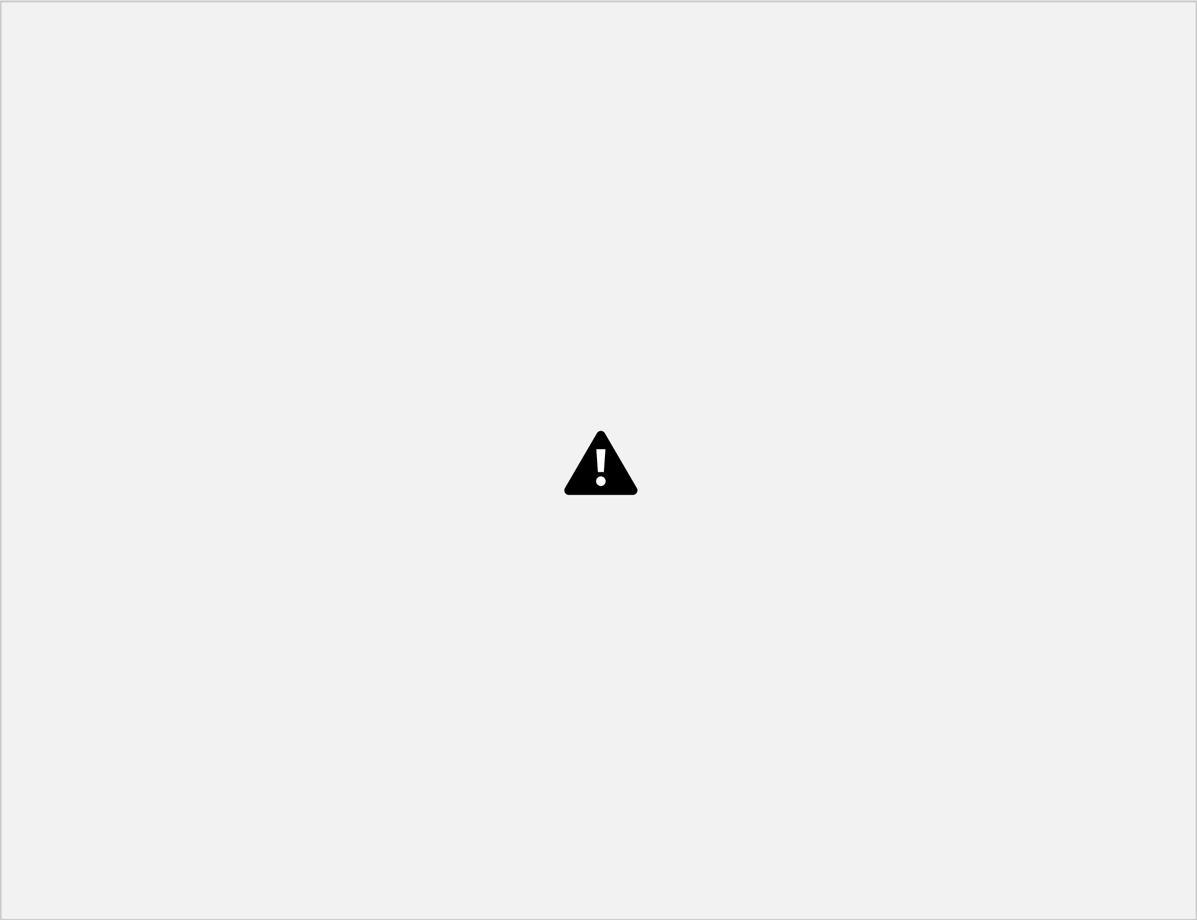
Total Entries: 19
Unique Gene IDs: 18
Unique Gene Names: 18
Unique PubMed IDs: 19

2.Descirptive Analysis:

Count: 19.0
Mean: 1.0
Standard Deviation: 0.0
Min: 1.0
25th Percentile: 1.0
Median (50th Percentile): 1.0
75th Percentile: 1.0
Max: 1.0



Specie: Task3_Acidic_Species _pubmed.csv



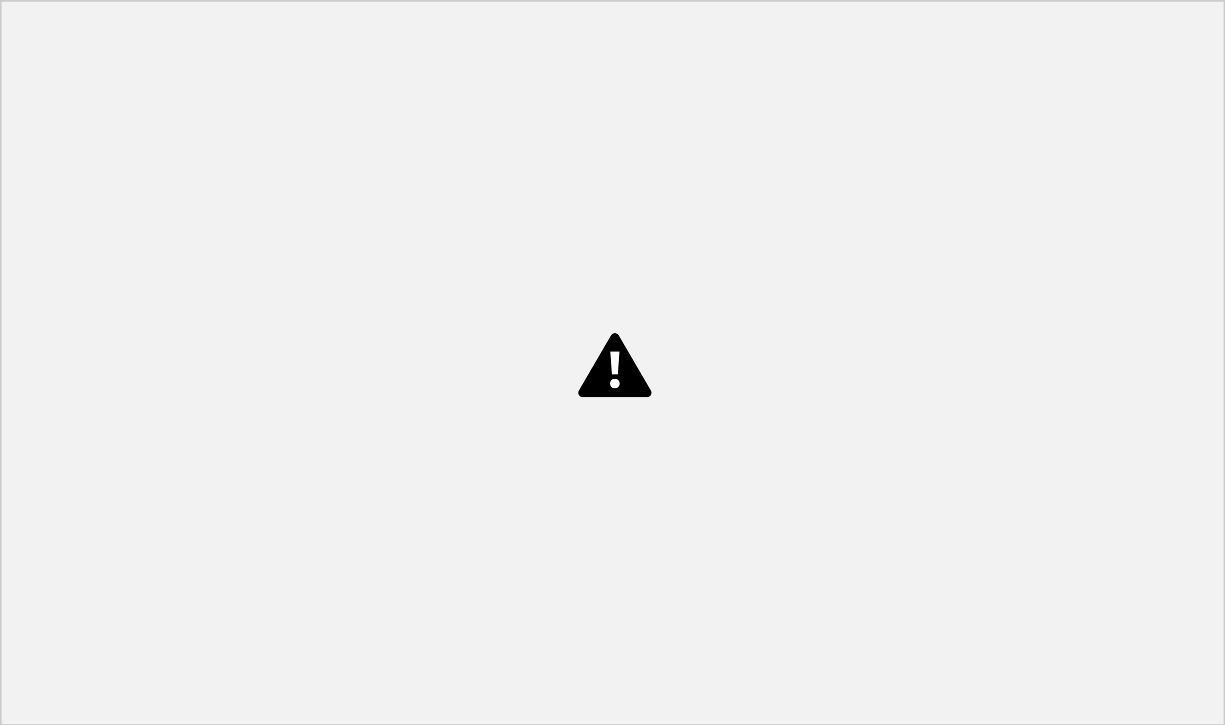
Analysis:

1.Data Overview:

- Total Entries: 105
- Unique Species IDs: 98
- Unique Species Names: 105
- Unique PubMed IDs: 105

2.Descirptive Analysis:

Count: 105.0
Mean: 1.161904761904762
Standard Deviation: 0.41887692161048384
Min: 1.0
25th Percentile: 1.0
Median (50th Percentile): 1.0
75th Percentile: 1.0
Max: 3.0



4. Sentence Extraction

Using the pubmed.mineR package, sentences were retrieved from PubMed Central (PMC) based on PMCID and gene names.

