STUDIES IN VIRAL POPULATION GENETICS AND BIOINFORMATICS

By

KSHITIJ WAGH

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Physics and Astronomy

written under the direction of

Gyan Bhanot

and approved by

_____

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2013

ABSTRACT OF THE DISSERTATION

Studies in Viral Population Genetics and Bioinformatics

By KSHITIJ WAGH

Dissertation Director:

Gyan Bhanot

This thesis consists of two studies pertaining to the evolution and genomic signatures of viruses. Viruses are obligate intracellular parasites that have a great impact on human, animal and plant health. The first study involves the human infecting Influenza A H5N1 viruses. H5N1 is an avian virus which occasionally infects humans, with a 50-60% mortality rate. Human-to-human transmission is limited, and most H5N1 infections are transmitted to humans from birds. Under such a transmission scheme, there can be a possibility of a biased transmission of H5N1 strains from birds to humans. Such a biased transmission could arise due to higher efficiency of some avian strains in infecting humans, an enhanced ability of the human immune response to clear some of the human-infecting avian strains, etc. We developed a novel strategy to identify such signatures and analyzed publicly available H5N1 hemagglutinin sequences from China, Egypt, and Indonesia. In each geographic region, it was found that human infecting strains arose from a subset of the avian viral pool characterized

by geography specific mutations. These mutations lie in functionally important regions of hemagglutinin proteins involved in viral attachment to cells, immune response etc. After correcting for this transmission bias, an absence of further widespread bias was observed. This research also showed that vaccine evasion mutant viruses are unlikely to infect humans, a finding with significant implications for rational vaccine design.

As a separate project, we developed a new method to detect novel capsid sequences. It is expected that a large part of the virosphere still remains uncharacterized. Viruses show remarkably high levels of sequence diversity. Hence, sequence similarity based methods have limited success in detection of novel viral sequences in metagenomic studies. However, in contrast to high sequence diversity, the capsid proteins from diverse families of icosahedral viruses show a conserved eight stranded beta barrel known as the "Jelly-roll" fold. Motivated by this structural conservation, we sought to classify such capsid protein sequences using a machine learning approach on alignment free features. The nature of the alignment free features suitable for the problem are first discussed. Using these alignment free features, a high-accuracy Support Vector Machine (SVM-Caps) was developed for classifying jelly-roll capsid proteins against other proteins. The predictive power of this classifier was compared to that of BLAST, a popular tool based on sequence similarity. SVM-Caps was found to have comparable but lower power to detect capsid sequences of known viral families, but significantly higher power in detection of capsid sequences from novel families. As an application of this method, the viral metagenomic data from the French Lake Bourget study were analyzed and many potential novel capsid sequences were found.

Acknowledgments

I would first like to thank my advisor, Gyan Bhanot. His encouragement and guidance have made my transition from theoretical physics to evolutionary biology a smooth and productive one. I found working with him very rewarding especially because of his focus on the clinical and biological relevance of our theoretical projects. I am also grateful to him for fostering a student-advisor relationship where I always felt comfortable to speak out my mind in discussions. His propensity for hard work and his drive to produce quality publications have been inspirational to me.

I would also like to thank Vlad Belyi, who over the past year, has been like a second adviser to me. In this short time I have learnt a lot from our interactions, which have introduced me to a plethora of topics that I was not familiar with. I was impressed by his clarity of thought and rigour for validation, which I hope to have imbibed in my research. I thank Vlad for also providing funding my research for the past year, when I needed it the most. I am also grateful to Natan Andrei with whom I worked on, what I feel is, one of the most interesting problems in theoretical condensed matter physics. His insights in non-equilibrium phenomena and strongly correlated systems, coupled with his encouragement for working on really hard but fundamental problems made my experience quite rewarding. I am also grateful to Emil Yuzbashyan for his guidance in my first research project at Rutgers. I am also thankful to my collaborators on various projects – Gabriela Alexe (for introducing me to bioinformatics), Aatish Bhatia, Sridhar Ganesan, Ben Greenbaum, Yao Ming, Haile Owusu and Vijay Ravikumar.

I am very thankful to my committee for being supportive of my research. Alex

Morozov, Anirvan Sengupta and Frank Zimmerman have been great in asking pertinent questions that have shaped both my research and presentation skills. Anirvan was also helpful in advising me on job opportunities. I am thankful to Siobain Duffy for agreeing to be on my committee and taking a keen interest in my research. It was great to get her perspectives, as a virologist, on my research projects.

My stay at Rutgers was amazing due to the comraderie and support of my awesome friends. I had a great time hanging out with Adina, Chioun, Chuck, Daniel, George, John, Lizzie, Manjul, Patrick, Roberto, Senia and Sinisa. I am especially thankful Aatish, Anindya, Darakhshan, Deepak, Purba, Sushmita and Vijay for providing a close, caring friendship and many memorable experiences. This dissertation could not have been possible without the support of my housemate Nida, whose inspiration, wisdom, encouragement and warm friendship through the tough final year helped me to a successful completion.

Finally, I continue to remain indebted to my parents for their love and support. They have always put my interests above everything else and have spared no effort in providing me a wholesome atmosphere to grow. Both of them are strong-willed, self-made people who made the most of the educational opportunities, which they often had to struggle for availing. Shaped by this experience, their thoughts have instilled in me a deep importance for education. This thesis is a culmination of their efforts, which I hope will please them. I am also grateful to my aunt Choti, who is like a second mother to me, and to my late grandmother Aai, who was the most perfect doting grandmother a kid could have hoped for. My sister Nupur's love and support has also been crucial to make me feel at home away from home.

Dedication

For my parents for their infinite love and support

&

Dr. Babasaheb Ambedkar whose inspirational leadership continues to lead Dalits out of the darkness of untouchability with the light of education

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

In this thesis, I discuss two projects on population genetics and bioinformatics of viruses. Viruses are obligate parasites that need a host to replicate and synthesize their proteins [11]. Viruses infect hosts from all domains of life (archea, bacteria and eukaryotes) [12], and many viruses are prominent human, animal and plant pathogens. In humans, some of the most notorious and damaging diseases, such as AIDS [13, 14], Influenza [15], some types of cancer [16], smallpox [17], polio[18], rabies [19] and Hepatitis [20] are of viral origin. Intense effort to understand their biology has led to remarkable advancements in molecular biology, immunology, and public health.

In this chapter, I first review the important facts about viruses. This will help set the stage for the research discussed in Chapter 3 on detecting novel capsid sequences from a diverse set of known and unknown viral families. I also review viral metagenomic sequencing, an emerging field which is revolutionizing our understanding of viral diversity and functions, and discuss the novel bioinformatics challenges presented by such analyses. Next, I focus on Influenza viruses and review key research relevant to our work on H5N1 Influenza A viruses, which is discussed in Chapter 2.

1.1 Diversity of the Viral Universe

Since their discovery in late 19^{th} century, intense research has led to an understanding of the life-cycles, structures, and pathogenicity of many diverse viruses infecting diverse hosts from all domains of life. Many viruses can now be cultured in the labs, and $\sim 5,000$ species of viruses are known. Through these studies, it is now clear that

viruses are an extremely diverse group of biological entities. This diversity stems from many different aspects. As viruses depend on their hosts for replication and translation of proteins, they have evolved different strategies to infect these diverse hosts. Even within the viruses infecting same types of hosts, there can be substantial differences in viral infection strategies due to the nature of their surface proteins (which governs the cellular channels they use to enter cells), their structures (which determine where in the cell they are located) and their genome-type and genomic content (which determine the host cellular machinery they need to interact with). I focus on the latter two aspects below.

1.1.1 Viral Genomes and Genes

Whereas the genomes of cellular organisms are composed of DNA, the genomes of viruses can be either DNA or RNA [11]. Genomes of DNA viruses can be further classified into single stranded DNA (ssDNA) or double stranded DNA (dsDNA) (all cellular organisms have double-stranded DNA genomes). Moreover, the viral DNA genomes can either be circular or linear. RNA genomes of viruses can be single (ssRNA) or double stranded (dsRNA) and are mostly linear. Single stranded RNA viral genomes can be further classified into positive or negative depending on whether the viral messenger RNA (which can be translated into proteins) is derived from the genome or its complement. Some viruses are ambisense and their proteins can be translated on both the RNA strand derived from the genome as well as its complement. RNA and DNA viruses can also have segmented genomes (similar to chromosomes in humans). Most of the viral genomes are haploid, i.e. they possess one homologous copy of each segment. In contrast, some viruses such as HIV are diploid and have two copies of each segment [11]. Each of these classes of viruses has distinct replication strategies: for DNA viruses, replication is DNA $\rightarrow$ DNA; for RNA viruses replication is RNA$\rightarrow$RNA. However, there also exist viruses whose replication strategy is either DNA $\rightarrow$ RNA $\rightarrow$ DNA or RNA$\rightarrow$ DNA $\rightarrow$ RNA. These viruses use the

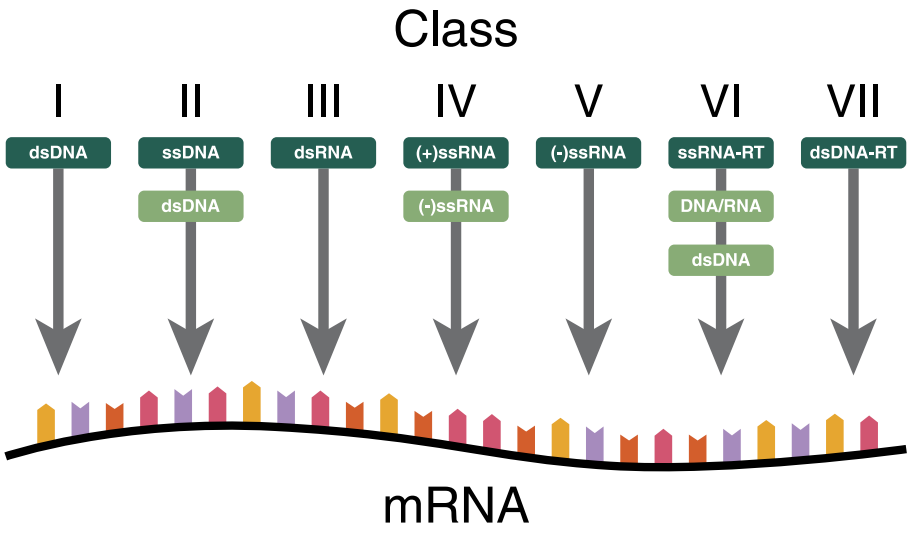

Figure 1.1: Virus classification based on genome type introduced by Baltimore [2]. The image was made available by Thomas Splettstoesser via the Creative Commons license.

enzyme Reverse Transcriptase (RT) to convert DNA to RNA, and form a separate class called RT utilizing DNA/RNA viruses. The classification of viruses based on genome types, messenger RNA production strategies and replication strategies was introduced by Baltimore [2] (Fig. 1.1), and is also acknowledged by the International Committee on Taxonomy of Viruses (ICTV) [21]. ICTV further classifies these classes into Orders, Families, Subfamilies, Genera, Species in decreasing order of hierarchy [21].

Viral genome sizes exhibit a huge range from a mere $1,680$ bases long for Deltavirus to $\sim 2,500,000$ base pairs long for Pandoraviruses [22]. The number of genes also show a similar variation from 1 to $\sim 2,500$ (Table 1.1). Such diversity in the gene content of viruses highlights the diverse nature of the viral life-cycles, wherein different proteins are utilized for different functions. Most of the viral proteins are family-specific, and few are homologous in higher taxonomic units. There are no genes which are universally present in all viruses. This is in contrast to cellular organisms, which possess several common genes such as the conserved ribosomal RNA and proteins [23, 24], etc. Notably, viruses do not possess ribosomes, the essential components of

cells which translate proteins from messenger RNA. Although there are few capsid-less viral families known [25], the most prevalent gene in viruses encodes the capsid protein, which is used to build protective shells around viral genomes. In spite of the prevalence of capsid proteins, there is a great diversity in their sequences and protein structures (see below for conservation of capsid structures). Other than this gene, there are also some other notable genes, which are conserved in many diverse viral families, but not in all. Such genes were called "viral hallmark genes" by Koonin et al. [12], and include Superfamily 3 Helicase (involved in DNA/RNA strand separation), Replicase (involved in replication of DNA), RNA dependent RNA Polymerase (involved in replication of RNA genomes), etc. However, apart from these few cases, there is considerable variability in the prevalence of viral genes, reflecting the need for specific host-dependent functions in different types of viruses.

Table 1.1: Properties of Viral Genomes (from ViralZone [1])

Class	Genome Size (kb)	Segments	Genes
dsDNA	4.5-2,500	1-105	5-2556
ssDNA	1.8-12.5	1-8	2-16
dsRNA	3.7-30.5	1-12	2-14
ssRNA(+)	2.3-31	1-5	1-15
ssRNA(-)	1.7-25.2	1-8	1-12
dsDNA-RT	3.0-8.3	1	3-8
ssRNA-RT	5.1-11.0	1	8

1.1.2 Viral Capsid Structures

Most viruses possess some sort of proteinaecious shells around their genomes and other contents [11]. These capsid shells are built out of multiple copies of a few proteins which self-assemble to form, in most cases, symmetric capsids. They can be broadly classified into two symmetry classes: a) helical and b) icosahedral (Fig. 1.2). Some viral capsids do not fall into either of these symmetry classes and are irregularly shaped, e.g. conical shaped HIV capsids, brick-shaped poxvirus capsids, bottle and droplet shaped archaeal virus capsids [26] etc. Nonetheless, all the exceptional shaped capsids are still built from multiple copies of few distinct proteins. Viruses of bacteria, called bacteriophages, often have an icosahedral or elongated icosahedral capsid attached to a helical tail with tail fibers. The majority of the viruses have icosahedral capsids, whose structures are further characterized by the number of building units they possess (a system due to Caspar and Klug [27]). Some viruses possess a lipid membrane outside the capsids, in which case they are referred to as enveloped. In case of enveloped viruses, the lipid layer is often obtained from the host cell when the virus exits the host cell membrane.

Since the principal contents of viral capsids are the viral genomes, the size of capsids can vary as dramatically as the size of the viral genomes. Capsid sizes can range from 20 nm for Circoviridae to $\sim 1\mu m$ for giant complex-shaped viruses such as the Pandoraviruses and long viruses like Filoviruses.

1.1.3 Conservation of the "Jelly-Roll Fold" in Capsid Subunit Proteins

Many viruses have icosahedral capsids and it was beleived that this common symmetry emerged from the constraint of building a symmetric shell with identical subunits. Indeed, the absence of sequence similarity of the subunit capsid proteins supported the notion of multiple routes in the evolution of icosahedral capsids. However, high-resolution structural studies since the late 1970s onwards began suggesting otherwise.

