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

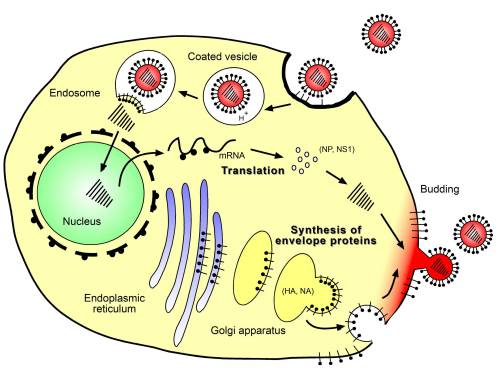
Abstract

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).

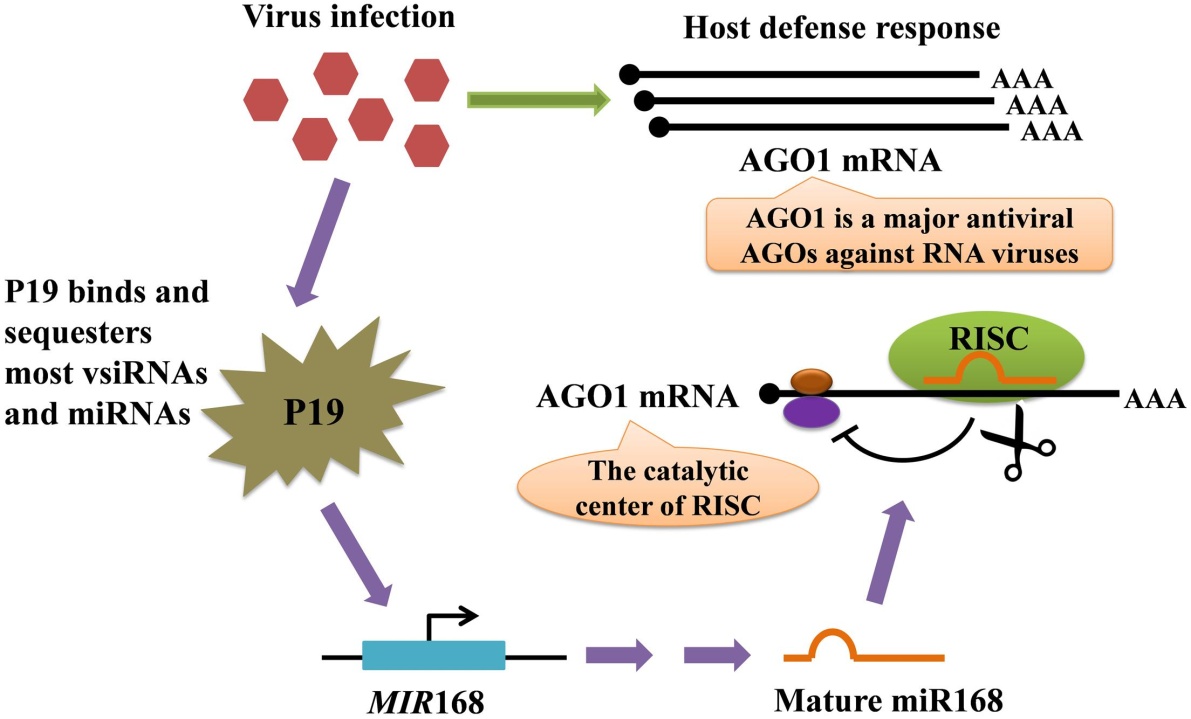


Figure 2: represents a model for the regulation of AGO mRNA level mediated by protein P19-induced miR168. The AGO mRNAs are made when a virus infection is detected within the cell. While the plant cell is creating mRNAs for AGO the virus produces P19 VSR which binds to the virus- encoded siRNAs and with 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.

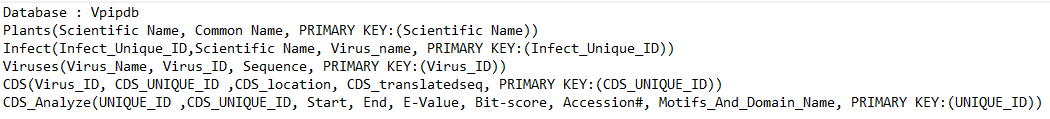


Figure 3: displays the schema of the proposed database.

1. 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.
2. 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.

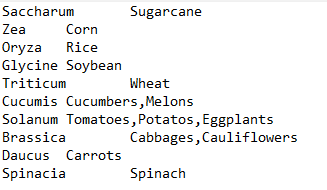


Figure 5: Displays the data that was obtained from the script.

1. 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.

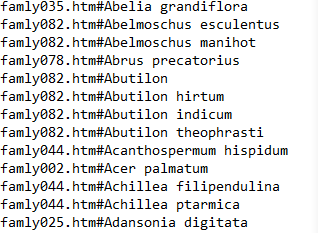


Figure 7: Shows the output from the script.