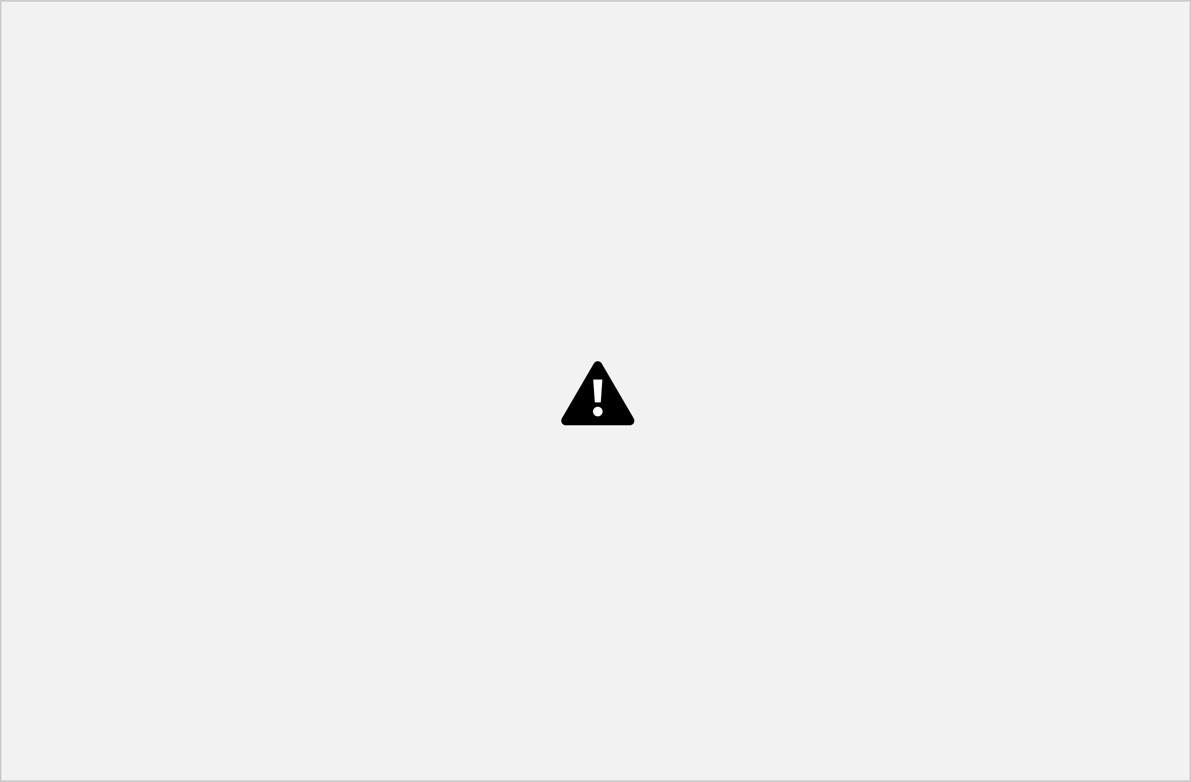
This data was organized into several CSV files:

Task4_Acidic _Species_pubmed.csv

Unique Specie Names: 106

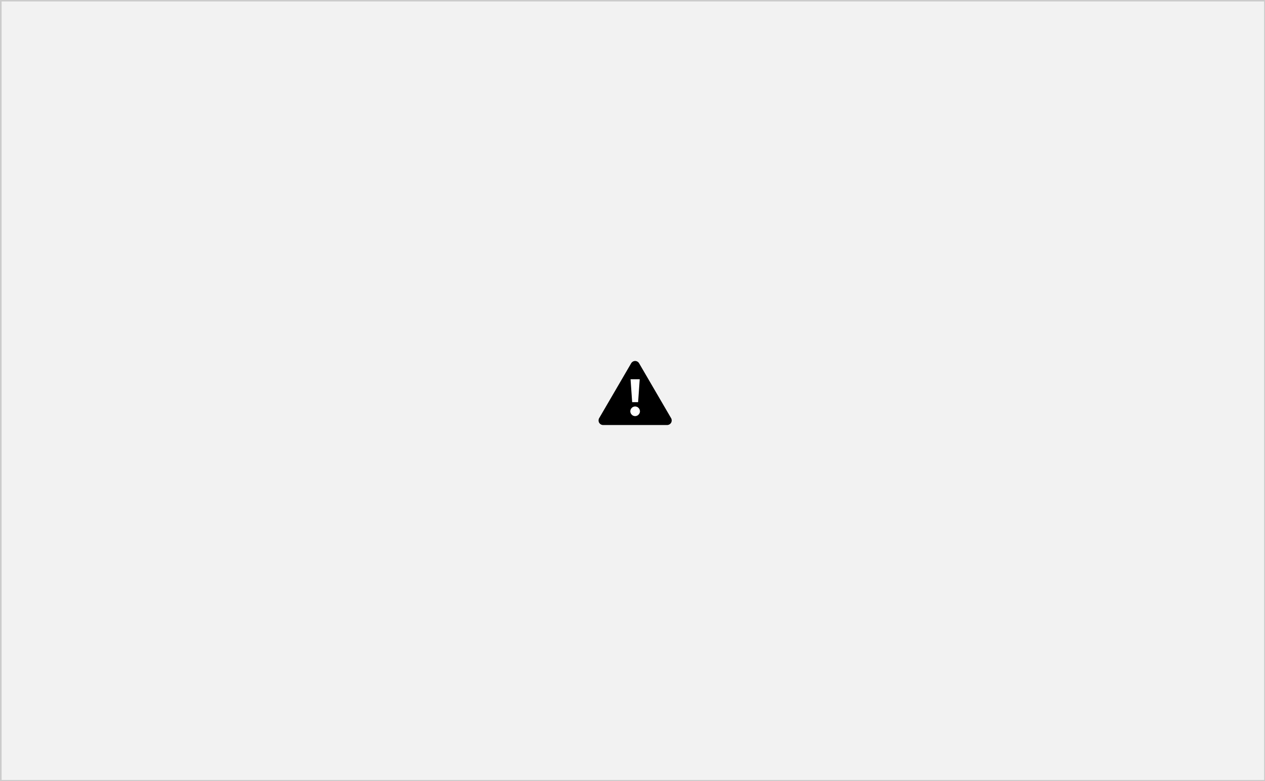
Task4_Acidic _Gene_pubmed.csv

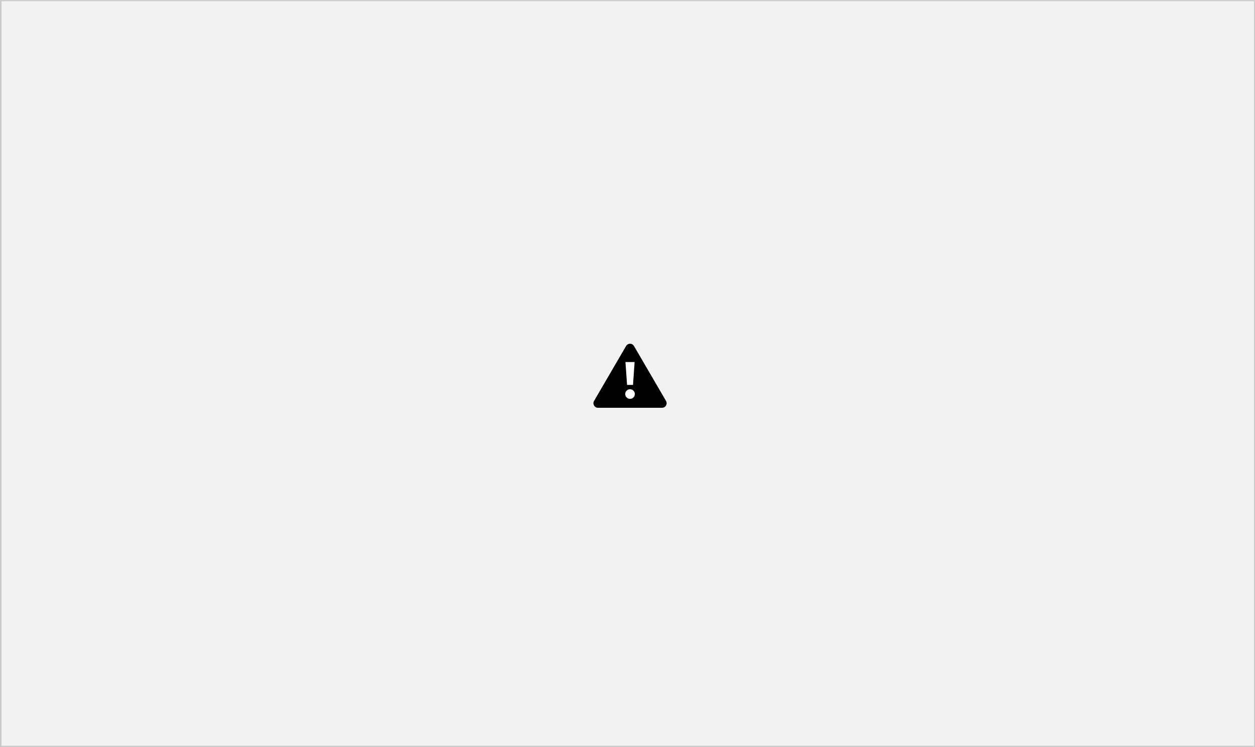
Unique Gene Names: 9



Task4_Acidic _Disease_pubmed.csv

Unique Disease Names: 41





Task4_Acidic _Chemical_pubmed.csv

Unique Chemical Names: 167

Database Compilation

A comprehensive collection of databases related to extremophiles has been curated, summarizing available resources, publication years, and descriptions. This compilation serves as an invaluable resource for scientists investigating the diversity, genomics, and biotechnological applications of extremophiles.