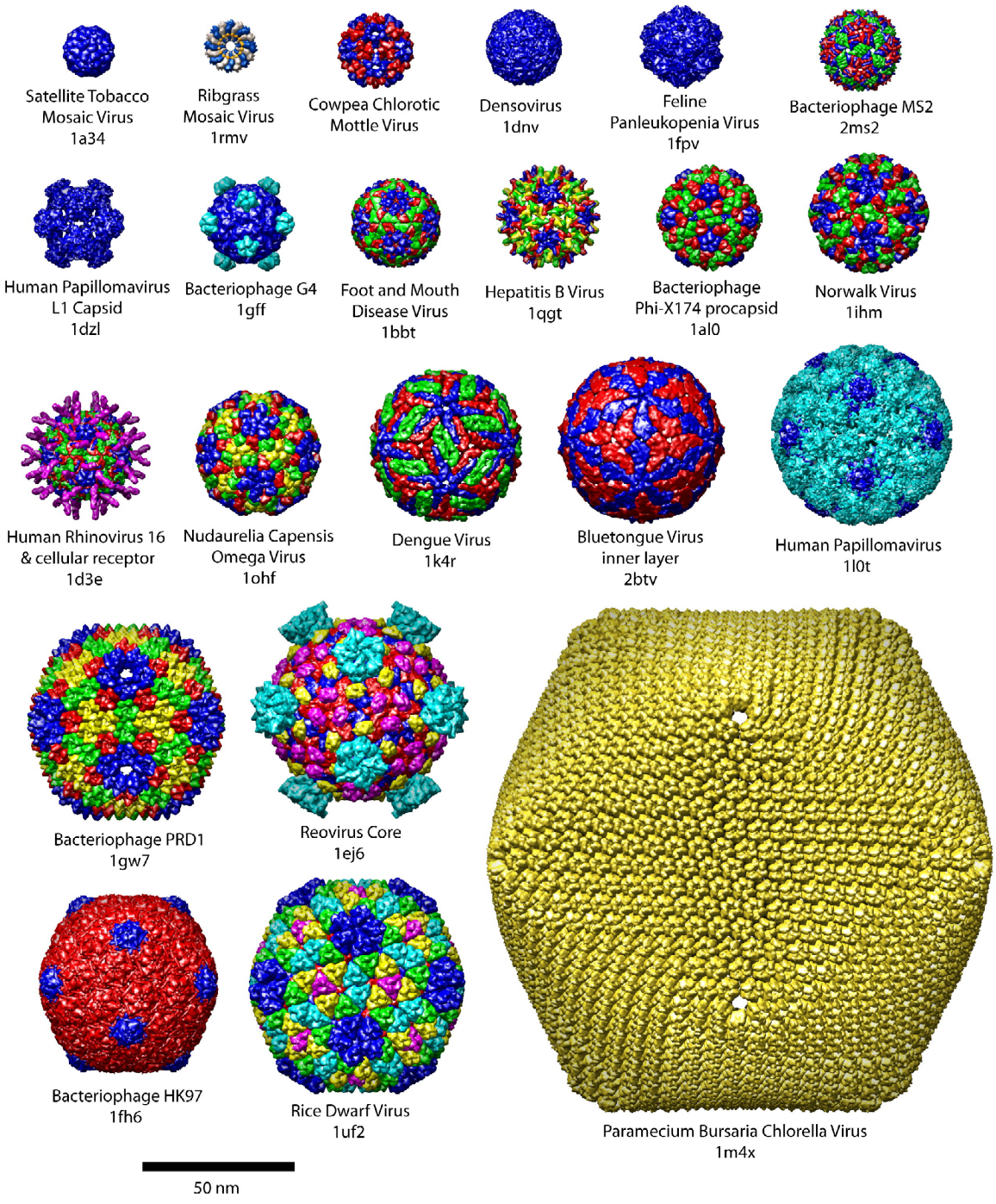

Figure 1.2: Viral Capsid structures. Except the helical capsid of Ribgrass Mosaic Virus (second in top row, transverse section shown), all other capsids have icosahedral symmetry. Figure reproduced with permission from Goddard et al. [3]. The codes below the names of viruses refer to the structure accession codes in PDB.

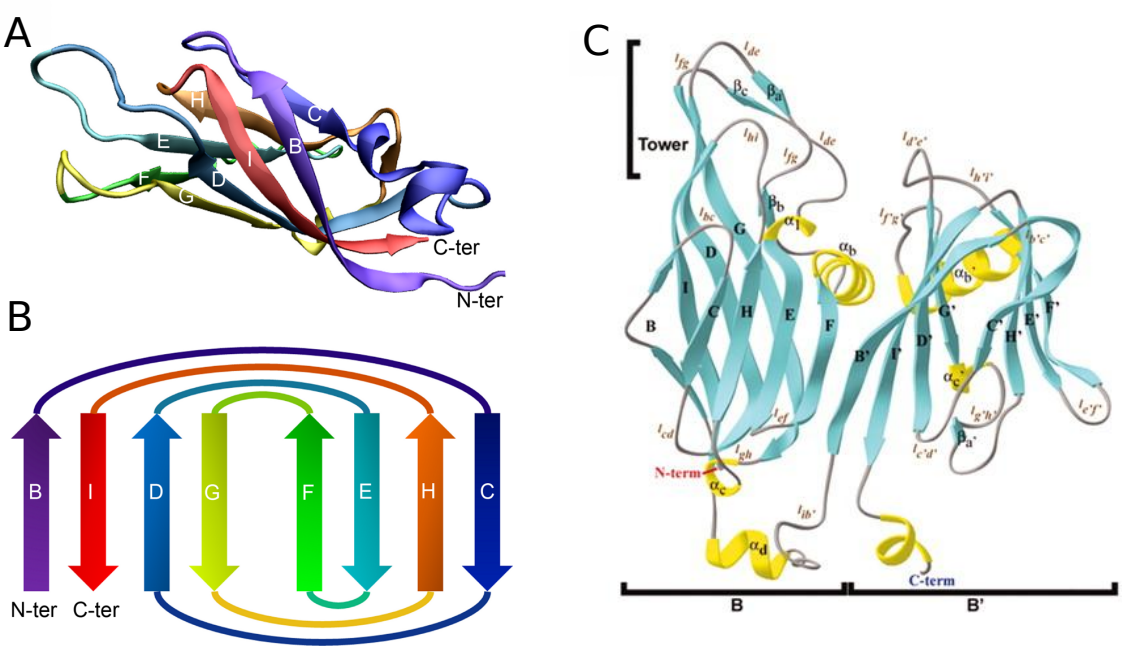

Figure 1.3: Single and Double Jelly-Roll containing Capsid Proteins. The canonical jelly-roll fold possessing 8 β strands (B-I) is shown in (A), with the fold structure schematically shown in (B). The capsid structure of STIV showing two jelly-roll folds is shown in (C). Sub-figures A and B are reproduced with permission from Cheng & Brooks [4], and (C) from Khayat et al. [5].

Initially, a plant and a human virus with icosahedral capsid were shown to have a similar fold in their capsid subunit structures – an eight beta-stranded "jelly-roll fold" (Fig. 1.3 A, B) [6]. This same fold was also later observed for the capsid subunit of an insect virus. As a variant on this theme, capsid subunits possessing two such folds, perhaps due to duplication, were found to be shared between viruses infecting different domains of life – the human infecting adenovirus, and the bacteriophage PRD1. The work of Khayat et al. [5] then extended this conservation to viruses infecting the third domain of life. They showed that the Sulfolobus Turreted Icosahedral Virus (STIV), a virus infecting the archaea species Sulofolobus, also contained the PRD1 like fold (Fig. 1.3 C). Currently, there are more than 20 families of viruses and many types of unclassified of RNA and DNA viruses known to carry either one or two copies of this conserved fold (Table 3.1) [6, 28, 29].

The evolutionary history and antiquity of viruses has always been disputed. Since

some viruses still use RNA genomes, one of the theories of origins of viruses proposes that viruses are relics of a pre-cellular era, the so-called RNA world [12]. Consistent with this theory, the structural conservation of the jelly-roll fold between viruses infecting all three domains of life suggests a common ancestry of these viruses, dating to before the separation of the evolutionary lineages of eukaryotes, bacteria and archaea, more than 3 billion years ago [6]. The structural conservation of the jelly-roll fold has motivated detailed research into conservation of structural motifs in capsid subunit proteins of other viruses. This has resulted in a phylogeny of viruses based on structural similarities in capsid proteins, and has identified 4 main lineages: a) Picorna-like – capsids containing one copy of jelly-roll fold, b) PRD1-like – capsids containing two copies of jelly-roll folds, c) HK97-like – this lineage includes tailed-bacteriophages (possessing icosahedral heads) and herpesviruses, and d) BTV-like – these are exclusively dsRNA viruses of eukaryotes and bacteria [6] (Fig. 1.4). In Chapter 3, I focus on developing a method to detect novel capsid sequences from the first two jelly-roll fold containing lineages.

1.2 Viral Metagenomics

1.2.1 Overview

Typically, viruses have been studied by isolating and culturing them in labs [11]. This involves first culturing their hosts, and then innoculating them with virus-containing solution to get plaques. Such a procedure is not amenable to large scale studies of viral diversity. In 1998, a new way of studying the genomic content of microbes in environmental samples was introduced by Handelsman et al. [30]. This novel method, termed "metagenomics", involved isolation and sequencing of the genomic content of organisms from environmental samples directly without the need to culture them. For viruses specifically, this method was first developed and applied to isolate viral DNA from marine water samples by Breitbart et al. [31] in 2002. Their method, which in

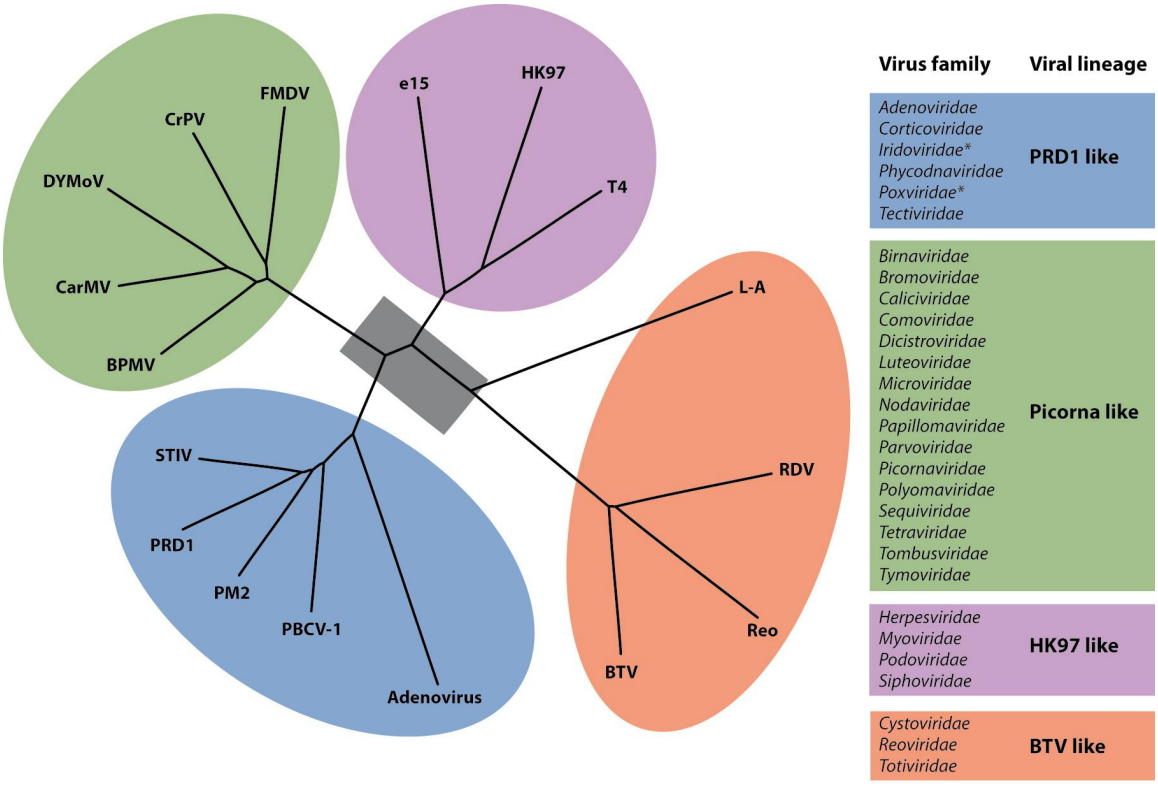

Figure 1.4: Capsid Structure Based Phylogeny of Viruses. Figure is reproduced with permission from Abrescia et al. [6]

essence has remained the same in more recent studies as well, involved filtration of virus-sized particles, extraction and amplification of the DNA from these virus-like particles, and sequencing of the amplified sequences using sequencing technologies. These studies found a huge diversity of viruses: 400-7000 different types of viruses. The majority of the viral genomic sequences identified could not be mapped to known viruses, indicating that most of the viruses sequenced were novel. In the last decade a variety of environments such as marine sediment, potable water, hot springs, stromatolites, human gut, infant feces, other animal tissues etc. have been analyzed to understand their viral content and diversity (see [32–34] for reviews) and many new viruses have been characterized. Consistent with the earliest studies, a huge diversity of viruses is found in most environmental samples, and a high fraction of this diversity is thought to originate from novel, uncharacterized viruses.

Apart from uncovering viral diversity in different environments, viral metagenomic studies have also provided novel insights into human disease and health [34]. The viromes of the human gut, saliva, respiratory and oropharyngaeal tracts have been characterized. It was found that the most numerous viruses were bacteriophages with the number of bacteriophages in the human body estimated at $10^{13} - 10^{15}$ (compared to 10^{13} human cells, and 10^{14} microbes in the human body) [33]. These bacteriophages are thought to play important roles in regulating the human bacterial microbiome as well as in the transfer of bacterial virulence genes. A number of novel viruses implicated in human diseases have also been discovered through metagenomics studies, such as a novel arenavirus responsible for fatal transplant-associated disease [35], novel bocaviruses and picornaviruses from stool samples of children with non-polio acute flaccid paralysis [36], etc. Furthermore, the viral diversity of herpesviruses and retroviruses were shown to be significantly different in the respiratory tracts of individuals with and without cystic fibrosis [37]. Thus, viral metagenomics provides not only a way to discover novel potential etiological agents for human diseases, but can also be used to understand the changes in viral communities associated with

different diseases. These studies are helping shed light on potentially novel processes present in some human diseases.

