EXPLORING FRAMEWORKS FOR RAPID VISUALIZATION OF VIRAL PROTEINS COMMON FOR A GIVEN HOST

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Rajesh Subramaniam

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

April 2019

Fargo, North Dakota

North Dakota State University
Graduate School

Title

EXPLORING FRAMEWORKS FOR RAPID VISUALIZATION OF VIRAL PROTEINS COMMON FOR A GIVEN HOST

By

Rajesh Subramaniam

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Anne Denton
Chair

Dr. Changhui Yan

Dr. Sangita Sinha

Approved:

May 14, 2019	Dr. Kendall Nygard
Date	Department Chair

ABSTRACT

Viruses are unique organisms that lack the protein machinery necessary for its propagation (like polymerase) yet possess other proteins that facilitate its propagation (like host cell anchoring proteins). This study explores seven different frameworks to assist rapid visualization of proteins that are common to viruses residing in a given host. The proposed frameworks rely only on protein sequence information. It was found that the sequence similarity-based framework with an associated profile hidden Markov model was a better tool to assist visualization of proteins common to a given host than other proposed frameworks based only on amino acid composition or other amino acid properties. The lack of knowledge of profile hidden Markov models for many protein structures limit the utility of the proposed protein sequence similarity-based framework. The study concludes with an attempt to extrapolate the utility of the proposed framework to predict viruses that may pose potential human health risks.

ACKNOWLEDGEMENTS

I deeply appreciate my adviser Dr. Anne Denton's guidance, patience with my research and understanding of my situation throughout my graduate life at the Department of Computer Science. I am ever so thankful to the faculty of the Department of Computer Science, especially Dr. Saeed Salem, Dr. Oksana Myronovych and Dr. Changhui Yan for their assistance in helping me transition to the new field of Computer Science at a later stage in my life (e.g. letting me attend their lectures on courses which I did not take for credits). I am so grateful to Dr. Kendall Nygard and the Department of Computer Science for providing me the opportunity and resources to pursue my master's in computer science at North Dakota State University. I would also like to thank Allegiance Fundraising Group, LLC for hiring me as a software developer while being a graduate student. I appreciate my daughter AnnaPoorna for being so understanding and kind to sacrifice her days of playing with her dad to let me pursue my goals. Finally, I cannot thank my wife Manjusha enough who has been central to everything that I may have accomplished in my life.

DEDICATION

Manjusha Saraswathiamma (My wife)

AnnaPoorna Rajesh (My daughter)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Pattern identification is central to several fields of study like drug discovery, forecasting, cybersecurity, network analyses, etc. Rising drug development costs and higher benefits to risk ratio expectations for newer drugs have led more pharmaceutical companies to adopt artificial intelligence to accelerate the drug discovery process (Fleming, 2018). The urgency for a rapid drug discovery process is also accentuated by uncertainties caused by global warming (Kurane, 2010) and evolution of drug-resistant microbes (Blair, 2018). While availability of newer modes of treatment like biosimilars, biologics, stem cells, gene therapy, etc. along with classical methods like vaccination and chemical synthesis (Mignani, Huber, Tomás, Rodrigues, & Majoral, 2016) expands drug development options, they also make accelerated drug development imperative besides other factors like patent expirations. The expectations for quicker drug discovery also stem from the fact that clinical trial data spanning decades are available (e.g. Project Data Sphere®) that can be mined to obtain crucial insights for drug development.

Data mining is the interdisciplinary study of extracting knowledge (correlations, patterns, associations, classes and/or clusters) from large seemingly unrelated datasets often utilizing prior subject knowledge. The field lies at the intersection of computer science, statistics and database design and applies the principles to other areas of study like bioinformatics, climatology, finance, social networks, etc. It encompasses several forms of study like frequent pattern identification, association rule mining, clustering, and classification. Data mining studies can be carried out to obtain:

- algorithms to efficiently prune large datasets by applying statistical principles and other subject matter knowledge (e.g. knowledge of amino acid sequence responsible for cellular localization can be used to classify protein datasets)

- graphs that summarize relevant information distilled out of large datasets (e.g. mining protein 3D structure motifs)
- identification of frequent item sets, patterns and association rules

In this dissertation, several frameworks are explored to graphically summarize the number of viral proteins that are common for a given host as a first step toward pattern identification of viral protein expressions for a given host.

1.1. Problem Statement

Drug development is a slow and costly enterprise. Rendering a protein that is crucial for a targeted pathology dysfunctional with a drug is one mode of treating the pathological condition. Viral infections are pathological conditions in which a virus recruits host organism’s protein machinery for its own propagation while utilizing their own proteins for entering other cells in the host. Identifying a viral protein that supports any aspect of a virus life cycle can provide a potential drug target to treat the viral infection. The said protein can be rendered dysfunctional with a drug disrupting the virus life cycle and be a treatment option for the viral infection. Identifying the targeted protein in several viruses can potentially increase the scope of viral infections that could be targeted with the same drug. In other words, identifying a viral protein or a set of proteins that are common amongst the viruses infecting a host can potentially accelerate drug development by expanding the scope of a drug developed for treating one viral infection to treat several viral infections.

A simple framework that can accurately classify related proteins (either based on sequence, structure, function or all) can be highly beneficial to identify common protein expression sets across different viruses for a given host. The challenge for developing such a framework is that these protein sets potentially have very different specific amino acid sequences (primary protein structure) as they are coded for by different gene sequences of the various viruses and the transcription/translation fate the transcribed messenger ribonucleic acids (mRNAs) could potentially undergo in the respective viruses.

This study proposes seven different simple frameworks to facilitate the identification and visualization of protein sets common for different viruses infecting a host. The proposed frameworks were calculated using only the protein amino acid sequence and can be classified either as sequence similarity-based, amino acid-composition-based, amino acid behavior-based (e.g., the retention time of a polypeptide containing the amino acid under consideration) or specific sequence-based frameworks. Protein structure and function were not considered for developing a framework in this initial study. This thesis explores the utility of the proposed frameworks in identifying protein expression sets common for various viruses infecting a host. This study also proposes a potential utility of the frameworks by predicting viruses residing in other hosts that could pose a potential health risk for humans.

Figure 1 summarizes the study method and objectives of the study. Briefly, the virus dataset was retrieved from National Institutes of Health (NIH) repository as a text file. The dataset was parsed to identify each virus, its host and the protein sequences expressed by the virus using Python. The retrieved information (virus, host and viral protein expression sequences reported) were saved in a MySQL database. Each protein sequence in the database was then parsed with various routines to generate the proposed protein identifiers (frameworks) that was used to identify protein expression sets common for various viruses residing in a given host. The identified common protein expression sets were charted using Matplotlib, a Python 2D plotting library. The methodologies used to generate the identifiers is summarized in Table 1.

Table 1. Methodology Used to Generate the Proposed Protein Identifiers.

Proposed Identifier	**Methodology to Generate the Protein Identifier (all codes were written in Python)**
Pfam-based (Pfam_Keys)	1. Query sequence in Pfam database to identify Pfam domains. 2. Concatenate the identified Pfam domains in the order they appear in the sequence separated by an underscore ("_").
Amino acid-based (AA_Type_Keys)	1. Count the various amino acid types (hydrophobic, neutral, hydrophilic and other) that appear in the protein sequence. 2. Concatenate the counts returned for each of the amino acid type returned in the order hydrophobic, neutral, hydrophilic and other separated by an underscore ("_"). An alternative terminology for hydrophobic, neutral, hydrophilic and other is used in this study (see below)
Hydrophobicity index (HI)-based (HI_Num)	1. Count the various individual amino acids that appear in the protein sequence. 2. Add the products of the HI reported for each amino acid with its respective count to obtain the cumulative HI_Num for the protein.
Hydrophobicity index (HI)-based (HI_Key)	The cumulative HI for the protein obtained above is concatenated with the protein sequence length separated by an underscore ("_").
Combo_Key	A string concatenation of HI_Key and AA_Type_Key for the protein sequence
MD5 Hash key -based (MD5_Key)	A hash is generated for the protein sequence using the MD5 hash algorithm.
SHA-512 Hash key -based (SHA512_Key)	A hash is generated for the protein sequence using the SHA512 hash algorithm.

The Venn diagram obtained in Step 3 (Figure 1) represents the protein expression sets that are common for any two viruses *i* and *j*. Such common protein expression sets, in principle, could be envisioned for multiple viruses. A drug (say Drug D in Steps 3 and 4, Figure 1) that was developed to treat a viral infection caused by a virus (say Virus *i*) by rendering a protein member of the common protein expression set amongst multiple viruses dysfunctional (indicated by the yellow circle in Step 3 of Figure 1), in principle, could be used to treat several infections caused by other viruses provided the other viruses expresses protein expression sets that are supersets of the common protein expression set represented by the Venn diagram in Figure 1. This study thus

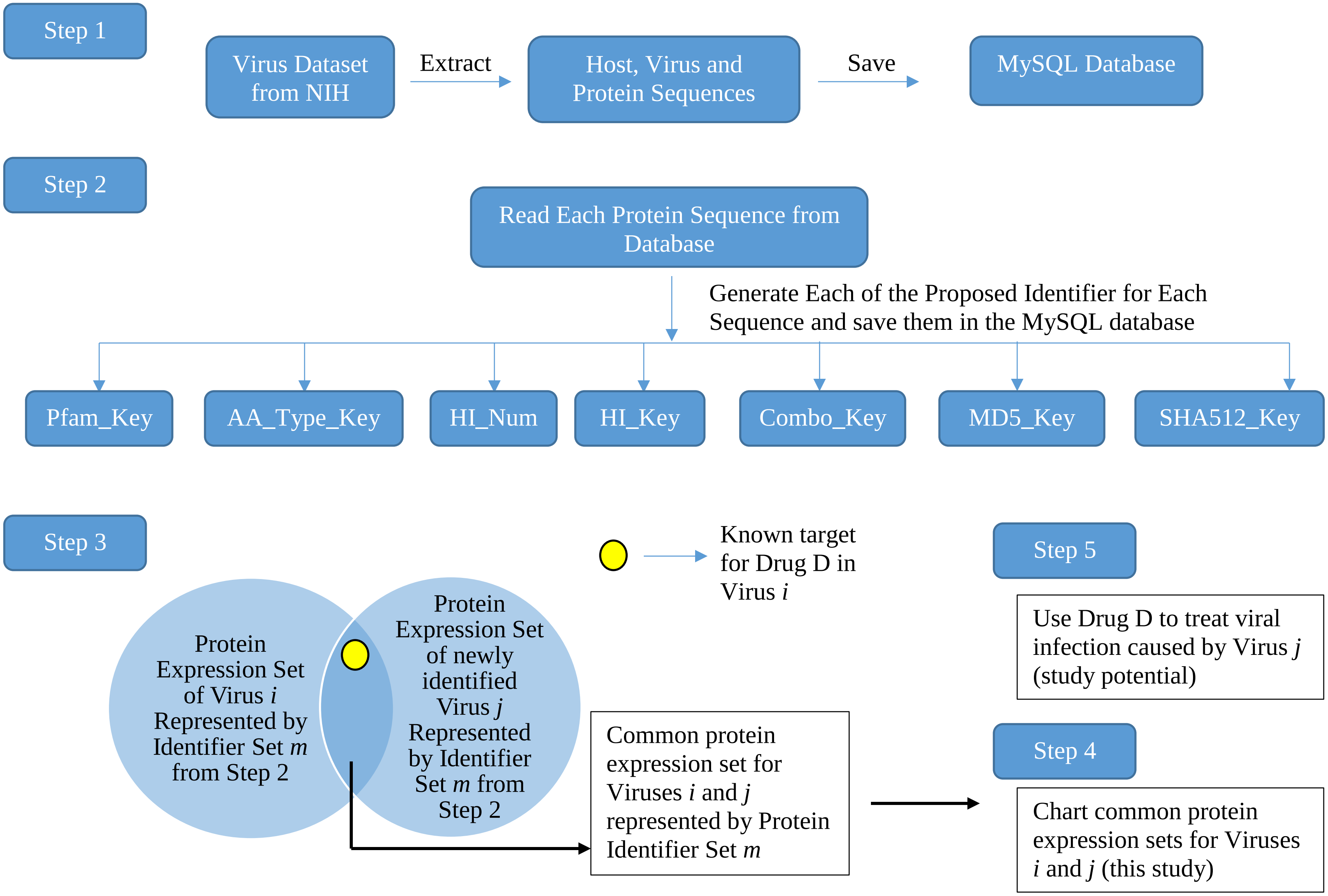

Step 1
Virus Dataset from NIH
Extract
Host, Virus and Protein Sequences
Save
MySQL Database
Step 2
Read Each Protein Sequence from Database
Generate Each of the Proposed Identifier for Each Sequence and save them in the MySQL database
Pfam_Key
AA_Type_Key
HI_Num
HI_Key
Combo_Key
MD5_Key
SHA512_Key
Step 3
Known target for Drug D in Virus *i*
Protein Expression Set of Virus *i* Represented by Identifier Set *m* from Step 2
Protein Expression Set of newly identified Virus *j* Represented by Identifier Set *m* from Step 2
Common protein expression set for Viruses *i* and *j* represented by Protein Identifier Set *m*
Step 5
Use Drug D to treat viral infection caused by Virus *j* (study potential)
Step 4
Chart common protein expression sets for Viruses *i* and *j* (this study)

Figure 1. Schematic Summary of the Study.

proposes a potential route to accelerate anti-viral drug development by increasing the scope of viral infections that could be treated with known drugs, which can potentially save the development costs for newer drugs that may require a longer time and higher costs from the bench to the market.

1.2. Viruses

Viruses are unique microscopic organisms that are considered by many to be the link between living and non-living worlds (Moreira & López-García, 2009). Unlike a living thing which possesses the abilities to propagate, derive energy through metabolism, and evolve during procreation, viruses rely on its host for their procreation and evolve depending on selective pressures exerted by the host for their optimal survival. On the other hand, they can remain dormant for ages outside a host without any need for metabolism. Viruses rely on the hosts they infest to propagate because they lack the protein machinery essential to make multiple copies of their genomes. In the process, viruses express its own proteins (e.g. express protein-based cell anchors to bind on to the next host cell) to make multiple copies of its genome.