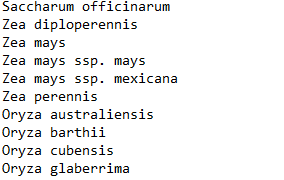


Figure 8: The data was then organized with the use of Excel.

1. 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.

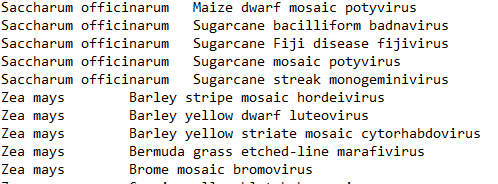


Figure 10: shows how the data looks like once it’s been organized in the tab delimited format. The first column shows the species of plant. The second column shows the virus that infects the plant.

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



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

1. 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.
2. The resulting data was then arranged using excel



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.

1. 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.

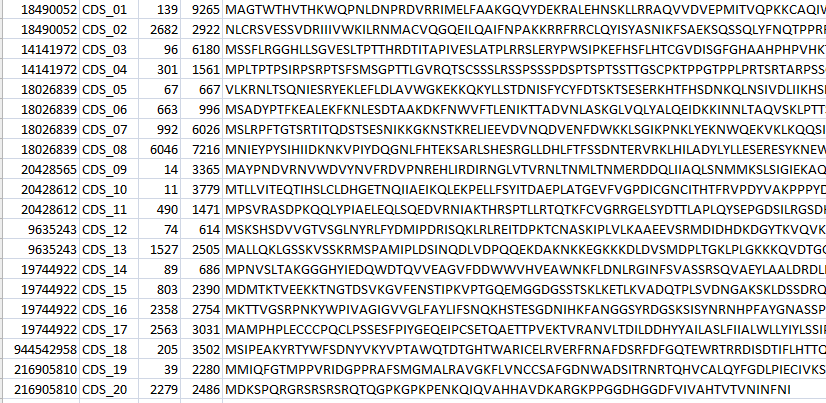


Figure 16: Shows how the data looks in excel. The first column shows the Virus\_ID. The second column shows the Unique\_ID given to every CDS region found. The third column shows the length/ region of CDS within the viral genome. Finally, the last column is the translated protein sequence.

1. 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:



Figure 17: Shows the CDD batch search.

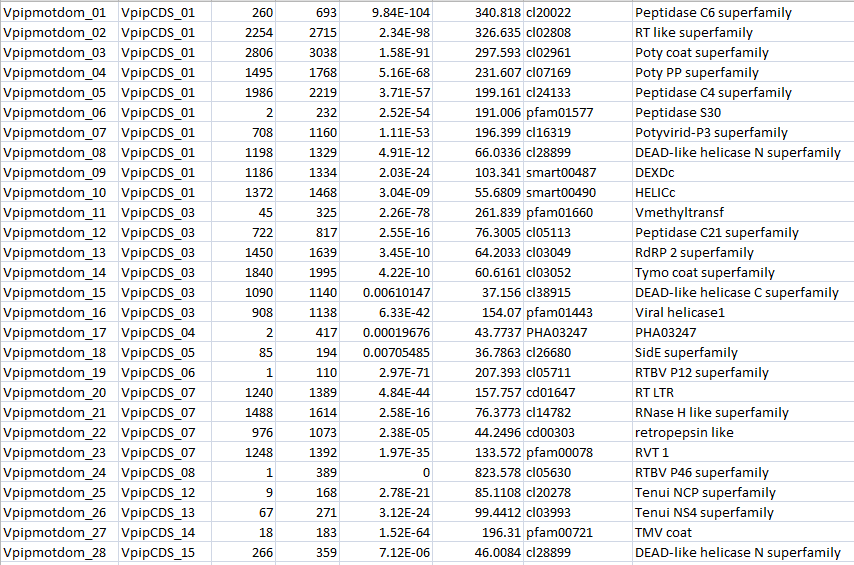


Figure 18: shows the final results that were obtained from the NCBI: CDD database. These results are edited and arranged based on the schema. Look at image 1 for reference.

1. 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..
2. 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.