1.2.2 Bioinformatics Challenges

Depending on the sequencing platform used, a typical viral metagenome project produces tens of megabases of data, in the form of short reads with average read-sizes of $100 - 500$ base pairs. These massive number of sequence reads are then subjected to bioinformatics analysis to further study the sequence properties. These analyses fall into four classes: a) pre-processing and filtering, b) characterization of viral species and gene of origin, c) assembly of reads into contigs, and d) abundance analysis (for a review see [34]). Each of these analyses are confronted with several challenges and confounding factors. The first class of analysis aims to ensure that only high-quality reads are retained and that there is no contamination in the dataset due to non-viral genomic sequences. The latter issue is complicated by the fact that some reads may be similar to non-viral organisms, but the similarity could be to inserted viral sequences in their genomes (e.g. prophages in bacteria). The second class is one of the most challenging aspects in the bioinformatics analysis of viral metagenomics reads, which I discuss in depth below. In the third class of analyses, short reads are assembled to form longer contigs, and possibly full viral segments. This can be challenging due to the existence of conserved sequence motifs between viruses, which could lead to assemblies of chimeric contigs. As a consequence, very stringent criteria are imposed on sequence overlap between reads for assembly. This in turn, depending on read coverage, can lead to incomplete viral segments. The fourth class of analyses involves characterizing the metagenome set in terms of the abundance of certain taxa or gene functions. This can be complicated by the uneven representation in terms of read coverage of the viral content of the metagenomic sample. Nonetheless, several publicly available bioninformatics tools and pipelines overcome these challenges in analyzing raw viral metagenomic read data [34].

The second class of analyses mentioned above involves the mapping of sequence reads to known viruses and genes. The most popular approach for this step uses sequence similarity of reads to the database of known organisms. The preferred tool used for this approach is tBLASTx [38], which compares the similarity of a translated DNA sequence (in all possible frames) to a database of proteins. Because, viral sequences in metagenomic datasets can be very different from the known viruses, use of translated reads, instead of actual reads, is preferred so as to eliminate the impact of silent mutations on sequence comparison. One of the main issues with such sequence similarity based methods, which arise in the context of viral metagenomics, is that a large fraction $(50 - 90\%)$ of the reads cannot be assigned to any known organism [32]. The reasons for this large discrepancy are believed to be the poor representation of actual viral diversity in our databases, the high diversity of viruses in the samples studied and the high rates of evolution of viruses.

To circumvent the problem of using sequence similarity to identify novel genomic sequences, several approaches based on alignment-free features have been studied (see [39] for a review of alignment-free features). These methods rely on species-specific signatures in the frequency of certain short sequence motifs in their genomes [40], which can be preserved even in the absence of evident sequence similarity of novel sequences to known database sequences. Examples of such methods are TETRA [41], PhyloPythia[42], Phymm [43], MGTAXA [44], etc. All these methods employ some form of machine learning algorithms (Support Vector Machines or Hidden Markov Models) to learn the signatures of different taxonomic groups in the alignment-free features used. Such algorithms are then used to predict the taxonomic classification of the sequence. Except MGTAXA, all of these methods have been developed for bacterial metagenomics, where a similar problem of large number of unassignable reads exists [40]. MGTAXA has been trained on viral genomes, and is currently the only available program using alignment-free features for identification of novel viruses [34]. Thus, our work in Chapter 3 on detection of novel capsid sequences

using alignment-free features is a crucial new addition.

1.3 Influenza Viruses

Influenza viruses are some of the most prevalent human and animal pathogens. They infect various species of mammals and birds [45], and in humans, cause millions of cases of "flu" each year with hundreds of thousands of fatalities [46]. They have also caused several pandemics resulting in huge losses of human lives. One such widely believed emerging pandemic threat is the Influenza A subtype H5N1 [47], the population genetics of which I discuss in Chapter 2.

1.3.1 Overview

Influenza viruses are single-stranded (negative sense) RNA viruses of the *Orthomyxoviridae* family. Their capsids are enveloped with the surface glycoproteins hemagglutinin (HA) and neuraminidase (NA) embedded in the lipid envelope. Their genomes are segmented into 8 segments carrying a total of 10 genes, although some strains have been shown to have some additional genes [48]. There are three genera of Influenza viruses, A-C (see [45, 49] for reviews). Of these three genera, the most commonly human-infecting viruses are from the genera A and B. Influenza B viruses are endemic to humans, and seldom cause significant disease in humans likely due to coevolution with humans for a long time [50]. In contrast, Influenza A viruses infect a variety of bird and mammal hosts, with wild aquatic birds being their natural reservoir [45]. They have higher mutation rates as compared to the genus B viruses, which enable them to cause seasonal epidemics inspite of vaccination as well as the occasional pandemics (discussed below).

The genus A is further classified into different subtypes based on the antigenic properties of the two surface glycoproteins HA and NA. There are currently 16 HA

(numbered H1-H16) and 9 NA (numbered N1-N9) subtypes are known. Human infections are predominantly caused by the subtypes H1N1 and H3N2 currently, and these are the viruses which cause the seasonal epidemics [50]. Within each subtype, there can be considerable genetic variety from year to year, requiring the development of new vaccines every one or two years [50]. The prime drivers of the evolutionary diversity of Influenza A viruses are point mutations and reassortment. Whereas the former refer to nucleotide mutations in the genomic sequence, the latter occurs when segments from different sub-types are packaged into the same virus particle during co-infection of a host cell by different subtypes. The high mutation rates of influenza viruses are caused by the error-prone RNA dependent RNA polymerase. This leads to mutations rates of around 10^{-3} amino acid/year, which are $\sim 10^6$ times those of humans. Due to the shuffling of viral segments, reassorted viruses can have substantially different antigenic properties and can cause pandemics in situations when the human immune system cannot effectively control such novel viruses [49].

1.3.2 Emergence of Novel Strains and Pandemics

Influenza viruses have caused three pandemics in the 20^{th} century [51], and one in the 21^{st} century [52]. These pandemics are the 1918 "Spanish Flu" (H1N1), the 1957 "Asian Flu" (H2N2), the 1968 "Hong Kong Flu" (H3N2), and the 2009 "Swine Flu" (H1N1, but different than the seasonal variety). These pandemics have cumulatively caused millions of human deaths worldwide. The Spanish Flu alone was reponsible for $50 - 100$ million deaths. Apart from these well-characterized pandemics, there is also evidence of ten influenza pandemics in historical documents since the late 16^{th} century, which have occured at 10-70 year intervals [53].

The strains which have caused the last four pandemics are now characterized. Unlike the strains which cause seasonal epidemics, currently the H3N2 and H1N1 subtypes, these pandemic causing strains were novel human-infecting strains where some of the eight segments originated from viral subtypes that infect other animals.

The 1918 H1N1 pandemic strain was thought to be a purely avian influenza virus (all eight segments from avian infecting viruses) and was introduced into humans just before the onset of the pandemic [54]. The 1957 H2N2 pandemic was due to a reassortant of an avian H2N2 virus with the descendants of the H1N1 human viruses from the 1918 pandemic. The 1968 H3N2 pandemic was due to a further reassortant between the 1957 H2N2 and an avian H3 (N subtype unknown) virus [47]. Finally, the 2009 H1N1 was a triple reassortant between a human H3N2, a swine H1N1 and an avian H1N1 virus [55]. It is believed that such introductions of influenza viruses with HA and NA of viruses infecting other species cause pandemics due to the lack of successful neutralization by human immune response of these antigenically novel influenza viruses [45]. The internal genes in such pandemic viruses are often from human infecting viruses, since the mechanisms involved in Influenza replication are different in humans from other hosts such as swine and birds.

Besides the few subtypes which currently infect humans, a large number of Influenza A subtypes circulate in wild birds, and it is possible that any of these through reassortment or mutation could give rise to a pandemic [15]. Indeed, there are already reports of human infections involving influenza subtypes such as H5N1, H7N7, H9N2 and most recently H7N9, which had so far not been known to infect humans.

1.3.3 H5N1 Influenza Viruses

Among the above mentioned novel Influenza A subtypes to infect humans, the H5N1 viruses have had the most number of human infections. H5N1 is an avian Influenza A virus, which occassionally infects humans. Similar to the 1918 H1N1 pandemic strain, all the segments of H5N1 are from avian infecting influenza viruses. There are now hundreds of reported human cases with a high mortality rate of around 60% [56]. The pandemic potential of H5N1 is currently limited due to the lack of human-to-human transmission and almost all the human infections arise from contact with infected birds [47] (see [57] for a probable human-to-human transmission chain).

Avian influenza viruses typically prefer attachment to particular cellular receptors that in humans are present in the lower respiratory tract. In contrast, human influenza viruses prefer attachment to cellular receptors in the upper respiratory tract. Since humans mainly transmit influenza viruses through airborne water droplets coughed or sneezed out, this receptor specificity of H5N1 is now believed to be the main reason behind the lack of human-to-human transmission [58]. However it was recently shown in experimental evolution studies [59, 60], that transmission between mammals can be achieved by 4-5 mutations in the genes HA and PB2 (RNA polymerase subunit). Thus, H5N1 overcoming the barriers to human-to-human transmission and giving rise to a pandemic still remains a looming possibility [61].

H5N1 viruses have a complicated evolutionary history. The precursor to current H5N1 viruses originated in migratory wildfowl, which then spread to domestic birds and poultry and subsequently diversified to produce different genotypes [48, 62]. The first outbreak of avian cases was in domestic geese in Guandong, China in 1996. The first outbreak of human cases was in Hong Kong in 1997, which was caused by reassorted viruses between H5N1, H9N2 and H6N1 subtype viruses from China. Although this lineage of Hong Kong H5N1 viruses was eliminated through culling of poultry, the H5N1 lineages in China continued to diversify through reassortment and to spread to many domestic and wild avian species, with domestic geese as their reservoir. The major genotype to emerge in Southeast Asia and China, was the genotype Z, which emerged in wild birds from Hong Kong in 2002. This genotype was more successful in infecting a large number of avian hosts. A reassortant virus between the genotypes Z and V led to the Qinghai Lake outbreak in wild geese [63]. The outbreak at this lake, which is an important sanctuary for a variety of migratory birds, was the precursor of the spread of H5N1 out of East and Southeast Asia into Eurasia and Africa. A case in point is Egypt, where the first outbreak in poultry and humans was observed in early 2006, soon after the Qinghai Lake outbreak [64]. H5N1 is now endemic in poultry and domestic ducks in Egypt, and more than a hundred

human cases have been reported so far.

To understand the pandemic potential of H5N1, recent studies have focussed on the mutations underlying the evolution of human-to-human transmissibility in H5N1. On the experimental side, the above mentioned studies on experimental evolution of transmissibility in ferrets uncovered 3-4 mutations in HA and a single mutation in PB2 [59, 60]. In addition, Watanabe et al. [65] identified mutations in H5N1 strains from Egypt, which in the course of natural evolution had acquired higher affinity for cellular receptors in upper respiratory tract in humans. On the computational side, several studies have looked for persistent sequence markers in human-to-human transmissible viruses versus avian-to-avian viruses [66–68], using a wide range of Influenza A subtypes. Although these studies shed light on mutations underlying human-to-human transmission, they have not looked at persistent markers associated with human H5N1 infections. This issue is the focus of the research discussed in Chapter 2.

1.4 Outline of the Dissertation

In Chapter 2, I discuss research on transmission bias of H5N1 viruses from birds to humans. Since human infections of H5N1 are transmitted from birds to humans, signs of transmission bias could potentially indicate mutations which are important for human infections of H5N1. We first develop a novel strategy which can detect transmission bias at residues at an annual level. This strategy is then applied to uncover several signs of transmission bias in human infections of H5N1 in publicly available HA sequences from China, Egypt and Indonesia – the countries with the highest number of human H5N1 infections. We find that in each geographic region, only a subset of avian H5N1 virusees characterized by specific mutations can infect humans. In Egypt, vaccination in poultry seems to have driven avian viruses away from this subset, suggesting that vaccination of poultry can be highly efficient in

reducing H5N1 human infections. After correcting for the transmission bias in each region, we find that human infecting H5N1 viruses are not substantially different from the aforementioned subset of avian viruses.

In Chapter 3, I present a novel method for the identification of novel jelly-roll containing capsid sequences using machine learning algorithms on alignment free features. As discussed above, using sequence similarity based methods majority of sequences in viral metagenomics studies that are expected to be of viral origin are not identifiable as known viruses. In such scenarios, methods based on alignment free features are expected to perform better than sequence similarity based methods and have been investigated in the context of microbial metagenomics. The motivation for this project comes from the paucity of algorithms using alignment-free features for detection of novel viral sequences. We focus on detecting novel capsid sequences with the jelly-roll structural motif, which in spite of the sequence divergence in viral sequences, is conserved in a variety of families. Using counts of amino-acid motifs, which are robust to sequence evolution, a machine learning algorithm is shown to classify known jelly-roll capsid sequences with high accuracy against virtually all other proteins. Next, the performance of this method is compared with the most popular sequence-similarity based method to show an improved performance in detecting novel capsid sequences from unknown families. As an application, this method is applied to a viral metagenomic dataset to find several potentially novel jelly-roll capsid sequences.

I conclude the dissertation with discussion on some promising future lines of research.

Chapter 2

Transmission Bias of Influenza A H5N1 Viruses from Birds to Humans and Vaccine Evasion

> "As long as H5N1 is out there in the world," Webster said, "there is the possibility of disaster. Thats really the bottom line with H5N1. So long as its out there in the human population, there is the theoretical possibility that it can acquire the ability to transmit human-to-human." He paused. "And then God help us".
>
> *David Quammen quoting Robert Webster in "Spillover"*

2.1 Introduction

The H5N1 Influenza A avian virus is widely believed to be a pandemic threat [69–71]. Although human H5N1 infections occur rarely, they are usually accompanied by severe respiratory complications with high morbidity. Of the 633 confirmed cases world-wide, there have been 377 deaths, with a mortality rate approaching 60% ([72], WHO report, July 5, 2013). Infections in humans occur almost exclusively from direct human contact with infected wild birds or poultry. Currently, the poor human-to-human transmission efficiency of circulating H5N1 strains [48] limits their pandemic

potential. However, this can be overcome by the rapid evolution of H5N1 [61]. Laboratory studies of experimentally evolved H5N1 strains have shown that current strains can transmit efficiently between mammals (ferrets) with only 4-5 substitutions at specific residues in Hemagglutinin (HA) and Polymerase Basic 2 (PB2) proteins [59, 60]. These results coupled with the high mortality rate of human infections from currently circulating strains highlight the urgent need to understand and control human infections of H5N1. Vaccination against H5N1 in birds has already been undertaken as a strategy to curb H5N1 outbreaks in humans. For effective vaccination strategies, it is crucial to identify which avian H5N1 strains are most likely to infect humans. It is also important to understand how the H5N1 virus is evolving under vaccination induced selection pressure. In this work, we investigate the nature of H5N1 strains most likely to infect humans and find that there are significant signs of transmission bias of H5N1 from birds to humans. The interplay between identified transmission bias and vaccine-induced selection pressures on evolution of H5N1 is also discussed.

