Ramapo College of New Jersey

Viral Protein Interactions in Plants Database

By

Ritwik Katiyar

In partial fulfillment of the requirement
for the degree of the Bachelor of Science
May 10th 2019

Abstract

Understanding viral proteins and small interfering RNAs (siRNAs) is crucial for developing effective strategies to combat viruses. A comprehensive database that systematically catalogues viral proteins, coupled with information on the specific hosts they infect, holds significant promise in advancing our comprehension of viral capabilities. Such a resource could provide a valuable foundation for proactive measures in countering viral threats, particularly in safeguarding agricultural crops against viral diseases. Establishing a robust database of this nature is instrumental in elevating our ability to anticipate and address viral challenges, thereby contributing to the development of innovative solutions for virus control and crop protection.

Background And Introduction

In 1892, the Russian botanist Dimitrii Ivanovsky conducted an experiment, revealing the presence of a pathogen distinct from bacteria infecting tobacco plants. This organism exhibited the capability to traverse porcelain filters designed to retain bacteria. Subsequently, in 1898, Marcus Beijerinck made analogous observations, leading him to conclude that the pathogen represented a distinct organism. Beijerinck officially coined the term "virus" (Lecoq). Since these groundbreaking discoveries, the study of viruses has evolved into its own subfield of virology. Numerous viruses have been identified, named, and sequenced, prompting the creation of various viral databases. These databases serve to record potential viral hosts, viral sequences, or simply document the existence of specific viruses. Following the revelation of viruses, substantial progress has been made in developing numerous vaccinations, leading to the near eradication of major viral diseases in humans. Vaccinations have also been successfully created for domestic animals, wild animals, and even plants, with the latter commonly referred to as pesticides. The application of pesticides has played a pivotal role in ensuring the safety of our crops. Nevertheless, despite these advancements, there remains a considerable amount that we do not comprehend about viruses (Smith).

Viruses are essentially simple parasitic entities that depend on a host organism for their survival. Lacking a proper cellular structure of their own, viruses necessitate a host to facilitate their replication. They exhibit a diverse array of shapes and sizes, predominantly comprising either DNA or RNA as their genetic material. The nucleic acid sequence in viruses can exist in forms such as single or double-stranded DNA or RNA. Additionally, viruses possess the capability to encode various proteins, ranging from a minimal count of 3-4 proteins to a more substantial range of 100-200 proteins (Lodish). The mechanism by which viruses infect their hosts is relatively straightforward. Initially, a virus infiltrates the membrane or membranes of the host cell. Given their diminutive size and uncomplicated structure, a single virus contains a limited amount of RNA or DNA, translating to a restricted number of genes. These genes play a pivotal role not only in coding for proteins necessary for the virus's replication within the host cell but also in influencing the production of proteins by the host cell itself (Smith) (Lodish). This intricate interaction enables the virus to commandeer the host cell's machinery, fostering its replication. Ultimately, the virus replicates sufficiently to breach the host cell and proceed to infect neighboring cells in a cascading fashion.

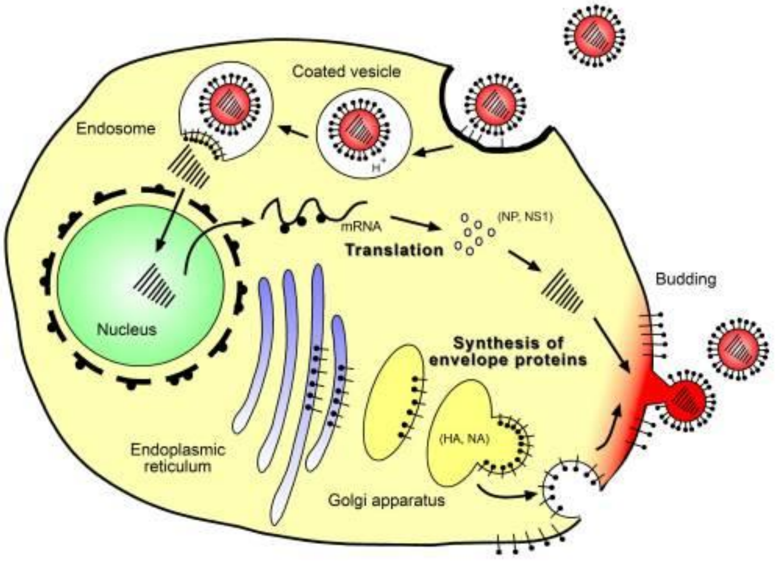


Figure 1: Replication cycle of influenza-A virus. Binding and entry of the virus, fusion with endosomal membrane and release of viral RNA, replication within the nucleus, synthesis of structural and envelope proteins, budding and release of virions capable of infecting neighboring epithelial cells. (Kamps)

Evidently, various plants, animals, and even certain single-celled organisms have evolved natural antiviral defense systems. However, many viruses have evolved countermeasures to thwart these host antiviral defenses. A crucial strategy employed by plants in defense against viral infections is known as si-RNA-mediated (small interfering RNA-mediated) gene silencing. This mechanism not only allows the defensive signal to propagate to neighboring cells but also shares similarities with mi-RNA (micro-RNA)-mediated gene silencing. In the realm of plant antiviral defense, mi-RNAs serve two key functions: they target viral RNA/DNA, preventing viral reproduction, and they stimulate the synthesis/biogenesis of si-RNA, the primary antiviral response in plant cells. Despite these natural defenses, viruses have developed a counter-defense system against plant cells known as viral suppressors of RNA silencing (VSRs). These VSRs interfere with host RNA silencing through various mechanisms, including their involvement in viral replication, encapsidation, and movement. VSRs primarily employ two methods to counteract plant defense systems. Firstly, they suppress and assemble AGOs (Argonautes, a protein family crucial in RNA silencing processes) into RISCs (RNA-induced silencing complex), designed to silence viral DNA/RNA. Secondly, VSRs interact with AGOs to degrade the proteins involved in the antiviral response. The schematic representation below illustrates the dynamic interplay between these viral countermeasures and the plant's antiviral mechanisms (Rui) (Kamthan).

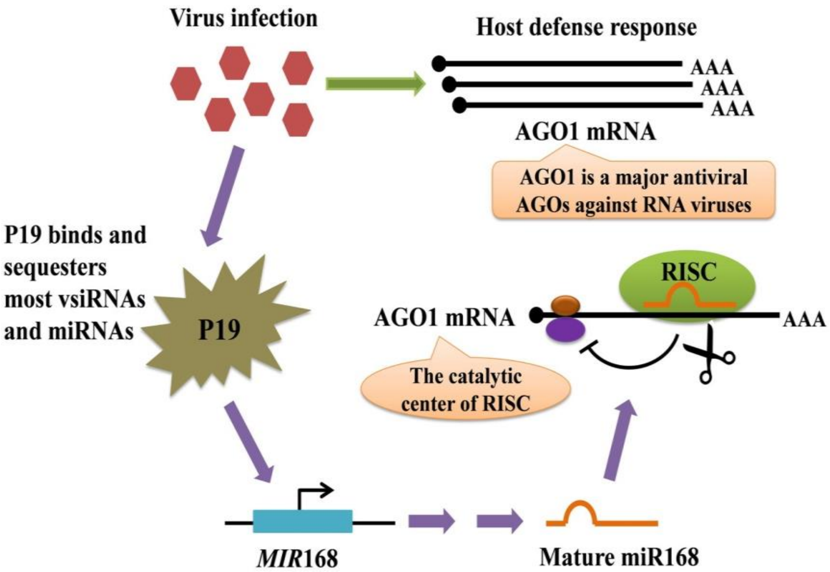

Virus infection
Host defense response
AAA
AAA
AAA
AGO1 mRNA
AGO1 is a major antiviral AGOs against RNA viruses
P19 binds and sequesters most vsiRNAs and miRNAs
P19
RISC
AGO1 mRNA
AAA
The catalytic center of RISC
MIR168
Mature miR168

Figure 2: represents a model for the regulation of AGO mRNA level mediated by protein P19-induced miR168. The AGO mRNAs are made when a virus infection is detected within the cell. While the plant cell is creating mRNAs for AGO the virus produces P19 VSR which binds to the virus- encoded siRNAs and with host cell miRNAs which prevents the miRNA loading into AGO. The P19 protein however, does not bind to MIR168 micro RNA (MIR168, represses the AGO1 mRNA); resulting in low production of AGO1 mRNA. (Rui) (Pratt)

In light of the presented information, it becomes evident that comprehending viral proteins and viral si-RNAs is crucial for devising effective strategies to counteract viruses and safeguard agricultural crops from viral diseases. Therefore, the establishment of a comprehensive database capable of recording viral proteins in conjunction with the type of host the virus infects emerges as a valuable resource.

Such a database would prove beneficial by enhancing our understanding of the capabilities of various viruses. It would provide essential insights into the molecular mechanisms employed by viruses during infection, aiding in the development of advanced countermeasures. By cataloging viral proteins and their interactions with specific hosts, researchers and agricultural professionals could gain a strategic advantage in anticipating and combatting viral threats. This proactive approach holds the potential to prevent significant crop losses resulting from viral outbreaks.

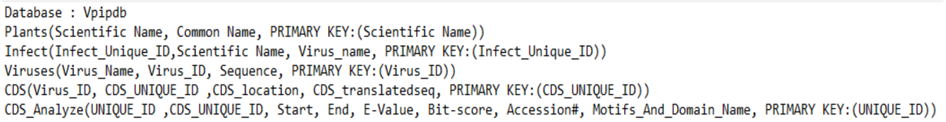
In essence, a dedicated viral protein database would contribute to the formulation of targeted interventions, fostering a more resilient and secure agricultural landscape. The proactive management of viral diseases in crops could thereby mitigate the economic and environmental impact of viral infections on agriculture.