1.3. Justifications for the Study

As noted earlier, there is a growing expectation for accelerated drug discovery process fueled by a variety of factors. These include, but are not limited to

i. economic threats like looming patent expirations and global economy (patent issued by one country can be enforced globally, thus initiating a global race to be the first in the market to reap economic benefits)

ii. global threats like climate change and drug resistance

iii. regulatory expectations of higher benefits to risk ratio

iv. scientific advances made in discovery of newer treatment modes (biosimilars, gene therapy, etc.) and

v. technological advances (artificial intelligence and data mining in drug discovery process)

This dissertation attempts to explore frameworks that can potentially be employed for quicker identification of viral protein expression patterns for a given host. As a first step, protein expression patterns identified using the Pfam database for viral protein expressions for a given host is compared with those identified using other frameworks proposed in the thesis. The Pfam database-based expression patterns are considered the standard against which other frameworks are compared because the Pfam database is manually curated and widely accepted amongst the scientific community (Sonnhammer, Eddy, & Durbin, 1997). The frameworks proposed here can be considered as part of initial studies toward various machine learning efforts that can be pursued to identify viral protein expression patterns for a given host.

A second justification for the study is the following. As was noted earlier, living beings evolve from one generation to another. Studies have shown that viruses evolve to optimize their survival within a host (Ali, Amroun, de Lamballerie, & Nougairède, 2018). Thus, Chikungunya virus has very minimal mutation when cultured in mosquito cells but exhibit higher mutation rates when propagated in vertebrate cells (Ali et al., 2018). This behavior has been attributed to the virus responding to selective pressures for optimal survival in the given host. Identification of protein expression patterns amongst different hosts can thus be helpful to identify infection potential and treatment opportunities for a yet to be identified virus. For example, consider a virus V1 expressing protein sets A and B in two different hosts because of selective pressures the virus is subjected to in the two hosts. Let us say protein set C is the intersection of protein sets A and B. In principle, another virus V2 with a different protein expression set X that is equal to protein set C or is a superset of protein set C can potentially be hosted by the same hosts.

A third justification is discovery of newer viral species from arctic permafrost (Legendre, et al., 2015). As glaciers and permafrost melt due to global warming, the scientific community anticipates discovery of dormant viruses that can cause new infectious diseases for which a treatment modality may not exist. Identifying the viral protein expression pattern for a given host can potentially help identify existing drug treatments that may prove efficacious against the new virus. For example, an existing antiviral drug D that is used to treat a viral infection caused by virus V3 which expresses protein set P can be used, *in principle*, to treat an infection caused by virus V4 that expresses protein set S which is a superset of protein set P expressed by virus V3.

1.4. Background Studies

Protein structures (primary, secondary, tertiary and quaternary) and function are closely intertwined. Understanding of both protein structures and function are crucial for several studies including the development of new therapeutics that offer higher benefit to risks ratio. Strategies employed to develop such potential therapeutics include targeted drug delivery to achieve high local concentration (Srivivasarao & Low, 2017) and/or reversible or irreversible protein binding to inhibit protein function (Lin, Meng, Jiang, & Roux, 2013). Understanding of protein function often aid in the design of these potential therapeutics. However, the structure and function of many proteins are still unknown.

Modeling studies are widely employed to understand both protein structure and function. Such studies have attempted to understand these fundamental protein attributes either using small molecule datasets or using datasets of several protein sequences. Modeling studies that employ small molecule datasets attempt to understand protein morphology at potential interaction sites of the small molecules for a given protein. Some of the techniques employed for studying protein morphologies using large datasets of small molecules include molecular docking studies (Pagadala, Syed, & Tuszynski, 2017), Quantitative Structure Activity Relationship studies (Damale, Harke, Khan, Shinde, & Sangshetti, 2014), Comparative Molecular Field Analysis, and others (Damale et al, 2014). On the other hand, modeling studies that employ protein datasets attempt to classify proteins into families of known structure or function. Techniques employed to classify and/or cluster related protein structures include supervised and unsupervised machine learning methodologies (Cheng, Tegge, & Baldi, 2008). For example, SVM-Prot is a webserver that employs machine learning algorithms to predict protein functional families independent of the protein sequence (Li et al., 2016). The technique relies on classifying proteins into functional families based on sequence-derived structural and physicochemical properties like amino acid composition, hydrophobicity, polarity, polarizability, etc. (Li et al., 2016, Han et al, 2004).

The present work, however, attempts to explore frameworks that can be utilized to visualize viral proteins common for a given host. This work is thus the next step of modeling studies on protein datasets as applied to viruses. The major challenge associated with such a study is to identify a protein classifier or clustering technique applicable to all viral proteins so that viral protein expression patterns can be appropriately identified.

Comparative protein expressions between species has been studied. In one such study, employing two-dimensional gel electrophoresis and microarray techniques (Enard et al., 2002), the authors compared protein expressions of chimpanzees with the protein expressions in humans from multiple cell types (blood leukocytes, liver and brain). The study revealed that despite having a high genomic similarity (98.7%) between humans and chimpanzees, the species had the greatest differences in their respective protein expressions in their brain cells.

Research laboratories have investigated comparative protein expressions between cancer cell lines to understand mechanisms of differential resistance expressed by the cell lines to oncolytic viruses. The goal of the study was to identify cancer types susceptible for oncolytic viruses-based cancer therapeutics (Tarasova et al., 2018). Using a shotgun LC-MS/MS based label-free quantitation of identified proteins, the authors were able to identify differences in interferon signaling pathways of the tumor cells that helped explain the sensitivity of one tumor cell line to oncolytic viruses as opposed to the other.

The relevance of comparative protein expression studies that identify differences or similarities in protein expressions across species and/or cell types cannot be overstated as the two studies briefly mentioned above show. In addition to comparative protein expression studies, other experimental comparative "omic" studies have also been studied extensively to differentiate, characterize and understand the molecular mechanism of several cancer progressions (Cao et al., 2019), telomere biology (Schrumpfová, Fojtová, & Fajkus, 2019), etc.

Computationally, protein clusters from microbial genomes have been studied (Zaslavsky, Ciufo, Fedorov, & Tatusova, 2016). The goal of the study was to develop an adequate sampling strategy to construct meaningful groups of similar proteins that are useful for analysis and functional annotation. As part of the sampling strategy, the authors created protein clusters at three levels:

i. tight clusters (species-level clades) in groups of closely related genomes taking sequence similarity and genome context considerations.

ii. conservative clustering of the clusters obtained in (i) into clustroids that are seed global clusters and

iii. clusters that were built around seed global clusters.

The authors acknowledged that non-conservative or unique proteins and/or rapidly evolving proteins from rare genomes did not group well under the clustering strategies delineated above ((i)-(iii)) and noted that processing of these proteins required significant computational resources and produced questionable clusters.

Similarly studies to classify bacterial proteins on their subcellular localizability prediction have been carried out to facilitate genome annotation, vaccine development and to identify drug targets (Gardy & Brinkman, 2006). Several methods have been proposed and rely on supervised learning algorithms that uses prior knowledge of sequence motifs, signal peptides, etc. to predict subcellular localization of protein sequences. The earliest proposed method was PSORT I that relies on several aspects of protein structure like amino acid composition, sequence motifs, signal peptides and trans-membrane α-helical structures to predict protein sub-cellular localization. PSORT I evolved over the years to PSORTb. PSORTb is based on a Support Vector Machine algorithm that incorporates frequent subsequence identification and motif- and profile-matching modules, in addition to the protein classification tools employed in PSORT I like signal peptides, amino acid composition, etc.

The goal of this study is to explore frameworks that could be used to visualize viral protein expressions for a given host that are common to more than one virus. The study can potentially help facilitate expand the number of anti-viral targets and vaccine development studies provided the surface glycoproteins like hemagglutinin have very similar structure.

2. STUDY DESCRIPTION

The protein primary sequence along with the folding kinetics and energetics that a protein experiences during its biosynthesis is the basis for a protein's final structure and its cellular function. As stated before, this study explores different frameworks that can potentially be used to quickly visualize protein similarities for different viruses that propagate in a given host. Identification of such a framework can help extract knowledge from large protein datasets rapidly. Potential benefits of such a framework identification include accelerated identification of frequent protein sets for viruses residing in a given host, capturing relationship between proteins in a dataset as association rules, clustering and classification studies besides implications toward understanding of protein networks and potential cell signaling.

2.1. Virus Dataset

The viral genome dataset was downloaded from National Institutes of Health (NIH) (viral dataset). The curated dataset was semi-structured and presented information on a virus's name, host's name when known, genome type (DNA, RNA, etc.), protein expressions and their associated genes, protein function when known, and the literature citation that reported the virus's characterization besides other information. The downloaded dataset was available as a single file that contained the information for all the viruses (viral.1.genomic.gbff).

The database included viruses that contained a single genome (the genetic information for the virus coded by a single genome sequence) or segmented genome (the genetic information for the virus were coded by multiple genome sequences that are not linked together). The segmented viral genomes were listed separately with the same name for the virus that was appended to a counter indicating the genome segment which encodes for the proteins listed under that entry. For example, the Candiru virus hosted by humans contains three separate genome sequences identified by a large (L), medium (M) and small (S) segments. Each of these segments were listed separately in the dataset file obtained from NIH (Candiru virus segment L, complete genome, Candiru virus segment M, complete genome, and Candiru virus segment S, complete genome, respectively). Similarly, other viruses like Wallal virus isolate 927 hosted by *Anopheles annulipes* had 10 genome segements and were listed by count one through 10 (e.g. Wallal virus isolate 927 segment 1, complete sequence, Wallal virus isolate 927 segment 2, complete sequence, and so on).

The NIH database used for this study identified 1493 hosts which included the same hosts identified under different names (e.g. Human, Human being and *Homo sapiens* all referring to the human hosts) and a group of viruses for which no hosts were explicitly identified that was caught by the code. The latter group of viruses was grouped as “host_unknown” host in this study. The situation of providing different names to the same host was not anticipated and was not handled in the code. For example, the host information for human viruses were listed as human, *Homo sapiens, Homo sapiens*; Child and *Homo sapiens*; *Bovine*. Thus, viruses hosted by humans that were listed under different names for the host appear under different host names in the database. However, except for human viruses, the viruses listed for other hosts were not collated in the charts as viruses for the same host.

As a first step of the study, the dataset was parsed to collect information pertaining to each virus in separate text files that were organized under folders named after the host for the given virus. During this step, all the proteins pertaining to the same virus, but those that were listed separately in the NIH dataset (e.g. the viruses with segmented genomes) were collated into a single text file. The protein sequences expressed by each virus was simultaneously parsed to generate the different unique identifiers that were used in this study as the folder organization reported above was being executed using Python code. The unique identifiers for the proteins that were being generated individually constitutes the frameworks that is being explored as part of this study to uniquely identify and visualize viral proteins of a given host.

2.2. Exploration of Unique Protein Identifiers

This study explores seven different protein identifiers on their feasibility to be used as a framework for rapid visualization of viral proteins for a given host. The motivation for the selection of these identifiers was based on the need for uniquely identifying every viral protein sequence in the NIH database such that the same protein sequence as determined by the identifier can be tracked on separate viruses. The unique identifiers proposed in this study were primarily based on the amino acid sequence for a given protein (primary structure). The identifiers proposed in this study, the bases for their selection and the rationale behind the selection of an identifier are summarized in Table 2. Briefly, the proposed protein identifiers and their methods of generation, respectively, are:

1. Pfam_Keys: The Pfam database is a large collection of protein domain families built off the UniProt database and is represented by multiple sequence alignments and hidden Markov models (HMMs) (http://pfam.xfam.org/help). The manually curated sequence alignment of a small set of representative family members yields a seed called the Pfam-A entry. The associated HMMs are searched against the UniProt database, and sequences that exceed the previously set threshold are included in the full sequence alignment. The Pfam-A entries have a proper HMM name assigned (e.g. RNA_helicase, Peptidase_C3, etc.). The Pfam keys used in this study were generated by submitting a protein sequence to a local install of the Pfam software (Pfam 27.0). Pfam 27.0 outputted Pfam-B entries as well when the software identified potential seeds and Pfam-A seeds were not known

for the submitted sequence. The HMM names for these seeds preceded with Pfam-B (e.g. Pfam-B_10762). The Pfam keys used for protein identification in this study were generated by concatenating the HMM names identified for a given sequence in the order they appeared in the sequence.

Table 2. Proposed Protein Identifiers, Bases and Rationale for Identifier Selection.

Proposed Identifier	**Bases**	**Rationale**
Pfam-based (Pfam_Keys)	Protein primary structure, multiple sequence alignment and profile hidden Markov models	Pfam is a curated database built on multiple sequence alignment and homology modeling. Pfam analysis assigns a protein sequence to a protein family that is representative of its function based on the domains identified for the sequence.
Amino acid-based (AA_Type_Keys)	Primary protein structure, amino acid type	Protein biosynthesis often involves point mutations in which one amino acid can be substituted by another of similar physicochemical properties. For example, a hydrophobic amino acid such as leucine can be substituted by another hydrophobic amino acid (e.g. isoleucine)
Hydrophobicity index-based (HI-based)	Primary protein structure, amino acid type and the retention time of specific polypeptides containing the amino acid under consideration	HI has been used to predict subcellular localization. Two key types were explored: one in which the protein sequence length was considered and the other, in which it was not considered as part of the key. The rationale for considering the length along with HI was to prune the dataset, if needed, to group similar proteins.
Combination key-based (Combo_Key)	Primary protein structure, amino acid type	This key was considered to prune the dataset to group similar proteins, if needed.
MD5 Hash key - based (MD5_Key)	Primary protein structure	The purpose of this key was to generate a unique identifier for a given protein so that repeated sequences in different viral genomes could be identified. MD5 hash algorithm generates a 32-digit hexadecimal sequence.
SHA-512 Hash key -based (SHA512_Key)	Primary protein structure	The purpose of this key was to generate a unique identifier for a given protein so that repeated sequences in different viral genomes could be identified. SHA-512 hash algorithm generates a 128-digit hexadecimal sequence.