Since almost all human H5N1 infections so far were transmitted from avian hosts, any observed signature of biased transmission from birds to humans could represent enhanced/diminished efficiency of certain H5N1 strains to infect human hosts. Selection in H5N1 viruses infecting humans has been studied previously [73–75] using differences in the rates of synonymous and non-synonymous mutations in human isolates as a signature of selection in humans. However, because H5N1 is transmitted from birds to humans, such analyses cannot distinguish between selection pressures on H5N1 from avian or human hosts. A hypothetical scenario of a mutation in H5N1 which is beneficial (to H5N1) for avian infections, but selectively neutral in human infections, illustrates this point (Fig. 2.1 A). Such a mutation would also show a rise in the frequency in human isolates, which could be interpreted as a sign of positive selection in human infections i.e. beneficial for H5N1 in human infections. In general, an analysis involving only human H5N1 isolates cannot identify mutations which are specifically important in human infections, and a comparative analysis between both

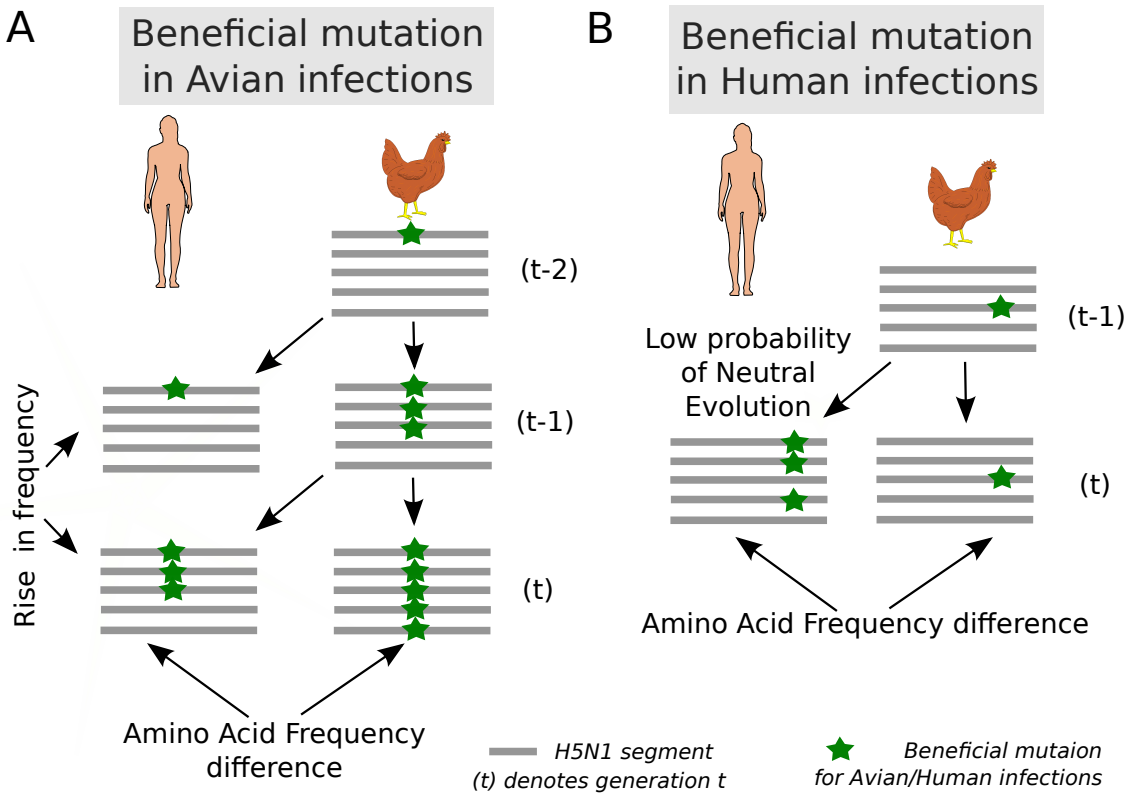

Figure 2.1: Hypothetical scenarios of mutations in H5N1 which are specifically beneficial only in avian infections (A) and only in human infections (B). A) A mutation which is beneficial for avian infections but selectively neutral for human infections will show an increase in frequency human isolates, since H5N1 infections are transmitted to humans from birds. B) A mutation which is beneficial in human infections, but selectively neutral in avian infections, will show: i) amino-acid frequency difference between human and avian isolates of the same year, and ii) low probability to neutrally evolve from the avian viral pool of the previous year. These are the methods used to detect the transmission bias of H5N1 infections from birds to humans.

human and avian isolates is required.

The hypothetical scenario of a mutation conferring enhanced efficiency in human infections but neutral in avian infections suggests a strategy to identify mutations important for human H5N1 infections (Fig. 2.1B). Such a mutation would be over-represented in human isolates as compared with avian isolates, leading to a transmission bias of H5N1 from birds to humans. In general, mutations which either enhance/diminish efficiency of H5N1 infections in humans can be identified using

the two signatures of transmission bias: a) a significant difference in amino-acid frequencies in human isolates compared to avian isolates from the same year, and b) a significantly low probability of neutral evolution of the human isolates from the avian viral pool of the previous year. These criteria were applied to detect mutations important for human infections using Hemagglutinin (HA) protein sequences of H5N1 avian and human isolates from 1996-2011 collected in China, Egypt and Indonesia. For each geographic region, several residues that show transmission bias on an annual resolution were identified. These results show that, in each geographic region, strains infecting humans are significantly more likely to originate from a subset of the avian viral pool rather than the entire avian viral pool. The residues which represent this transmission bias lie in immunologically relevant regions of HA, such as the epitope regions, the receptor binding site, the polybasic cleavage site and the trans-membrane site. In Egypt, we find that human isolates are significantly different from vaccine resistant avian isolates. This suggests that vaccine-resistant avian strains are not likely to infect humans.

2.2 Methods

2.2.1 Sequence Data

Aligned amino acid and nucleotide sequences for Hemagglutinin of H5N1 isolates were downloaded from the NCBI Influenza Virus Resource database: http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi, on August 8, 2012 (Egyptian isolates) and October 18, 2012 (Asian isolates). Alignment was performed using the program MUSCLE [76] using default parameters. Identical strains were removed using both the web resource's option and additional programming (to account for identity up to missing residues) with human isolates preferably retained from a set of identical isolates. Host, region, and year information for all isolates were also downloaded from the above website. The resulting dataset comprised of 1209 (153 human, 1056 avian)

isolates in Egypt, China and Indonesia from years 1996-2011. Human derived isolates for all geographical regions combined were from years 2005-2010.

2.2.2 Principal Component Analysis (PCA)

Principal Component Analysis is a general method of feature reduction used to capture and visualize high-dimensional data in few most important variables. The method amounts to diagonalizing the covariance matrix and representing the high-dimensional data in the subspace of top eigenvectors with the highest eigenvalues. In the context of population genetics, PCA is routinely used to understand the relatedness of different genomic sequences, referred to as population structure [77]. We performed PCA on Hemagglutinin amino acid and nucleotide sequence data for isolates from both avian and human hosts to understand population structure of H5N1 in each geographic region. Both amino acid and nucleotide sequences in the dataset had sites with more than two variants. To encode these amino acids or nucleotides into numerical values, we used the following prescription. Amino acids at each residue were assigned values 0,1,2,...,19, with the most common variant assigned to 0, the next frequent 1, and so on. In all isolates in each geographic region, we excluded residues with a missing amino-acid, which could indicate a deletion or missing sequence. The numerical data for each residue was normalized by subtracting the mean. However, we did not divide the result by the standard deviation to ensure that the more variant sites carry higher weight in the PCA analysis. The PCA analysis was done using the module for Singular Value Decomposition in SciPy [78].

2.2.3 Clustering H5N1 strains using a distance cutoff in PCA space

PCA on H5N1 isolates from each geographic region revealed that human isolates from each region cluster together with a subset of avian isolates (see Results below). To understand this population sub-structure and its relevance to transmission bias, we constructed clusters in PCA space by first clustering human isolates that were close to

one other, and then clustering avian isolates that were close to these human isolates. More specifically, we first retained only those PCA components which accounted for $> 4\%$ variance. Clusters were constructed of all human isolates closer than a distance corresponding to 4% variation of the total variation in each local region. It was found that by using this distance cutoff, almost all ($> 80\%$) of the human isolates in each region clustered together.

The following algorithm was used for clustering human isolates. Initially all human derived isolates were placed in the un-clustered list. Because each human isolate belongs to a cluster (albeit of size one), a randomly chosen isolate was chosen to seed the first cluster and was removed from the list of unclustered isolates. In the next step, all isolates within the distance cutoff from this initial isolate were included in the cluster, and removed from the list of unclustered isolates. If the cluster size was greater than one, then new unclustered isolates were added to this cluster if they were closer than the distance cutoff to at least one of the cluster isolates. This step was iterated until there were no isolates in the unclustered list that were within the distance cutoff to any of the cluster isolates. To construct the next cluster, an isolate was randomly chosen from the list of unclustered isolates, and the same algorithm was repeated. The construction of clusters ended when the continuously updated list of unclustered isolates was exhausted. For each geographic region we found that most ($> 80\%$) human derived isolates formed a single cluster using the distance cutoff of 4% of total variance.

Avian isolates that fell within a distance corresponding to 2% variation from all the human isolates in the identified cluster were also added to the cluster. This subset of the avian isolates was then used as the set of avian isolates closest to the human isolates. Schematic representation of this algorithm using data on isolates from Egypt is shown in Fig. (2.2), and PCA plots showing human and closely clustering avian isolates for each region are shown in Fig. (2.3). The list of all closely clustering avian and human isolates from each geographical region is given in Appendix A.

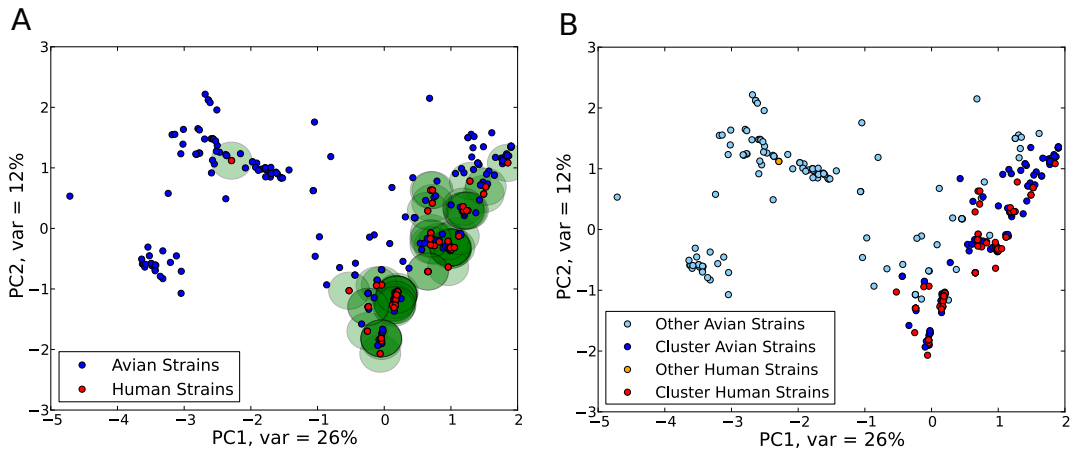

Figure 2.2: A) Schematic representation of the clustering algorithm used to cluster most similar human and avian isolates from Egypt using the first two principal components. Discs of proximity of radius corresponding to 1% of the total variance around each human isolate (red circle), and those human isolates whose discs of proximity overlap are said to cluster together. Next, all the avian isolates (blue circles) that fall in the discs of proximity of clustering human isolates are retained in the cluster. B) The results of the actual implementation using a distance cutoff of 4% of the total variance using the top 4 principal components.

2.2.4 Detection of residues in the human isolates with significant amino acid frequency differences from the avian isolates

For isolates from each region and year, we computed the significance of differences in amino acid frequencies at each residue between the human and avian isolates using the multinomial distribution. Since human infections of H5N1 are derived from viruses transmitted from birds, we expect the amino-acid frequencies at residues in human isolates to be similar to those of avian isolates upto sampling bias. Thus, amino acids at a given residue in human isolates from a given year were treated as samples drawn from the distribution of amino acids present at the same residue in the avian isolates of the same year. The multinomial formula for sampling was used to evaluate the likelihood of sampling the human amino acid configuration from the amino acid distribution from the avian isolates. At a given residue in isolates from a given year and region, let $\{n_1, n_2, n_3, ...\}$ be the observed counts for amino-acids $\{aa_1, aa_2, ...\}$ in the human isolates, and let $\{p_1, p_2, p_3, ...\}$ be the corresponding amino-acid frequencies

in the avian isolates. The likelihood of observing these counts in human isolates, given that they are sampled randomly from the avian isolates, is given by

$$L(\{n_1, n_2, n_3, ...\}; \{p_1, p_2, p_3, ...\}) = \frac{N!}{n_1! n_2! ....} (p_1^{n_1} p_2^{n_2} p_3^{n_3} ...) \tag{2.1}$$

where $N = n_1 + n_2 + n_3 + ...$ is the total number of human derived isolates from a given year and region.

An empirical p-value for this likelihood was computed by drawing 10^8 random sample sets of equal size to the human isolates from the distribution of amino-acid residues in avian isolates, and counting the fraction of such realizations with a lower likelihood than observed. To correct for population structure differences between human and avian H5N1 isolates from each geographic region, we repeated this analysis only on the subset of avian isolates that clustered closest to human isolates, as described in the previous section.

2.2.5 Detection of residues in human derived isolates with low probability of evolving neutrally from the avian H5N1 isolates

We adapted the method introduced by Pan and Deem [79] to compute the probabilities of neutrally evolving the observed amino-acid configurations at each site of human isolates from the avian isolates of the previous year. In this method, the process of neutral mutation is modeled as Poisson process. We deviate from this method in two important ways. First, since H5N1 strains infecting humans were transmitted to humans from birds, we compute the probability of neutral evolution of human isolates from the avian isolates. Second, for the rate matrix, we used an Influenza specific protein evolution model called "FLU", which was developed by Dang et al. [80]. This model was constructed using maximum likelihood analysis on thousands of influenza virus protein sequences.

If the amino-acid frequencies in the avian isolates for the previous year (say y-1) were observed to be $\{p_1, p_2, p_3, ...\}$, then the theoretically evolved frequencies

$\{p_{e,1}, p_{e,2}, p_{e,3}, ...\}$ can be computed using the protein evolution model $\mathbf{Q}$ as

$$[p_{e,1}, p_{e,2}, p_{e,3}, ...] = [p_1, p_2, p_3, ...] \cdot \exp(\mathbf{Q}t) \tag{2.2}$$

where [..] is a row vector, $\mathbf{Q}$ is a 20×20 matrix, and t is measured in units of mutation rate, which I assume to be the substitution rate of $4.77 \times 10^{-3}/site/year$ [62]. This assumption is exact for infinite effective population size, and should be accurate for the micro-evolution of H5N1, where the population size has been estimated to be $\sim 10^3 - 10^4$ [81]. Using these evolved frequencies, the likelihood of observing the amino-acid configuration of human isolates of year y can be computed as before:

$$L\left(\{n_1, n_2, n_3, ...\}; \{p_{e,1}, p_{e,2}, p_{e,3}, ...\}\right) = \frac{N!}{n1!n2!....} p_{e,1}^{n_1} p_{e,2}^{n_2} p_{e,3}^{n_3} ... \tag{2.3}$$

To compute the significance (p-value) of this likelihood value, randomly generated 10^8 sets of samples from the evolved distribution $\{p_{e,1}, p_{e,2}, p_{e,3}, ...\}$ were obtained and the likelihoods were calculated for each of these sets using the above formula. The empirical p-value of the observed likelihood value is the fraction of these 10^8 samples with lower likelihoods than the one observed.