Methodology

Requirements & Materials	Summary
Visual Studio Code	Used to write all the code
Python and Bio-python	Python was used as a base to bio-python and postgresql. Bio-python was used to grab viral sequence information and viral IDs
Flask, HTML and CSS	To visualize all the data within the data and create an app to showcase all the data
Postgresql (Elephant SQL)	Database where the data was stored
R, R-studio and Taxize	Used to grab scientific names and common names of plants
Beautiful Soup	Used to grab information from Plant viruses online website
NCBI and Conserved Domains Database	NCBI was used from bio-python and CDD was used to search for various motifs and domains within a sequence of translated CDS's
Microsoft Excel	To better organize the raw data gathered from various sources
Matplotlib and Pandas	To create the graph on the home page of the website.

Table 1: Shows a comprehensive list of requirements for the following methods of setting up the database

1. The first step to creating any such database is to develop a schema for the database.



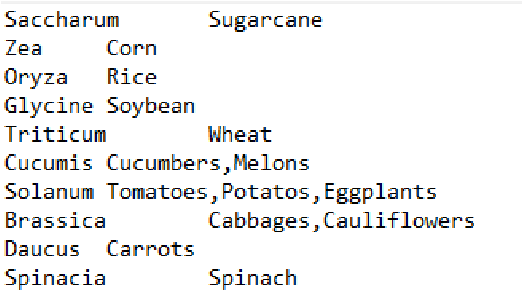
```
Database : Vpipdb
Plants(Scientific Name, Common Name, PRIMARY KEY:(Scientific Name))
Infect(Infect_Unique_ID,Scientific Name, Virus_name, PRIMARY KEY:(Infect_Unique_ID))
Viruses(Virus_Name, Virus_ID, Sequence, PRIMARY KEY:(Virus_ID))
CDS(Virus_ID, CDS_UNIQUE_ID ,CDS_location, CDS_translatedseq, PRIMARY KEY:(CDS_UNIQUE_ID))
CDS_Analyze(UNIQUE_ID ,CDS_UNIQUE_ID, Start, End, E-Value, Bit-score, Accession#, Motifs_And_Domain_Name, PRIMARY KEY:(UNIQUE_ID))
```

Figure 3: displays the schema of the proposed database.

2. Next to select the plants for the database. Since there are numerous agricultural crops grown around the world and a lack of data linking a specific virus to a particular crop; it is important to focus on the top grown crops around the world. To do that, the data about the amount of agricultural crops grown around the world was taken from Food and

agricultural organization (FAO). This data was then sorted and only the top 10 plants were kept.

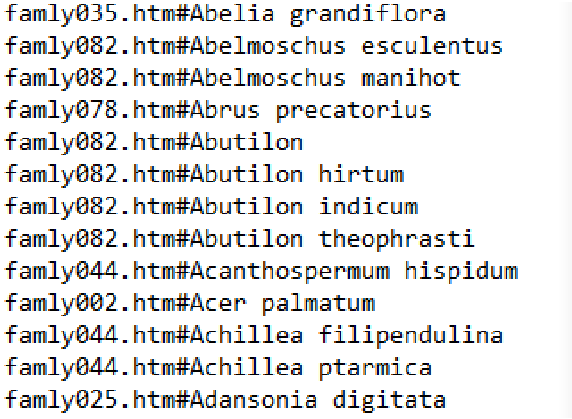
3. FAO presents the data with the use of common names hence to get the scientific names of the data a simple R script was used that would grab the genus of the plants based on the common name.



```
Saccharum       Sugarcane
Zea     Corn
Oryza   Rice
Glycine Soybean
Triticum        Wheat
Cucumis Cucumbers,Melons
Solanum Tomatoes,Potatos,Eggplants
Brassica        Cabbages,Cauliflowers
Daucus  Carrots
Spinacia        Spinach
```

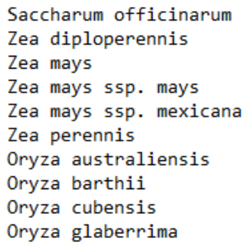
Figure 5: Displays the data that was obtained from the script.

4. With this information in hand a script was written using beautiful soup to gather information from the Plant and virus's online website. The data that was extracted from Plant and viruses online website was the data on plant susceptibility to a specific virus , all the plant species that the data was available for and all the viruses that were recorded by the website.



```
famly035.htm#Abelia grandiflora
famly082.htm#Abelmoschus esculentus
famly082.htm#Abelmoschus manihot
famly078.htm#Abrus precatorius
famly082.htm#Abutilon
famly082.htm#Abutilon hirtum
famly082.htm#Abutilon indicum
famly082.htm#Abutilon theophrasti
famly044.htm#Acanthospermum hispidum
famly002.htm#Acer palmatum
famly044.htm#Achillea filipendulina
famly044.htm#Achillea ptarmica
famly025.htm#Adansonia digitata
```

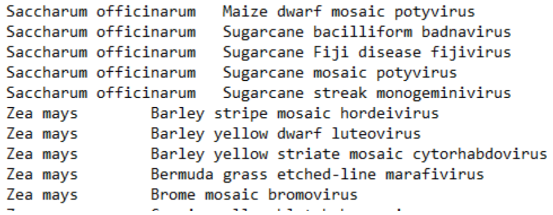
Figure 7: Shows the output from the script.



```
Saccharum officinarum
Zea diploperennis
Zea mays
Zea mays ssp. mays
Zea mays ssp. mexicana
Zea perennis
Oryza australiensis
Oryza barthii
Oryza cubensis
Oryza glaberrima
```

Figure 8: The data was then organized with the use of Excel.

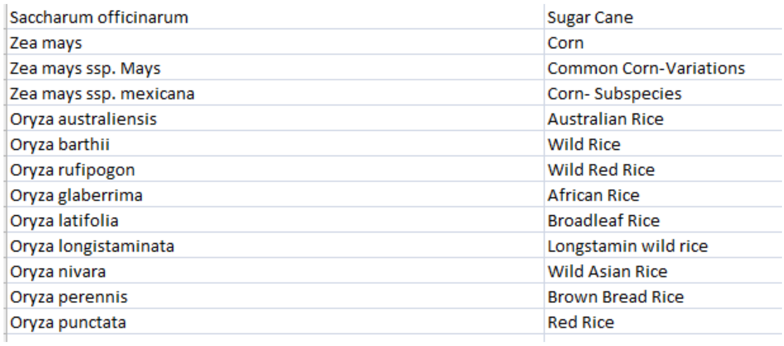
5. Once the list of all plants and the viruses used by the Plant Viruses Online website was gathered, the list was then organized and arranged using excel to contain appropriate data. The viruses were also then linked with their host plants; figure 10 shows the final version of what the data looked like.



```
Saccharum officinarum   Maize dwarf mosaic potyvirus
Saccharum officinarum   Sugarcane bacilliform badnavirus
Saccharum officinarum   Sugarcane Fiji disease fijivirus
Saccharum officinarum   Sugarcane mosaic potyvirus
Saccharum officinarum   Sugarcane streak monogeminivirus
Zea mays        Barley stripe mosaic hordeivirus
Zea mays        Barley yellow dwarf luteovirus
Zea mays        Barley yellow striate mosaic cytorhabdovirus
Zea mays        Bermuda grass etched-line marafivirus
Zea mays        Brome mosaic bromovirus
```

Figure 10: shows how the data looks like once it’s been organized in the tab delimited format. The first column shows the species of plant. The second column shows the virus that infects the plant.

6. With the use of a similar script to that shown in figure 4, the common names were determined for each and every plant.

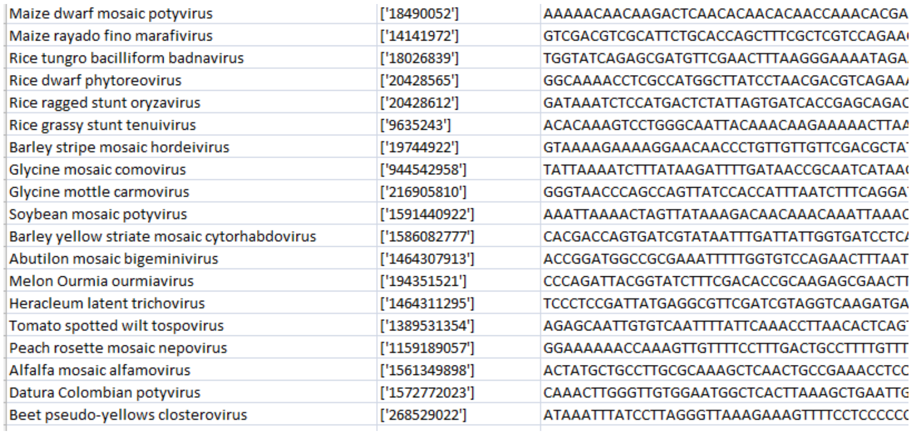


Saccharum officinarum	Sugar Cane
Zea mays	Corn
Zea mays ssp. Mays	Common Corn-Variations
Zea mays ssp. mexicana	Corn- Subspecies
Oryza australiensis	Australian Rice
Oryza barthii	Wild Rice
Oryza rufipogon	Wild Red Rice
Oryza glaberrima	African Rice
Oryza latifolia	Broadleaf Rice
Oryza longistaminata	Longstamin wild rice
Oryza nivara	Wild Asian Rice
Oryza perennis	Brown Bread Rice
Oryza punctata	Red Rice

Figure 11: The data for the plants table mentioned in the schema in figure 3. Scientific Names on the left and the common names on the right

7. Once the viruses were connected to their respective host plants the NCBI viruses ID’s was determined using a script. Then another script was used to grab the sequences for the viruses based on the determined ID.