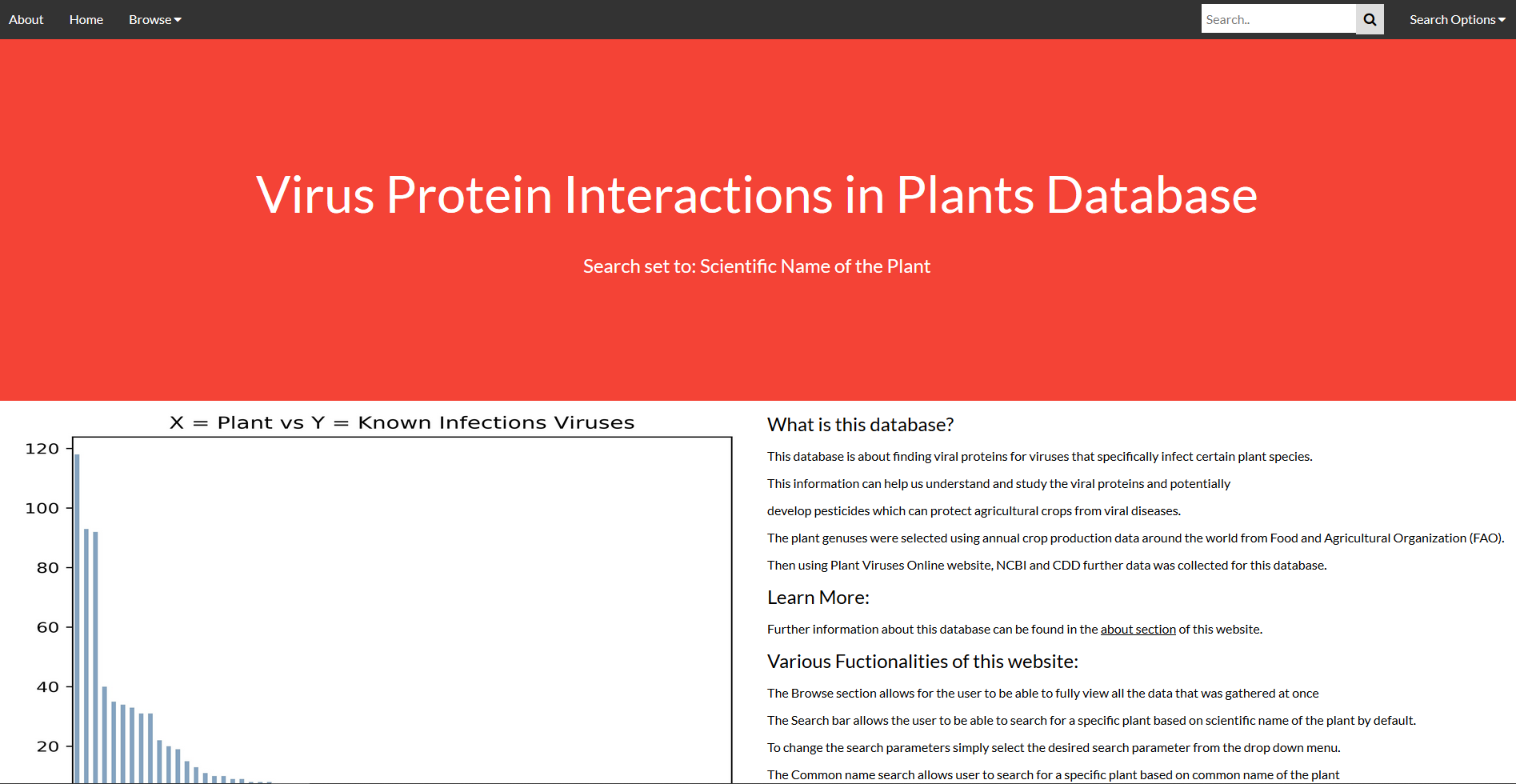


Figure 25: Shows the home page of the application. The graph showcases all the recorded infections for a particular plant species. The home page of the application goes in more detail about the database and about the graph itself.