2.2.6 Sensitivity to Ascertainment Bias

The database contained far fewer human isolates than avian isolates, with some years having only ~ 10 human isolates. In such a scenario, a few outlier samples could bias the results. To understand this, the sensitivity of results when only a subset of the full dataset is used (known as "jackknifing test") was studied. Randomly chosen 1,000 subsets containing 75% of human and 1,000 subsets of 75% of avian isolates in each year were generated. The above two analyses were repeated on the $1,000 \times 1,000 = 10^6$ combinations of these subsets of human and avian isolates. The mean and standard deviations for the log likelihoods of amino acid frequency difference and of neutral evolution were calculated using the methods described above for all the combinations. In the final results, only those residues were retained that

either had a mean likelihood of amino acid frequency difference $< 10^{-5}$ and neutral evolution likelihood $< 10^{-3}$, or vice versa.

2.2.7 Sensitivity to Mutation Rate Variation

As the computation of probabilities of neutral evolution of human derived isolates from the avian viral isolates uses mutation rate as an input parameter, the sensitivity to site-to-site variation in mutation rates was also studied. First, the program PhyML [82] was used to generate maximum likelihood phylogenetic trees for amino-acid data of H5N1 isolates from Egypt by using a popular model for modeling variable mutation rates, the discrete $\Gamma 4$ model [83]. In this model, mutation rates are assumed to be distributed according to the Gamma distribution. Instead of using the computationally more expensive continuous variation of mutation rates, discrete mean values for four equally weighted intervals of mutation rates are used. PhyML calculates the maximum likelihood values for mutation rates of the 4 classes of the discrete $\Gamma 4$ model. Both human and avian isolates from Egypt were analyzed using PhyML using the rate model FLU and other default parameters and obtained the maximum likelihood values for 4 classes of the discrete $\Gamma 4$ model to be $\{0.0288, 0.2353, 0.8012, 2.9346\}$, which was multiplied with the mean rate of $4.77 \times 10^{-3}/site/year$ [62] to get the 4 classes of mutation rates. The likelihood of neutral evolution for all the significant sites using each of these rates and all the significant residues were found to have a mean likelihood of neutral evolution $< 10^{-5}$.

2.3 Results

2.3.1 Human H5N1 Isolates derive from a subset of avian viruses with geography specific epitope profiles

We analyzed 1209 HA sequences of H5N1 isolates from avian (n=1056) and human hosts (n=153) from China, Indonesia and Egypt, collected from 1996-2011. Principal

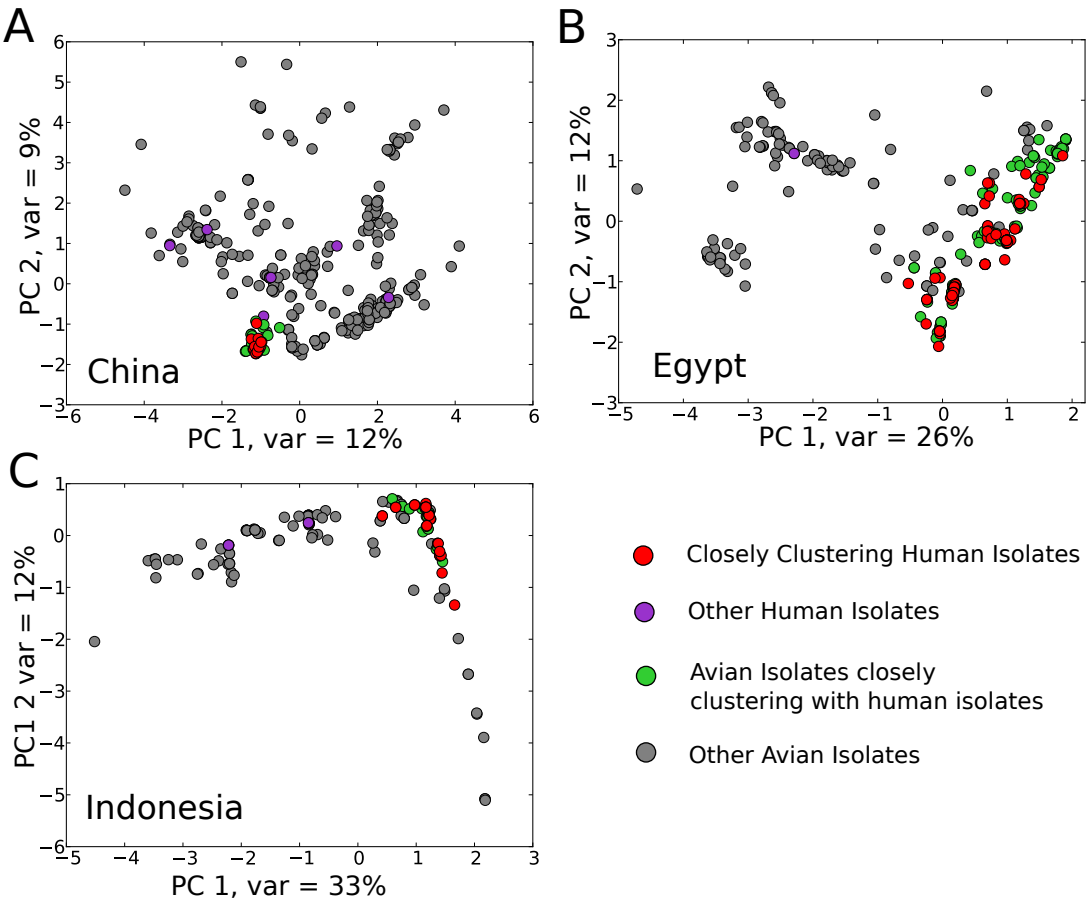

Figure 2.3: PCA of Hemagglutinin amino-acid sequences from H5N1 isolates from China (A), Egypt (B) and Indonesia (C).

Component Analysis (PCA) was used to study population structure (Methods). PCA plots for HA sequences from each geographic region are shown in Fig. 2.3. In each geographic region, human isolates cluster with subsets of avian isolates, suggesting a transmission bias in H5N1 infections from avian hosts to humans. To characterize the subsets of avian isolates most likely to infect humans, we identified clusters of closely related human and avian isolates, using a distance cutoff in PCA space (Methods, Fig. 2.2). The identified clusters consist of most of the human isolates in each region: 30 out of 36 in China, 70 out of 71 in Egypt, and 46 out of 50 in Indonesia.

Two signatures were used to identify transmission bias in avian to human infections: a) the residues should have a significant difference in amino-acid frequency in human isolates compared to avian isolates from the same year, and b) the residues

in human isolates should have a significantly low probability to derive from neutrally evolved avian viral isolates from the avian viral pool of the previous year. To evolve the avian pool neutrally from one year to the next, we adapted the method of Pan and Deem [79] (Methods). The expected frequencies of amino acids at a given residue in a given year were obtained by neutrally evolving the observed amino acid frequencies at this residue in the previous year, using an amino acid transition probability matrix from the influenza specific substitution model of Dang et al. [80]. Using the avian isolate frequencies (either actual or expected under neutral evolution from the previous year) as a-priori expected human isolate amino-acid frequencies, the multinomial formula was used to estimate the probabilities of the observed human isolate amino-acid frequencies. A jackknife test was used to determine significantly low probabilities (Methods). Significant residues thus identified are listed in Table 2.1 by year and geography. Amino acid frequencies for these residues are in Table 2.2.

Table 2.1: Significant residues identified by comparison of human isolates with all avian isolates in each region and year

Year	Position	P-value of neutral evolution	Significance of amino acid frequency difference	Jackknifing mean log-likelihood for neutral evolution	Jackknifing mean log likelihood for amino acid frequency difference
Egypt					
2009	43	$< 1.00 \times 10^{-8}$	1.00×10^{-8}	21.44 ± 3.58 [1]	5.53 ± 1.01

[1] log likelihoods are expressed as negative $\log_{10}$.

2009	74	4.28×10^{-6}	1.42×10^{-6}	6.59 ± 0.68	4.33 ± 0.51
2009	97	2.60×10^{-6}	2.80×10^{-7}	6.6 ± 0.66	4.59 ± 0.52
2009	110	1.10×10^{-5}	3.10×10^{-7}	6.36 ± 0.65	4.61 ± 0.57
2009	120	$< 1.00 \times 10^{-8}$	2.00×10^{-8}	19.7 ± 3.63	5.59 ± 1.05
2009	123	1.43×10^{-5}	6.77×10^{-6}	6.36 ± 0.66	3.74 ± 0.47
2009	141	2.96×10^{-5}	2.85×10^{-5}	6.49 ± 0.93	4.9 ± 0.8
2009	144	2.64×10^{-6}	1.19×10^{-6}	6.57 ± 0.65	4.38 ± 0.5
2009	151	$< 1.00 \times 10^{-8}$	3.36×10^{-6}	22.13 ± 4.63	5.84 ± 1.01
2009	162	4.00×10^{-7}	6.44×10^{-4}	9.2 ± 1.43	3.55 ± 0.69
2009	165	8.83×10^{-6}	1.90×10^{-6}	6.35 ± 0.68	3.94 ± 0.47
2009	226	1.00×10^{-8}	9.00×10^{-8}	9.32 ± 0.9	6.05 ± 0.66
China					
2005	140	$< 1.00 \times 10^{-8}$	2.64×10^{-3}	9.5 ± 2.74	3.71 ± 0.24
2005	174	$< 1.00 \times 10^{-8}$	$< 1.00 \times 10^{-8}$	8.94 ± 1.65	3.37 ± 0.22
2005	181	$< 1.00 \times 10^{-8}$	$< 1.00 \times 10^{-8}$	9.47 ± 2.57	3.47 ± 0.23
2005	322	$< 1.00 \times 10^{-8}$	$< 1.00 \times 10^{-8}$	20.82 ± 0.03	3.45 ± 0.23
Indonesia					
2005	86	$< 1.00 \times 10^{-8}$	$< 1.00 \times 10^{-8}$	13.9 ± 1.61	7.01 ± 1.01
2005	140	$< 1.00 \times 10^{-8}$	1.12×10^{-3}	19.64 ± 1.6	3.26 ± 0.4
2005	200	4.20×10^{-7}	$< 1.00 \times 10^{-8}$	7.12 ± 3.13	7.03 ± 1.02
2005	325	$< 1.00 \times 10^{-8}$	$< 1.00 \times 10^{-8}$	17.63 ± 1.97	7.13 ± 1.04
2006	86	$< 1.00 \times 10^{-8}$	1.00×10^{-8}	35.03 ± 12.55	5.98 ± 1.31
2006	94	1.00×10^{-7}	1.15×10^{-3}	8.78 ± 1.66	4.04 ± 0.84
2006	140	2.60×10^{-7}	2.13×10^{-3}	9.99 ± 1.7	3.86 ± 0.76

2006	200	$< 1.00 \times 10^{-8}$	6.08×10^{-5}	32.71 ± 9.35	4.27 ± 1.04
2006	325	$< 1.00 \times 10^{-8}$	$< 1.00 \times 10^{-8}$	39.08 ± 18.59	8.32 ± 1.56
2007	184	2.25×10^{-6}	$< 1.00 \times 10^{-8}$	5.66 ± 2.51	4.45 ± 0.45

Table 2.2: Amino-acid frequency for major amino acid at significant residues

Position	Major Amino-Acid	Average Frequency (in %)				Year of first report of amino acid		Other Amino Acids	Region of HA
		Human Isolates		Avian Isolates					
		Cluster	Other	Cluster	Other	Avian Isolates	Human Isolates		
Egypt		(n=70) [2]	(n=1)	(n=195)	(n=140)	(2005)[3]	(2006)		
43	N	57.1	0	63.1	8.6	2007	2007	D,S,del	Epi. E[4]
74	P	100	0	100	32.9	2005	2006	S	Near Epi. E
97	D	98.6	0	99	29.3	2005	2006	N, E, del	-
110	H	100	0	100	33.6	2005	2006	R, G	Epi. A
120	N	51.4	0	34.9	8.6	2007	2007	S,D,G	Near Epi. A

[2]Total number of cluster/other samples from each host
[3]Years of first reported avian/human cases
[4]Epitope E

123	S	100	0	99.5	37.9	2005	2006	P, L	Near Epi. A & RBS[5]
141	S	97.1	0	94.9	23.6	2005	2006	P, L	Epi. B
144	F	100	0	100	31.4	2005	2006	Y, C	Epi. B
151	T	57.1	0	64.1	12.1	2007	2007	I,L,V	Epi. D
162	R	72.9	0	44.1	13.6	2005	2006	K, I, E	Near Epi. D
165	N	100	0	100	36.4	2005	2006	H	GS[6]
226	M	97.1	0	89.2	22.9	2005	2006	V, I	Epi. D
Indonesia		(n=42)	(n=4)	(n=50)	(n=128)	(2003)	(2005)		
86	T	100	0	90	19.5	2005	2005	A, N	Epi. D
94	S	100	0	100	36.7	2004	2005	N, D, M, del, K	Near Epi. B

[5]Receptor Binding Site
[6]Glycosylation Site

140	S	100	0	100	32.0	2005	2005	K, T, Q, R, D, N, del	Epi. B
184	A	76.2	100	96	86.7	2003	2005	E, V, D, G, del	Near Epi. D
200	I	100	0	100	25.0	2003	2005	V, del	Epi. D
325	S	100	0	98	19.5	2005	2005	R, A, G	PBS[7]
China		(n=30)	(n=6)	(n=66)	(n=394)	(1996)	(2003, 2005 [8])		
140	T	100	0	100	17.5	2004	2005	R, K, S, N, E, M, A, V	Epi. B
174	I	86.7	33.3	97	24.4	1999	2005	V	Epi. B, near RBS
181	S	100	16.7	95.5	23.4	2004	2005	P, F	Epi. B

[7]Polybasic Cleavage Site
[8]First human case in China was reported in 2003, but there have been contiguous human cases since 2005

322	L	100	16.7	100	21.6	2005	2005	Q, P, del, H, K, R, S	Near PBS

Several of these residues have a high frequency ($> 80\%$) amino-acid in the human isolates in each region (Fig. 2.4 and Table 2.2). Most of the human isolates in each geographical region cluster together ($80 - 99\%$) (Fig. 2.3) and these amino acids are virtually conserved in these closely clustering human isolates. These amino acids are also almost conserved (frequencies $> 89\%$) in closely clustering avian isolates, but have low to intermediate frequencies ($18 - 38\%$) in other avian isolates (see Methods for identification of closely clustering avian isolates). Moreover, at these residues, we found that the human isolates show much higher probability to neutrally evolve from the closely clustering avian isolates of the previous year (data not shown). These results, taken together, suggest that for each geographic region, human infections are significantly more likely to arise from an identifiable subset of avian isolates, characterized by specific amino acids at identified residues, rather than from the entire avian viral pool.