2. AA_Type_Keys: These keys are generated from the protein sequence reported for the virus. The individual standard amino acids in the virus were classified either as Group A, Group B, Group C or Group D as shown in Table 3. Group D was included to account for non-standard amino acids that may be found. The number of each amino acid type were counted and the key was generated for each sequence by concatenating the counts for Groups A-D (in that order) with an underscore ("_") character separating two counts. Thus, an AA_Type_Key "821_687_618_1" reported for virus Duvenhage virus isolate 86132sa in humans implies that the protein is comprised, respectively, of 821 Group A, 687 Group B, 618 Group C and 1 non-standard (Group D) amino acids in the sequence.

The amino acids were grouped into Groups A-D based on the HI that was reported by Sereda, Mant, Sönnichsen, & Hodges (1994) and Monera, Sereda, Zhou, Kay & Hodges (1995) as shown in Table 3 (see below). Even though an initial glance may suggest the Groups A-D correspond with hydrophobic, neutral, hydrophilic and other amino acid types, the terminology of Groups A-D was chosen because of discrepancy with HI reported for proline that suggests proline to be highly hydrophilic contrary to accepted scientific consensus that considers proline to be a hydrophobic amino acid.

3. HI_Num: These keys were generated by summing the products of hydrophobicity indices (HI) reported for the standard amino acids (Table 3) and their frequency of appearance in a specific protein sequence (equation 1). The HI for the standard amino acids were based on the retention times of a nine amino acid polypeptide containing the specific amino acid under consideration at position 5 of the polypeptide (Sereda et al., (1994) and Monera et al (1995)). The HI values used for each amino acid (Table 3) is the value reported for each amino acid individually and is not that for the amino acid within

a protein (Rose, Geselowitz, Lesser, Lee, & Zehfux, 1985). The HI for non-standard amino acid were not known and was taken as zero.

Table 3. Amino Acid Classification Scheme

Amino Acid Type	Amino Acid	Single-Letter Amino Acid Abbreviation	Hydrophobicity Index
Group A	Alanine	A	41
	Isoleucine	I	99
	Leucine	L	97
	Phenylalanine	F	100
	Tryptophan	W	97
	Tyrosine	Y	63
	Valine	V	76
Group B	Asparagine	N	-28
	Cysteine	C	49
	Glutamine	Q	-10
	Glycine	G	0
	Methionine	M	74
	Serine	S	-5
	Threonine	T	13
Group C	Arginine	R	-14
	Aspartic acid	D	-55
	Glutamic acid	E	-31
	Histidine	H	8
	Lysine	K	-23
	Proline	P	-46
Group D	Selenocysteine	U	0
	Ornithine	O	0
	All others	-	0

$$\text{HI}_{\text{protein}} = \sum_n \prod_i c_i AA_i \quad (1)$$

$where\ c_i\ represents\ the\ total\ number\ of\ amino\ acid\ AA_i\ at\ position\ i\ of a$

$protein\ chain\ of\ length\ n$

4. HI_Keys: These keys were generated by calculating the HI_Num key for each protein sequence as noted above and concatenating the calculated value with the sequence length. The two values were separated with an underscore ("_"). Thus, a HI_Key of 53674_2433 is an identifier of a protein with 2433 amino acid residues and a calculated HI_Num of 53674.
5. Combo_Keys: These keys represent combination keys that were generated by concatenating the corresponding HI_Key and AA_Type_Key the reported above for the protein under consideration. Thus, a Combo_Key 53674_2433_949_818_666_0 represents a protein with 2433 amino acid residues and a calculated HI_Num of 53674. The 2433 amino acid residues comprised of 949 Group A, 818 Group B and 666 Group C and 0 Group D amino acid residues.
6. MD5_Keys: These keys represent the MD5 hash for the protein sequence.
7. SHA512_Keys: These keys represent the SHA512 hash for the protein sequence.

2.3. Identifier Selection Criteria

Protein structure and function are closely intertwined (Hou, Jun, Zhang, & Kim, 2005). It has been observed that sequence-level homology of protein sequences is less conserved. On the other hand, protein evolution has remarkably conserved structure-level homology. It seems that Nature has strived to maintain protein structure that may have been initially developed for a certain function during protein evolution. Thus, the observed lack of protein sequence conservation during biosynthesis may be thought of as Nature's experimentation to optimize protein structure for a given function.

Thus, the Pfam-based (Pfam_Keys) protein identifiers are proposed to identify common viral proteins for a given host to account for the structural homology of proteins as each domain represents a structural unit (http://pfam.xfam.org/help). On the other hand, the amino acid type-based (AA_Type_Keys) protein identifiers are proposed to account for sequence variability that may result due to mutation effects on the viral genome because of evolution that may cause codon variability for an amino acid at any given location of a protein sequence (missense mutation). Thus, the Pfam_Keys and AA_Type_Keys are employed to assist potential grouping of proteins based on structure and sequence similarities.

Protein hydrophobicity index (HI) has been used to predict protein cellular localization propensities (Feng & Zhang, 2001). The HI-based identifiers are proposed to capture such localization propensities of viral proteins. Two types of HI-based identifiers are proposed: one in which the protein sequence length is also a part of the identifier and the other in which the sequence length is not considered. The former identifier is called HI_Key and the latter HI_Num (numeric) in this study respectively. The sequence length modifier was considered as a potential protein identifier to assist grouping of proteins if the numeric HI_Num identifier presented a continuous protein space devoid of any clear demarcation.

Although several methods have been proposed to calculate HI of amino acids (Wolfenden, Lewis, Jr., Yuan & Carter, Jr. 2015), the study uses the HI reported for amino acids based on retention time (*vide supra*). The reasoning for this approach is two-fold:

i. With the large number of viral protein sequences that were available from NIH dataset, it was hoped that employing equation 1 on reported HI of individual amino acids would yield a wider range of cumulative calculated protein hydrophobicities, which could potentially be useful to better group proteins of similar hydrophobicities.

ii. It has been reported in the literature that the microenvironment of an amino acid can influence its hydrophobicity (Bandyopadhyay & Mehler, 2008). A simple approach to calculate protein HI was selected since this study is only an initial exploration of potential protein identifiers that could be used as frameworks for rapid visualization of viral proteins common for a given host.

The next protein identifier proposed in this study is called the Combo_Key. A protein Combo_Key is simply a combination of two protein identifiers proposed in this study, namely the HI_Key and AA_Type_Key identifiers. The goal of the Combo_Key identifier, like others, was to help group similar viral proteins.

Finally, two protein identifiers based on hash algorithms are proposed in this study. The goal of these identifiers was to uniquely identify specific protein sequences that may be expressed by two different viruses. Two hash algorithms were considered: The MD5 and SHA512 hash algorithms. The justification for the two hash-based protein identifiers is to identify any potential collisions (two sequences generating the same hash key). Thus, the MD5 hash algorithm which generates a 32- digit hexadecimal sequence has a higher risk of running into collisions than the SHA512 hash algorithm since the latter generates a longer (128 digit) hash key.

2.4. Database Schema Description

The information gathered for the viruses were stored in a MySQL 8.1 database. The table names and their purpose in this study are summarized in Table 4. The information for the hosts, viruses, proteins and the seven proposed keys were all stored in separate tables with the same name as the record they stored. Thus, the hosts table stored information about the hosts, the viruses table stored data for viruses, proteins table for the proteins and so on. Apart from these10 tables, other cross tables were also created storing information of relationship between these tables. A total of 28 tables were constructed to store the data collected from the original NIH dataset (viral.1.genomic.gbff viral dataset) and the relationship between the data.

Table 4. Table Names and Purpose

Table Name	Purpose
Hosts	Store basic information about the hosts like host name, number of viruses hosted, etc. The primary key (PK) for the table is named host_no.
Viruses	Store basic information about the viruses like virus name and genome type among other information. The PK for the table is named virus_no.
Proteins	Store basic information about the proteins like the sequence, chain length, etc. The primary key (PK) for the table is protein_no. This table is a cross-reference to other tables that stores the information about the seven proposed protein identifiers in this study. The PKs for proposed identifiers are the foreign keys (FK) in this table.
AA_Type_Keys	Stores the information pertaining to amino acid type-based protein identifier proposed in this study. The PK for the table is AA_Type_key_no and the table stores information about the calculated AA_Type_key, the host_no(s), virus_no(s) and protein_no(s) associated with a given AA_Type_key.
Combo_Keys	Stores the information pertaining to combination key-based protein identifier proposed in this study. The PK for the table is Combo_key_no and the table stores information like AA_Type_Keys table for a given Combo_key.
HI_Keys	Stores the information pertaining to hydrophobicity index-based protein identifier proposed in this study. The PK for the table is HI_key_no and the table stores information like AA_Type_Keys table for a given HI_key.
HI_Num	Stores the information pertaining to hydrophobicity index-based protein identifier proposed in this study. The PK for the table is HI_no and the table stores information like AA_Type_Keys table for a given HI_no.
MD5_Keys	Stores the information pertaining to MD5 hash-based protein identifier proposed in this study. The PK for the table is MD5_key_no and the table stores information like AA_Type_Keys table for a given MD5_key_no.
Pfam_Keys	Stores the information pertaining to Pfam-based protein identifiers proposed in this study. The PK for the table is pfam_key_no and the table stores information about the identified Pfam domains, number of domains, the host_no(s), virus_no(s) and protein_no(s) associated with a given pfam_key_no.
SHA512_Keys	Stores the information pertaining to SHA512 hash-based protein identifier proposed in this study. The PK for the table is SHA512_key_no and the table stores information like AA_Type_Keys table for a given SHA512_key_no.

3. RESULTS AND DISCUSSION

The viral genome dataset from NIH was downloaded as a single zip file, that was extracted and parsed to collect individual virus information into separate text files that were stored in folders named after the host reported for the virus in the data file. The protein sequences reported for each virus simultaneously parsed to generate the identifiers proposed in this study and recorded in the same text file for the virus.

The test files that were generated as described above were then read and the information saved into a MySQL 8.1 database. There was a total of 4840 distinct viruses reported for 1493 hosts. Some of these hosts were duplicates as they were reported under different names as was noted earlier (e.g. human and *Homo sapiens*).

3.1. Viral Proteins Identified and Proposed Identifier Performance

The 4840 viruses had a total of 202777 proteins distributed amongst them, of which 179273 protein sequences were unique. The number of unique identifiers calculated using each of the proposed identifiers are summarized in Table 5.

Table 5. Summary of Protein Identifiers for All Hosts and All Viruses

Protein Identifier	Number of Unique Identifiers Identified	Number of Proteins the Unique Identifier Accounts for
Pfam_Keys	9454	80024
AA_Type_Keys	121249	202386
HI_Keys	161922	202619
HI_Num	23243	199170
Combo_Keys	178768	202670
MD5_Keys	179273	202742
SHA512_Keys	179273	202742

The proteins generated a total of 179273 hash-based identifiers (MD5_Keys and SHA512_Keys) of which 12651 protein sequences appeared more than once in different viruses.

Similarly, both hash-based identifiers reported that only the sequence 'MHKPLTQEHADPDKPEEALAWAFWGLPHPSGGHSLSNPVMAKYWSKHFTELGIVHVDSLRRLADENGNIHVSKLPQQTKKFQAPARGPRSHYNPAAQWVPSDTPEPPKFRVQDPRTLTQQEQQAQLDIYKQMGLIPTAPLPQHQAAVE' specifically of the 202777 proteins appeared the highest number of times amongst all the viruses. The specific sequence appeared in 30 different viruses. In general, the two hash-based identifiers provided identical counts for the number of viruses expressing above a certain count of the sequence identified by these identifiers. This along with the fact that the number of hash-based keys and number of unique protein sequences are equal indicate that there were no hash collisions during the generation the hash-based protein identifiers employed in this study. Furthermore, the hash-based protein identifiers provided the highest number of unique identifiers for all the viral proteins that were reported in the NIH data file suggesting that the hash-based protein identifiers were appropriate choice to uniquely identify every protein sequence in the study.

A surprising observation in this study is that the Combo_Keys proposed in this study were also remarkably selective in providing unique identifiers (178768 unique identifiers for the 179273 reported unique protein sequences). As was noted earlier, the Combo_Keys were generated by concatenating the HI_Num calculated for the protein sequence using individual amino acid HI (equation 1), the sequence length and the amino acid composition of the sequence (number of Group A-D amino acid type residues). The individual measures (protein HI, sequence length, number of Group A-D amino acid type residues) were separated using an underscore (“_”).

The number of unique identifiers reported by Pfam_Keys, AA_Type_Keys, HI_Keys and HI_Num identifiers are 9454, 121249, 161922 and 23243, respectively (Table 5). As can be seen from Table 5, the Pfam_Keys were available only for a total of 80024 proteins of a total of 179273 unique protein sequences reported in the NIH data file unlike the other identifiers that accounted for more than the number of reported unique sequences (Table 5). This discrepancy is because of the poor reporting of protein sequences in the NIH data file (*vide infra*) besides the fact that protein structure is not yet known for all proteins nor can they be predicted accurately.

For example, the sequence MTTTHDTNTKKLKYQFHTIHSQRIMTTVTQKPFTASPYI FSTTLRTTQTDGNNAINSHSHTQAGYNNSSERFLYLICTYIT appears twice for the virus Acidianus bottle-shaped virus, complete genome (Virus number 40) which is hosted by Acidianus convivator (Host number 18). This was an unanticipated data entry. The code expected identical protein sequences to appear and was written to query the database for every sequence encountered. The fact that the same sequence appeared multiple times for the same virus in the same file might have resulted in a situation where the database commit was not completed for the prior entries and was not included in subsequent query results. This seems very much likely to be the cause for the observed discrepancy where the number of proteins accounted for by the unique identifiers exceeds the number of unique protein sequences in the entire dataset. The reasoning is based on the following observation. The database connection was opened in code once when reading of a virus file started. The connection was left open for all queries while the file was open and was closed only after the entire file was read. It seems that all database entries must have remained in memory and not committed to disk which caused subsequent queries to not retrieve the memory data since the queries are executed for data committed in disk and not to uncommitted data that is still in memory. It is most likely that the uncommitted data in memory is written to disk when the database connection is closed in code. This seems to be a logical reason because the code did identify duplicate protein identifiers and sequences when they appeared in separate virus files. Opening and closing of database connections are expensive operations and were kept to a minimum due to performance considerations.