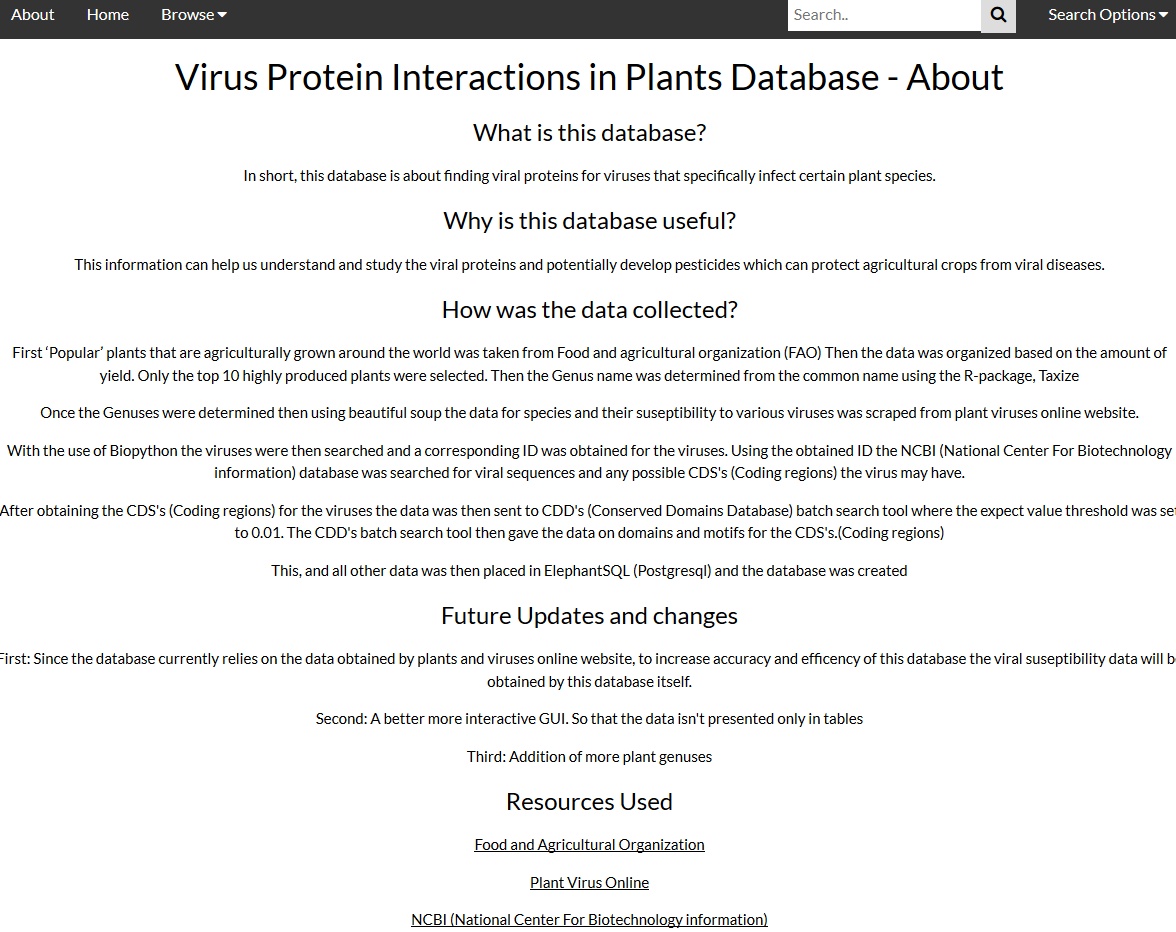


Figure 26: Shows the about page of the application. The about page summarizes how the data was obtained and what recourses were used for the creation of the application.

The Browse section

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

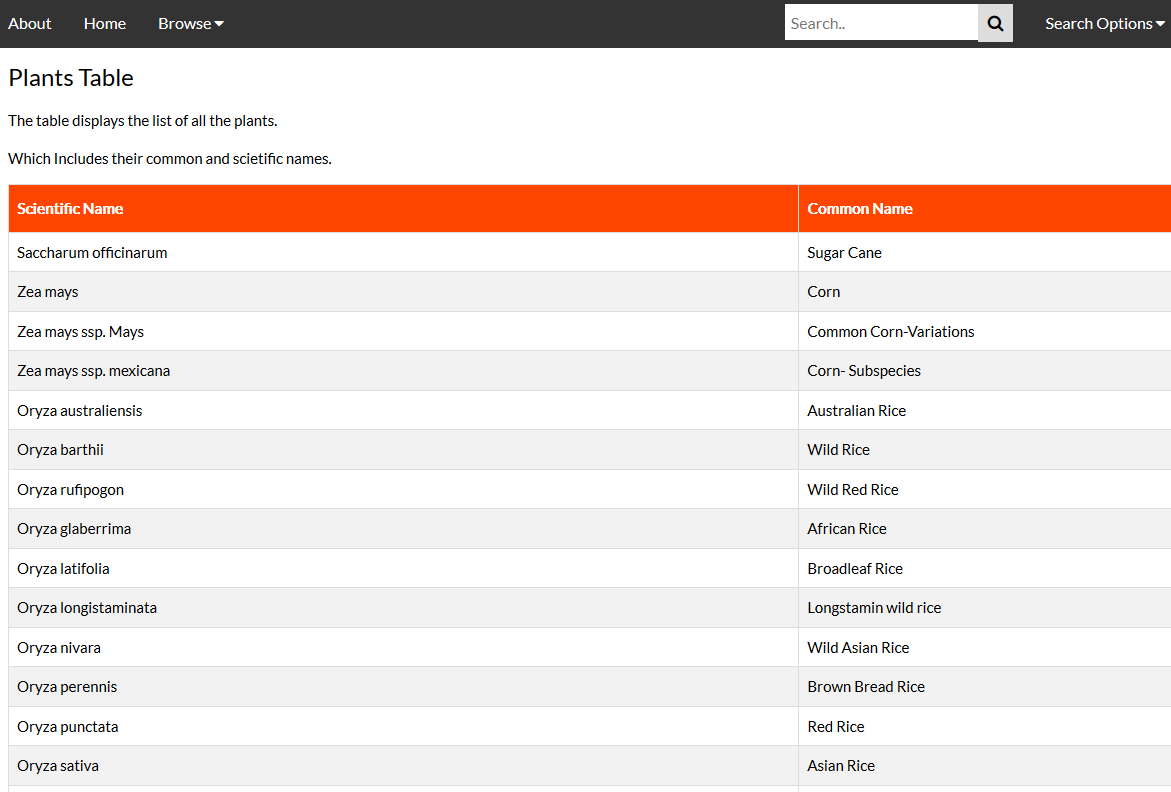


Figure 27: Shows the first table within the browse section.