Many of the loci associated with transmission bias are located in or near functional regions of HA, such as the epitope regions (corresponding to epitopes B, D and E in H3 HA), the receptor binding site, the polybasic cleavage site, and the trans-membrane region (Table 2.2). The mapping of these residues on the protein structure of H5N1 HA [7] shows that most of these residues are in the head region of the HA protein structure (Fig. 2.5). We also found that all except one of the high frequency residues identified arose in the avian viral pool of the region in either the same year or the year previous to when the corresponding human infections were reported, which further suggests their relevance to human infections (Table 2.2).

2.3.2 In Egypt, H5N1 isolates exhibiting Vaccine-induced Antigenic Drift are less likely to infect Humans

Some avian strains circulating in Egypt have undergone diversification in response to vaccine induced selection pressure in poultry [84, 85]. These antigenically drifted avian isolates are now classified as a variant group within the sub-clade 2.2.1 (group I

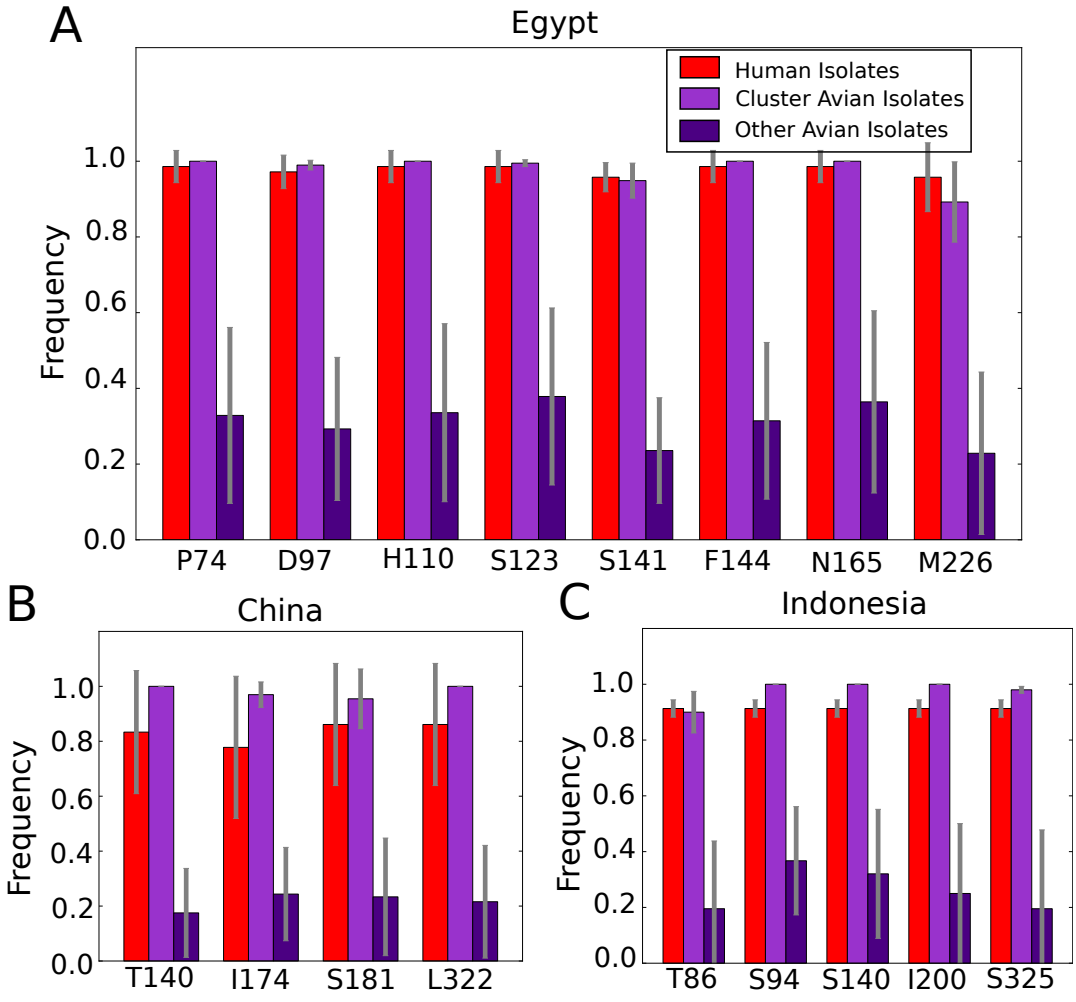

Figure 2.4: Average annual amino-acid frequencies for significant residues that have a high-frequency amino-acid in human isolates. Averaging was performed by weighing each annual frequency with the number of isolates in that year, and grey bars span two standard deviations.

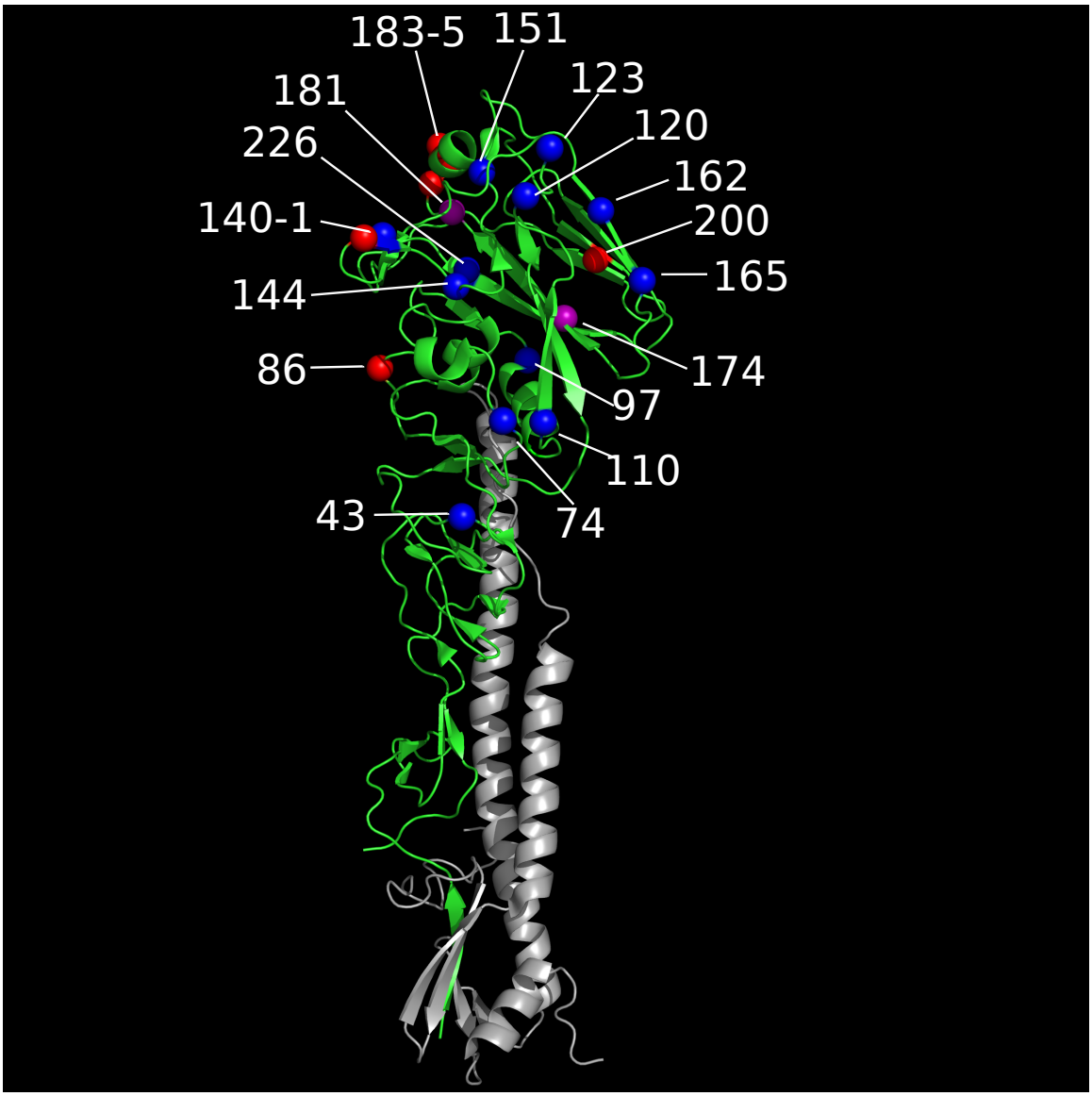

Figure 2.5: Mapping of significant residues on the protein structure of H5N1 Hemagglutinin. Color coding for residues identified is: Blue - Egypt, Red - Indonesia, Purple - China. For this figure, the protein structure from Yamada et al. [7] (pdb code: 2IBX) was analyzed using the program Pymol [8].

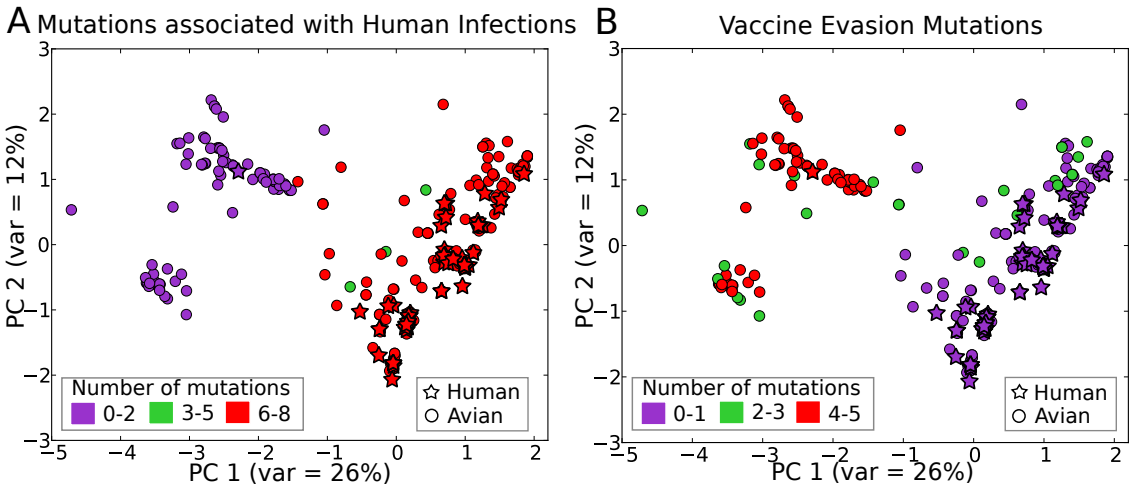

Figure 2.6: PCA plot of human and avian isolates from Egypt showing the number of mutations associated with human infections (A) and vaccine-evasion mutations (B) possessed by each isolate. A) The mutations associated with human infections are those significant mutations identified in Table 2.1, and which have $> 80\%$ average frequency in human isolates. These mutations are P-74, D-97, H-110, S-123, S-141, F-144, N-165 and M-226 (total = 8). B) Vaccine-evasion mutations are taken from Cattoli et al. [9] and are S-74, G-140, P-141, Y-144 and K-162 (total = 5).

in Abdelwhab et al. [85]). The mutations characteristic of these isolates are S-74, N-97, R-110, P-123, G-140, P-141, Y-144, K/E-162, H-165, E-184, and V-226 [85]. Our results identified all but one (G-140) of the residues characterizing this vaccine resistant avian H5N1 group as characteristic of the transmission bias of H5N1 infections from avians to humans (Table 2.1). However, the mutations characterizing human isolates at these residues were *distinct* from those characteristic of the escape mutant group. Specifically, residues 74, 97, 110, 123, 144, and 165 have virtually conserved amino-acids in closely clustering human and avian isolates, which are different from those characterizing the variant group of avian isolates (Fig. 2.4). Since 70 out of 71 human isolates from our Egypt dataset cluster closely, these findings suggest that the variant group of avian isolates from vaccinated birds are significantly unlikely to infect humans.

A serological study using reverse genetically designed viruses carrying the above

mentioned variant group specific mutations showed that the mutations S-74, G-140, P-141, Y-144 and K-162 are involved in escape from neutralization due to Mexican H5N2 vaccine induced antibodies in chickens [9]. A comparative PCA analysis of the H5N1 isolates from Egypt carrying the high frequency transmission bias mutations versus the vaccine evasion mutations (Fig. 2.6) found that the closely-clustering human and avian isolates carry 0-1 out of the 5 vaccine evasion mutations, whereas the more divergent avian isolates carry 3-5 mutations. The cluster of vaccine escape mutant isolates is distinct from the cluster containing human isolates, and carries 0-1 of the 8 high frequency transmission bias mutations (P-74, D-97, H110, S-123, S-141, F-144, N-165, and M-226). These results suggest that mutations involved in vaccine evasion, at least in Egypt, have led to inefficient transmission of avian H5N1 viruses to humans. In other words, the potential of human infections for avian H5N1 viruses was effectively neutralized using the Mexican-derived H5N2 vaccine on poultry in Egypt.

2.3.3 Residues associated with human H5N1 isolates after correcting for Biased Transmission

We investigated whether there are any residues associated with human infection after correcting for the transmission biases described above. Such loci should display a) significant amino acid frequency differences between human isolates and the subset of *closely clustering avian isolates* of the same year, and b) significantly low probabilities of having evolved neutrally from the subset of *closely clustering avian isolates* of the previous year. The residues which have these properties are listed in Table 2.3 for each geographic region and year. The identified residues are in the epitope D region of the H3 Hemagglutinin (residues 184-186 in isolates from Indonesia), and near the trans-membrane site (residue 513 in isolates from Egypt). The T513I mutation in Egypt arose in 2006, and for the years 2007 and 2008 showed enrichment in human isolates (8 out of 19 and 3 out of 7 human isolates respectively) as compared with the

closest avian isolates (frequencies 0.07-0.15), but not for the year 2009 (Fig. 2.7 A). The mutation N-184 arose in 2005 in isolates circulating in Indonesia, and showed enrichment in the reported human isolates in 2007 (5 out of 5) over the avian isolates (frequency $< 20\%$) (Fig. 2.7 3B). The mutations E-185 and E-186 are in close linkage with the N-184 mutation and show similar enrichment in human isolates as compared with avian isolates from Indonesia from the year 2007 (data not shown).