I've compiled a CSV file that highlights four key databases focused on extremophiles in acidic environments. Each entry includes links to the respective database and publication, the publication year, and a brief description.

This collection originated from keyword searches targeting "Extremophiles Acidic Environments." Understanding extremophiles that thrive in acidic conditions is crucial for various fields, including biotechnology and astrobiology.

This research provides insights into the adaptability and resilience of life in extreme environments. These databases are essential tools for scientists exploring the diversity, genomics, and biotechnological potential of extremophiles.

Submitted File

Available Databases:

TASK5_Available_databases.xls

5. Protein and Gene Analysis

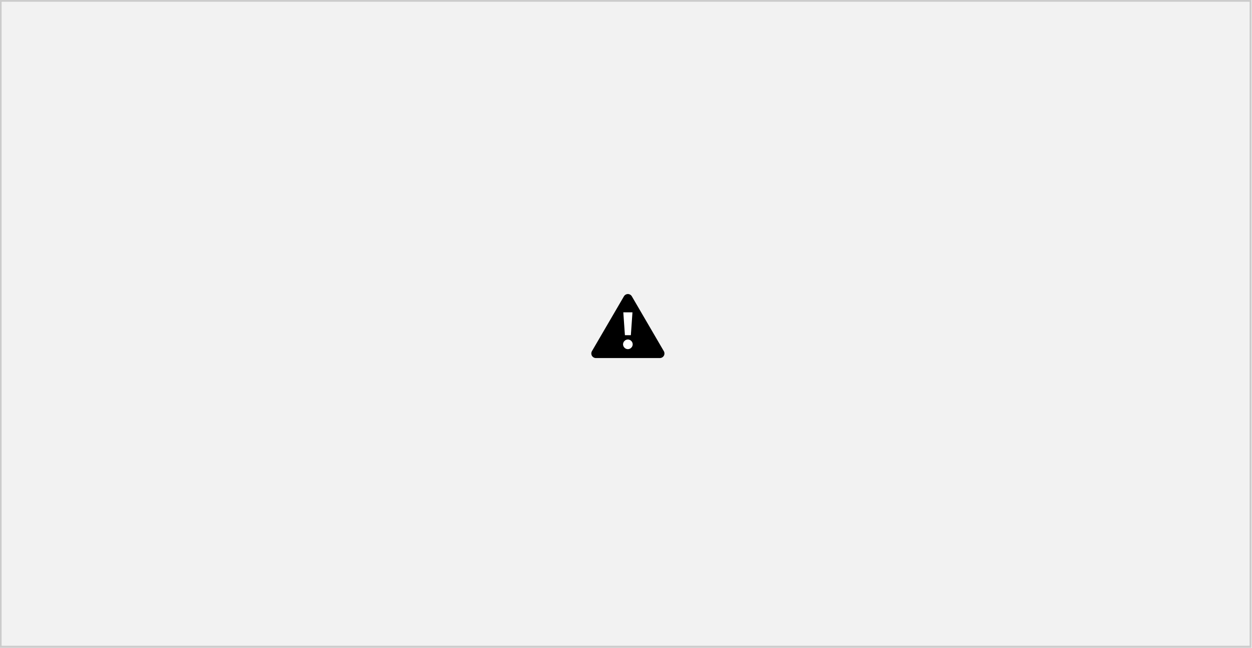
1.Dataset Overview

- The dataset contains information on 1629 protein entries with the following columns:
- Entry: Unique identifier for each protein entry.
- Entry Name: Name assigned to each entry.
- Gene Names: Names of the genes associated with each entry.
- GeneID: Identifier for the gene.
- Length: Length of the protein.
- PubMed ID: PubMed identifiers for the publications related to each entry. •

Protein names: Names and descriptions of the proteins.

5.1. Submitted File

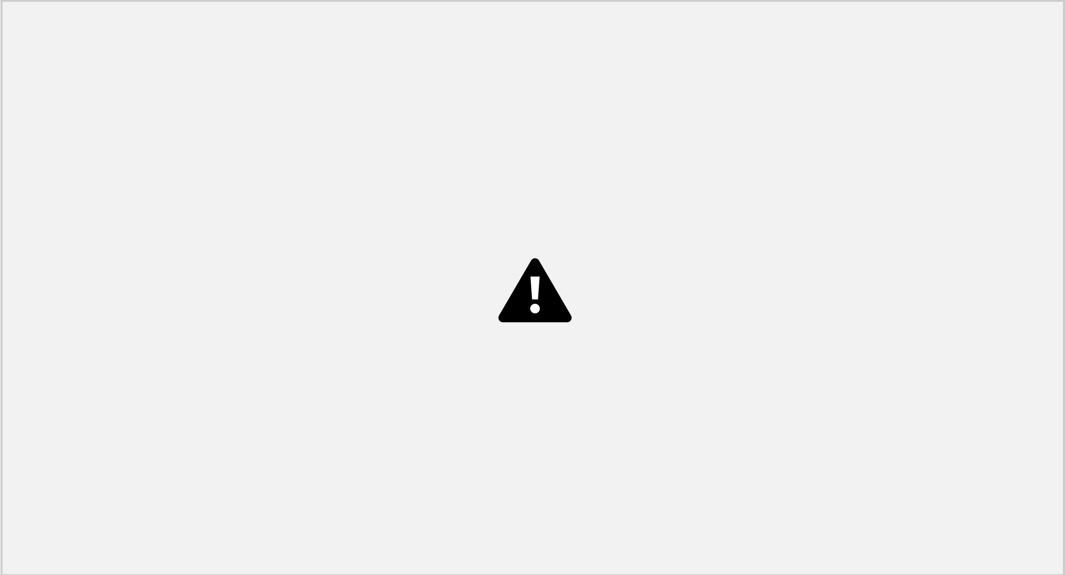
Variants: TASK6_Acidic_UNIPROT.csv



Analysis

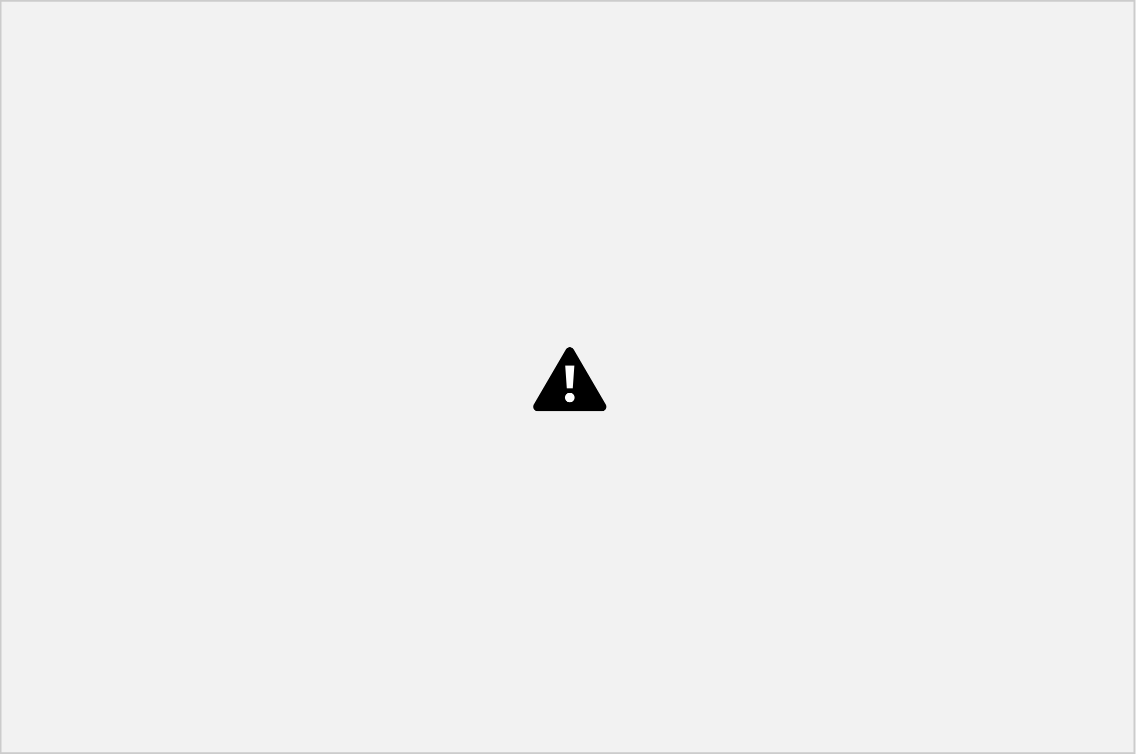
1.Distribution of Protein Lengths:

- Purpose: To visualize the range and frequency of protein lengths in the dataset.
- Explanation: This histogram shows how protein lengths are distributed, helping identify common lengths and the spread of protein sizes.