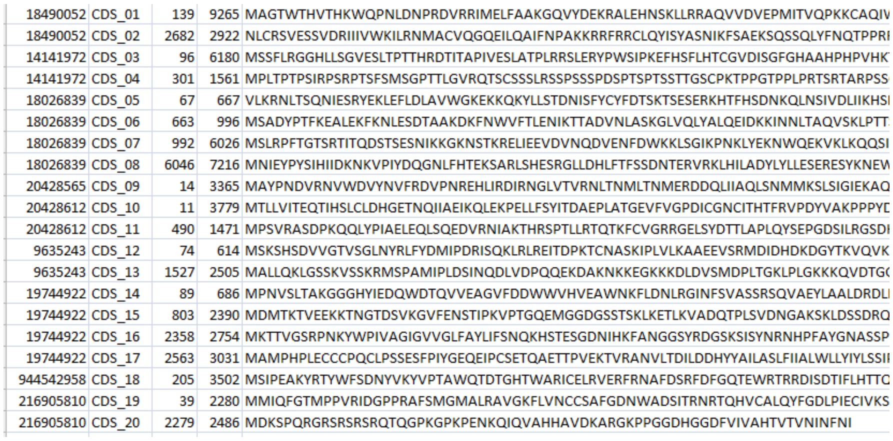
8. The resulting data was then arranged using excel



Maize dwarf mosaic potyvirus	['18490052']	AAAAACAACAAGACTCAACACAACACAACCAAACACGA
Maize rayado fino marafivirus	['14141972']	GTCGACGTCGCATTCTGCACCAGCTTTCGCTCGTCCAGAA
Rice tungro bacilliform badnavirus	['18026839']	TGGTATCAGAGCGATGTTCGAACTTTAAGGGAAAATAGA
Rice dwarf phytoreovirus	['20428565']	GGCAAAACCTCGCCATGGCTTATCCTAACGACGTCAGAA
Rice ragged stunt oryzavirus	['20428612']	GATAAATCTCCATGACTCTATTAGTGATCACCGAGCAGA
Rice grassy stunt tenuivirus	['9635243']	ACACAAAGTCCTGGGCAATTACAAACAAGAAAAACTTA
Barley stripe mosaic hordeivirus	['19744922']	GTAAAAGAAAAGGAACAACCCTGTTGTTGTTCGACGCTA
Glycine mosaic comovirus	['944542958']	TATTAAAATCTTTATAAGATTTTGATAACCGCAATCATAA
Glycine mottle carmovirus	['216905810']	GGGTAACCCAGCCAGTTATCCACCATTTAATCTTTCAGGA
Soybean mosaic potyvirus	['1591440922']	AAATTAAAACTAGTTATAAAGACAACAAACAAATTAAA
Barley yellow striate mosaic cytorhabdovirus	['1586082777']	CACGACCAGTGATCGTATAATTTGATTATTGGTGATCCTC
Abutilon mosaic bigeminivirus	['1464307913']	ACCGGATGGCCGCGAAATTTTTGGTGTCCAGAACTTTAAT
Melon Ourmia ourmiavirus	['194351521']	CCCAGATTACGGTATCTTTCGACACCGCAAGAGCGAACTT
Heracleum latent trichovirus	['1464311295']	TCCCTCCGATTATGAGGCGTTCGATCGTAGGTCAAGATGA
Tomato spotted wilt tospovirus	['1389531354']	AGAGCAATTGTGTCAATTTTATTCAAACCTTAACACTCAG
Peach rosette mosaic nepovirus	['1159189057']	GGAAAAAACCAAAGTTGTTTTCCTTTGACTGCCTTTTGTTT
Alfalfa mosaic alfamovirus	['1561349898']	ACTATGCTGCCTTGCGCAAAGCTCAACTGCCGAAACCTC
Datura Colombian potyvirus	['1572772023']	CAAACTTGGGTTGTGGAATGGCTCACTTAAAGCTGAATTG
Beet pseudo-yellows closterovirus	['268529022']	ATAAATTTATCCTTAGGGTTAAAGAAAGTTTTCCTCCCCC

Figure 14: shows how the data looked in excel. The first column shows the name of the virus. The second column shows the ID found. The third column shows the complete sequence of the virus.

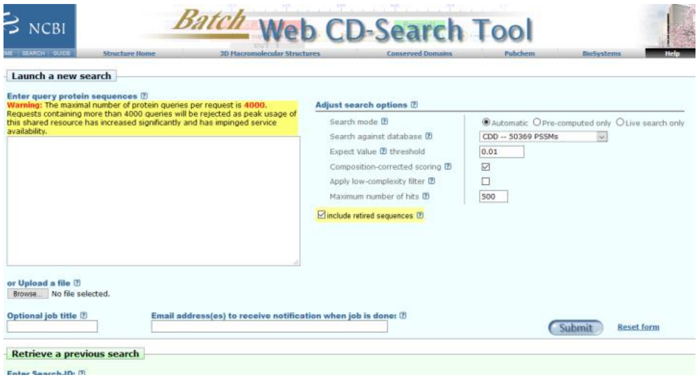
9. After the sequences were obtained the next step was to obtain information on potential CDS regions for each virus. Not all viruses contain a CDS region and for viruses without a CDS region no further data was necessary to obtain.



18490052	CDS_01	139	9265	MAGTWTHVTHKWQPNLDNPRDVRRIMELFAAKGQVYDEKRALEHNSKLLRRAQVVDVEPMITVQPKKCAQI
18490052	CDS_02	2682	2922	NLCRSVESSVDRIIIVWKILRNMACVQGQEILQAIFNPAKKRRFRRCLQYISYASNIKFSAEKSQSSQLYFNQTPPR
14141972	CDS_03	96	6180	MSSFLRGGHLLSGVESLTPTTHRDTITAPIVESLATPLRRSLERYPWSIPKEFHSFLHTCGVDISGFGHAAHPHPVHK
14141972	CDS_04	301	1561	MPLTPTPSIRPSRPTSFSMSGPTTLGVRQTSCSSSLRSSPSSSPDSPTSPTSSTTGSCPKTPPGTPPLPRTSRTARPSS
18026839	CDS_05	67	667	VLKRNLTSQNIESRYEKLEFLDLAVWGKEKKQKYLLSTDNISFYCYFDTSKTSESERKHTFHSDNKQLNSIVDLIIKHS
18026839	CDS_06	663	996	MSADYPTFKEALEKFKNLESDTAAKDKFNWVFTLENIKTTADVNLASKGLVQLYALQEIDKKINNLTAQVSKLPTT
18026839	CDS_07	992	6026	MSLRPFTGTSRTITQDSTSESNIKKGKNSTKRELIEEVDVNQDVENFDWKKLSGIKPNKLYEKNWQEKVKLKQQSI
18026839	CDS_08	6046	7216	MNIEYPYSIHIIDKNKVPIYDQGNLFHTEKSARLSHESRGLLDHLFTFSSDNTERVRKLHILADYLYLLESERESYKNE
20428565	CDS_09	14	3365	MAYPNDVRNVWDVYNVFRDVPNREHLIRDIRNGLVTVRNLTNMLTNMERDDQLIIAQLSNMMKSLSIGIEKAQ
20428612	CDS_10	11	3779	MTLLVITEQTIHSLCLDHGETNQIIAEIKQLEKPELLFSYITDAEPLATGEVFVGPDICGNCITHTFRVPDYVAKPPPY
20428612	CDS_11	490	1471	MPSVRASDPKQQLYPIAELEQLSQEDVRNIAKTHRSPTLLRTQTKFCVGRRGELSYDTTLAPLQYSEPGDSILRGSD
9635243	CDS_12	74	614	MSKSHSDVVGTVSGLNYRLFYDMIPDRISQKLRLREITDPKTCNASKIPLVLKAAEEVSRMDIDHDKDGYTKVQVK
9635243	CDS_13	1527	2505	MALLQKLGSSKVSSKRMSPAMIPLDSINQDLVDPQQEKDAKNKKEGKKKDLDVSMDPLTGKLPLGKKKQVDTG
19744922	CDS_14	89	686	MPNVSLTAKGGGHYIEDQWDTQVVEAGVFDDWWVHVEAWNKFLDNLRGINFSVASSRSQVAEYLAALDRDL
19744922	CDS_15	803	2390	MDMTKTVEEKKTNGTDSVKGVFENSTIPKVPTGQEMGGDGSSTSKLKETLKVADQTPLSVDNGAKSKLDSSDRQ
19744922	CDS_16	2358	2754	MKTTVGSRPNKYWPIVAGIGVVGLFAYLIFSNQKHSTESGDNIHKFANGGSYRDGSKSISYNRNHPFAYGNASSP
19744922	CDS_17	2563	3031	MAMPHPLECCCPQCLPSSESFPIYGEQEIPCSETQAETTPVEKTVRANVLTDILDDHYYAILASLFIIALWLLYIYLSSI
944542958	CDS_18	205	3502	MSIPEAKYRTYWFSDNYVKYVPTAWQTDTGHTWARICELRVERFRNAFDSRFDFGQTEWRTRRDISDTIFLHTT
216905810	CDS_19	39	2280	MMIQFGTMPPVRIDGPPRAFSMGMALRAVGKFLVNCCSAFGDNWADSITRNRTQHVCALQYFGDLPIECIVKS
216905810	CDS_20	2279	2486	MDKSPQRGRSRSRSRQTQGPKGPKPENKQIQVAHHAVDKARGKPPGGDHGGDFVIVAHTVTVNINFNI

Figure 16: Shows how the data looks in excel. The first column shows the Virus_ID. The second column shows the Unique_ID given to every CDS region found. The third column shows the length/ region of CDS within the viral genome. Finally, the last column is the translated protein sequence.

10. Using **NCBI: CDD** data on possible Motifs and Domains within the Translated Protein Sequence was obtained. NCBI:CDD allows for batch sequence searches which looked like:



Launch a new search

Enter query protein sequences

Warning: The maximal number of protein queries per request is 4000. Requests containing more than 4000 queries will be rejected as peak usage of this shared resource has increased significantly and has impinged service availability.

Adjust search options

Search mode ● Automatic ○ Pre-computed only ○ Live search only

Search against database CDD -- 50369 PSSMs

Expect Value threshold 0.01

Composition-corrected scoring ☑

Apply low-complexity filter □

Maximum number of hits 500

☑ include retired sequences

or Upload a file

No file selected.