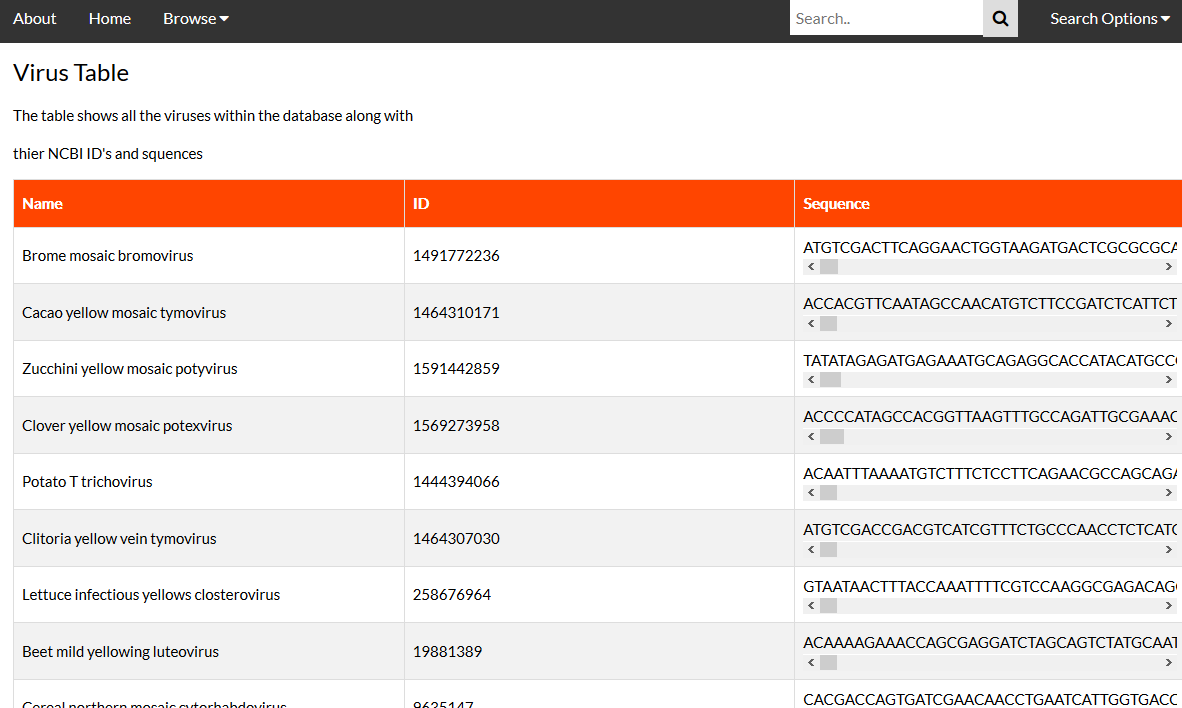


Figure 28: Shows the second table within the browse section. The scroll under the sequences allows the user to view the whole sequence without the need to scroll horizontally on the page itself.