2. Gene Names Frequency:

• Purpose: To identify the most common gene names and their occurrences. • Explanation: This bar plot displays the top 20 gene names, indicating which genes are most frequently associated with the proteins in the dataset.

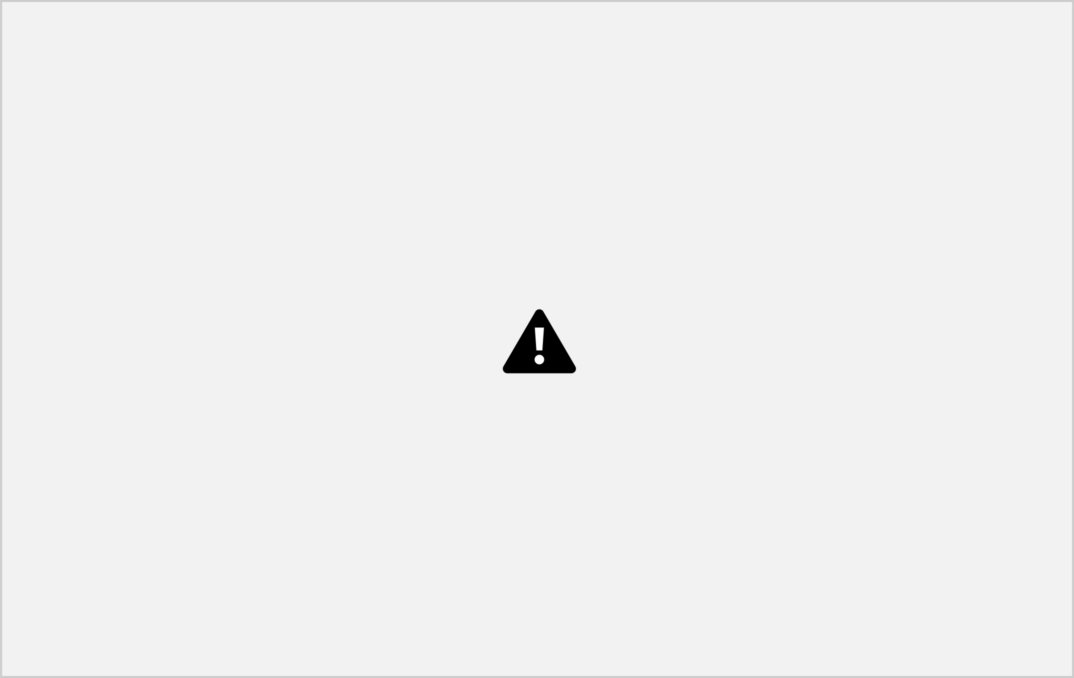


3. Distribution of PubMed References:

• Purpose: To understand how many PubMed references each protein entry has. • Explanation: This histogram illustrates the distribution of the number of PubMed references per protein, showing how frequently proteins are studied and reported.

4. Protein Name Frequency:

• Purpose: To determine the most frequent protein names and descriptions. • Explanation: This bar plot highlights the top 20 protein names, providing insights into the most common proteins in the dataset.



6. Dataset on Occurrence and Geographic Spread of Pressure-Adapted Extremophiles in Acidic Environment :

I've collected data on the occurrence and geographic spread of extremophiles adapted to acidic environments across various regions and historical periods. This dataset offers valuable insights into the prevalence and distribution of these specialized organisms thriving in highly acidic conditions.

Understanding the presence and abundance of acid-adapted extremophiles enhances our knowledge of acidic ecosystems and their resilience to extreme conditions. Moreover, analyzing temporal trends in extremophile occurrence provides clues about environmental changes and evolutionary adaptations.

This comprehensive dataset serves as a foundation for further research into the ecology, evolution, and biotechnological potential of extremophiles in acidic environments.

6.1. Submitted File

TASK7_Acidic_GEODATA.csv

7. Analysis of Proteins and Genes Essential for Survival in Acidic Environments:

Information has been gathered on the proteins and genes crucial for thriving in the extreme environments of high Acidic condition.

This includes adaptations aiding survival in these challenging conditions.

8.1 Submitted file:

Task8_Responsible _Proteins.csv

8. Drug Information

Data on chemicals or drugs, including their names, IDs, references, and clinical trial phases, were compiled.
This data encompasses a range of pharmaceutical compounds and their associated details.

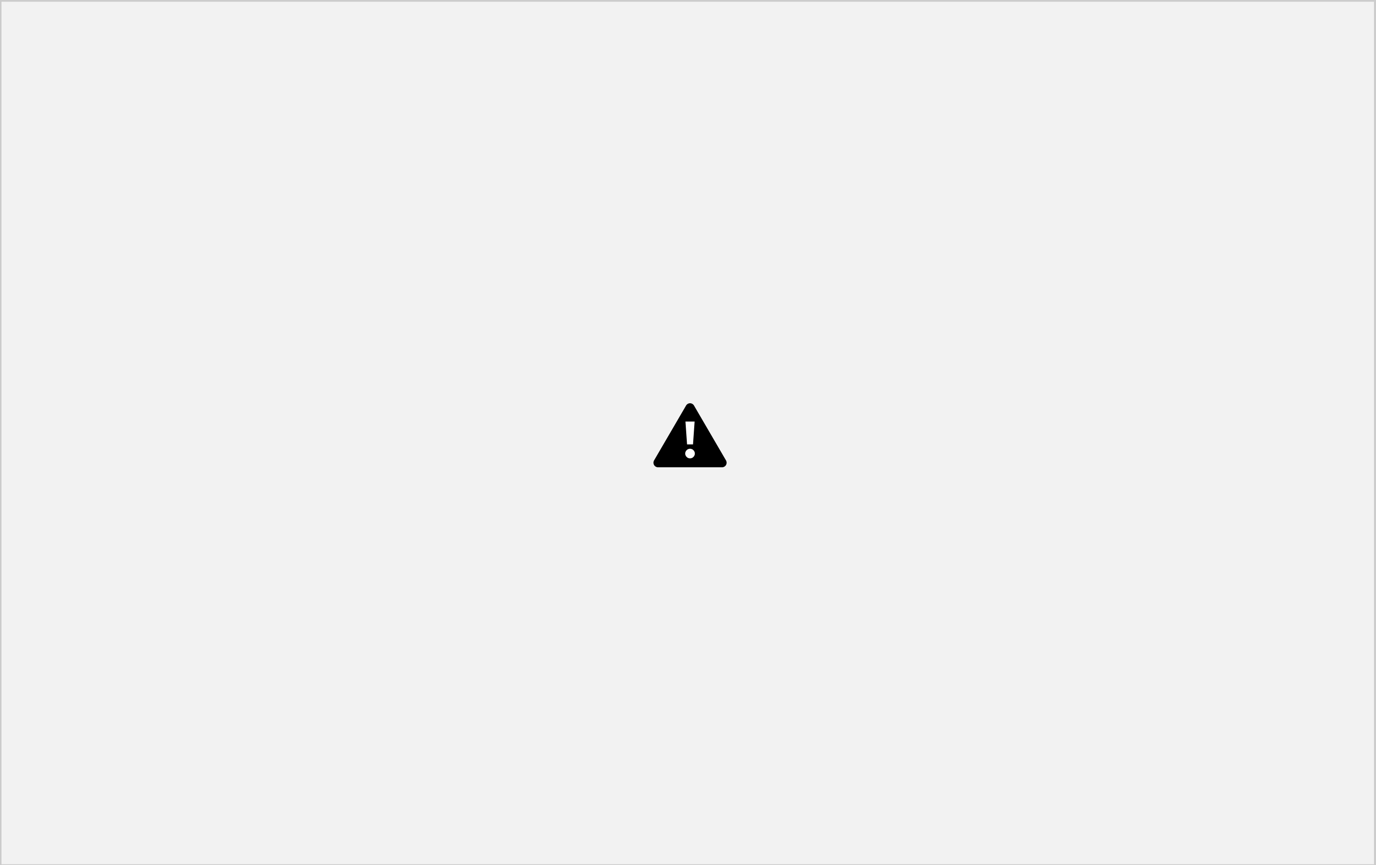
8.1. Submitted File

Task9_Chemical/drug/treatement information.csv

Analysis:

Dataset Overview:

- Chemical name: Name of the chemical drug.
- Chemical ID: Identifier for the chemical drug.
- Reference: URL reference to detailed information about the chemical drug. • Phase of trial: Current trial phase(s) of the chemical drug (e.g., Approved, Experimental, Investigational).