Table 2.3: Significant residues identified by comparison of human isolates with *closely clustering avian isolates* in each region and year

Year	**Position** [9]	**P-value of neutral evolution**	**Significance of amino acid frequency difference**	**Jackknifing mean log-likelihood for neutral evolution** [10]	**Jackknifing mean log likelihood for amino acid frequency difference**
China					
-	-	-	-	-	-
Egypt					
2007	513	$< 10^{-8}$	2.07×10^{-5}	14.92 ± 2.75	3.34 ± 1.19
Indonesia					
2007	183	$< 10^{-8}$	4.71×10^{-6}	7.69 ± 10^{-11}	3.30 ± 0.46
2007	184	1.01×10^{-3}	$< 10^{-8}$	4.51 ± 3.71	4.29 ± 10^{-11}

[9] H5 numbering
[10] Log-likelihoods are reported as negative $\log_{10}$

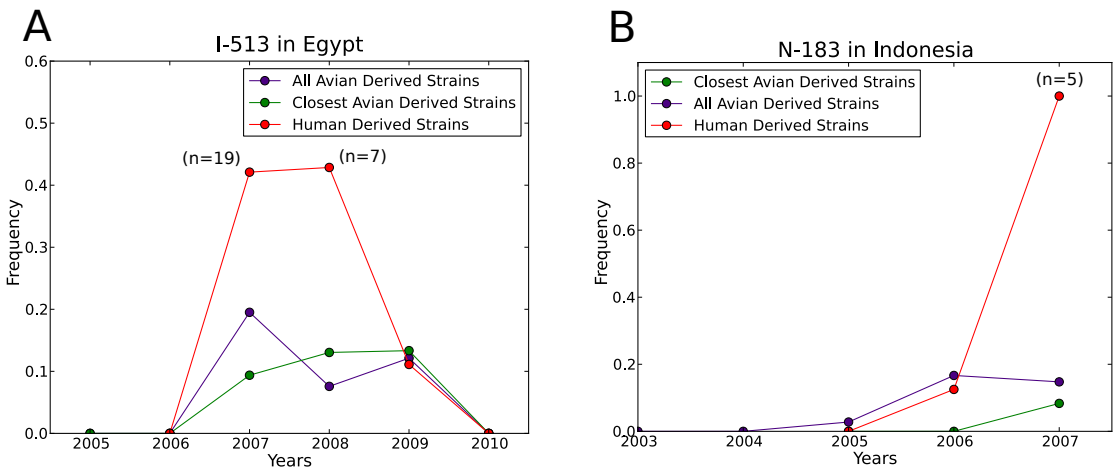

Figure 2.7: Annual frequencies for major amino acid at significant residues in Table 2.3 in human isolates.

2007	185	$< 10^{-8}$	5.50×10^{-7}	7.79 ± 1.88	2.68 ± 0.89

2.4 Discussion

We found a pronounced population substructure in the H5N1 strains in each geographical region studied, with human isolates clustering together with a subset of avian isolates (Fig. 2.3). A comparative analysis of all avian and human isolates in each region identified residues characterizing the subsets of avian isolates with increased potential for infecting humans in each geographic region (Table 2.1). These residues are in or near the epitope regions, the receptor binding site, and the polybasic cleavage site of the HA proteins (Fig. 2.5). Several of these residues have high frequency of an amino-acid in human and closely clustering avian isolates but significantly lower frequency in other avian isolates (Fig. 2.4, Table 2.2). This suggests that not all avian strains can efficiently infect humans. Instead, only an identifiable subset, with specific amino acids at identified residues, can do so. The amino-acids at these identified residues seem to have been important for the H5N1 viruses to infect

humans in the specific geographic regions.

A possible biological reason for this transmission bias is that the identified mutations are involved in efficient binding to receptors on human epithelial cells. It is known from the study of Yamada et al. [7] that HA from human isolates has the ability to bind to cells with both the avian-type ($\alpha 2,3$) and human-type ($\alpha 2,6$) sialic acids, whereas HA from avian isolates can bind only to the avian type sialic acid. Watanabe et al. [65] also studied the mutations responsible for receptor binding properties of human infecting H5N1 isolates circulating in Egypt and found that isolates with increased affinity for human-type sialic acid binding also retained binding to the avian-type sialic acid. They showed that mutations at residue 192 and at residues 129 in combination with 151 (also identified in Table 2.1) enhanced the binding to the human-type sialic acid, while still retaining binding to avian-type sialic acid. Thus, these studies suggest that some human strains possess affinity to both human and avian type receptors, indicating that these strains can infect both human and avian hosts. This is consistent with our claim that only a subset of avian strains can infect humans, since they possess HA proteins that are compatible with human type cell receptors also.

Counter to the above claim, Watanabe et al. [65] showed that an older reference avian strain with the amino acids identified in Table 2.1 does not bind efficiently to human-type sialic acid. They showed that reverse-genetically designed isolates with specific mutations at the residues 129, 151 and 192, in the background of the reference strain, increased the virulence of H5N1 in mice as compared to the original reference avian-derived strain. These results suggest that the identified mutations in Table 2.1 may not be directly responsible for increased human-type sialic acid binding or increased virulence. Given that the mutations identified in Table 2.1 are almost conserved within human isolates and closely clustering avian isolates, but are at low frequencies in the other avian isolates, and that these mutations were present in the isolates in the study by Watanabe et al. [65], these mutations may be a pre-requisite

for higher human-type receptor binding and/or higher virulence.

Another possible reason for transmission bias of H5N1 strains from birds to humans could be vaccine-induced diversification of avian viruses. In Egypt, some of the identified residues conserved in the closely clustering human and avian isolates have also been shown to be involved in vaccine evasion [84, 85]. Cattoli et al. [9] studied the effect of Mexican H5N2 strain induced antibodies in chickens on a divergent clade of avian-derived H5N1 isolates from commercial poultry farms in Egypt, where the Mexican H5N2 vaccines are used intensively. Using reverse genetics and serological studies, they found the mutations at residues 74, 140, 141, 144 and 162 to be important for the vaccine-resistance of the divergent clade of H5N1 isolates. Our results identified all the above residues, except the residue 140. The amino acids at the residues 74, 141 and 144 were found to be almost conserved in human isolates and are different from the ones involved in vaccine-resistance (Figs. 2.4, 2.6). The mutual exclusivity of the vaccine-evading mutations (in poultry) and the the high frequency transmission bias mutations associated with human infections (Fig. 2.6) suggests that during acquisition of vaccine-evasion, the divergent avian isolates lost the ability to infect humans.

This situation could arise either due to: a) low viral loads in vaccinated poultry [86] leading to reduced transmission to humans, b) the escape mutant virus is a poor transmitter in general, or c) vaccine induced molecular changes in HA make the mutant strains transmit inefficiently to humans. Although it is currently difficult to distinguish between these possible scenarios, our results show that vaccination with the Mexican H5N2 strain evolved the virus away from human infectivity. Cattoli et al. [9] showed that avian-derived strain with mutations that are associated with human infections can be neutralized by antibodies induced in chickens by vaccination with Mexican H5N2. Thus, although the intensive use of the Mexican H5N2 vaccine has led to the development of vaccine-resistant avian H5N1 isolates, this vaccine could prove beneficial to control human infections in Egypt. To our knowledge, this

is the first observation that selection pressure from vaccination of poultry may be driving H5N1 away from being able to infect humans. Our analysis also suggests that appropriate vaccination of poultry against specific epitopes may significantly mitigate the risk of human infections.

After correcting for transmission bias of H5N1 isolates from the avian viral pool to humans, we found that certain amino acids at identified residues have a higher frequency in human isolates compared to the closely clustering avian isolates. This suggests that these residues increase the likelihood of human infectivity in the particular genetic background of avian H5N1 strains which are most likely to infect humans (Table 2.3, Fig. 2.7). The residues identified to have this property in Indonesia are in the region corresponding to epitope D of H3 HA, suggesting selective pressure by the human immune response. The residue 513 identified in Egypt lies close to the transmembrane site of HA, whose function is involved in later stages of cell-entry [87]. Intriguingly, these residues do not have significant scores when human isolates are compared with all the avian isolates after correcting for ascertainment bias. Although this scenario could arise due to small sample size bias (18 human isolates in Egypt in year 2007 and 5 in Indonesia in 2007), it could also arise from the similarity of the amino-acid frequencies at these sites in the human isolates to that of the entire avian pool rather than the closely clustering avian isolates to the human isolates, which suggests that these mutations have originated multiple times on different genetic backgrounds. In any event, after correcting for the biased transmission of human isolates from the avian viral pool, my analysis suggests that natural selection in human H5N1 infections is not very widespread.

In summary, the main results of this chapter are: 1) that in each geography, only certain identifiable subgroups of avian-derived H5N1 isolates seem able to infect humans, and 2) selection pressure from vaccination has created escape mutants which are unable to infect humans efficiently. Experimental investigation of these results would provide additional insights into the biological mechanisms underlying enhanced

human infectivity of certain H5N1 strains as well as on how vaccination pressure impacts the ability of H5N1 avian viruses to infect humans.

Chapter 3

Detection of Novel Viral Capsid Sequences using a Machine Learning Approach on Alignment-Free Features

> There are a million virus particles per milliliter of seawater for a global total of 10^{30} virions! Lined up end to end, they would stretch 200 million light years into space.
>
> *Vincent Racaniello of "Virology Blog"*

3.1 Introduction

Viruses are now believed to be the most numerous and diverse biological organisms in the biosphere with an estimated $10^{30} - 10^{31}$ virus particles on the Earth [88]. With the advances in sequencing and other molecular biology techniques, it is now possible to isolate virus-sized particles from an environmental sample, extract the DNA/RNA content of these virus-like particles, sequence and assemble these genomes, and identify viruses in the environmental sample. Such viral "metagenomic" studies have provided a new way of analyzing the viral content of numerous environmental and biotic microcosms by circumventing the traditional approach of culturing viruses [32] and have tremendous potential to uncover the enormous diversity of viruses [33]. The viral composition for different environments such as seawater, hot springs,

marine sediment, potable water, etc. and organism-derived samples such as tissues, feces, etc. are now known (see [33] for a review). Metagenomic studies have also elucidated the biological functions of the viruses in the microcosms studied [89], and have the potential to discover novel disease causing viruses [35, 36].

In spite of their success, such studies have also pointed to limitations of the existing methods in uncovering the full genomic content of environments [33, 34]. One such limitation, arising consistently in all viral metagenomic studies, is that a large number of sequences (44-99 % in the studies reviewed in [33]) are not homologous to any known sequences in databases. Most metagenomic studies use sequence similarity with known organisms to identify novel sequences, with Basic Local Alignment Search Tool (BLAST) being the most popular tool [38]. BLAST evaluates the sequence similarity between the query sequence and a target database using a protein evolution model to assess the significance. Unlike the conserved 16S rRNA used to detect the presence of prokaryotic genomes [42], viruses do not share a single conserved genomic sequence and viral sequences can be quite diverse. The immense diversity of the virosphere, the fast rates of viral evolution, and the comparatively small number of known viral sequences ($\sim 5,000$ viral species with complete genomes) are believed to be the reasons which limit the identification of novel viral sequences using BLAST and other sequence similarity based methods. Thus, viral metagenomic studies could benefit from a method which could identify novel viral sequences in a sequence-similarity independent manner. In this chapter, I present our work on such a method to identify novel capsid sequences.

Capsid proteins are the building blocks of virion shells and are essential features of viruses [90]. In spite of having considerable variability in sequences, capsid proteins from a large set of diverse viral families have been shown to have a conserved structural motif known as the jelly-roll fold, composed of eight beta strands forming a wedge [6, 28, 29]. This structural conservation has been seen across some families of bacteriophages, viruses of archea, and eukaryotic RNA and DNA viruses (Table 3.1).

The existence of the conserved jelly-roll motif possibly points to the common origin of these diverse viral families, and viruses are now believed by some to predate the origin of cellular life [12, 91]. Another explanation for the conservation of the jelly-roll fold could be convergent evolution, thus pointing to the advantage of adopting this fold. In either case, the conservation of this motif could be useful in identification of novel viruses, although the sequence properties underlying this strucutural motif remain poorly understood [92]. Such sequence properties are likely to be shaped by biophysical constraints such as self-organization into icosahedral capsids, maintaining the structural stability of such capsids in variety of environments and the ease of disembly in host cells. The imprint of such constraints can be obscured and scattered in the primary sequences of capsid proteins, which may not be evident in sequence similarity based comparisons. Machine learning algorithms have been used in numerous such contexts, where useful information for the problem at hand is present in a cryptic fashion in the various features of data samples. We therefore chose to use a machine-learning approach to learn classification of the jelly-roll motif containing capsid sequences against other viral and non-viral proteins using alignment-free features.

Alignment-free features offer an alternative characterization of DNA/protein sequences [39]. A typical class of such features consists of counts of certain short motifs in a given sequence (e.g. di-, tri-, and tetra-nucleotide frequencies). Sequence-homology based methods are limited in their applicability to divergent sequences and in the relatively high computation time. Alignment-free features can address these limitations, and have been used for fast taxonomic binning (classification of DNA sequences according to organisms of origin) [40] and evolutionary relationships between divergent organisms such as viruses [93], among other uses. Several programs exist for taxonomic binning using alignment-free features such as TETRA (tetranucleotide frequency [41]), Phylopythia (5-6 nucleotide frequencies with gaps [42]), Phymm (subset of 1-12 nucleotide frequencies [43]), Metacluster (tetranucleotide frequencies [94]),

but none of these have been trained and validated on viral genomes [34]. The only program using alignment-free features and trained on viral sequences is MGTAXA [44] – a program based on Phymm that has been used to predict host-taxonomy for phages. One of the problems with using oligonucleotide frequencies is that viruses can possess similar composition to their hosts [34], and can confound the prediction of taxonomy for novel viruses. Moreover the use of oligonucleotide frequencies does not utilize the information of evolutionary conservation of protein structures, such as the jelly-roll fold, which is more likely to be present in the amino-acid sequences rather than nucleotide sequences of genes. Thus, to capture the structural conservation of jelly-roll motifs in capsid proteins for identification of novel capsids, a feature space composed of frequencies of short amino-acid motifs was used to characterize capsid and other proteins.

In this chapter, I first describe the nature of the alignment free features used and the rationale behind their choice. The performance of the Support Vector Machine (SVM) trained to classify jelly-roll containing capsid proteins from human proteins and viral polymerases and reverse transcriptases is discussed next. The performance of this SVM classifier (SVM-Caps from hereon) is compared to that of the program BLAST, which is the most popular tool used in viral metagenomics for annotation based on sequence-similarity. I find that this SVM-Caps can outperform BLAST in situations mimicking the detection of novel viral families. Finally, SVM-Caps is used to detect novel putative viral capsids in the viral metagenomic data from a French freshwater lake, Lake Bourget.