On the other hand, the number of proteins accounted for by the proposed Pfam_Key unique identifier (80024) is significantly less than the number of unique protein sequences in the NIH dataset (179273) (Table 5). This suggests that the Pfam entries for all the viral proteins are not yet completely known, which can be a drawback for employing the Pfam_Keys as unique identifiers for the viral proteins.

3.2. Proposed Identifier Performance on Human Viruses

As a first step in the study, the proposed identifiers listed in Table 1 were calculated for human viruses. The downloaded NIH data file contained 115 viruses hosted by humans. The number of unique identifiers calculated using each of the proposed identifiers for these viruses are summarized in Table 6. A total of 1656 distinct protein sequences were reported for the viruses hosted by humans.

Table 6. Summary of Protein Identifiers for Viruses Hosted by Human Beings

Protein Identifier	Number of Unique Identifiers Identified	Number of Proteins the Unique Identifier Accounts for
Pfam_Keys	434	1332
AA_Type_Keys	1643	1683
HI_Keys	1656	1683
HI_Num	1564	1683
Combo_Keys	1656	1683
MD5_Keys	1656	1683
SHA512_Keys	1656	1683

It can be seen from Table 6 that Pfam_Keys accounted for only 1332 proteins of a total of 1656 distinct protein sequences suggesting that Pfam_Keys were not available for all the viral proteins expressed by the viruses that are hosted by human beings. It is also evident that the number of unique identifiers based on amino acid type (AA_Type_Key) and HI (HI_Num) are less than the number of unique protein sequences reported (1643 and 1564, respectively. Table 6), which suggests that some of these keys appeared in more than one virus. The other keys (HI_Key, Combo_Key, MD5_Key and SHA512_Key) were more specific to the individual sequences as the count for these unique identifiers equal the number of unique viral protein sequences in human beings (Table 6).

However, as was noted earlier for the viral proteins listed for all viruses (*vide supra*), a total of 1683 proteins were reported for these viruses indicating that some of the protein sequences were duplicated within the same virus file that caused the sequence (and the proposed identifiers associated with these sequences) to be not captured by the queries before new entries were to be saved to the database (*vide supra*). There were 27 protein sequences on the files pertaining to the viruses that were hosted by human beings with duplicate entries (Table 7).

Table 7. List of File Names for Viruses Hosted by Human Beings with Duplicate Entries in the NIH Data

Virus Name	**Number of Pairs of Duplicate Entries**
Human herpesvirus 1, complete genome	3
Human herpesvirus 2, complete genome	3
Human herpesvirus 3, complete genome	3
Human herpesvirus 6A, complete genome	2
Human herpesvirus 6B, complete genome	6
Human herpesvirus 7, complete genome	2
Bufavirus-3 genes for NS1, putative VP1, hypothetical protein, VP2, complete cds, strain: BTN-63	4
Candiru virus segment L, complete genome	1
Candiru virus segment M, complete genome	1
Candiru virus segment S, complete genome	2

The situation of the Candiru virus files reported in Table 7 is unique because the sequences were not duplicated in these files, but the files themselves were duplicated as the virus was reported to be hosted by *Homo sapiens* and Human being.

3.3. Exploring the Identifiers as Visualization Aid of Common Proteins

The goal of the current study is to explore unique identifiers that could be used to facilitate rapid visualization of proteins that are common to the viruses hosted by a species. The proposed identifiers were initially explored for viruses hosted by human beings. As Table 6 shows, only the Pfam_Keys seemed to be a reliable identifier as the other proposed identifiers turned out to be very specific a given protein sequence. The AA_Type_Key and HI_Num identifiers were less specific compared to HI_Key, Combo_Key, MD5_Key and SHA512_Key identifiers, but still was not generic enough like the Pfam_Key (Table 6).

The Pfam_Key covered only about 80% (1332/1656 = 0.8043, Table 5) of the proteins expressed by viruses in humans with only about 434 unique identifiers. The other keys offered 100% coverage of all the reported proteins and was relatively easy to calculate knowing only the protein sequence. However, these identifiers were unique to at least 94% of the proteins (HI_Num identifier, 1564/1656 = 0.9444, Table 5) and could not be employed to identify common proteins amongst viruses that were hosted by a species. However, these identifiers because of their relative uniqueness were useful to identify duplicate virus entries in the NIH database under different names (*vide infra*).

The relatively lower coverage for Pfam_Key while disappointing is not unexpected since these identifiers are based on protein structure unlike protein sequence. Solving protein structure and assigning the solved structure to a protein family is an incredibly slow process, and software that could predict protein structure are still evolving and are less reliable or are computationallyexpensive (Lee, Freddolino, & Zhang, 2017).

Figure 2 shows a plot of the Pfam_Key identifiers calculated for viruses hosted by humans against itself. The diagonal elements thus represent the intersection of the set of proteins represented by their respective Pfam_Key identifiers expressed by a virus with itself, which means that the diagonal elements are always the complete set of Pfam_Key identifiers for the proteins expressed by the virus. The off-diagonal data points, similarly, represent the intersection of Pfam_Key identifiers for proteins expressed by one virus against those expressed by other proteins. Thus, the off-diagonal data points are either equal to the set of Pfam_Keys on the diagonal data point or a subset of it.

Table 8. Genome Types for Viruses Hosted by Human Beings

Genome Type	**Virus Count**
Double-stranded DNA	44
Single-stranded DNA	26
Single-stranded RNA – positive strand	25
Single-stranded RNA – negative strand	18
Unknown DNA type	2

There are 114 unique viruses reported as being hosted by human beings in the NIH dataset. These viruses predominantly had their genetic information encoded by DNA (Table 8). The color coding in Figure 2 represents instances when the intersection of the set of Pfam_Keys of proteins expressed by a virus of one genome type yielded a non-null set with the set of Pfam_Keys of proteins expressed by a second virus of a different genome type (red color). It can be seen that, at least amongst viruses hosted by humans, the Pfam_Keys are very rarely common between viruses of different genome types. Figure 2 shows only four instances where two viruses of different genome types have common Pfam_Key identifier between the viruses.

Despite Figure 2 appearing to be really “crowded” and unable to convey any specific information, it is evident “clusters” of viruses exist that expresses the same Pfam_Key set between viruses. These clusters are primarily centered between virus numbers 3-10 (Astrovirus type, Figure 3), 20-23 (Gyrovirus type), 26-32 (Adenovirus type, Figure 4), 34-36 (Bocavirus type), 39-42 (Cosavirus type), 47-55 (Herpesvirus type, Figure 4), 56-73 (Papillomavirus type, Figure 5), and 94-100 (Torque teno mini virus type). Figures 3-5 show the prominent clusters among those listed above. Figure 4 also shows that the adenovirus type and herpesvirus type have common Pfam_Key protein identifier types.

3.4. Potential Utility of the Proposed Identifiers

The performance of Pfam_Key identifier was encouraging for viruses hosted by human beings. However, the Pfam_Key identifier covered less than 50% (80024/179273 = 0.4464) of all the viral proteins reported in the NIH data file (Table 5). The low coverage of proteins by the Pfam_Key identifier is because the Pfam_Key identifier is based on protein structural similarity and the structure for many of the proteins seems to be unknown in the Pfam database that was used in this study (Pfam 27.0).

On the other hand, the other proposed identifiers were based on protein sequence and accounted for all the proteins reported in the NIH data file. The MD5_Keys and SHA512_Keys were useful to generate unique and specific identifiers for a protein based on its sequence alone. The HI_Keys and Combo_Keys were less specific than MD5_Keys and SHA512_Keys but were not very useful as they generated nearly unique and specific identifier for every viral protein reported (161922/179273 = 0.9032 for HI_Keys, Table 5 and (178768/179273 = 0.9972 for

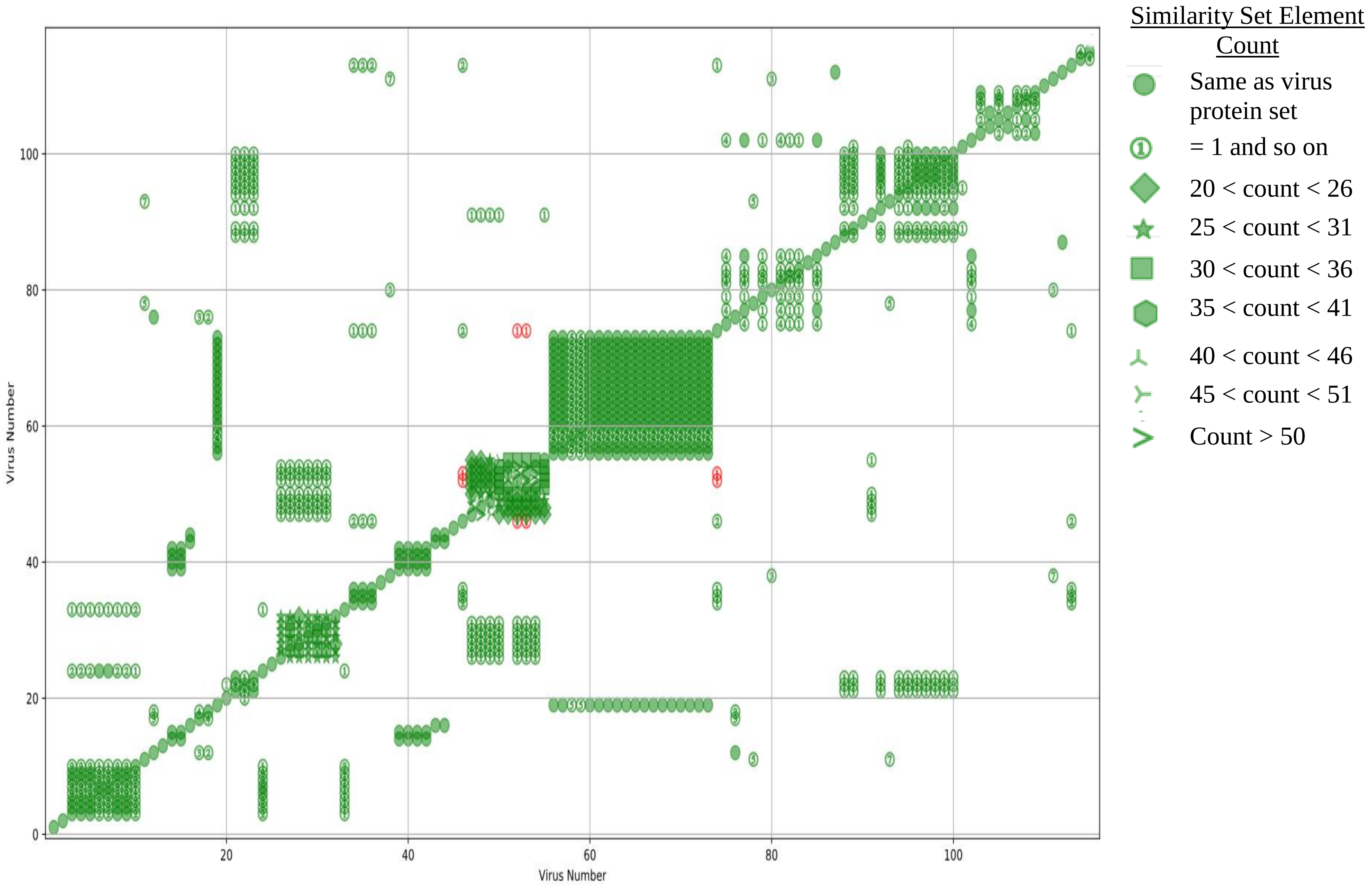


Figure 2. Pfam_Key Similarities for Viruses Hosted by Human Beings.