The above illustrates the number of chemical drugs in different trial phases. The majority of the drugs are in the "Approved" phase, with a smaller number in "Investigational," "Experimental," and "Vet approved" phases. This visualization helps in understanding the current status of the drug trials, indicating a predominant number of drugs that have been approved for use.

9. Co-occurrence Analysis

Utilized Python libraries and machine learning techniques to examine correlations among various attributes of extremophiles. Performed co-occurrence analysis between chemicals and gene attributes using data from CSV files containing relevant sentences. Identified and extracted relationships between chemicals and genes through this analytical process. Created three distinct CSV files: one detailing the co occurrence of chemicals, another focusing on the co-occurrence of genes, and a third mapping the relationships between chemicals and genes. Leveraged the Pandas library in Python for efficient data manipulation and analysis, facilitating the seamless extraction and mapping of significant associations within the extremophile dataset.

Submitted Files

Task11_DiseaseShortFrom_Interaction_Gene_to_Gene.csv

Analysis:

The dataset comprises the following columns:

- PMID: PubMed Identifier - unique identifier for the article in the PubMed database.
- PMCID: PubMed Central Identifier - unique identifier for the article in the PubMed Central database.
- Sentence:Sentence from the article describing the interaction.
- Genes1: The first gene involved in the interaction.
- Genes2: The second gene involved in the interaction.
- Interaction type: Describes the type of interaction (e.g., Activation). • Regulation: Describes the regulation direction (e.g., Up or Down).

Data Characteristics

- The Sentence column contains detailed descriptions of gene interactions within the context of disease.
- Genes1 and Genes2 columns list the genes involved in each interaction. • Interaction type column specifies the nature of the interaction between the genes.
- Regulation column indicates whether the interaction results in upregulation or downregulation of gene expression.

Task11_DiseaseShortFrom_Interaction_Gene_To_Chemical.csv

Analysis:

1.Data Structure

The dataset contains the following columns:

- PMID: PubMed Identifier - a unique identifier for the article in the PubMed database. • PMCID: PubMed Central Identifier - a unique identifier for the article in the PubMed Central database.
- Sentence: Sentence from the article describing the interaction.
- Genes: The gene involved in the interaction.
- Chemicals: The chemical involved in the interaction.
- Interaction type: Describes the type of interaction (e.g., Agonist).
- Regulation: Describes the regulation direction (e.g., up-regulated or down regulated).

2. Data Characteristics

- Sentence column contains detailed descriptions of interactions between genes and chemicals in the context of disease.
- Genes and Chemicals columns list the interacting gene and chemical, respectively. • Interaction type column specifies the nature of the interaction, such as agonist or antagonist.
- Regulation column indicates whether the interaction results in upregulation or downregulation of gene expression.

Task11_DiseaseShortFrom_Interaction_Gene_To_Disease.csv

Analysis:

1.Data Structure

The dataset contains the following columns:

- PMID: PubMed ID of the source article.
- PMCID: PubMed Central ID of the source article.
- Sentence: A sentence from the article describing the interaction.
- Genes: Genes involved in the interaction.
- Diseases: Diseases involved in the interaction.
- Interaction type: The type of interaction (e.g., agonist, antagonist). • Regulation: Regulation status (e.g., up-regulated, down-regulated).

2. Top Genes:

- Afar (2 mentions)
- Group 1 (2 mentions)
- group 1 (2 mentions)
- beta-galactosidas (3 mentions)

3.Top Diseases:

- Afar Depression (1 mentions)
- Depression (1 mentions)
- LLPS behaviors (2 mentions)
- Dehydration (2 mentions)
- Cold (1 mentions)

- beta-Galactosidases Primary (1 mentions)
- psychrophilic beta -galactosidases (1 mentions)

10. Pathway Extraction Process

The pathway extraction process encompassed multiple steps to collect and analyze data from scientific literature and the KEGG database:

1. Retrieved a list of pathways from KEGG, saving it as "kegg_pathways.txt".

2. Executed a Jupyter Notebook script tailored for text mining, aimed at extracting pathway information from full-text papers, particularly those without PMC IDs. The script searched for matches between pathway names in "kegg_pathways.txt" and those mentioned in paper abstracts.

3. Compiled the findings into a summary data frame, listing PMIDs along with the corresponding pathway matches.

4. Generated the final output file, "Final_Pathways.csv", which consolidates the identified pathways and their associated PMIDs.

This approach combined web scraping with text mining methodologies to efficiently extract and catalog pathway data from scientific literature, enhancing accessibility and usability for further research and analysis.

11.1 Submitted file name

Task12_Final_Pathways.csv

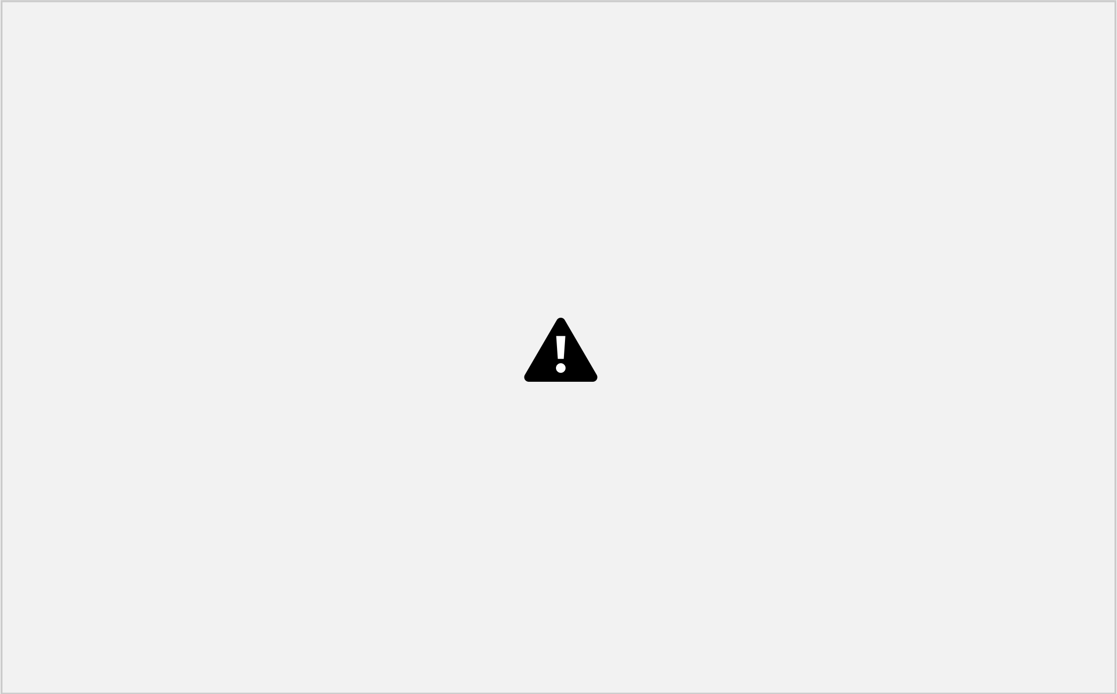
11. Verification of Extremophile-Related Sentences Using BIOBERT

1. **CSV File Specification:**
    - Used CSV files from Task 4, which contained sentences for validation.
2. **Model Setup and Execution:**
    - Implemented Python scripts to handle the verification.
    - Downloaded and set up the BIOBERT model and tokenizer for processing sentences.
3. **Verification Process:**
    - Evaluated each sentence using BIOBERT with a predefined threshold. ○ Sentences with scores above the threshold were classified as True (relevant to extremophiles), while those below were marked False.
4. **Output Handling:**
    - Documented the verification results in an output CSV file.
    - The file listed the original sentences along with their verification status (True/False).