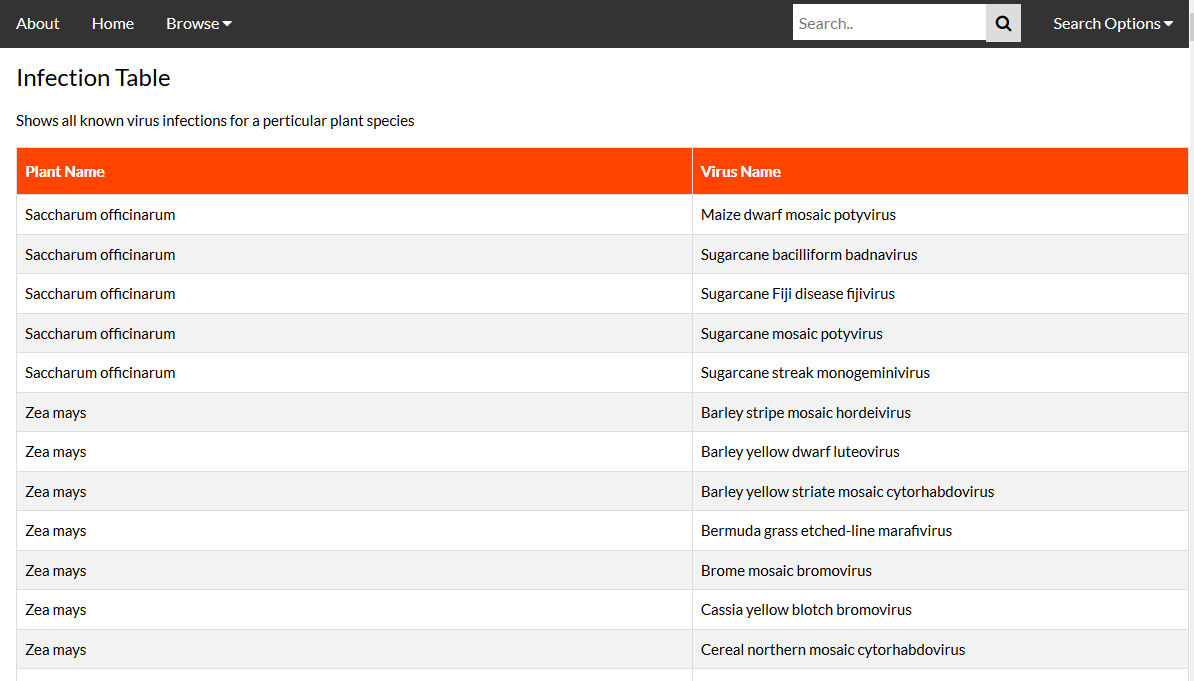


Figure 29: Shows the third table within the browse section.

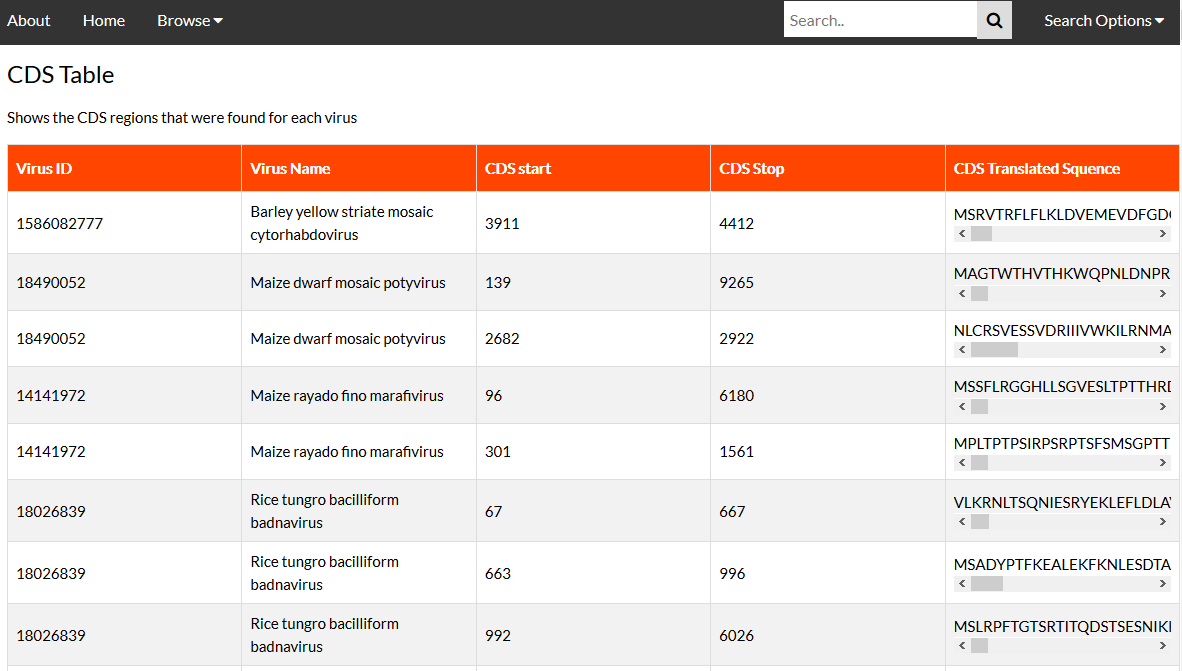


Figure 30: Shows the forth table within the browse section. The scroll under the sequences allows the user to view the whole sequence without the need to scroll horizontally on the page itself.