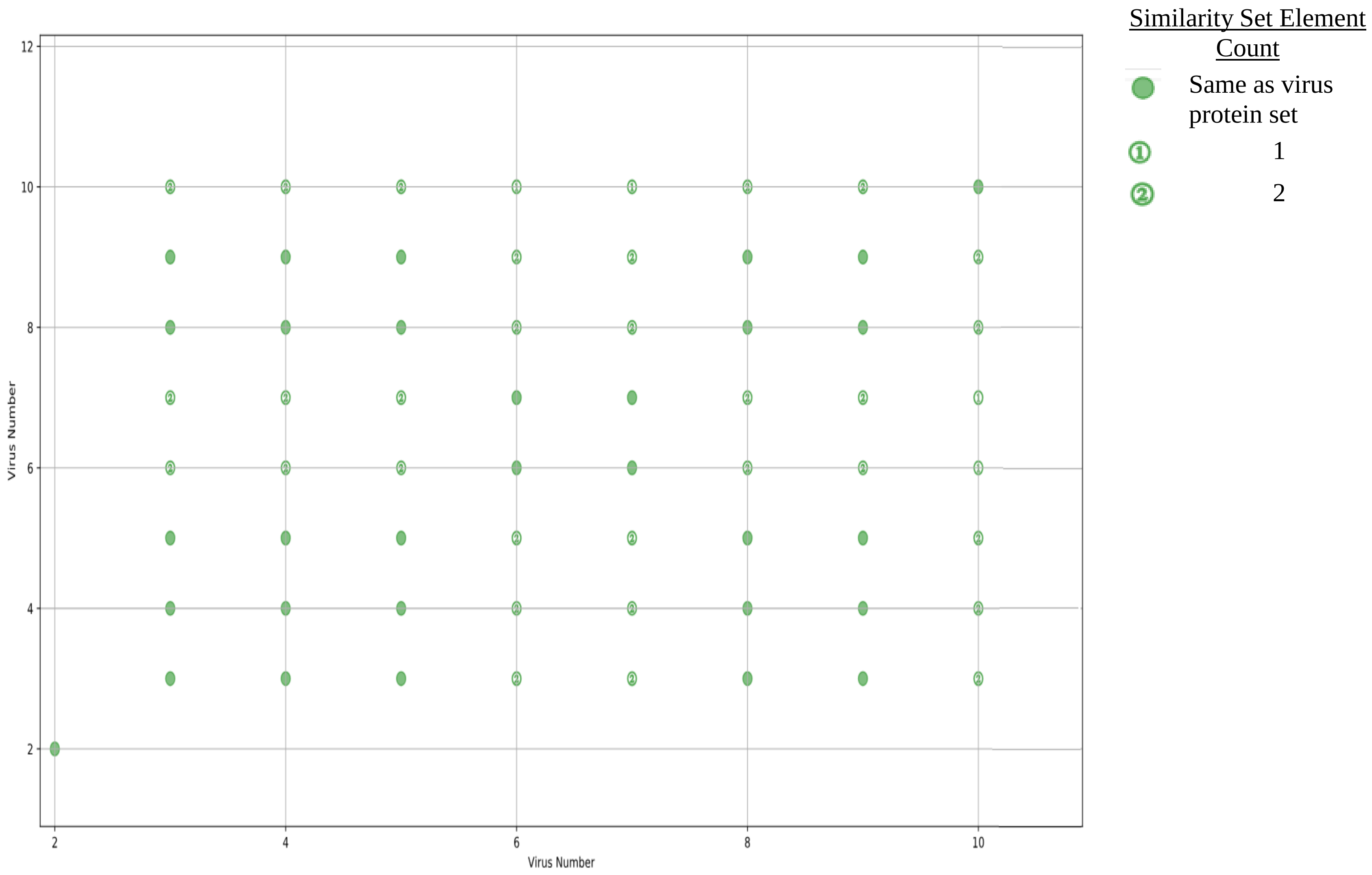


Figure 3. Pfam_Key Similarities for Astroviruses Hosted by Human Beings.

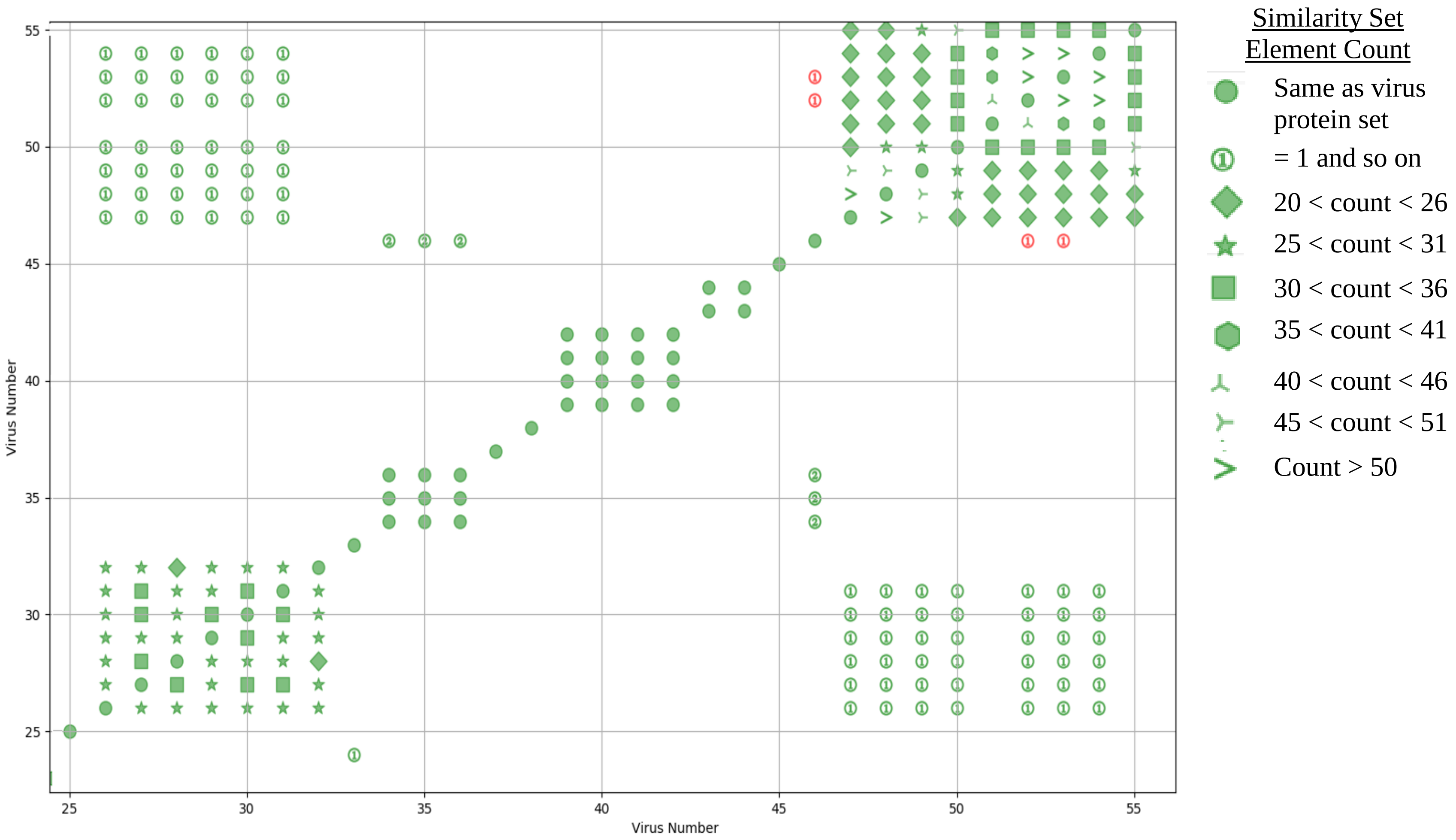
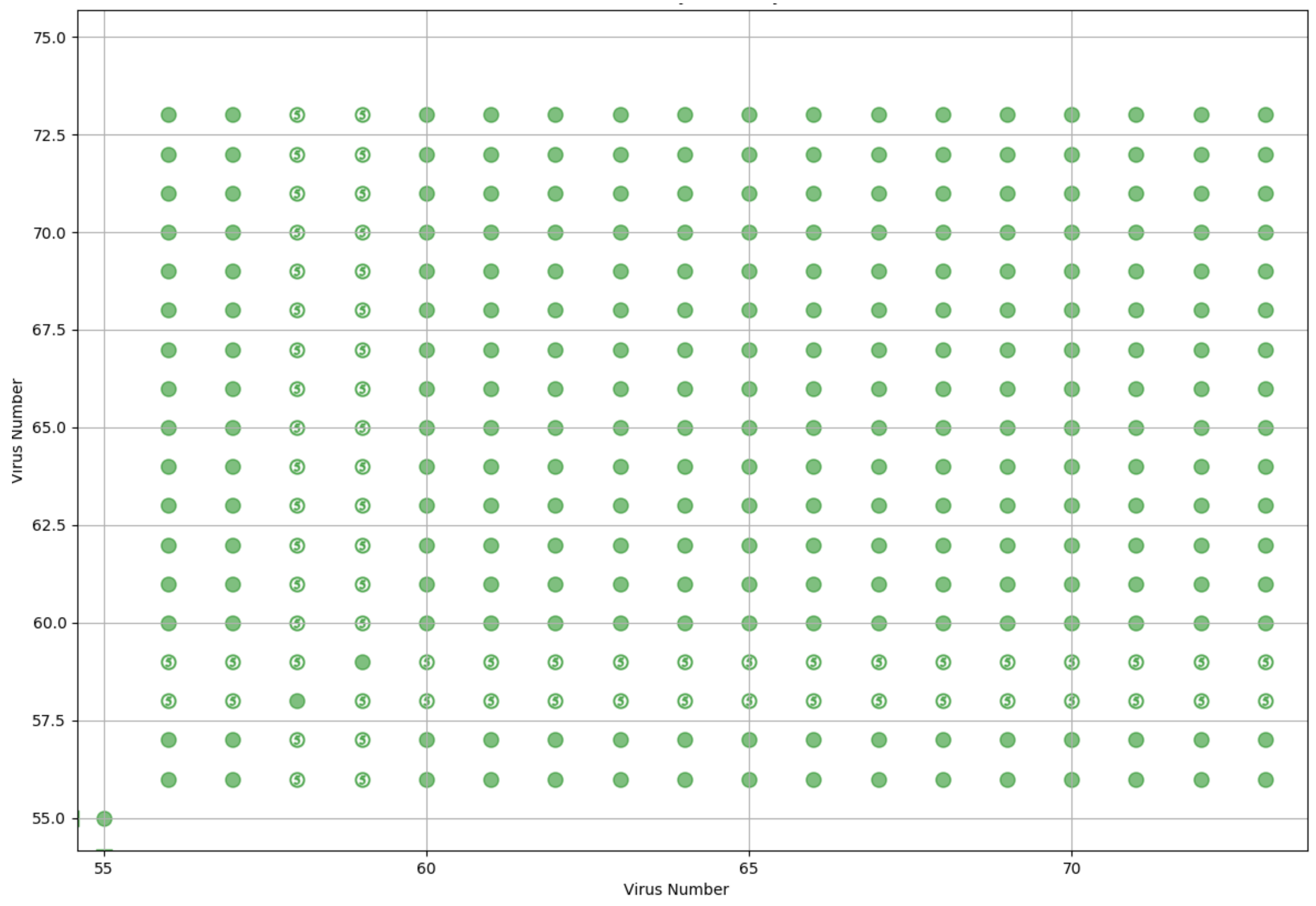


Figure 4. Pfam_Key Similarities for Adenoviruses and Herpesviruses Hosted by Human Beings.



Similarity Set Element Count

● Same as virus protein set

① = 1 and so on

Figure 5. Pfam_Key Similarities for Papillomaviruses Hosted by Human Beings

.

Combo_Keys, Table 5). The AA_Type_Keys and HI_Num identifier provided less specific protein identifiers based on the protein sequence for the proteins. Since these identifiers were not useful in identifying common protein sequences between viruses hosted by human beings, the utility of these keys was not further explored when all hosts and their hosted viruses were considered.

The proposed identifiers could not be employed to identify common proteins in viruses for various hosts other than humans due to lack of knowledge of protein structures. The Pfam_Key based protein identifier was useful for human hosts because the structure for a majority the viral proteins hosted by humans were known. The host information for many viruses from the NIH data file could not be captured in code because of inconsistent data entry, and these viruses were collected together under “host_unknown”.

One of the potential benefits that was mentioned of this study is that a drug that can treat a viral infection caused by a virus expressing a protein set P can be used to treat another infection caused by a different virus as long as the latter virus expresses a protein set that is a superset of the protein set P expressed by the former virus. The Pfam_Key identifier set for viruses hosted by all the other hosts except humans were then inspected to check if any of these viruses could be a superset of any of the studied human virus’ Pfam_Key protein identifier set. This analysis revealed that a total of 380 viruses that were reported to be hosted by hosts including “host_unknown” other than human beings has protein expression sets that could be considered a superset of the Pfam_Key identifier set calculated for the viruses hosted by humans in this study.

The 380 viruses that can be a potential human health risk are listed in Appendix. It must be noted that the 380 viruses are *only* potential human health risks and not necessarily *real* human health risks. This distinction needs to be understood because one of the key characteristics of viruses as they propagate in a host is that they express proteins to optimize their own survival as they respond to selective pressures the virus experience in the given host.

Another factor that needs to be considered as well in evaluating the potential health risk is that the virus capable of posing human health risk may get cleared out by human immune system as soon as the virus makes an entry into a human body because of its surface characteristics (e.g. surface-bound glycoproteins on the virus may trigger a spontaneous immune response and offer instance immunity).

Finally, the risk analyses posed by the viruses listed in Appendix only considers protein expression sets that are accounted for by the Pfam_Key identifiers. The other proteins that are still unaccounted may make human environment inhabitable for the virus (e.g. these proteins may not be able to fold to their native state in a human cell).

3.5. Conclusions

Seven protein identifiers were proposed in this study to uniquely identify, potentially classify and assist in rapid visualization of common proteins between viruses. The initial goal of the study was to identify common proteins for viruses that propagate in a given host. Of the seven proposed identifiers, one of the identifiers relied on protein structure (Pfam_Key identifier), while the remaining six relied on protein sequence (AA_Type_Key, HI_Key, HI_Num, Combo_Key, MD5_Key and SHA512_Key). The reason for exploring six different identifiers that relied on protein sequence alone is the goal of the study, which is to explore frameworks to assist rapid visualization of common proteins expressed by viruses that propagate in a given host. Elucidation and/or prediction of protein structure and its subsequent classification into a protein family is a slow process. Hence, an ideal unique identifier should rely on protein sequence alone since the protein sequence can be quickly predicted from the open reading frame for the protein in the genome. However, this study found that identifiers relying on protein sequence alone are not efficient in generating unique identifiers that are useful to identify common viral proteins for a given host. On the other hand, identifiers that relied on protein structure were much better in providing unique identifiers that can be used in developing rapid visualization frameworks of common viral proteins. However, lack of structure information for many proteins remains a drawback of this approach.