This approach leverages advanced NLP tools to accurately identify and classify sentences related to extremophiles, supporting precise scientific analysis and interpretation.

12.1 Submitted file name

Task13_Genes_Extremophiles_results.csv



12. Summary of Sentences

1. **Implementation of BERT Summarizer:**
    - Utilized the BERT Summarizer, a model based on BERT (Bidirectional Encoder Representations from Transformers), tailored for creating abstractive summaries.
2. **Capabilities Exploration:**

- Investigated the model's ability to condense input text while preserving semantic relevance and coherence in the summaries.

3. **Execution of Summarization Tasks:**
   - Applied the BERT Summarizer to various text inputs to assess its effectiveness in generating clear and informative summaries.
4. **Performance Assessment:**
   - Evaluated the model’s performance in capturing the essence and key points of the original text through its abstractive summarization technique.

13.1 Submitted file name

Task14_data_with_summary.csv

13. Protein Information

Retrieval and Processing of UniProt Data:
Initial Phase: Accessing and Processing Protein Data from UniProt: • API Query Construction: Construct URLs to interact with UniProt's REST API, setting parameters for specific organisms and desired data fields.

- **Data Download:** Obtain compressed TSV files with comprehensive protein data according to the constructed queries.
- **Data Extraction and Compilation:** Parse and consolidate the extracted data from TSV files into a unified CSV format, facilitating further analysis and integration with other datasets.

Error Handling: Any encountered errors during the retrieval process are logged and managed to maintain data integrity and pipeline reliability.

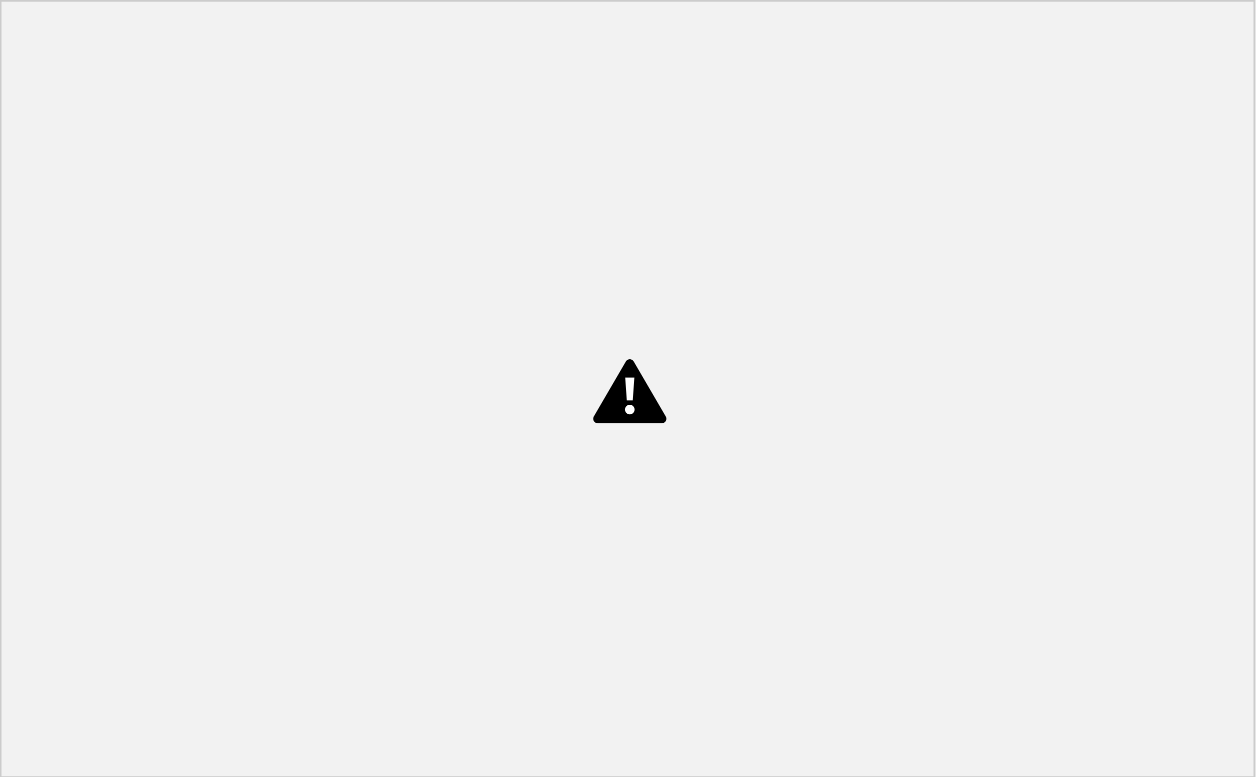
14.1 Submitted file name

Task15_merged_data.csv

Protein Analysis

Input files for Protein Analysis:

TASK_10_CLUBBED_SPECIES_OUTPUT.csv



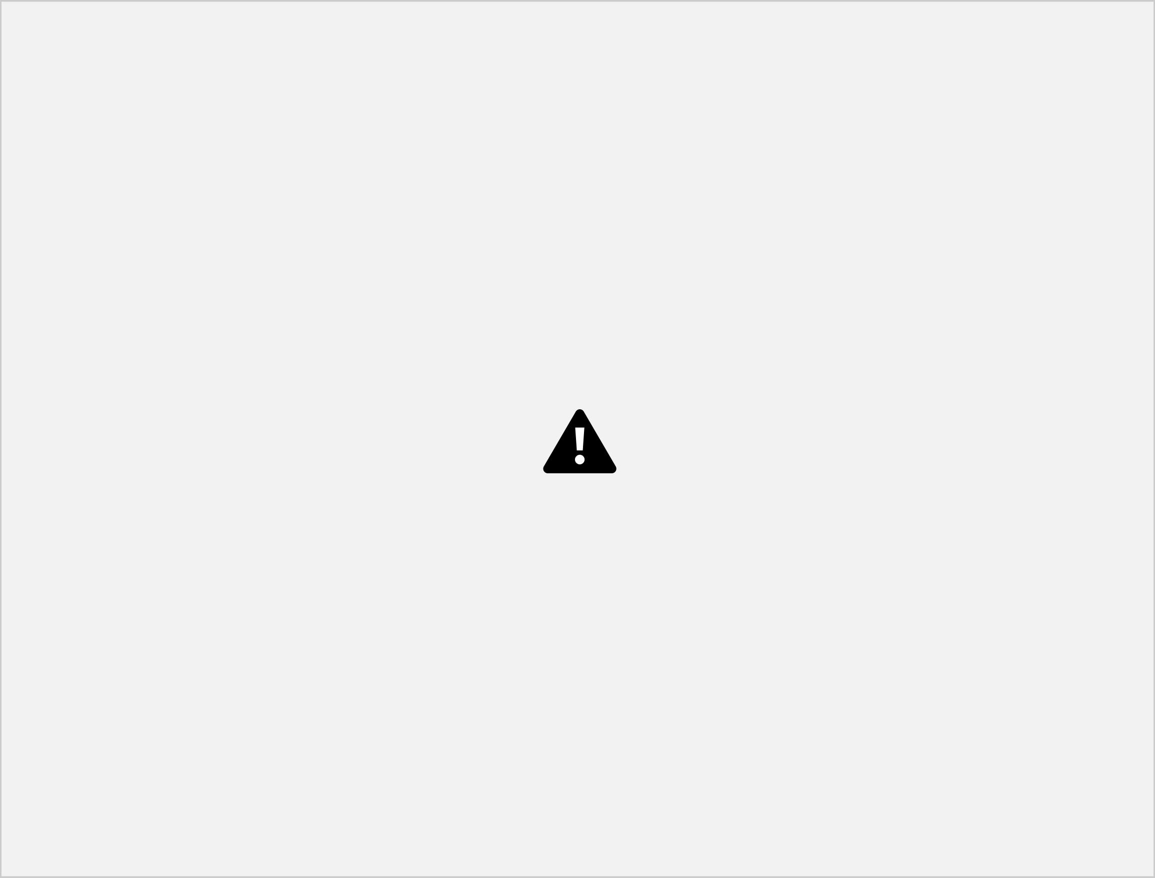
List Of Proteins/Genes (Helps in Survival Of Organisms) :-

Result : Out of the 181 high temperature extremophile organisms,we are able to retrieve the protein information of only 341 in TASK_10_CLUBBED_SPECIES_OUTPUT.csv and failed to get the information of 77 organisms that have been saved in failed_organism.csv.