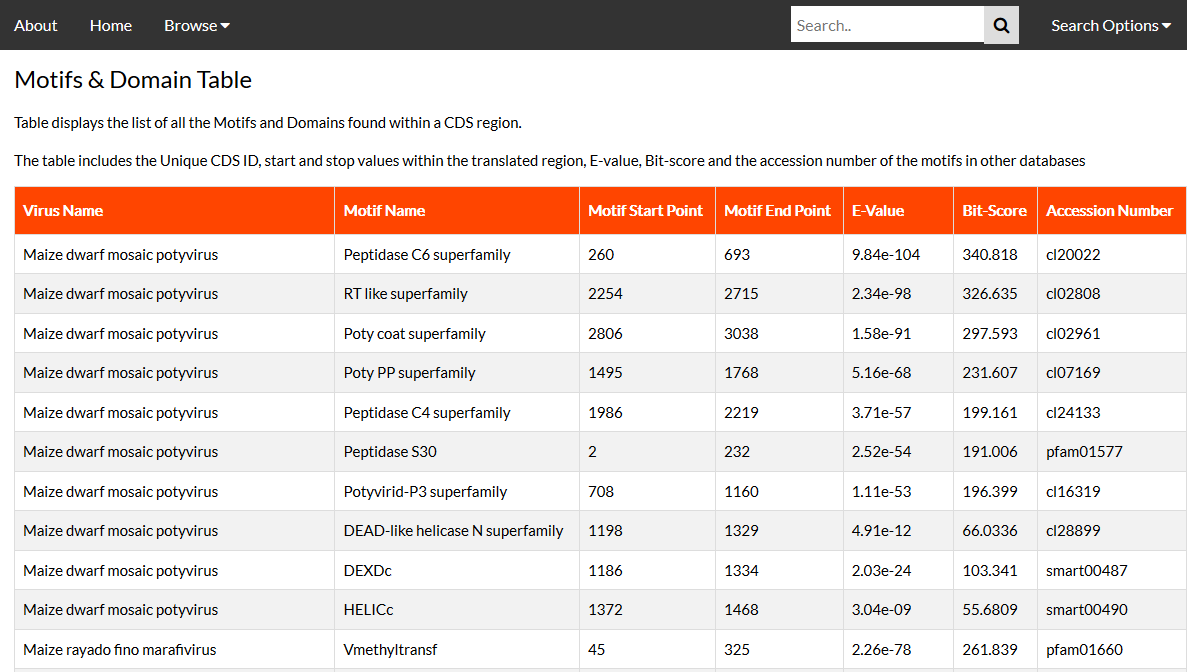


Figure 31: Shows the fifth table within the browse section. The accession numbers for the motifs can be used in the CDD website to search for more information on the motifs that were found.

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.

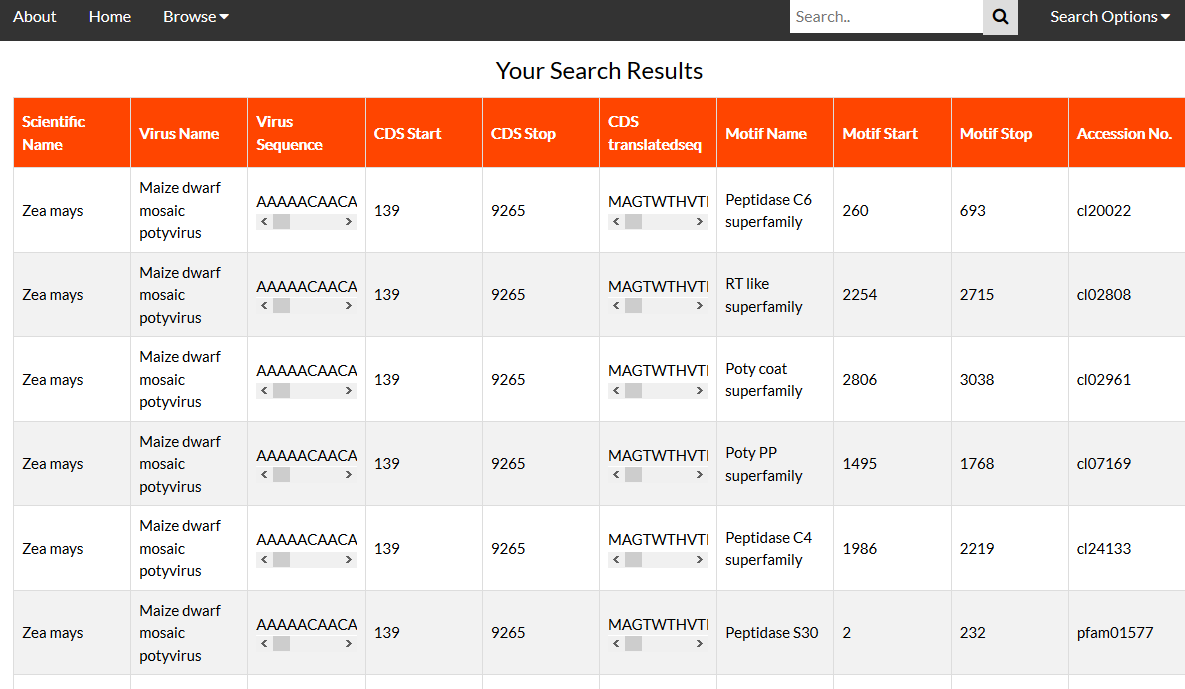


Figure 32: Shows an example of the default search results where “Zea Mays” plant was searched.