REFERENCES

Ali, S. M., Amroun, A., de Lamballerie, X., & Nougairède, A, (2018). Evolution of Chikungunya virus in mosquito cells. *Scinetific Reports*, *8*, 16175. https://doi.org/10.1038/s41598-018-34561

Amino acid hydrophobicities used in this study were retrieved from https://www.sigmaaldrich.com/life-science/metabolomics/learning-center/amino-acid-reference-chart.html on 4/14/2019

Bandyopadhyay, D., & Mehler, E. L. (2008). Quantitative expression of protein heterogeneity: response of amino acid side chains to their local environment. *Proteins: Structure, Function, Bioinformatics*, *72*, 646-659. https://doi.org/10.1002/prot.21958

Blair, J. M. A. (2018). A climate for antibiotic resistance. *Nature Climate Change*, *8*, 460-461. https://doi.org/10.1038/s41558-018-0183-0

Cao, Y., Li, Z., Mao, L., Cao, H., Kong, J., Yu, B., …Liao, W., (2019). The use of proteomic technologies to study molecular mechanisms of multidrug resistance in cancer. *European Journal of Medicinal Chemistry*, *162*, 423-434. https://doi.org/ 10.1016/j.ejmech.2018.10.001

Cheng, J., Tegge, A. N., & Baldi, P. (2008). Machine learning methods for protein structure prediction. *IEEE Reviews in Biomedical Engineering*, *2008*, 1, 41-49. https://doi.org/10.1109/RBME.2008.2008239

Damale, M. G., Harke, S. N., Khan, F. A. K., Shinde, D. B., & Sangshetti, J. N. (2014). Recent advances in multidimensional QSAR (4D-6D): a critical review. *Mini-Reviews in Medicinal Chemistry*, *14*, 35-55.

Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P., Pääbo, S. (2002).Intra- and interspecific variation in primate gene expression patterns. *Science*, *296*, 340-343. https://doi.org/10.1126/science.1068996

Feng, Z-P., & Zhang, C-T. (2001). Prediction of subcellular location of prokaryotic proteins based on the hydrophobicity index of amino acids. *International Journal of Biological Macromolecules*, *28*, 255-261. https://doi.org/10.1016/S0141-8130(01)00121-0

Fleming, N. (2018). Computer-calculated compounds. *Nature*, *557*, S55-S57. https://doi.org/10.1038/d41586-018-05267-x

Gardy, J. L., & Brinkman, F. S. L. (2006). Methods for predicting bacterial subcellular localization. *Nature Reviews Microbiology*, *4*, 741-751. https://doi.org/10.1038/nrmicro1494

Han, L. Y., Cai, C, Z., Ji, Z. L., Cao, Z. W., Cui, J., & Chen, Y. Z. (2004). Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Research*, *32*, 6437-6444. https://doi.org/10.1093/nar/gkh984

Hou, J., Jun, S-R., Zhang, C., & Kim, S-H. (2005). Global mapping of the protein structure space and application in structure-based inference of protein function. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 3651-3656. https://doi.org/10.1073/pnas.0409772102

Wolfenden, R., Lewis, Jr., C. A., Yuan, Y., & Carter, Jr., C. W. (2015). Temperature dependence of amino acid hydrophobicities. *Proceedings of the National Academy of the United States of America*, *112*, 7484-7488. https://doi.org/10.1073/pnas.1507565112

Kurane, I. (2010). The effect of global warming on infectious diseases. *Public Health Research Perspectives*, *1*, 4-9, https://doi.org/10.1016/j.phrp.2010.12.004

Lee, J., Freddolino, P. L., & Zhang, Y. (2017). Ab Initio Protein Structure Prediction. In D. J. Rigden (Ed.), From Protein Structure to Function with Bioinformatics, 2nd edition (3-35). Springer Dordrecht, Netherlands. https://doi.org/ 10.1007/978-94-024-1069-3_1

Legendre, M., Lartigue, A., Bertaux, L., Jeudy, S., Bartoli, J., Lescot, M., …Claverie, J-M. (2015). In-depth study of *Mollivirus sibericum*, a new 30,000-y-old giant virus infecting *Acanthamoeba*. *Proceedings of the National Academy of Sciences of the United States of America, 112,* E5327-E5335. https://doi.org/10.1073/pnas.1510795112

Li, Y. H., Xu, J. Y., Tao, L., Li, X. F., Li, S., Zeng, X., …Chen, Y. Z. (2016). SVM-Prote 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. *PLOS One*, *11*, e0155290. https://doi.org/10.1371/journal.pone.0155290

Lin, Y-L., Meng, Y., Jiang, W., & Roux, B. (2013). Explaining why Gleevec is a specific and potent inhibitor of Abl kinase. *Proceedings of the National Academy of Sciences of the United States of America*, *29*, 1664-1669. https://doi.org/10.1073/pnas.1214330110

Mignani, S., Huber, S., Tomás, H., Rodrigues, J., & Majoral, J-P. (2016). Why and how have drug discovery strategies in pharma changed? What are the new mindsets? *Drug Discovery Today*, *21*, 239-249. https://doi.org/10.1016/j.drudis.2015.09.007

Monera, O. D., Sereda, T. J., Zhou, N. E., Kay, C. M., & Hodges, R. S. (1995). Relationship of sidechain hydrophobicity and α-helical propensity on the stability of the single-stranded amphipathic α-helix. *Journal of Peptide Science*, *1*, 319-329. https://doi.org/10.1002/psc.310010507

Moreira, D., & López-García, P., (2009). Ten reasons to exclude viruses from the tree of life. *Nature Reviews Microbiology*, *7*, 306-311. https://doi.org/10.1038/nrmicro2108

Pagadala, N. S., Syed, K., & Tuszynski, J. (2017). Software for molecular docking: a review. *Biophysics Review*. *9*, 91-102. https://doi.org/10.1007/s12551-016-0247-1

Pfam 27.0 Retrieved from ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam27.0/

Rose, G. D., Geselowitz, A. R., Lesser, G. J. Lee, R. H., & Zehfux, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science*, *229*, 834-838. https://doi.org/10.1126/science.4023714

Schrumpfová, P. P., Fojtová, M., & Fajkus, J., (2019). Telomeres in plants and humans: not so different, not so similar. *Cells*, *8*(1), 58. https://doi.org/10.3390/cells8010058

Sereda, T. J., Mant, C. T., Sönnichsen, F. D., & Hodges, R. S. (1994). Reversed-phase chromatography of synthetic amphipathic α-helical peptides as a model for ligand/receptor interactions: Effect of changing hydrophobic environment on the relative hydrophobicity/hydrophilicity of amino acid side chains. *Journal of Chromatography A*, *676*, 139-153. https://doi.org/10.1016/0021-9673(94)00371-8

Sonnhammer, E. L. L., Eddy, S. R., & Durbin, R. (1997). Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins: Structure, Function, and Genetics*, *28*, 405-420. https://doi.org/10.1002/(SICI)1097-0134(199707)28:3<405::AID-PROT10>3.0.CO;2-L

Srinivasarao, M. & Low, P. S. (2017). Ligand-targeted drug delivery. *Chemical Reviews*, *117*, 12133-12164. https://doi.org/10.1021/acs.chemrev.7b00013

Tarasova, I. A., Tereshkova, A. V., Lobas, A. A., Solovyeva, E. M., Sidorenko, A. S., Gorshkov, V., …Gorshkov, M. V. (2018). Comparative proteomics as a tool for identifying specific alterations within interferon response pathways in human glioblastoma multiforme cells. *Oncotarget*, *9*, 1785-1802. https://doi.org/10.18632/oncotarget.22751

Viral dataset retrieved in 2015. Retrieved from ftp://ftp.ncbi.nih.gov/refseq/release/viral/

Zaslavsky, L., Ciufo, S., Fedorov, B., & Tatusova, T., (2016). Clustering analysis of proteins from microbial genomes at multiple levels of resolution. *BMC Bioinformatics*, *17*(Suppl 8): 276. https://doi.org/10.1186/s12859-016-1112-8

APPENDIX. LIST OF VIRUSES RESIDING IN OTHER HOSTS WITH POTENTIAL FOR HUMAN HEALTH RISK

A1.	AGERATUM_CONYZOIDES_SYMTOMLESS_ALPHASATELLITE
A2.	BHENDI_YELLOW_VEIN_MOSAIC_VIRUS-ASSOCIATED_ALPHASATELLITE
A3.	MESTA_YELLOW_VEIN_MOSAIC_VIRUS-ASSOCIATED_ALPHASATELLITE
A4.	CROTON_YELLOW_VEIN_MOSAIC_ALPHASATELLITE
A5.	ACARTIA_TONSA_COPEPOD_CIRCOVIRUS_ISOLATE_154_D11
A6.	ACHETA_DOMESTICUS_VOLVOVIRUS_ISOLATE_ADVVV-JAPAN
A7.	AGERATUM_YELLOW_VEIN_SINGAPORE_ALPHASATELLITE- [SINGAPORE; 1998]
A8.	ANGUILLA_ANGUILLA_CIRCOVIRUS_ISOLATE_BA1
A9.	GOOSE_CIRCOVIRUS
A10.	MILK_VETCH_DWARF_C10_ALPHASATELLITE_GENE_FOR_REPLICATION_INITIATIONPROTEIN,
A11.	BARBEL_CIRCOVIRUS
A12.	CLEOME_LEAF_CRUMPLE_VIRUS_ASSOCIATED_DNA_1
A13.	RAVEN_CIRCOVIRUS
A14.	CYCLOVIRUS_ZM36A_DNA
A15.	CYANORAMPHUS_NEST_ASSOCIATED_CIRCULAR_K_DNA_VIRUS
A16.	CYANORAMPHUS_NEST_ASSOCIATED_CIRCULAR_X_DNA_VIRUS
A17.	DRAGONFLY_CYCLOVIRUS_3_ISOLATE_FL2-5E-2010
A18.	DRAGONFLY-ASSOCIATED_ALPHASATELLITE_ISOLATE_PR_NZ48_2009
A19.	DRAGONFLY_CYCLOVIRUS_5_ISOLATE_PR-6E-2010
A20.	STARLING_CIRCOVIRUS
A21.	FELINE_CYCLOVIRUS
A22.	GOSSYPIUM_DAVIDSONII_SYMPTOMLESS_ALPHASATELLITE_DNA-ALPHA-B
A23.	GOSSYPIUM_MUSTILINUM_SYMPTOMLESS_ALPHASATELLITE_DNA-ALPHA-B
A24.	FABA_BEAN_NECROTIC_STUNT_ALPHASATELLITE_1_ISOLATE_PESHTATUEK_12B
A25.	FABA_BEAN_NECROTIC_STUNT_ALPHASATELLITE_2_ISOLATE_PESHTATUEK_12B
A26.	DRAGONFLY_CYCLICUSVIRUS_ISOLATE_FL1-NZ37-2010
A27.	DRAGONFLY_CYCLOVIRUS_2_ISOLATE_FL1-NZ38-2010
A28.	DUCK_CIRCOVIRUS

A29.	BLACK_MEDIC_LEAFROLL_ALPHASATELLITE_1_ISOLATE_LERIK-XALIFA_ 47
A30.	DRAGONFLY_LARVAE_ASSOCIATED_CIRCULAR_VIRUS-1_ISOLATED FLACV-1_NZ-PG11-LD,
A31.	DRAGONFLY_LARVAE_ASSOCIATED_CIRCULAR_VIRUS-2_ISOLATED FLACV-2_NZ-PG8-LS,
A32.	DRAGONFLY_LARVAE_ASSOCIATED_CIRCULAR_VIRUS-6_ISOLATED FLACV-6_NZ-PG9-LD,
A33.	DRAGONFLY_LARVAE_ASSOCIATED_CIRCULAR_VIRUS-7_ISOLATED FLACV-7_NZ-PG5-LH,
A34.	BAT_CIRCOVIRUS_ISOLATE_XOR7
A35.	DRAGONFLY_CYCLOVIRUS_4_ISOLATE_BG-NZ46-2007
A36.	PO-CIRCO-LIKE_VIRUS_41
A37.	PO-CIRCO-LIKE_VIRUS_51
A38.	PORCINE_CIRCOVIRUS_TYPE_1-2A
A39.	DRAGONFLY_CYCLOVIRUS_ISOLATE_DFCYV-A1_TO-6NZ21-TT- 2010_ REPLICATIONASSOCIATED PROTEIN AND CAPSID PROTEINGENES,
A40.	SUBTERRANEAN_CLOVER_STUNT_C2_ALPHASATELLITE
A41.	SUBTERRANEAN_CLOVER_STUNT_C6_ALPHASATELLITE
A42.	VERNONIA_YELLOW_VEIN_FUJIAN_VIRUS_ALPHASATELLITE
A43.	CYCLOVIRUS_BAT-USA-2009
A44.	BAT_CIRCOVIRUS_POA-2012-II
A45.	CYCLOVIRUS_NGCHICKEN15-NGA-2009
A46.	FINCH_CIRCOVIRUS
A47.	CYCLOVIRUS_PKGOAT11-PAK-2009
A48.	CYCLOVIRUS_PKGOAT21-PAK-2009
A49.	GULL_CIRCOVIRUS
A50.	BEAK_AND_FEATHER_DISEASE_VIRUS
A51.	CANARYPOX_VIRUS
A52.	CIRCOVIRIDAE_10_LDMD-2013
A53.	CIRCOVIRIDAE_11_LDMD-2013
A54.	CIRCOVIRIDAE_13_LDMD-2013
A55.	CIRCOVIRIDAE_14_LDMD-2013
A56.	CIRCOVIRIDAE_15_LDMD-2013
A57.	CIRCOVIRIDAE_21_LDMD-2013
A58.	CIRCOVIRIDAE_2_LDMD-2013
A59.	CIRCOVIRIDAE_5_LDMD-2013
A60.	CIRCOVIRIDAE_8_LDMD-2013