Uniprot:

Code:

Output files for Uniprot:



2. Retrieving Protein Structures from the RCSB Protein Data Bank (PDB) Sending Requests to the RCSB Search API:

- **Create Search Queries:** Formulate search queries to find protein structures related to the target organisms.
- **Send Requests:** Use these queries to send requests to the RCSB search API.
- **Save Results:** Save the JSON responses with search results to individual JSON files for each query.

Extracting Identifiers:

- **Parse Responses:** Read the JSON responses to extract unique identifiers (like PDB IDs) for the relevant protein structures.

- **Compile List:** Create a list of these identifiers for the next steps of data retrieval.

Fetching Detailed Data:

- **Construct URLs:** Use the extracted identifiers to build URLs for fetching detailed protein structure data from the RCSB API.
- **Send Requests:** Request the detailed data using these URLs.
- **Extract Information:** Parse the JSON responses to get details such as atomic coordinates, secondary structure, and experimental methods.

Data Retrieval and Storage

Constructing URLs for Data Fetching:

- **Build URLs:** Use the list of identifiers to construct URLs for fetching detailed protein structure data from the RCSB API.

Fetching and Parsing JSON Responses:

- **Send Requests**: Request detailed data using the constructed URLs.
- **Extract Data:** Parse the JSON responses to get comprehensive information about each protein structure.

Storing and Reporting:

- **Store Data:** Save the parsed data in a structured format, like a database or a set of CSV files, for easy access and analysis.
- **Handle Errors:** Implement error handling to report any issues during data retrieval, such as missing data or failed requests.
- **Maintain Logs**: Keep logs of the retrieval process to ensure reproducibility and help troubleshooting.

with

14. Knowledge -

Interaction graph

Objective: The goal of this project is to visualize and analyze interactions between chemicals and genes using an interactive graph visualization tool.

1. **Data Acquisition and Preprocessing:**
   - **Data Sources:** Load information from three CSV files (chemicals_file_path, genes_file_path, interactions_file_path) into Pandas dataframes, which contain details on chemicals, genes, and their interactions.

   - **Text Preprocessing:** Standardize entity names (chemicals and genes) by removing punctuation and converting to lowercase to ensure uniform data representation.

2. **Graph Construction:**
   - **NetworkX Graph:** Use NetworkX, a Python library for network creation and analysis, to build a directed graph (G).

   - **Edges and Attributes:** Add edges to the graph to represent interactions between chemicals and genes, including interaction types as annotations from the interactions_df.

3. **Node and Edge Representation:**
   - **Node Information:** Nodes represent chemicals and genes, with additional metadata such as IDs and PubMed references stored in dictionaries (chemical_info, gene_info).

   - **Edge Information:** Edges show interaction types as labels to clarify the relationships between chemicals and genes.

4. **Streamlit App Configuration:**
   - **Page Configuration:** Set up Streamlit's interface (set_page_config) for an improved user experience, featuring an expanded sidebar and an "About" section to explain the graph's purpose.

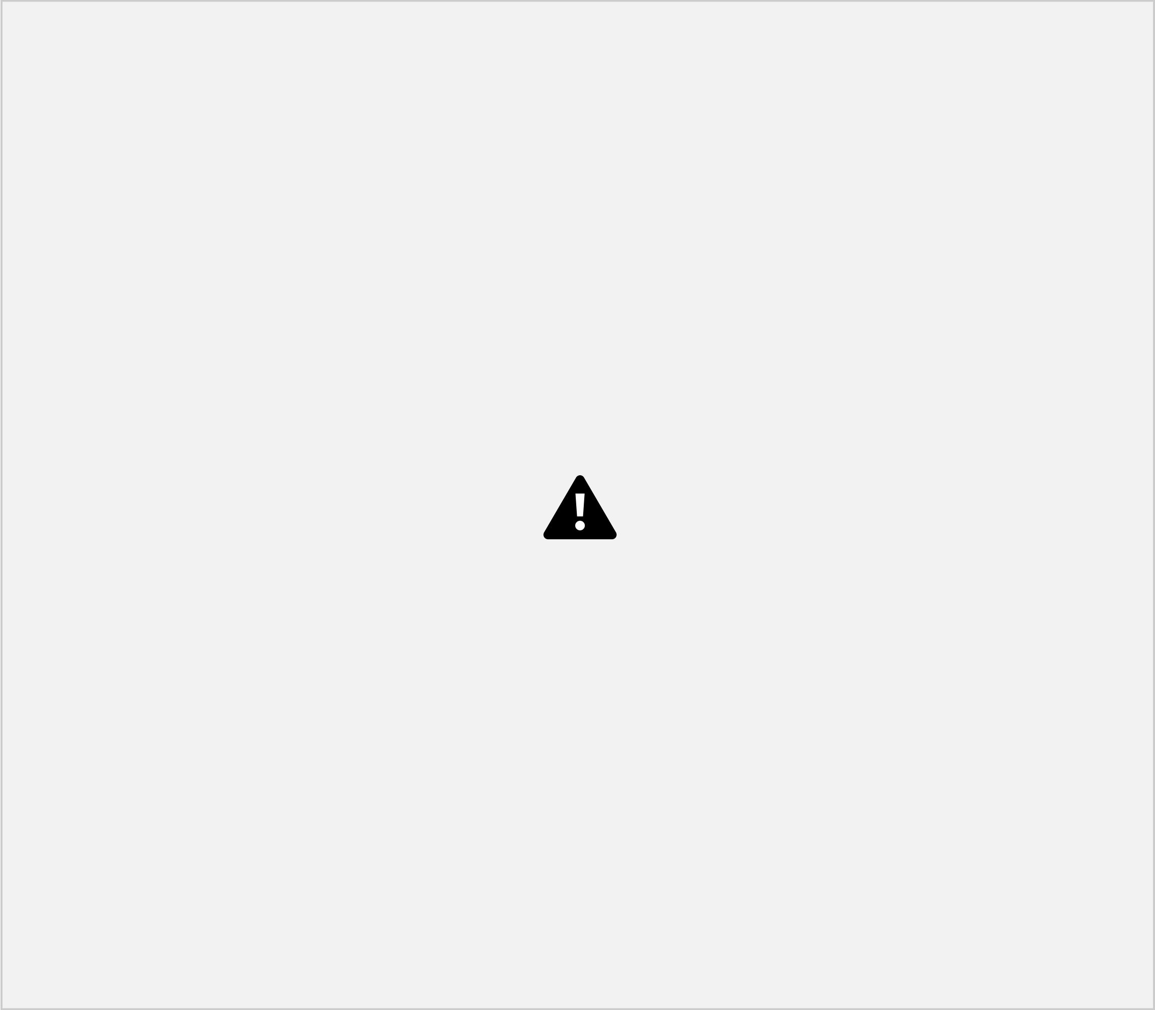
   - **Visual Theme:** Implement a light theme using custom CSS (st.markdown) to enhance readability and visual appeal.

5. **Interactive Graph Visualization:**
   - **Streamlit agraph Component:** Display the interactive graph using streamlit_agraph, allowing users to dynamically explore chemical-gene interactions.

6. **Future Development Potential:**
   - Investigate integrating advanced analytics for predictive modeling of interactions.

   - Improve user interactivity with features like tooltips for detailed node and edge information.
   - Expand the application to include additional biological data sources for a more comprehensive analysis.

Deployed link : https://knowledgegraphrepoforextremophiles
nb7fytzy98mcr2abrkvnr.streamlit.app/



1. Nodes:

   - Green Nodes: These likely represent Gens.

   - Blue Nodes: These might represent Chemicals.

2. Edges:

   - The lines (edges) connecting the nodes represent the relationships or interactions between these entities.

3. Labels:

   - Each node is labeled with the name of the entity it represents. For example, "bevacizumab," "AMP," "CO2," etc.