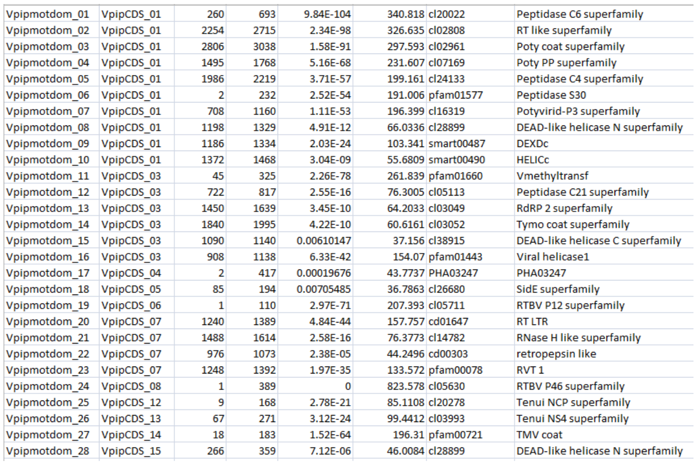
Optional job title

Email address(es) to receive notification when job is done:

Retrieve a previous search

Enter Search-ID:

Figure 17: Shows the CDD batch search.



Vpipmotdom_01	VpipCDS_01	260	693	9.84E-104	340.818	cl20022	Peptidase C6 superfamily
Vpipmotdom_02	VpipCDS_01	2254	2715	2.34E-98	326.635	cl02808	RT like superfamily
Vpipmotdom_03	VpipCDS_01	2806	3038	1.58E-91	297.593	cl02961	Poty coat superfamily
Vpipmotdom_04	VpipCDS_01	1495	1768	5.16E-68	231.607	cl07169	Poty PP superfamily
Vpipmotdom_05	VpipCDS_01	1986	2219	3.71E-57	199.161	cl24133	Peptidase C4 superfamily
Vpipmotdom_06	VpipCDS_01	2	232	2.52E-54	191.006	pfam01577	Peptidase S30
Vpipmotdom_07	VpipCDS_01	708	1160	1.11E-53	196.399	cl16319	Potyvirid-P3 superfamily
Vpipmotdom_08	VpipCDS_01	1198	1329	4.91E-12	66.0336	cl28899	DEAD-like helicase N superfamily
Vpipmotdom_09	VpipCDS_01	1186	1334	2.03E-24	103.341	smart00487	DEXDc
Vpipmotdom_10	VpipCDS_01	1372	1468	3.04E-09	55.6809	smart00490	HELICc
Vpipmotdom_11	VpipCDS_03	45	325	2.26E-78	261.839	pfam01660	Vmethyltransf
Vpipmotdom_12	VpipCDS_03	722	817	2.55E-16	76.3005	cl05113	Peptidase C21 superfamily
Vpipmotdom_13	VpipCDS_03	1450	1639	3.45E-10	64.2033	cl03049	RdRP 2 superfamily
Vpipmotdom_14	VpipCDS_03	1840	1995	4.22E-10	60.6161	cl03052	Tymo coat superfamily
Vpipmotdom_15	VpipCDS_03	1090	1140	0.00610147	37.156	cl38915	DEAD-like helicase C superfamily
Vpipmotdom_16	VpipCDS_03	908	1138	6.33E-42	154.07	pfam01443	Viral helicase1
Vpipmotdom_17	VpipCDS_04	2	417	0.00019676	43.7737	PHA03247	PHA03247
Vpipmotdom_18	VpipCDS_05	85	194	0.00705485	36.7863	cl26680	SidE superfamily
Vpipmotdom_19	VpipCDS_06	1	110	2.97E-71	207.393	cl05711	RTBV P12 superfamily
Vpipmotdom_20	VpipCDS_07	1240	1389	4.84E-44	157.757	cd01647	RT LTR
Vpipmotdom_21	VpipCDS_07	1488	1614	2.58E-16	76.3773	cl14782	RNase H like superfamily
Vpipmotdom_22	VpipCDS_07	976	1073	2.38E-05	44.2496	cd00303	retropepsin like
Vpipmotdom_23	VpipCDS_07	1248	1392	1.97E-35	133.572	pfam00078	RVT 1
Vpipmotdom_24	VpipCDS_08	1	389	0	823.578	cl05630	RTBV P46 superfamily
Vpipmotdom_25	VpipCDS_12	9	168	2.78E-21	85.1108	cl20278	Tenui NCP superfamily
Vpipmotdom_26	VpipCDS_13	67	271	3.12E-24	99.4412	cl03993	Tenui NS4 superfamily
Vpipmotdom_27	VpipCDS_14	18	183	1.52E-64	196.31	pfam00721	TMV coat
Vpipmotdom_28	VpipCDS_15	266	359	7.12E-06	46.0084	cl28899	DEAD-like helicase N superfamily

Figure 18: shows the final results that were obtained from the NCBI: CDD database. These results are edited and arranged based on the schema. Look at image 1 for reference.

11. With the completion of the data gathering process, the subsequent step involved the systematic incorporation of this gathered information into a database. This crucial stage laid the foundation for executing advanced queries, enabling a more profound analysis of the data.The data was meticulously organized and structured within the database, ensuring its integrity and accessibility for future inquiries. This organized repository not only facilitated efficient storage but also paved the way for the development of complex queries to extract specific insights, trends, and patterns from the amassed dataset. By placing the data in the database, researchers, analysts, or any relevant stakeholders gained the capability to conduct in-depth analyses, generate meaningful reports, and derive valuable conclusions. This structured database framework enhanced the overall

manageability of the data, fostering a conducive environment for informed decision-making and comprehensive exploration of the acquired dataset..

12. Following the acquisition of data, a systematic organization was implemented by structuring the information into a table within the database. Subsequently, a Flask application was established to facilitate seamless interaction with the data. Queries were meticulously crafted to empower users with the ability to harness the gathered information effectively. These queries, embedded within the Flask application, provided users with a dynamic and user-friendly means to access and manipulate the stored data. The establishment of a web interface served as a pivotal component of this process, enhancing user engagement and accessibility. Through the Flask application, users were granted an intuitive platform to interact with the data, enabling them to retrieve specific information, perform analyses, and explore the intricacies of the database. The web interface not only streamlined the user experience but also contributed to the democratization of data utilization, ensuring that stakeholders could leverage the wealth of information housed within the database. In essence, the integration of Flask, coupled with strategically designed queries, transformed the static data into a dynamic resource. This not only empowered users to explore and exploit the gathered data but also facilitated the creation of an interactive and user-centric environment for efficient data utilization.

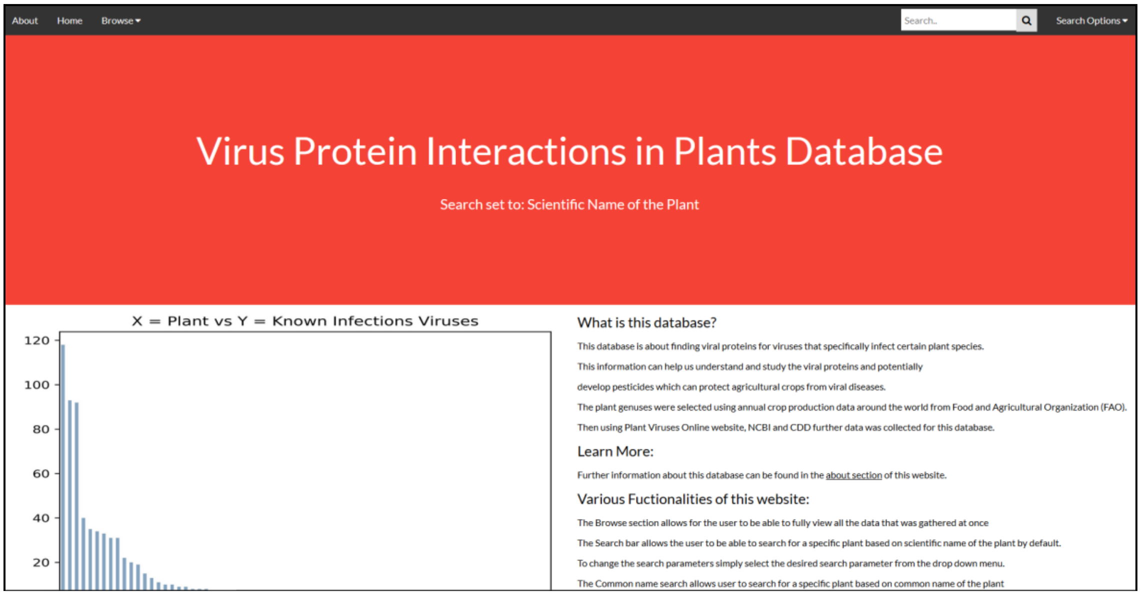
Results

Over All Statistics for the application

10 Plant genuses and 73 Plant species were recorded
193 Viruses were recorded that infect at-least one plant species
Over all 764 Recorded infections
486 Coding regions identified for various viruses
717 Various motifs identified from the coding regions

Home and About pages

The application contains a home page, about page, a browse section (allows the user to browse the all the data that was gathered for the project) and a search bar (which can be set according to what the user wants to search for.



Virus Protein Interactions in Plants Database

Search set to: Scientific Name of the Plant

What is this database?

This database is about finding viral proteins for viruses that specifically infect certain plant species.

This information can help us understand and study the viral proteins and potentially

develop pesticides which can protect agricultural crops from viral diseases.

The plant genuses were selected using annual crop production data around the world from Food and Agricultural Organization (FAO).

Then using Plant Viruses Online website, NCBI and CDD further data was collected for this database.

Learn More:

Further information about this database can be found in the about section of this website.