A61.	CIRCOVIRUS-LIKE_GENOME_BBC-A
A62.	CIRCOVIRUS-LIKE_GENOME_RW-A
A63.	CIRCOVIRUS-LIKE_GENOME_RW-B
A64.	CIRCOVIRUS-LIKE_GENOME_RW-C
A65.	CIRCOVIRUS-LIKE_GENOME_RW-D
A66.	CIRCOVIRUS-LIKE_GENOME_RW-E
A67.	COCONUT_FOLIAR_DECAY_ALPHASATELLITE
A68.	COLUMBID_CIRCOVIRUS
A69.	MCMURDO_ICE_SHELF_POND-ASSOCIATED_CIRCULAR_DNA_VIRUS-3_ ISOLATEALG49-39,
A70.	MCMURDO_ICE_SHELF_POND-ASSOCIATED_CIRCULAR_DNA_VIRUS-6_ ISOLATEALG49-69,
A71.	MILK_VETCH_DWARF_C1_ALPHASATELLITE_GENE_FOR_VIRAL REPLICATION-ASSOCIATED PROTEIN,
A72.	MILK_VETCH_DWARF_C2_ALPHASATELLITE_GENE_FOR_VIRAL REPLICATION-ASSOCIATED PROTEIN,
A73.	MILK_VETCH_DWARF_C3_ALPHASATELLITE_GENE_FOR_VIRUS REPLICATION-ASSOCIATED PROTEIN,
A74.	MULARD_DUCK_CIRCOVIRUS
A75.	MUSCOVY_DUCK_CIRCOVIRUS
A76.	OKRA_YELLOW_CRINKLE_CAMEROON_ALPHASATELLITE_ [CM%3ALYS1SP2%3A09]
A77.	PORCINE_CIRCOVIRUS_1
A78.	PORCINE_CIRCOVIRUS_2
A79.	SILURUS_GLANIS_CIRCOVIRUS_ISOLATE_H5
A80.	CARDAMOM_BUSHY_DWARF_VIRUS_SATELLITE_CLONE_FR-X7
A81.	MINK_CIRCOVIRUS_STRAIN_MICV-DL13
A82.	CYGNUS_OLOR_CIRCOVIRUS_ISOLATE_H51
A83.	FLORIDA_WOODS_COCKROACH-ASSOCIATED_CYCLOVIRUS_ISOLATE_ GS140
A84.	CANARY_CIRCOVIRUS
A85.	BHENDI_YELLOW_VEIN_DELHI_VIRUS_[2004%3ANEW_DELHI]DNA-A
A86.	BHENDI_YELLOW_VEIN_BHUBHANESWAR_VIRUS_DNA-A
A87.	BHENDI_YELLOW_VEIN_INDIA_VIRUS_[INDIA%3ADHARWAD_OYDWR2%3A2006]_DNA-A
A88.	COTTON_LEAF_CURL_ALLAHABAD_VIRUS_[INDIA%3AKARNAL%3AOY77%3A2005]_DNA-A
A89.	OKRA_ENATION_LEAF_CURL_VIRUS_[INDIA%3AMUNTHAL_EL37%3A2006] DNA-A

A90.	OKRA_LEAF_CURL_INDIA_VIRUS_[INDIA%3ASONIPAT_EL14A%3A2006] DNA-A
A91.	TOMATO_LEAF_CURL_CAMEROON_VIRUS_-_[CAMEROON%3ABUEA%3A OKRA%3A2008]
A92.	AGERATUM_ENATION_VIRUS
A93.	AGERATUM_LEAF_CURL_VIRUS_-_[G52]
A94.	AGERATUM_YELLOW_VEIN_CHINA_VIRUS
A95.	AGERATUM_YELLOW_VEIN_TAIWAN_VIRUS
A96.	PAPAYA_LEAF_CURL_CHINA_VIRUS_-_[G8]
A97.	HOLLYHOCK_YELLOW_VEIN_MOSAIC_VIRUS
A98.	ALLAMANDA_LEAF_MOTTLE_DISTORTION_VIRUS_ISOLATE_AL-K1
A99.	ALLAMANDA_LEAF_CURL_VIRUS_DNA-A
A100.	ASYSTASIA_BEGOMOVIRUS_1
A101.	PEPPER_LEAF_CURL_YUNNAN_VIRUS-[YN323]
A102.	PEPPER_LEAF_CURL_LAHORE_VIRUS-[PAKISTAN%3ALAHORE1%3A2004]
A103.	TOMATO_YELLOW_LEAF_CURL_VIRUS
A104.	PAPAYA_LEAF_CRUMPLE_VIRUS-PANIPAT_8_[INDIA%3APANIPAT%3A PAPAYA%3A2008]DNA-A,
A105.	PAPAYA_LEAF_CURL_GUANDONG_VIRUS_-_[GD2]DNA_A
A106.	CLERODENDRUM_GOLDEN_MOSAIC_CHINA_VIRUS_DNA_A
A107.	COCCINIA_MOSAIC_TAMIL_NADU_VIRUS_ISOLATE_TN_TDV_COC_1
A108.	CORCHORUS_YELLOW_VEIN_MOSAIC_VIRUS_ISOLATE_CEA8
A109.	CRASSOCEPHALUM_YELLOW_VEIN_VIRUS_-_JINGHONG
A110.	CROTON_YELLOW_VEIN_MOSAIC_VIRUS
A111.	CROTON_YELLOW_VEIN_VIRUS
A112.	SQUASH_LEAF_CURL_PHILIPPINES_VIRUS
A113.	CATHARANTHUS_YELLOW_MOSAIC_VIRUS
A114.	ECLIPTA_YELLOW_VEIN_VIRUS_CLONE_ECYVV-[PK_FAI_06]
A115.	EMILIA_YELLOW_VEIN_VIRUS-[FZ1]
A116.	EUPHORBIA_LEAF_CURL_VIRUS_DNA_A
A117.	GOSSYPIUM_DARWINII_SYMPTOMLESS_VIRUS_DNA-A
A118.	GOSSYPIUM_PUNCTATUM_MILD_LEAF_CURL_VIRUS_DNA_A
A119.	COTTON_LEAF_CURL_BUREWALA_VIRUS_-[INDIA%3AVEHARI %3A2004]
A120.	COTTON_LEAF_CURL_VIRUS_DNA-A
A121.	HEMIDESMUS_YELLOW_MOSAIC_VIRUS_CLONE_H1
A122.	KENAF_LEAF_CURL_VIRUS_DNA_A
A123.	MESTA_YELLOW_VEIN_MOSAIC_BAHRAICH_VIRUS-[INDIA%3A BAHRAICH%3A2007]_DNAA,

A124.	MESTA_YELLOW_VEIN_MOSAIC_VIRUS_DNA-A
A125.	JATROPHA_LEAF_CRUMPLE_INDIA_VIRUS_[J._CURCAS%3A_JODHPUR] ISOLATESKJ2,
A126.	JATROPHA_LEAF_CRUMPLE_VIRUS_ISOLATE_SKJ1
A127.	JATROPHA_MOSAIC_NIGERIAN_VIRUS_ISOLATE_2
A128.	JATROPHA_YELLOW_MOSAIC_INDIA_VIRUS_DNA-A
A129.	HONEYSUCKLE_YELLOW_VEIN_VIRUS-[UK1]
A130.	LUDWIGIA_YELLOW_VEIN_VIRUS_DNA-A
A131.	LOOFA_YELLOW_MOSAIC_VIRUS_DNA_A
A132.	TOBACCO_LEAF_CURL_KOCHI_VIRUS
A133.	TOMATO_LEAF_CURL_CHINA_VIRUS_-_[G32]
A134.	TOMATO_LEAF_CURL_NEW_DELHI_VIRUS_DNA_A
A135.	TOMATO_LEAF_CURL_GUANGDONG_VIRUS_DNA-A
A136.	TOMATO_LEAF_CURL_MADAGASCAR_VIRUS-MENABE [MADAGASCAR %3AMORONDOVA%3A2001],
A137.	TOMATO_LEAF_CURL_MAYOTTE_VIRUS
A138.	TOMATO_YELLOW_LEAF_CURL_GUANGDONG_VIRUS_DNA-A
A139.	MALVASTRUM_LEAF_CURL_GUANGDONG_VIRUS
A140.	MALVASTRUM_LEAF_CURL_VIRUS_-_[G87]
A141.	MALVASTRUM_YELLOW_VEIN_BAOSHAN_VIRUS_DNA-A
A142.	MALVASTRUM_YELLOW_VEIN_CHANGA_MANGA_VIRUS
A143.	MALVASTRUM_YELLOW_VEIN_YUNNAN_VIRUS
A144.	MALVASTRUM_LEAF_CURL_PHILIPPINES_VIRUS_ISOLATE_MC1
A145.	CASSAVA_MOSAIC_MADAGASCAR_VIRUS_DNA_A
A146.	EAST_AFRICAN_CASSAVA_MOSAIC_KENYA_VIRUS_DNA_A
A147.	MIRABILIS_LEAF_CURL_INDIA_VIRUS
A148.	TOBACCO_LEAF_CURL_PUSA_VIRUS_DNA-A
A149.	AGERATUM_YELLOW_VEIN_CHINA_VIRUS_-_OX1
A150.	TOMATO_LEAF_CURL_CHINA_VIRUS_-_OX2
A151.	FRENCH_BEAN_LEAF_CURL_VIRUS-KANPUR_ISOLATE_FBLCV-KANPUR _SEGMENTDNA-A,
A152.	POUZOLZIA_GOLDEN_MOSAIC_VIRUS_ISOLATE_TY01
A153.	SENECIO_YELLOW_MOSAIC_VIRUS
A154.	SIDA_YELLOW_MOSAIC_CHINA_VIRUS_-_[HAINAN_8]
A155.	SIEGESBECKIA_YELLOW_VEIN_VIRUS-[GD13]
A156.	TOMATO_LEAF_CURL_LIWA_VIRUS_ISOLATE_LW1
A157.	TYLCAXV-SIC1-[IT%3ASIC2-2%3A04]
A158.	TOMATO_LEAF_CURL_PALAMPUR_VIRUS

A159. TOMATO_LEAF_CURL_SEYCHELLES_VIRUS
A160. TOMATO_YELLOW_LEAF_CURL_AXARQUIA_VIRUS_ISOLATE_ HOMRA
A161. TOMATO_YELLOW_LEAF_CURL_YUNNAN_VIRUS_ISOLATE_YN2013_ CLONE_10SEGMENT DNA-A,
A162. TOMATO_LEAF_CURL_KERALA_VIRUS
A163. TOMATO_YELLOW_LEAF_CURL_SAUDI_VIRUS_ISOLATE_HAIL1
A164. TOMATO_LEAF_CURL_OMAN_VIRUS
A165. STACHYTARPHETA_LEAF_CURL_VIRUS
A166. MIMOSA_YELLOW_LEAF_CURL_VIRUS_DNA-A
A167. VERNONIA_YELLOW_VEIN_VIRUS_DNA-A
A168. SIDA_YELLOW_VEIN_VIETNAM_VIRUS_DNA-A
A169. BITTER_GOURD_YELLOW_VEIN_VIRUS_ISOLATE_BD12C8
A170. EAST_AFRICAN_CASSAVA_MOSAIC_ZANZIBAR_VIRUS_DNA-A
A171. LINDERNIA_ANAGALLIS_YELLOW_VEIN_VIRUS_DNA-A
A172. ERECTITES_YELLOW_MOSAIC_VIRUS_DNA-A
A173. CLERODENDRUM_GOLDEN_MOSAIC_VIRUS_DNA-A
A174. HOLLYHOCK_LEAF_CRUMPLE_VIRUS
A175. AGERATUM_LEAF_CURL_CAMEROON_VIRUS
A176. AGERATUM_YELLOW_VEIN_VIRUS
A177. BHENDI_YELLOW_VEIN_MOSAIC_VIRUS
A178. CHILLI_LEAF_CURL_VIRUS
A179. CLERODENDRON_YELLOW_MOSAIC_VIRUS
A180. COTTON_LEAF_CURL_ALABAD_VIRUS
A181. COTTON_LEAF_CURL_GEZIRA_VIRUS
A182. COTTON_LEAF_CURL_KOKHRAN_VIRUS
A183. COTTON_LEAF_CURL_MULTAN_VIRUS
A184. EAST_AFRICAN_CASSAVA_MOSAIC_CAMEROON_VIRUS_DNA_A
A185. EAST_AFRICAN_CASSAVA_MOSAIC_VIRUS_DNA_A
A186. EUPATORIUM_YELLOW_VEIN_VIRUS
A187. HONEYSUCKLE_YELLOW_VEIN_MOSAIC_VIRUS-[KAGOSHIMA]
A188. HONEYSUCKLE_YELLOW_VEIN_MOSAIC_VIRUS
A189. INDIAN_CASSAVA_MOSAIC_VIRUS_DNA_A
A190. MALVASTRUM_YELLOW_MOSAIC_VIRUS_DNA-A
A191. MALVASTRUM_YELLOW_VEIN_VIRUS
A192. OKRA_LEAF_CURL_CAMEROON_VIRUS
A193. OKRA_YELLOW_VEIN_MOSAIC_VIRUS
A194. PAPAYA_LEAF_CURL_VIRUS