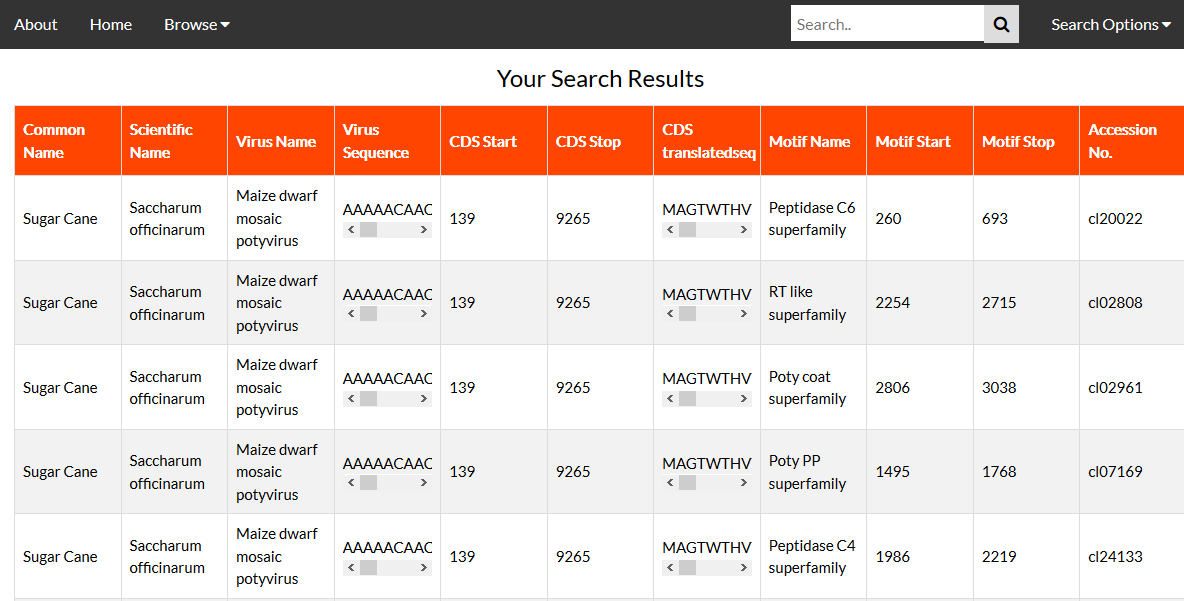


Figure 33: Shows an example where “Sugar Cane” was searched by setting the search settings to “common name” by using the dropdown menu next to the search bar.

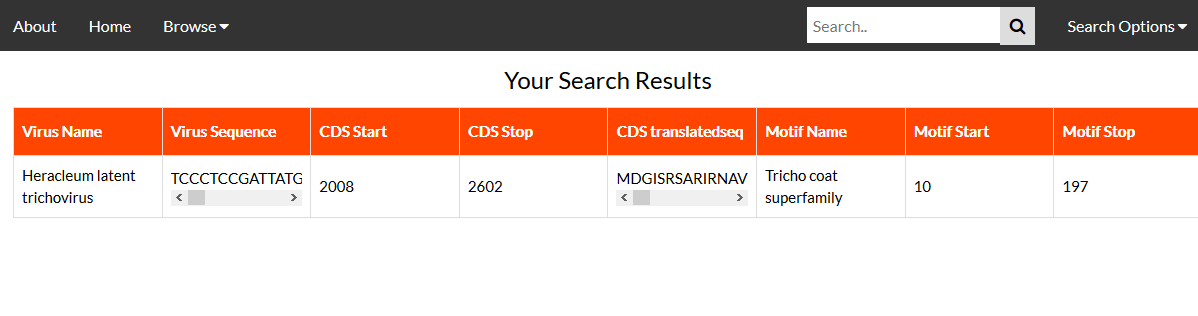


Figure 34: Shows an example where “Heracleum latent trichovirus” was searched by setting the search settings to “viruses” by using the dropdown menu next to the search bar.

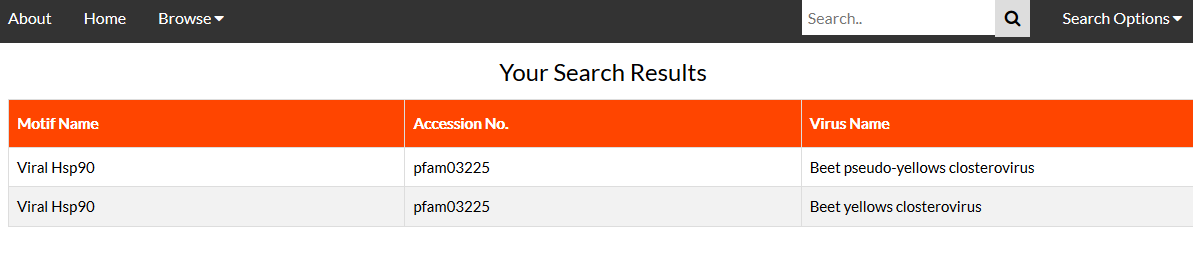


Figure 35: Shows an example where “Viral Hsp90” Motif was searched by setting the search settings to “Motif” by using the dropdown menu next to the search bar.

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