Various Fuctionalities of this website:

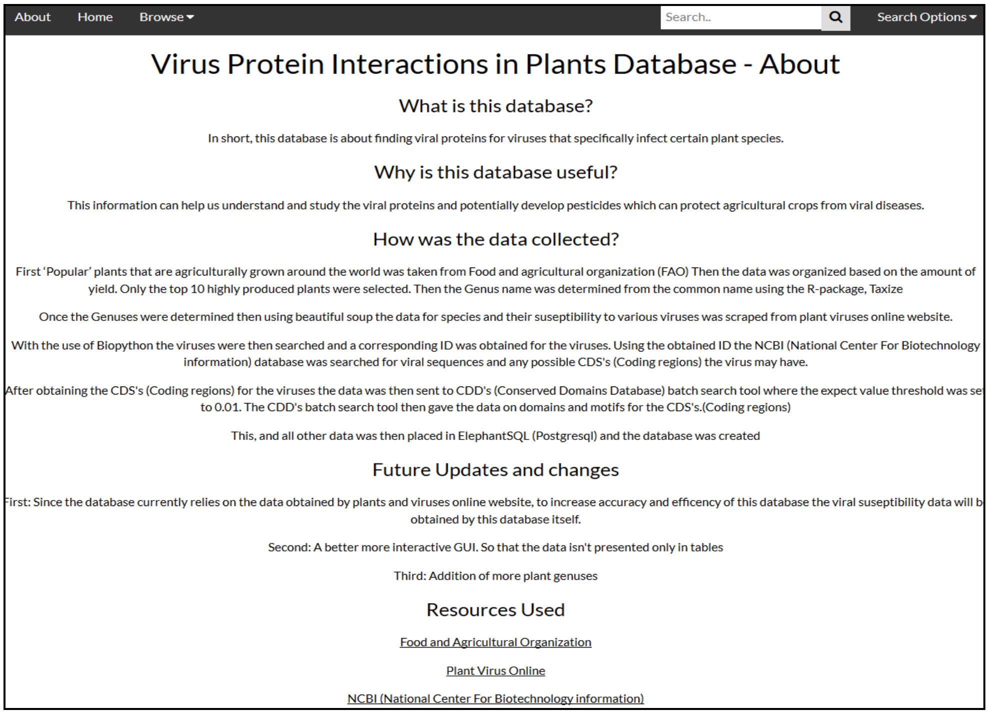
The Browse section allows for the user to be able to fully view all the data that was gathered at once

The Search bar allows the user to be able to search for a specific plant based on scientific name of the plant by default.

To change the search parameters simply select the desired search parameter from the drop down menu.

The Common name search allows user to search for a specific plant based on common name of the plant

Figure 25: Shows the home page of the application. The graph showcases all the recorded infections for a particular plant species. The home page of the application goes in more detail about the database and about the graph itself.



Virus Protein Interactions in Plants Database - About

What is this database?

In short, this database is about finding viral proteins for viruses that specifically infect certain plant species.

Why is this database useful?

This information can help us understand and study the viral proteins and potentially develop pesticides which can protect agricultural crops from viral diseases.

How was the data collected?

First 'Popular' plants that are agriculturally grown around the world was taken from Food and agricultural organization (FAO) Then the data was organized based on the amount of yield. Only the top 10 highly produced plants were selected. Then the Genus name was determined from the common name using the R-package, Taxize

Once the Genuses were determined then using beautiful soup the data for species and their suseptibility to various viruses was scraped from plant viruses online website.

With the use of Biopython the viruses were then searched and a corresponding ID was obtained for the viruses. Using the obtained ID the NCBI (National Center For Biotechnology information) database was searched for viral sequences and any possible CDS's (Coding regions) the virus may have.

After obtaining the CDS's (Coding regions) for the viruses the data was then sent to CDD's (Conserved Domains Database) batch search tool where the expect value threshold was se to 0.01. The CDD's batch search tool then gave the data on domains and motifs for the CDS's.(Coding regions)

This, and all other data was then placed in ElephantSQL (Postgresql) and the database was created

Future Updates and changes

irst: Since the database currently relies on the data obtained by plants and viruses online website, to increase accuracy and efficency of this database the viral suseptibility data will b obtained by this database itself.

Second: A better more interactive GUI. So that the data isn't presented only in tables

Third: Addition of more plant genuses

Resources Used

Food and Agricultural Organization

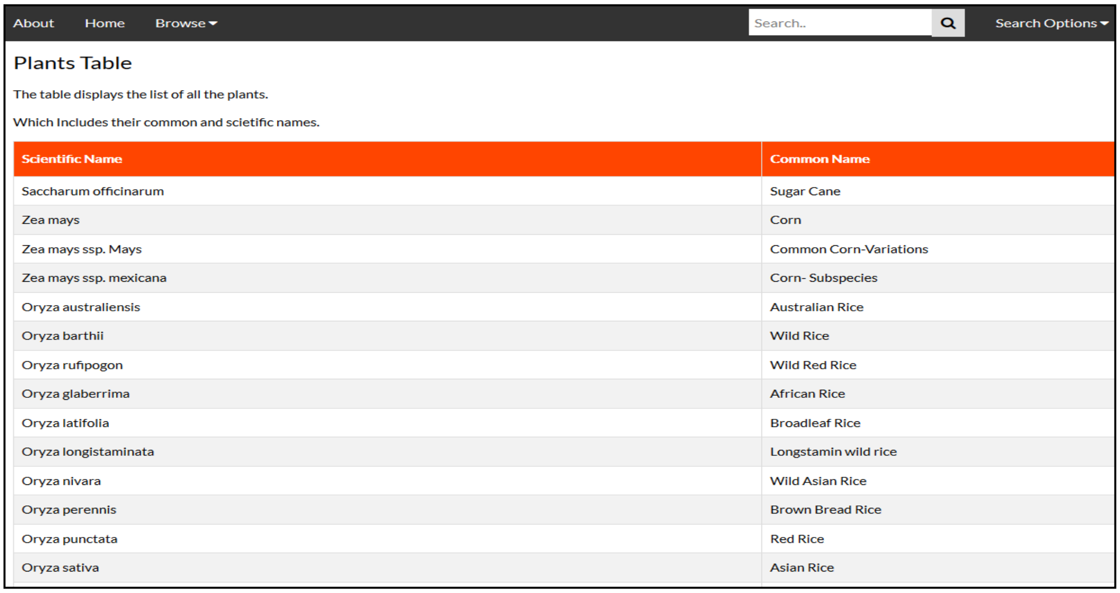
Plant Virus Online

NCBI (National Center For Biotechnology information)

Figure 26: Shows the about page of the application. The about page summarizes how the data was obtained and what recourses were used for the creation of the application.

The Browse section

Contains the combination of all five various tables that showcase all the data from the database.



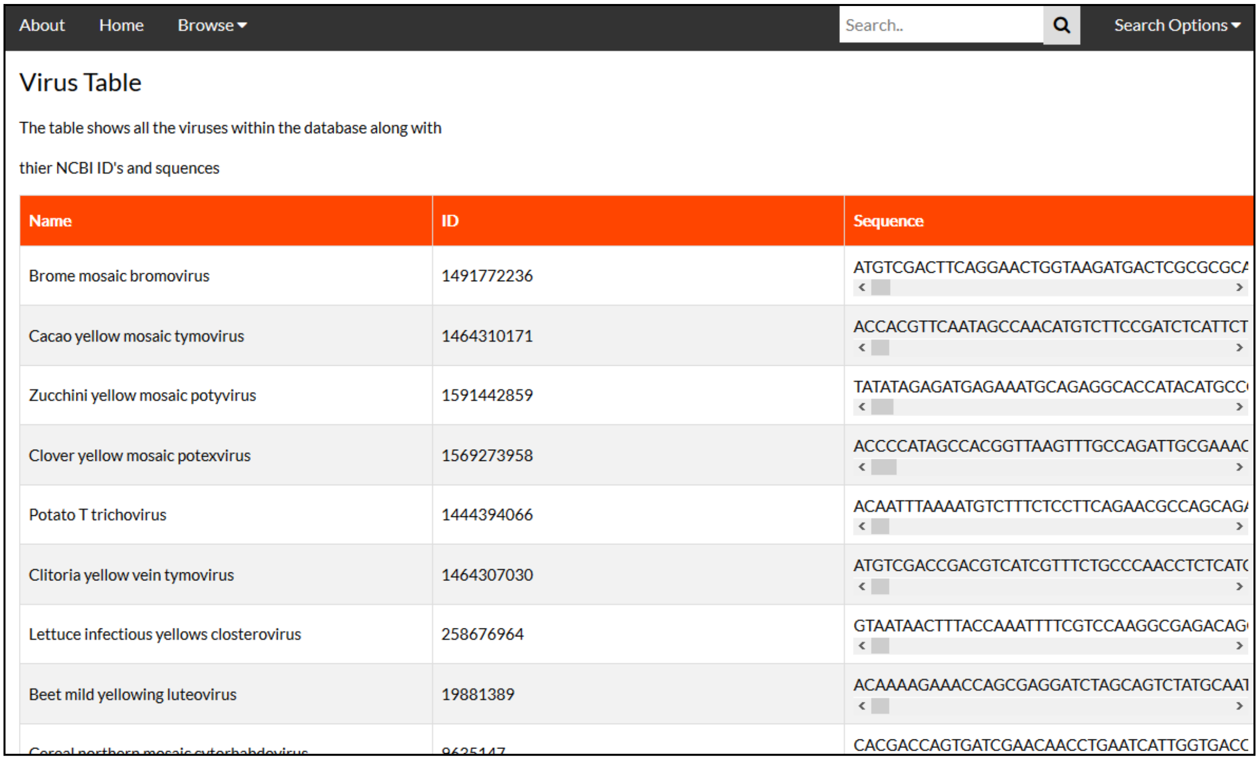
Plants Table

The table displays the list of all the plants.

Which Includes their common and scietific names.

Scientific Name	Common Name
Saccharum officinarum	Sugar Cane
Zea mays	Corn
Zea mays ssp. Mays	Common Corn-Variations
Zea mays ssp. mexicana	Corn- Subspecies
Oryza australiensis	Australian Rice
Oryza barthii	Wild Rice
Oryza rufipogon	Wild Red Rice
Oryza glaberrima	African Rice
Oryza latifolia	Broadleaf Rice
Oryza longistaminata	Longstamin wild rice
Oryza nivara	Wild Asian Rice
Oryza perennis	Brown Bread Rice
Oryza punctata	Red Rice
Oryza sativa	Asian Rice

Figure 27: Shows the first table within the browse section.