- Each edge is labeled with the type of interaction it represents, such as "Inhibition" or "Agonist."

4. Cluster/Groups:

- The nodes are grouped based on their relationships..

5. Central Nodes:

- Some nodes appear to be more central with many edges connecting to them

Graph To Knowledge Graph

Gene To Gene Graph:

Gene To Gene Interactive Knowledge Graph:

Gene-gene interactions were represented to explore regulatory networks and pathways. These interactions can indicate how genes influence each other's expression and activity, crucial for understanding genetic regulation and cellular function.

Implementation Steps:

Data Extraction: Extracted gene-gene interaction data from scientific literature and databases.

Edge Representation: Created edges between gene nodes to represent interactions such as co-expression, suppression, or enhancement. **Pathway Analysis**: Analyzed interaction patterns to identify key regulatory pathways.

Gene To Chemical Graph:

Gene To Chemical Interactive Knowledge Graph:

This section of the knowledge graph aimed to illustrate the links between specific genes and chemicals. By visualizing these interactions, researchers can gain insights into genetic predispositions and the molecular basis of various diseases.

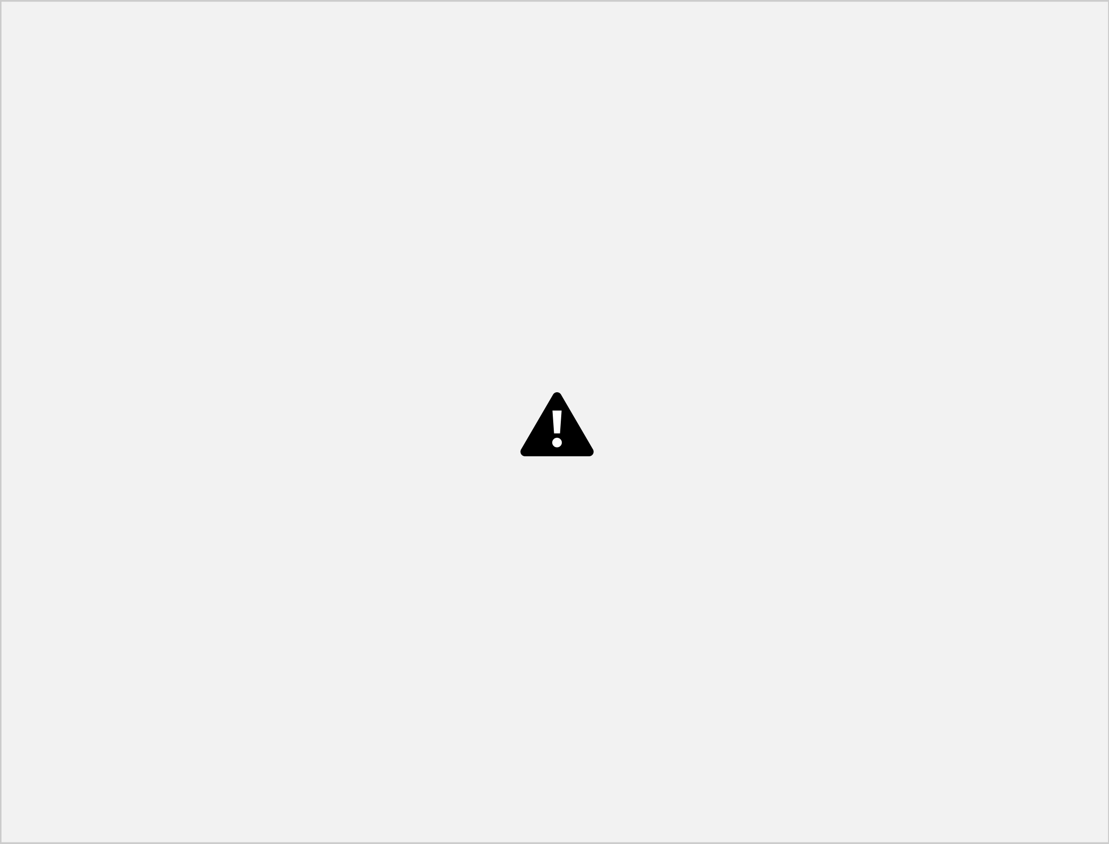
Implementation Steps:

Disease Association: Mapped genes to diseases based on known associations from biomedical research.

Visualization: Created connections between gene nodes and disease node highlighting significant associations.

Contextual Information: Provided detailed information on the nature of gene-disease relationships.

Gene To Disease Graph:



Gene To Disease Interactive Knowledge Graph:

This section of the knowledge graph aimed to illustrate the links between specific genes and diseases. By visualizing these interactions, researchers can gain insights into genetic predispositions and the molecular basis of various diseases.

Implementation Steps:

Disease Association: Mapped genes to diseases based on known associations from biomedical research.

Visualization: Created connections between gene nodes and disease node highlighting significant associations.

Contextual Information: Provided detailed information on the nature of gene-disease relationships.

Database Download and Verification

Resources for Studying Acidophilic Microbes

Acidophile Database (AD)

- **Navigation:** AD features a user-friendly design that simplifies finding information related to acidophilic organisms.

- **Distributed Information:** Data is segmented across various categories such as genomic sequences, research articles, and environmental conditions. Users must navigate through multiple sections to gather comprehensive data.

Acidic Microbiome Project (AMP)

- **Ease of Access:** The AMP offers a straightforward layout with tools that make it easier for researchers to access information on microbes thriving in acidic environments.

- **Segmented Data:** The platform organizes data into distinct sections focusing on different acidophiles or specific acidic habitats. Comprehensive data collection may require exploring various sections.

Acidophile Genomes Database (AGD)

- **Searchability:** The AGD features a user-friendly search interface that facilitates finding genomic information related to acidophilic organisms. Filters are available based on organism type, habitat, and project status.
- **Diverse Data Sources:** Data is dispersed across multiple projects and studies, requiring users to integrate information from different entries to get a complete view.

Acidic Enzyme Database (AED)

- **Interface Design:** AED is designed with accessibility in mind, providing an easy-to-use interface for retrieving data on enzymes from acidophilic microbes.
- **Fragmented Information:** Data on enzyme properties, functions, and related research are spread out, necessitating thorough exploration to gain a complete understanding.

These resources offer valuable insights into extremophiles that thrive in acidic conditions, though users may need to navigate through multiple sections or databases to gather comprehensive information.

Conclusion

I have devoted my efforts to working on Overleaf, focusing on developing a detailed database and knowledge graph of extremophiles in extreme environments. This project comprises several key elements:

- **Methodology:** Outlines the methods used for data collection and processing from scientific literature, with a particular emphasis on text mining from PubMed.
- **Data Collection:** Details the keywords and search strategies for gathering relevant studies on extremophiles and the criteria for study inclusion or exclusion.
- **Data Processing:** Describes the process of cleaning and organizing extracted data to make it suitable for database integration.
- **Database Development:** Explains how the organized data is compiled into a comprehensive database covering taxonomy, habitats, genetic information, and biochemical pathways of extremophiles.
- **Knowledge Graph Construction:** Involves creating a 3D knowledge graph to illustrate the interactions between chemicals and genes, improving the understanding of complex relationships in extremophiles.

This thorough documentation serves as a valuable resource for researchers exploring extremophiles' diversity, genomics, and biotechnological potential. Additionally, I use Mendeley for managing references and citations, and Overleaf for maintaining an organized and detailed record of my work to support future research and collaboration.

Future Goals

1. **Enhanced Data Integration**: Develop methods to integrate additional data sources, such as proteomic and metabolomic data, to

enrich the extremophile database and provide a more comprehensive view of their biological functions.

2. **Advanced Analytical Tools:** Implement advanced analytical tools and machine learning algorithms to predict extremophiles' responses to various environmental conditions, potentially discovering novel extremophiles or new applications in biotechnology.

3. **Interactive Visualization:** Expand the 3D knowledge graph into an interactive platform that allows users to explore relationships and interactions between chemicals, genes, and extremophiles dynamically.

4. **Collaborative Platform Development:** Create a collaborative platform for researchers to contribute data, share insights, and refine the database, fostering a community-driven approach to extremophile research.

5. **Automated Literature Mining:** Enhance text mining capabilities with automated systems for real-time literature updates, ensuring the database remains current with the latest research developments.

These

steps will advance the research capabilities and applications of extremophiles, furthering our understanding and utilization of these unique organisms.