A195.	PEDILANTHUS_LEAF_CURL_VIRUS-PEDILANTHUS_[PAKISTAN%3A MULTAN%3A2004]
A196.	PEPPER_LEAF_CURL_VIRUS_DNA-A
A197.	PEPPER_YELLOW_LEAF_CURL_INDONESIA_VIRUS_DNA-A
A198.	PEPPER_YELLOW_VEIN_MALI_VIRUS
A199.	PUMPKIN_YELLOW_MOSAIC_MALAYSIA_VIRUS_DNA_A
A200.	SIDA_LEAF_CURL_VIRUS
A201.	SOUTH_AFRICAN_CASSAVA_MOSAIC_VIRUS_DNA_A
A202.	SOYBEAN_CRINKLE_LEAF_VIRUS
A203.	SQUASH_LEAF_CURL_CHINA_VIRUS_-_[B]_DNA-A
A204.	SQUASH_LEAF_CURL_YUNNAN_VIRUS
A205.	SRI_LANKAN_CASSAVA_MOSAIC_VIRUS_DNA_A
A206.	TOBACCO_LEAF_CURL_JAPAN_VIRUS
A207.	TOBACCO_LEAF_CURL_THAILAND_VIRUS
A208.	TOBACCO_LEAF_CURL_YUNNAN_VIRUS_-_[Y136]
A209.	TOBACCO_LEAF_CURL_ZIMBABWE_VIRUS
A210.	TOMATO_CURLY_STUNT_VIRUS
A211.	TOMATO_LEAF_CURL_BANGALORE_VIRUS
A212.	TOMATO_LEAF_CURL_BANGLADESH_VIRUS
A213.	TOMATO_LEAF_CURL_HAINAN_VIRUS
A214.	TOMATO_LEAF_CURL_IRAN_VIRUS
A215.	TOMATO_LEAF_CURL_JAVA_VIRUS
A216.	TOMATO_LEAF_CURL_KARNATAKA_VIRUS
A217.	TOMATO_LEAF_CURL_LAOS_VIRUS
A218.	TOMATO_LEAF_CURL_MALAYSIA_VIRUS
A219.	TOMATO_LEAF_CURL_MALI_VIRUS
A220.	TOMATO_LEAF_CURL_PHILIPPINES_VIRUS
A221.	TOMATO_LEAF_CURL_PUNE_VIRUS
A222.	TOMATO_LEAF_CURL_SUDAN_VIRUS_-_[GEZIRA]
A223.	TOMATO_LEAF_CURL_TAIWAN_VIRUS
A224.	TOMATO_LEAF_CURL_VIETNAM_VIRUS_DNA_A
A225.	TOMATO_LEAF_CURL_VIRUS
A226.	TOMATO_YELLOW_LEAF_CURL_KANCHANABURI_VIRUS_DNA_A
A227.	TOMATO_YELLOW_LEAF_CURL_MALAGA_VIRUS
A228.	TOMATO_YELLOW_LEAF_CURL_SARDINIA_VIRUS
A229.	TOMATO_YELLOW_LEAF_CURL_THAILAND_VIRUS_DNA_A
A230.	TOMATO_LEAF_CURL_GHANA_VIRUS

A231. TOMATO_LEAF_CURL_GUJARAT_VIRUS_-_[VARANASI]
A232. TOMATO_LEAF_CURL_PAKISTAN_VIRUS
A233. SPILANTHES_YELLOW_VEIN_VIRUS_DNA-A
A234. PEPPER_YELLOW_LEAF_CURL_CHINA_VIRUS_ISOLATE_YN65-1
A235. RADISH_LEAF_CURL_VIRUS
A236. RAMIE_MOSAIC_VIRUS_DNA-A
A237. ROSE_LEAF_CURL_VIRUS_ISOLATE_AS24
A238. AGERATUM_YELLOW_VEIN_HUALIAN_VIRUS-[TAIWAN%3AHSINCHU%3ATOM%3A2003]_DNA_A
A239. TOBACCO_CURLY_SHOOT_VIRUS
A240. TOMATO_LEAF_CURL_ARUSHA_VIRUS_DNA-A
A241. TOMATO_LEAF_CURL_BARKA_VIRUS_ISOLATE_TOM-55
A242. TOMATO_LEAF_CURL_CEBU_VIRUS_DNA-A
A243. TOMATO_LEAF_CURL_COTABATO_VIRUS_DNA-A
A244. TOMATO_LEAF_CURL_GUANGXI_VIRUS
A245. TOMATO_LEAF_CURL_HANOI_VIRUS
A246. TOMATO_LEAF_CURL_HSINCHU_VIRUS_-_[TAIWAN%3 AHSINCHU%3A2005]_DNA_A
A247. TOMATO_LEAF_CURL_JOYDEBPUR_VIRUS_DNA-A
A248. TOMATO_LEAF_CURL_MINDANAO_VIRUS_DNA-A
A249. TOMATO_LEAF_CURL_NIGERIA_VIRUS-[NIGERIA%3A2006]
A250. TOMATO_LEAF_CURL_PATNA_VIRUS_DNA-A
A251. TOMATO_LEAF_CURL_RANCHI_VIRUS_DNA-A
A252. TOMATO_LEAF_CURL_SULAWESI_VIRUS_DNA-A
A253. TOMATO_LEAF_CURL_TOGO_VIRUS-[TOGO%3A2006]
A254. TOMATO_YELLOW_LEAF_CURL_CHINA_VIRUS
A255. TOMATO_YELLOW_LEAF_CURL_INDONESIA_VIRUS-[LEMBANG]
A256. TOMATO_LEAF_CURL_GANDHINAGAR_VIRUS_ISOLATE_PTOGNAX15
A257. TOMATO_LEAF_CURL_KUMASI_VIRUS
A258. TOMATO_YELLOW_LEAF_CURL_VIETNAM_VIRUS_DNA-A
A259. TOMATO_LEAF_CURL_SRI_LANKA_VIRUS
A260. WATERMELON_CHLOROTIC_STUNT_VIRUS_DNA_A
A261. AGERATUM_YELLOW_VEIN_SRI_LANKA_VIRUS
A262. COTTON_LEAF_CURL_BANGALORE_VIRUS
A263. COTTON_LEAF_CURL_RAJASTHAN_VIRUS
A264. OKRA_YELLOW_CRINKLE_VIRUS
A265. PEPPER_LEAF_CURL_BANGLADESH_VIRUS

A266.	VELVET_BEAN_SEVERE_MOSAIC_VIRUS_DNA_A
A267.	MUNGBEAN_YELLOW_MOSAIC_INDIA_VIRUS_DNA_A
A268.	MUNGBEAN_YELLOW_MOSAIC_VIRUS_DNA_A
A269.	SOYBEAN_CHLOROTIC_BLOTCH_VIRUS_DNA_A
A270.	TOMATO_MOTTLE_WRINKLE_VIRUS_ISOLATE_AR%3APICHANAL%3A400
A271.	OKRA_LEAF_CURL_VIRUS-[CAMEROON]
A272.	TURNIP_CURLY_TOP_VIRUS
A273.	CAPRARIA_YELLOW_SPOT_YUCATAN_VIRUS
A274.	DOLICHOS_YELLOW_MOSAIC_VIRUS_ISOLATE_DA
A275.	SWEET_POTATO_GOLDEN_VEIN_ASSOCIATED_VIRUS
A276.	SWEET_POTATO_LEAF_CURL_BENGAL_VIRUS_-_[INDIA%3AWEST_BENGAL%3A2008]SEGMENTA,
A277.	SWEET_POTATO_LEAF_CURL_CANARY_VIRUS
A278.	SWEET_POTATO_LEAF_CURL_CHINA_VIRUS_[CHINA%3ASICHUAN14%3A2012]
A279.	SWEET_POTATO_LEAF_CURL_LANZAROTE_VIRUS
A280.	SWEET_POTATO_LEAF_CURL_SAO_PAULO_VIRUS_ISOLATESPLCSPV-[BR%3AALVM%3A09],
A281.	SWEET_POTATO_LEAF_CURL_SHANGHAI_VIRUS_ISOLATE_CHINA%3AJILIN1%3A2012
A282.	SWEET_POTATO_LEAF_CURL_SOUTH_CAROLINA_VIRUS
A283.	SWEET_POTATO_LEAF_CURL_SPAIN_VIRUS
A284.	SWEET_POTATO_LEAF_CURL_VIRUS_ISOLATE_CHINA%3ASHANDONG11%3A2012
A285.	IPOMOEA_YELLOW_VEIN_VIRUS
A286.	SWEET_POTATO_LEAF_CURL_UGANDA_VIRUS-[UGANDA%3AKAMPALA%3A2008]
A287.	JATROPHA_LEAF_CURL_VIRUS_DNA_A
A288.	KUDZU_MOSAIC_VIRUS_DNA-A
A289.	ALTERNANTHERA_YELLOW_VEIN_VIRUS_DNA-A
A290.	HORSEGRAM_YELLOW_MOSAIC_VIRUS
A291.	SIEGESBECKIA_YELLOW_VEIN_GUANGXI_VIRUS
A292.	SOLANUM_MOSAIC_BOLIVIA_VIRUS
A293.	TOMATO_YELLOW_MOTTLE_VIRUS
A294.	BEET_CURLY_TOP_VIRUS_-_CALIFORNIA_[LOGAN]
A295.	CHAYOTE_YELLOW_MOSAIC_VIRUS
A296.	SIDA_YELLOW_VEIN_MADURAI_VIRUS
A297.	SWEET_POTATO_LEAF_CURL_CHINA_HENAN_VIRUS

A298. SWEET_POTATO_LEAF_CURL_GEORGIA_VIRUS
A299. SWEET_POTATO_LEAF_CURL_VIRUS
A300. TOMATO_PSEUDO-CURLY_TOP_VIRUS
A301. SIDA_MICRANTHA_MOSAIC_VIRUS
A302. SOYBEAN_MILD_MOTTLE_VIRUS
A303. SWEET_POTATO_LEAF_CURL_GUANGXI_VIRUS_ISOLATE_CHINA%3AGUANGXI5%3A2011
A304. SWEET_POTATO_LEAF_CURL_HENAN_VIRUS_ISOLATE_CHINA%3AHENAN10(2)%3A2012
A305. BEAN_GOLDEN_YELLOW_MOSAIC_VIRUS_DNA_A
A306. SIDA_MOTTLE_VIRUS
A307. SIDA_YELLOW_MOSAIC_VIRUS
A308. TOMATO_LEAF_DEFORMATION_VIRUS_ISOLATE_EA-LE3-5K
A309. LAUSANNEVIRUS
A310. MELBOURNEVIRUS_ISOLATE_1
A311. MARSEILLEVIRUS_MARSEILLEVIRUS_STRAIN_T19
A312. ACANTHAMOEBA_POLYPHAGA_MIMIVIRUS
A313. ACANTHAMOEBA_POLYPHAGA_MOUMOUVIRUS
A314. AEROMONAS_PHAGE_44RR2.8T
A315. AEROMONAS_PHAGE_PHIAS4
A316. ALTEROMONAS_PHAGE_VB_AMAP_AD45-P1
A317. ANOMALA_CUPREA_ENTOMOPOXVIRUS_DNA
A318. UNVERIFIED%3A_ANOPHELES_MINIMUS_IRODOVIRUS_ISOLATE_AMIV
A319. CAFETERIA_ROENBERGENSIS_VIRUS_BV-PW1
A320. CAMPYLOBACTER_PHAGE_CP21
A321. CRONOBACTER_PHAGE_VB_CSAM_GAP32
A322. DICKEYA_PHAGE_VB_DSOM_LIMESTONE1
A323. EDWARDSIELLA_PHAGE_PEI21
A324. ENTEROBACTERIA_PHAGE_EPS7
A325. ENTEROBACTERIA_PHAGE_T5
A326. ENTEROBACTERIA_PHAGE_VB_ECOM-VR7
A327. ESCHERICHIA_PHAGE_BV_ECOS_AKFV33
A328. ESCHERICHIA_PHAGE_PHAXI
A329. ESCHERICHIA_PHAGE_VB_ECOS_FFH1
A330. KLEBSIELLA_PHAGE_JD001
A331. ENTEROBACTERIA_PHAGE_VB_KLEM-RAK2
A332. MICROMONAS_SP._RCC1109_VIRUS_MPV1

A333. MYCOBACTERIUM_PHAGE_LLIJ
A334. MYXOCOCCUS_PHAGE_MX8
A335. OSTREOCOCCUS_LUCIMARINUS_VIRUS_OLV1
A336. RHODOCOCCUS_PHAGE_REQ2
A337. SALMONELLA_PHAGE_C341
A338. SALMONELLA_PHAGE_EPSILON34
A339. SALMONELLA_PHAGE_PVP-SE1
A340. SERRATIA_PHAGE_PS2
A341. SPODOPTERA_FRUGIPERDA_ASCOVIRUS_1A
A342. VIBRIO_PHAGE_ICP1
A343. VIBRIO_PHAGE_PVP-1
A344. YERSINIA_PHAGE_PHIR201
A345. MEGAVIRUS_LBA_ISOLATE_LBA111
A346. ARMADILLIDIUM_VULGARE_IRIDESCENT_VIRUS
A347. ENTEROBACTERIA_PHAGE_HK106
A348. ESCHERICHIA_PHAGE_121Q
A349. HELIOTHIS_VIRESCENS_ASCOVIRUS_3E
A350. INVERTEBRATE_IRIDESCENT_VIRUS_30
A351. INVERTEBRATE_IRIDESCENT_VIRUS_6
A352. LYMANTRIA_DISPAR_MNPV
A353. MEGAVIRUS_CHILIENSIS
A354. MYCOBACTERIUM_PHAGE_AVANI
A355. MYCOBACTERIUM_PHAGE_BOBI
A356. MYCOBACTERIUM_PHAGE_JABBAWOKKIE
A357. MYCOBACTERIUM_PHAGE_SG4
A358. PECTOBACTERIUM_PHAGE_MY1
A359. SHEWANELLA_SP._PHAGE_1-4
A360. SULFITOBACTER_PHAGE_PCB2047-A
A361. SYNECHOCOCCUS_PHAGE_S-CAM8_STRAIN_S-CAM8_06008BI06
A362. MYCOBACTERIUM_PHAGE_WIVSMALL
A363. PARAMECIUM_BURSARIA_CHLORELLA_VIRUS_FR483
A364. MYCOBACTERIUM_PHAGE_OMEGA
A365. MYCOBACTERIUM_PHAGE_PMC
A366. MYCOBACTERIUM_PHAGE_TWEETY
A367. MYCOBACTERIUM_PHAGE_ARDMORE
A368. MYCOBACTERIUM_PHAGE_BOOMER
A369. MYCOBACTERIUM_PHAGE_DEADP

A370. MYCOBACTERIUM_PHAGE_FRUITLOOP
A371. MYCOBACTERIUM_PHAGE_GUMBIE
A372. MYCOBACTERIUM_PHAGE_HAMULUS
A373. MYCOBACTERIUM_PHAGE_REDNO2
A374. MYCOBACTERIUM_PHAGE_THIBAULT
A375. MYCOBACTERIUM_PHAGE_WANDA
A376. MYCOBACTERIUM_PHAGE_WEE
A377. PITHOVIRUS_SIBERICUM_ISOLATE_P1084-T
A378. CLOSTRIDIUM_PHAGE_C-ST
A379. ACIDIANUS_ROD-SHAPED_VIRUS_1
A380. SALMONELLA_PHAGE_SSU5