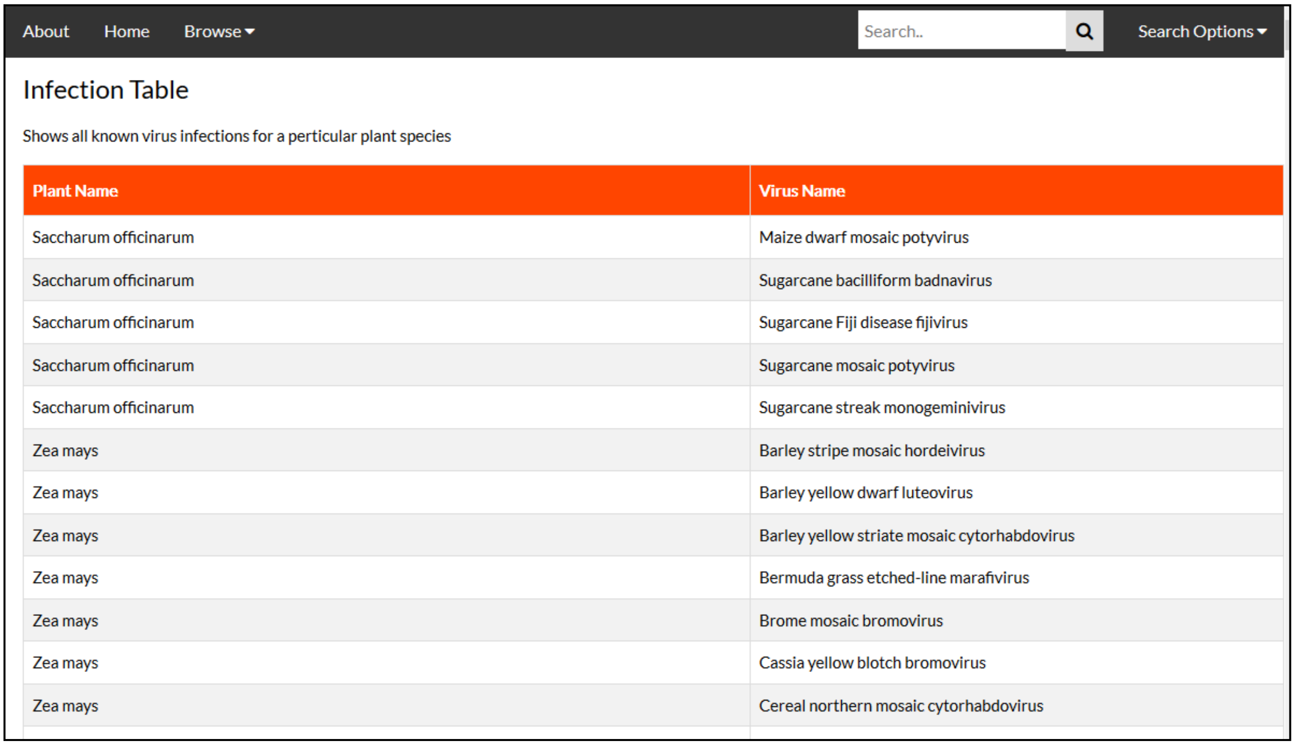
Virus Table

The table shows all the viruses within the database along with

thier NCBI ID's and squences

Name	ID	Sequence
Brome mosaic bromovirus	1491772236	ATGTCGACTTCAGGAACTGGTAAGATGACTCGCGCGCA
Cacao yellow mosaic tymovirus	1464310171	ACCACGTTCAATAGCCAACATGTCTTCCGATCTCATTCT
Zucchini yellow mosaic potyvirus	1591442859	TATATAGAGATGAGAAATGCAGAGGCACCATACATGCC
Clover yellow mosaic potexvirus	1569273958	ACCCCATAGCCACGGTTAAGTTTGCCAGATTGCGAAAC
Potato T trichovirus	1444394066	ACAATTTAAAATGTCTTTCTCCTTCAGAACGCCAGCAGA
Clitoria yellow vein tymovirus	1464307030	ATGTCGACCGACGTCATCGTTTCTGCCCAACCTCTCATC
Lettuce infectious yellows closterovirus	258676964	GTAATAACTTTACCAAATTTTCGTCCAAGGCGAGACAG
Beet mild yellowing luteovirus	19881389	ACAAAAGAAACCAGCGAGGATCTAGCAGTCTATGCAAT
Cereal northern mosaic cytorhabdovirus	9625147	CACGACCAGTGATCGAACAACCTGAATCATTGGTGACC

Figure 28: Shows the second table within the browse section. The scroll under the sequences allows the user to view the whole sequence without the need to scroll horizontally on the page itself.

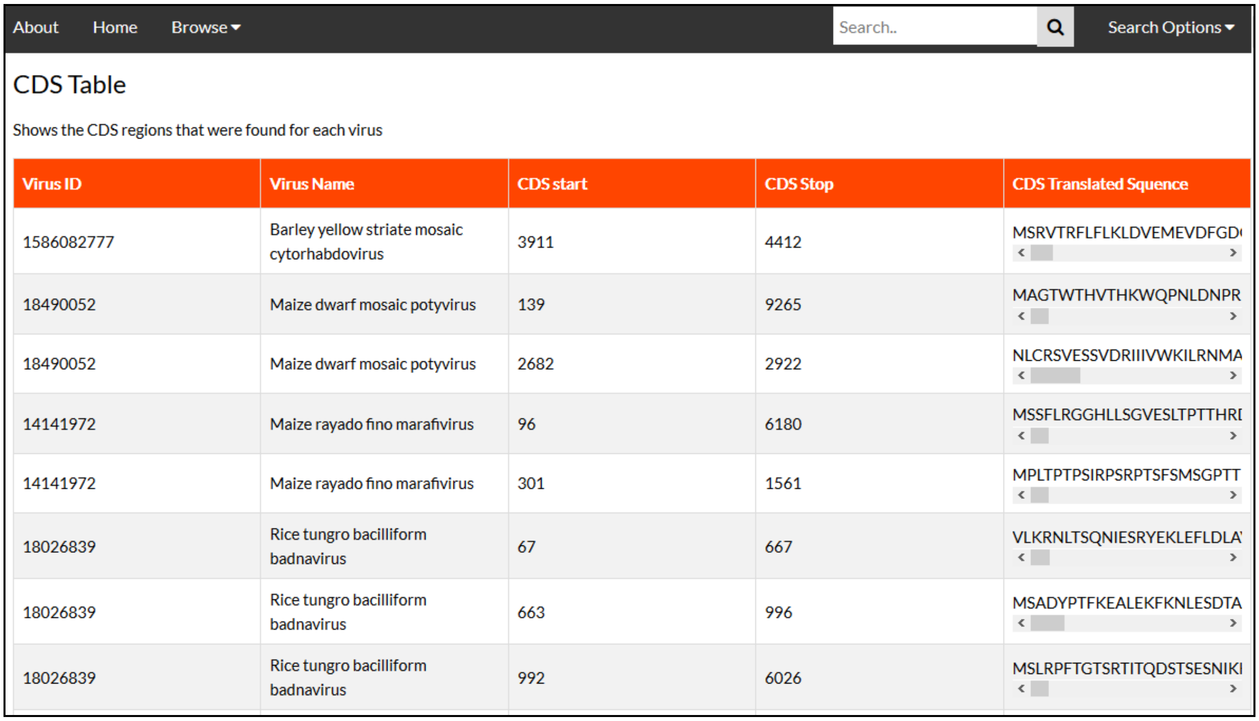


Infection Table

Shows all known virus infections for a perticular plant species

Plant Name	Virus Name
Saccharum officinarum	Maize dwarf mosaic potyvirus
Saccharum officinarum	Sugarcane bacilliform badnavirus
Saccharum officinarum	Sugarcane Fiji disease fijivirus
Saccharum officinarum	Sugarcane mosaic potyvirus
Saccharum officinarum	Sugarcane streak monogeminivirus
Zea mays	Barley stripe mosaic hordeivirus
Zea mays	Barley yellow dwarf luteovirus
Zea mays	Barley yellow striate mosaic cytorhabdovirus
Zea mays	Bermuda grass etched-line marafivirus
Zea mays	Brome mosaic bromovirus
Zea mays	Cassia yellow blotch bromovirus
Zea mays	Cereal northern mosaic cytorhabdovirus

Figure 29: Shows the third table within the browse section.

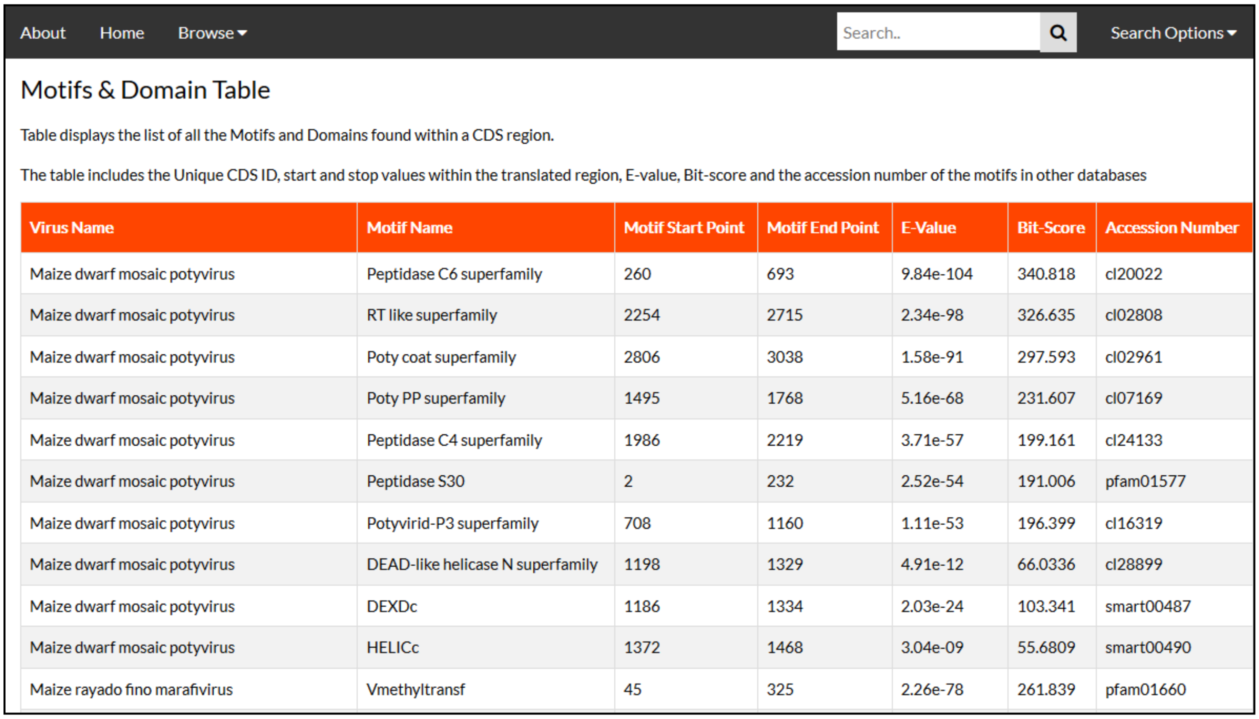


CDS Table

Shows the CDS regions that were found for each virus

Virus ID	Virus Name	CDS start	CDS Stop	CDS Translated Squence
1586082777	Barley yellow striate mosaic cytorhabdovirus	3911	4412	MSRVTRFLFLKLDVEMEVDFGD
18490052	Maize dwarf mosaic potyvirus	139	9265	MAGTWTHVTHKWQPNLDNPR
18490052	Maize dwarf mosaic potyvirus	2682	2922	NLCRSVESSVDRIIIVWKILRNMA
14141972	Maize rayado fino marafivirus	96	6180	MSSFLRGGHLLSGVESLTPTTHR
14141972	Maize rayado fino marafivirus	301	1561	MPLTPTPSIRPSRPTSFSMSGPTT
18026839	Rice tungro bacilliform badnavirus	67	667	VLKRNLTSQNIESRYEKLEFLDLA
18026839	Rice tungro bacilliform badnavirus	663	996	MSADYPTFKEALEKFKNLESDTA
18026839	Rice tungro bacilliform badnavirus	992	6026	MSLRPFTGTSRTITQDSTSESNIK

Figure 30: Shows the forth table within the browse section. The scroll under the sequences allows the user to view the whole sequence without the need to scroll horizontally on the page itself.



Motifs & Domain Table

Table displays the list of all the Motifs and Domains found within a CDS region.

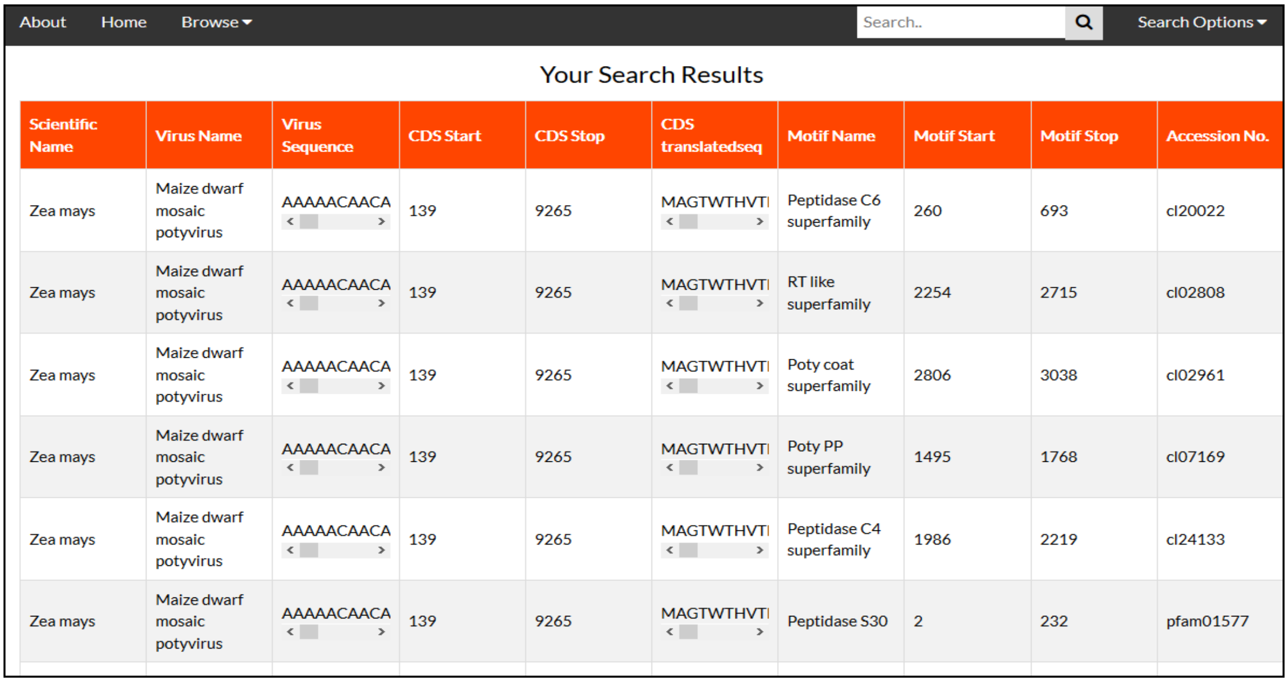
The table includes the Unique CDS ID, start and stop values within the translated region, E-value, Bit-score and the accession number of the motifs in other databases

Virus Name	Motif Name	Motif Start Point	Motif End Point	E-Value	Bit-Score	Accession Number
Maize dwarf mosaic potyvirus	Peptidase C6 superfamily	260	693	9.84e-104	340.818	cl20022
Maize dwarf mosaic potyvirus	RT like superfamily	2254	2715	2.34e-98	326.635	cl02808
Maize dwarf mosaic potyvirus	Poty coat superfamily	2806	3038	1.58e-91	297.593	cl02961
Maize dwarf mosaic potyvirus	Poty PP superfamily	1495	1768	5.16e-68	231.607	cl07169
Maize dwarf mosaic potyvirus	Peptidase C4 superfamily	1986	2219	3.71e-57	199.161	cl24133
Maize dwarf mosaic potyvirus	Peptidase S30	2	232	2.52e-54	191.006	pfam01577
Maize dwarf mosaic potyvirus	Potyvirid-P3 superfamily	708	1160	1.11e-53	196.399	cl16319
Maize dwarf mosaic potyvirus	DEAD-like helicase N superfamily	1198	1329	4.91e-12	66.0336	cl28899
Maize dwarf mosaic potyvirus	DEXDc	1186	1334	2.03e-24	103.341	smart00487
Maize dwarf mosaic potyvirus	HELICc	1372	1468	3.04e-09	55.6809	smart00490
Maize rayado fino marafivirus	Vmethyltransf	45	325	2.26e-78	261.839	pfam01660

Figure 31: Shows the fifth table within the browse section. The accession numbers for the motifs can be used in the CDD website to search for more information on the motifs that were found.

Search Section

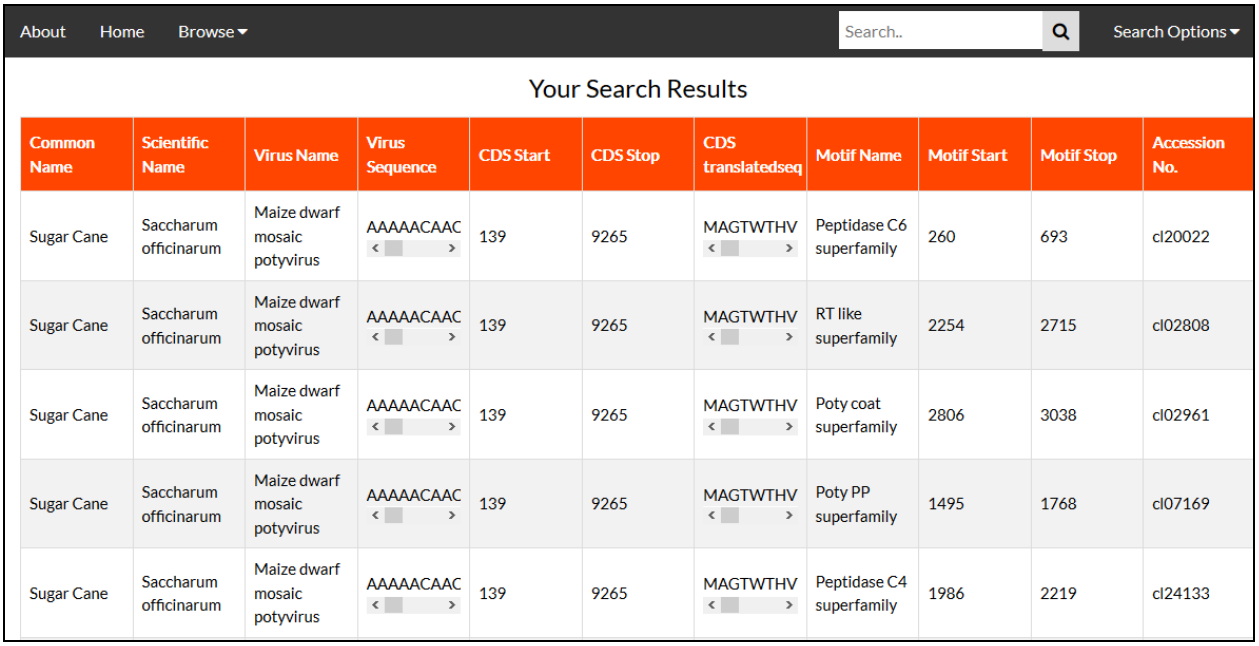
The search section allows the user to search by the scientific name by default. However, it also allows the user to search by common name, virus and motif as well. The search option for scientific and common name of the plant is for the user to see the search results for a specified plant rather than needing to browse through the browse section of the application. The Virus search allows the user to search for a specific virus and view the virus sequence, cds and motifs found for the cds that a virus might posses. The motif search results allows user to search for various viruses that may possess a specific motif; this search also provides the accession number for the motifs.



Your Search Results

Scientific Name	Virus Name	Virus Sequence	CDS Start	CDS Stop	CDS translatedseq	Motif Name	Motif Start	Motif Stop	Accession No.
Zea mays	Maize dwarf mosaic potyvirus	AAAAACAACA	139	9265	MAGTWTHVTI	Peptidase C6 superfamily	260	693	cl20022
Zea mays	Maize dwarf mosaic potyvirus	AAAAACAACA	139	9265	MAGTWTHVTI	RT like superfamily	2254	2715	cl02808
Zea mays	Maize dwarf mosaic potyvirus	AAAAACAACA	139	9265	MAGTWTHVTI	Poty coat superfamily	2806	3038	cl02961
Zea mays	Maize dwarf mosaic potyvirus	AAAAACAACA	139	9265	MAGTWTHVTI	Poty PP superfamily	1495	1768	cl07169
Zea mays	Maize dwarf mosaic potyvirus	AAAAACAACA	139	9265	MAGTWTHVTI	Peptidase C4 superfamily	1986	2219	cl24133
Zea mays	Maize dwarf mosaic potyvirus	AAAAACAACA	139	9265	MAGTWTHVTI	Peptidase S30	2	232	pfam01577

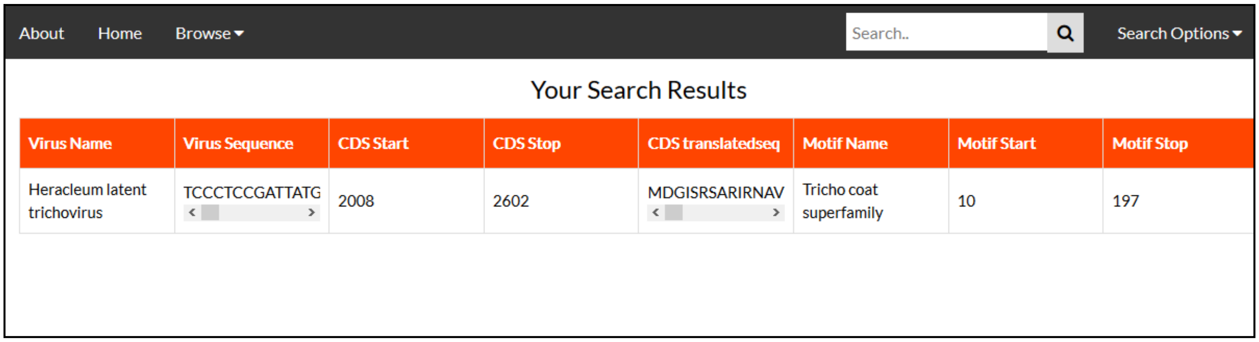
Figure 32: Shows an example of the default search results where “Zea Mays” plant was searched.



Your Search Results

Common Name	Scientific Name	Virus Name	Virus Sequence	CDS Start	CDS Stop	CDS translatedseq	Motif Name	Motif Start	Motif Stop	Accession No.
Sugar Cane	Saccharum officinarum	Maize dwarf mosaic potyvirus	AAAAACAAC	139	9265	MAGTWTHV	Peptidase C6 superfamily	260	693	cl20022
Sugar Cane	Saccharum officinarum	Maize dwarf mosaic potyvirus	AAAAACAAC	139	9265	MAGTWTHV	RT like superfamily	2254	2715	cl02808
Sugar Cane	Saccharum officinarum	Maize dwarf mosaic potyvirus	AAAAACAAC	139	9265	MAGTWTHV	Poty coat superfamily	2806	3038	cl02961
Sugar Cane	Saccharum officinarum	Maize dwarf mosaic potyvirus	AAAAACAAC	139	9265	MAGTWTHV	Poty PP superfamily	1495	1768	cl07169
Sugar Cane	Saccharum officinarum	Maize dwarf mosaic potyvirus	AAAAACAAC	139	9265	MAGTWTHV	Peptidase C4 superfamily	1986	2219	cl24133

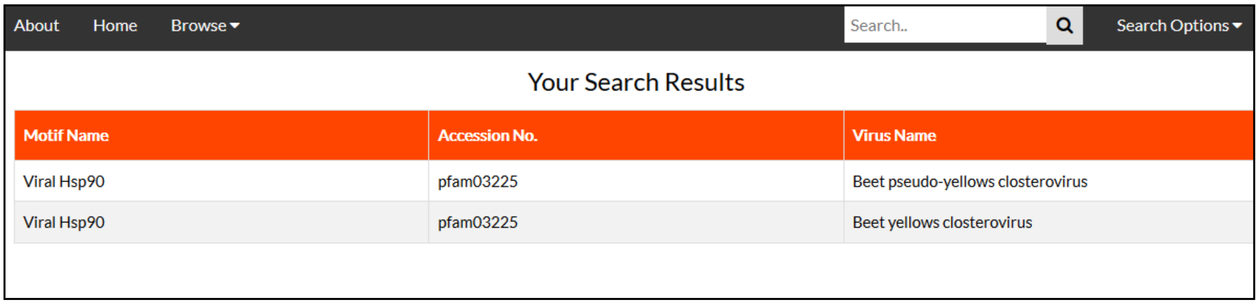
Figure 33: Shows an example where "Sugar Cane" was searched by setting the search settings to "common name" by using the dropdown menu next to the search bar.



Your Search Results

Virus Name	Virus Sequence	CDS Start	CDS Stop	CDS translatedseq	Motif Name	Motif Start	Motif Stop
Heracleum latent trichovirus	TCCCTCCGATTATG	2008	2602	MDGISRSARIRNAV	Tricho coat superfamily	10	197

Figure 34: Shows an example where "Heracleum latent trichovirus" was searched by setting the search settings to "viruses" by using the dropdown menu next to the search bar.



Your Search Results

Motif Name	Accession No.	Virus Name
Viral Hsp90	pfam03225	Beet pseudo-yellows closterovirus
Viral Hsp90	pfam03225	Beet yellows closterovirus

Figure 35: Shows an example where "Viral Hsp90" Motif was searched by setting the search settings to "Motif" by using the dropdown menu next to the search bar.

Discussion & Conclusion

The application has successfully established connectivity and retrieved data from the database as intended, fulfilling its overall purpose. Nevertheless, there exist notable opportunities for improvement in subsequent updates. Future iterations of the application will prioritize the creation of a refined Graphical User Interface (GUI) to enhance the user experience when viewing sequences and results. This refinement aims to eliminate redundancy in the displayed results, a concern currently arising from the application of join statements on various tables without subsequent edits.

One prominent aspect slated for enhancement is the user interface for the search functionality. The current design necessitates users to manipulate a drop-down menu to alter the search parameter. In the updated version, efforts will be directed towards making the search bar more user-friendly, allowing users to seamlessly modify search parameters directly. Furthermore, the case-sensitive nature of searches will be addressed to accommodate variations in user input, ensuring that search terms are not confined to exact case matching.

Expansion of the plant genus data in the database is also on the agenda for future updates. The current limitation of utilizing only ten genuses stems from the storage constraints imposed by Elephant SQL, which provides up to 128 MB of free usage. To overcome this limitation, the addition of more plant genuses is imperative, although this may necessitate purchasing additional database space.

Notably, the application's current dependence on data from the Plant and Viruses Online website poses a potential limitation due to the infrequent updates on the website. To mitigate this dependency, a proposed improvement involves the development of an algorithm within the application to generate viral infection data based on plant species. This algorithmic approach aims to reduce reliance on external data sources, ensuring greater autonomy and data accuracy.

Despite these identified areas for enhancement, the existing database effectively records viral proteins alongside the host types typically infected by the viruses. This functionality holds considerable promise in advancing our understanding of viral capabilities, thereby providing a strategic advantage in countering viruses and safeguarding agricultural crops.

References

1. (Lecoq) Lecoq, H. “Discovery of the First Virus, the Tobacco Mosaic Virus: 1892 or 1898?” *Comptes Rendus De L'Academie Des Sciences. Serie III, Sciences De La Vie*, U.S. National Library of Medicine, Oct. 2001, www.ncbi.nlm.nih.gov/pubmed/11570281.
2. (Smith) Smith ,H. “Mechanisms of Viral Pathogenicity” *Department of microbiology, University of Birmingham, Birmingham, England,* Vol.36,NO.3 Bacteriological Reviews. American Society of Microbiology, Sept. 1972, https://mmbr.asm.org/content/mmbr/36/3/291.full.pdf
3. (Kamps) Kamps, Bernd Sebastian, et al. *Influenza Book | Pathogenesis and Immunology*, www.influenzareport.com/ir/pathogen.htm.
4. (Lodish) Lodish H, Berk A, Zipursky SL, et al. Molecular Cell Biology. 4th edition. New York: W. H. Freeman; 2000. Section 6.3, Viruses: Structure, Function, and Uses. Available from:
5. (Rui) Rui, Liu Sheng, et al. “MicroRNA-Mediated Gene Silencing in Plant Defense and Viral Counter-Defense.” *Frontiers in Microbiology* , vol. 08, no. 1664-302x, 2017, p. 1801., doi:10.3389/fmicb.2017.01801.
6. (Pratt) Pratt, Ashley J and Ian J MacRae. “The RNA-induced silencing complex: a versatile gene-silencing machine” *Journal of biological chemistry* vol. 284,27 (2009): 17897-901.
7. (Kamthan) Kamthan, Ayushi et al. “Small RNAs in plants: recent development and application for crop improvement” *Frontiers in plant science* vol. 6 208. 2 Apr. 2015, doi:10.3389/fpls.2015.00208