3.2 Methods

3.2.1 Sequence Data

All protein sequences used in this study were downloaded from a curated database called RefSeq [95]. This curated database consists of a non-redundant collection of

genomes and protein sequences from broad range of organisms sequenced to date. From this database, we downloaded all proteins from viruses (n=134,031), non-mammal vertebrates (n=258,301), invertebrates (n=631,386), plants (n=566,219), fungi (n=734,575), and protozoa (n=430,365). For archea and bacteria, a reduced set of proteins (n=97,070) from 35 species evenly spaced on the phylogenetic tree were used, and for mammals human proteins (n=34,521) were used.

Because capsid proteins for the viruses in the dataset have different names, capsid proteins were isolated from the viral dataset by using the keywords 'capsid', 'coat', 'gp23', 'head protein', 'L1', 'VP1', 'VP2', 'VP3' and 'gag', followed by manual curation. This procedure resulted in 1823 capsid protein sequences. From these, 606 capsid sequences were extracted from viral families that are known to possess the jelly-roll fold, using the information in references [6, 29], and taxonomy information of viruses from ICTV [21] (Table 3.1). Information about these sequences can be found in Appendix B.

Similarly, 599 protein sequences were extracted for viral RNA and DNA Polymerases and Reverse Transcriptases using keywords such as 'Polymerase', 'RDRP', 'Pol', 'Reverse Transcriptase' and 'RT'.

Table 3.1: Viral Families with Jelly-Roll Capsid Proteins

Viral Family	Genome Type	Jelly-roll type	Host	Dataset Samples
Adenoviridae	dsDNA (L[1])	Double	Vertebrates	11
Ascoviridae	dsDNA (C [2])	Double	Invertebrates	4
Asfarviridae	dsDNA (L)	Double	Vertebrates	-
Birnaviridae	dsRNA	Single	Vert.[3], Invert.[4]	-

[1]Linear
[2]Circular
[3]Vertebrates
[4]Invertebrates

Bromoviridae	ssRNA (+ [5])	Single	Plants	26
Caliciviridae	ssRNA (+)	Single	Vertebrates	11
Corticoviridae	dsDNA (L)	Double	Bacteria	1
Comoviridae	ssRNA (+)	Single	Plants	-
Dicistroviridae	ssRNA (+)	Single	Invertebrates	2
Geminiviridae	ssDNA (C)	Single	Plants	222
Iridoviridae	dsDNA (L)	Double	Vert., Invert.	6
Luteoviridae	ssRNA (+)	Single	Plants	23
Microviridae	ssDNA (C)	Single	Bacteria	14
Mimiviridae	dsDNA (L)	Double	Protozoa	1
Nodaviridae	ssRNA (+)	Single	Vert., Invert.	13
Papillomaviridae	dsDNA (C)	Single	Vertebrates	112
Parvoviridae	ssDNA (L)	Single	Vert., Invert.	29
Phycodnaviridae	dsDNA (L/C)	Double	Algae	15
Picornaviridae	ssRNA (+)	Single	Vertebrates	28
Polyomaviridae	dsDNA (C)	Single	Vertebrates	14
Poxviridae	dsDNA (L)	Double	Vert., Invert.	-
Sequiviridae	ssRNA (+)	Single	Plants	-
Tectiviridae	dsDNA (L)	Double	Bacteria	6
Tetraviridae	ssRNA (+)	Single	Invertebrates	5
Tombusviridae	ssRNA (+)	Single	Plants	39
Tymoviridae	ssRNA (+)	Single	Plants, Invert.	24

[5] positive-stranded

3.2.2 Alignment Free Features

Genomic and protein sequences can be aligned for homologous proteins from different species when the sequences under study are similar. Such alignments can then be used for comparative analyses of differences among the species. For example, 16SrRNA is a ribosomal RNA which shows similarity across a number of eukaryotic, bacterial and archaeal species, and was used to understand the evolutionary relationship of species across these domains of life [23]. But if the sequences under study are divergent, alignment of sequences can fail to give any meaningful information. Alignment free features are useful in such scenarios for comparitive characterization of divergent sequences. Features such as frequencies of small sequence motifs do not require sequence alignment, and have been used to study the species-specific signatures in genomes [40], phylogeny of divergent viruses [93], etc. Because the capsid protein sequences from different viruses can show very little sequence similarity, we chose to characterize proteins using frequency of short amino-acid motifs for this study.

Given a protein sequence S, and a family of motifs $\{m_1, ..., m_k\}$, the protein can be represented by an N-dimensional vector $\{f_1, ..., f_k\}$, where f_i is the number of occurences of motif m_i in the sequence S. Some of the motifs used contain a variable gap, in which case each gapped motif is actually a tuple of simpler motifs. For example, in the case of the motif $\alpha\beta[gap \leq 3]\gamma\delta$, the number of occurences of this motif in a sequence is the sum of occurences of the motifs $\alpha\beta\gamma\delta$, $\alpha\beta * \gamma\delta$, $\alpha\beta * *\gamma\delta$ and $\alpha\beta * * * \gamma\delta$, where $*$ can be any alphabet. The exact nature of the short amino-acid motifs used is discussed below. A custom Python program was developed to calculate the motif counts of each protein sequence using string operations for pattern-matching. Since the number of occurences of all the motifs in a sequence scales linearly with the length of the sequence and the lengths of protein sequences used can vary dramatically, the counts for each sequence were normalized so that they add to unity.

3.2.3 Support Vector Machine Algorithm for Classification

Classification is one of the central problems in machine learning [96]. It involves "learning" of patterns separating the different classes of training samples to predict the class of test samples. One of the most popular algorithms for classification is the Support Vector Machine algorithm (SVM) [97]. Given two or more classes of training samples represented in a vector space, SVM constructs a hyper-surface separating the different classes. In its simplest form, it constructs a separating hyperplane with the normal to the hyperplane as a superposition of some of the sample vectors ("support vectors"). Of the many possible such hyperplanes, the algorithm choses the hyperplane which maximizes the distance of the nearest training sample of each class to the separating hyperplane (maximum margin). SVM can be extended to also construct more complicated separating hyper-surfaces using non-linear distance kernels. The mathematical formulation of the SVM is discussed in Appendix C. In this study, the SVM implementation for Python from the package Scikit Learn [98] was used.

3.2.4 BLASTP and DELTA-BLAST

Basic Local Alignment Search Tool (BLAST) [38] is the most popular tool widely used to compare nucleotide and protein sequence similarity. BLAST is used in most metagenomic studies to identify the organisms present in the environmental sample by finding sequence similarity between the genomic sequences present in the sample and the database of known organisms. This method consists of finding local similarities in query sequence and target database, with a similarity score based on a known substitution rate matrix (e.g. BLOSUM62 in case of protein sequences). The statistical significance of this similarity score is then evaluated by estimating the probability of this score arising due to a randomly generated target database and is given by the expect-value (E-value). The technical details about the implementation of this algorithm can be found in Altschul et al. [38]. The program BLASTP (protein BLAST)

version 2.2.27+ available from the NCBI software repository was used. When using BLASTP with a custom target database (such as the dataset of all jelly-roll containing datasets), the e-value reported for each pair of query and target sequence corrects for the lengths of the sequences, but not for the size of the database. Thus, to obtain an e-value corrected for multiple hypothesis testing, the e-value was divided by the number of sequences in the database.

DELTA-BLAST is an improved algorithm for detecting sequence similarity between a query sequence and target database in situations of lower sequence identity [99]. Whereas the original algorithm uses sequence similarity between the query sequence and a single target protein sequence, DELTA-BLAST compares the similarity of the query sequence to a number of protein sequences known to have a conserved domain. Using this approach a more accurate estimate of local evolutionary rate matrix can be obtained by using the sequence variation in a given family of proteins rather than using a simplified global rate matrix (such as the BLOSUM62 used by BLAST). For detection of conserved domains in the sequence, this algorithm relies on the Conserved Domain Database (CDD) [100], a manually curated database of families of proteins showing similar structures. This algorithm has been shown to be more sensitive than BLASTP [99], although this advantage is expected only for protein families existing in the CDD. Both the web interface at the NCBI website, as well as the stand-alone command-line version of DELTA-BLAST downloaded from the NCBI software repository were used.

3.2.5 Protein Structure Prediction

Because the SVM predicts whether a given sequence is a jelly-roll containing capsid sequence or not, the best proof of validity of the prediction will be to show that the predicted capsid sequence has the jelly-roll fold. Protein structure prediction is a highly complex problem and unfortunately protein structure prediction algorithms currently have only limited success. For most methods, similarity of sequences to

known proteins whose structures have been solved is one of the first steps in modeling, and thus prediction accuracy is crucially dependent on the sequence similarity of sequences to the known proteins. In community wide blind protein structure prediction competitions, it has been observed that depending on the similarity to known sequences and structures of test proteins, the accuracy of prediction algorithms ranges from 20-90% of amino acids correctly placed (within a distance cut-off) [101]. Thus, it is not currently possible to reliably model divergent novel protein sequences. The criterion used to deem a novel candidate sequence suitable for structural modeling was based on similarity of candidate sequences to known jelly-roll containing capsid sequences. When we applied our method to predict novel capsid sequences in the metagenomic data from Lake Bourget, we found that some of the novel candidate proteins have similar sequences to known jelly-roll containing capsid proteins using the more sensitive sequence-similarity algorithm DELTA-BLAST (see above). We used some of the best existing methods for protein structure modeling of these putative capsid sequences. First, the CPHmodels 3.2 web-server [102] was used to find a known template protein structures for the structure prediction of candidate sequences. Then using these as (optional) user-submitted templates, the I-TASSER web-server [10] was used for structure prediction and comparative analysis.

CPHmodels 3.2 webserver is one of the fastest algorithms for protein structure prediction that has been shown to have a cumulative accuracy of 74% on benchmark datasets [102]. This algorithm works in two modes. First, a sequence similarity to known protein structures is sought, and if found, this known protein structure is used as a template in modeling. If no significant similarity is found, then a second mode of detecting remote homology is used. In this mode, a secondary structure prediction is used to find similarities with known protein structures, with similar structures used as templates for model building. Once a template has been found then 3D protein structure is modeled, using the template as an initial backbone, based on ab-initio energy minimizing and sequence/structural similarity. CPHmodels 3.2 web-server

was used to find suitable templates for candidate capsid sequences with similarity to capsids sequences. The significance of the sequence, predicted secondary structure, and predicted solvent accessibility similarities of the query sequence to those template is calculated empirically as the chance of finding such similarities if randomly chosen templates were used. This significance is listed as a Z-score (range 0 to inf) and a threshold for accuracy determined using benchmark studies is $Z > 10$.

Using the highest scoring template generated by CPHmodels 3.2 web-server as a user-submitted template, the candidate capsid sequences were submitted for structural modeling at the I-TASSER web-server. I-TASSER algorithm has consistently been ranked the best performing algorithm in community-wide competitions [10]. This algorithm works in a similar way to the CPHmodels3.2 algorithm mentioned above, but the details of template finding and ab-initio modeling have some differences. The detailed algorithm can be found in Roy et al. [10] and references therein. The significance of the prediction for the structure is given by metric called the C-score. C-score, which ranges from -5 to 2, has been shown to be correlated with root mean square distance (RMSD) for the amino-acids of the predicted and actual protein structures in benchmark studies, and it has been observed that a C-score greater than -1.5 corresponds to the predicted structure having the same fold as the actual test structure.

3.3 Results

3.3.1 Support Vector Machine Classifier can Classify Jelly-roll Containing Capsid Sequences against Other Proteins with High Accuracy

A Support Vector Machine was trained to classify capsid protein sequences containing the jelly-roll fold against other proteins. We downloaded all viral protein sequences from RefSeq [95], and using search keywords isolated 1823 capsid sequences (Methods). From these capsid proteins, 606 belonged to viruses from families known to

possess the jelly-roll fold (both single and double) (Table 3.1). We next focussed on the nature of the alignment-free features which can be useful in the classification of such capsid protein sequences against other proteins.

Table 3.2: Classification of amino-acids based on physical properties

Amino Acids	**Encoded Class**	**Polar**	**Charge**	**Hydropathy**
Ala, Cys, Phe, Ile, J, Leu, Met, Val	0	No	Neutral	Hydrophobic
Asp, Glu	1	Yes	Negative	Hydrophilic
Gly, Pro, Trp	2	No	Neutral	Hydrophilic
His, Asn, Gln, Ser, Thr, Tyr	3	Yes	Neutral	Hydrophilic
Lys, Arg	4	Yes	Positive	Hydrophilic

The physical properties of the amino acids are most likely to be important, rather than the specific amino acid per se, for the conserved jelly-roll structural motif. Therefore, each amino acid was encoded in a reduced alphabet composed of five classes (denoted 0 to 4) based on combinations of charge, polarity, and hydropathy (Table 3.2). Each protein sequence, encoded in the reduced alphabet for amino acids, was then characterized by the occurences of sequence motifs. A priori, it is not clear what would be the best way of choosing these sequence motifs for the problem at hand. The choice for motifs was guided by two rules: a) the number of motifs should be smaller (or comparable) to the size of the capsid protein sequence (average size ~ 400

aa), and b) to capture universal properties of the jelly-roll forming capsid sequences, the motifs should be insensitive to the high sequence variability observed in viruses, i.e. they should be partially robust to amino acid substitutions, insertions and deletions (indels). By using the reduced alphabet discussed above, substitutions within the class of amino acids having same charge, polarity and hydropathy are tolerated. To account for indels, motifs with variable gaps were used. Several types of sequence motifs that fit the above criteria we explored. The best performance was achieved for the following type of sequence motifs: two letters followed by a variable gap of upto 10 characters followed by two more letters i.e. $\{\alpha\beta[gap \leq 10]\gamma\delta\}$ where variables α, β, γ, δ can take values of each class 0-4. This procedure resulted in each protein sequence being represented by a $5^4 = 625$ dimensional vector, where each entry of the vector corresponds to count of a certain motif in the sequence. Because the lengths of the proteins can vary and the total number of occurences of all motifs is proportional to the length of the sequences, the counts were normalized so as to sum up to unity.

By representing each protein with its profile of counts of the above mentioned motifs, I used a linear SVM (Methods, Appendix C) to learn classification between jelly-roll containing capsid proteins and the outgroup of human proteins, and viral polymerases and reverse transcriptases. For the outgroup training dataset, 600 randomly chosen human proteins and 400 randomly chosen viral polymerases and reverse transcriptases (Methods) were used. Since the number of capsid sequences for each viral family in the capsid dataset is variable ($n = 1 - 222$, Table 3.1), a more balanced representation of different families was ensured by using randomly chosen 75% of capsid sequences for each family, with a maximum of 10 used sequences from each family. This procedure resulted in 150 sequences in the capsid training dataset. By testing the predictions on hold out data, the true positive rate of this SVM classifier (SVM-Caps) was found to be $76.5 \pm 3.5\%$ for capsids (n=456), and false-positive rates were $6.3 \pm 0.8\%$ for human proteins (n=1400) and $1.3 \pm 3.5\%$ for viral polymerases and reverse transcriptases (n=199). The Receiver Operating Characteristic (ROC) curve

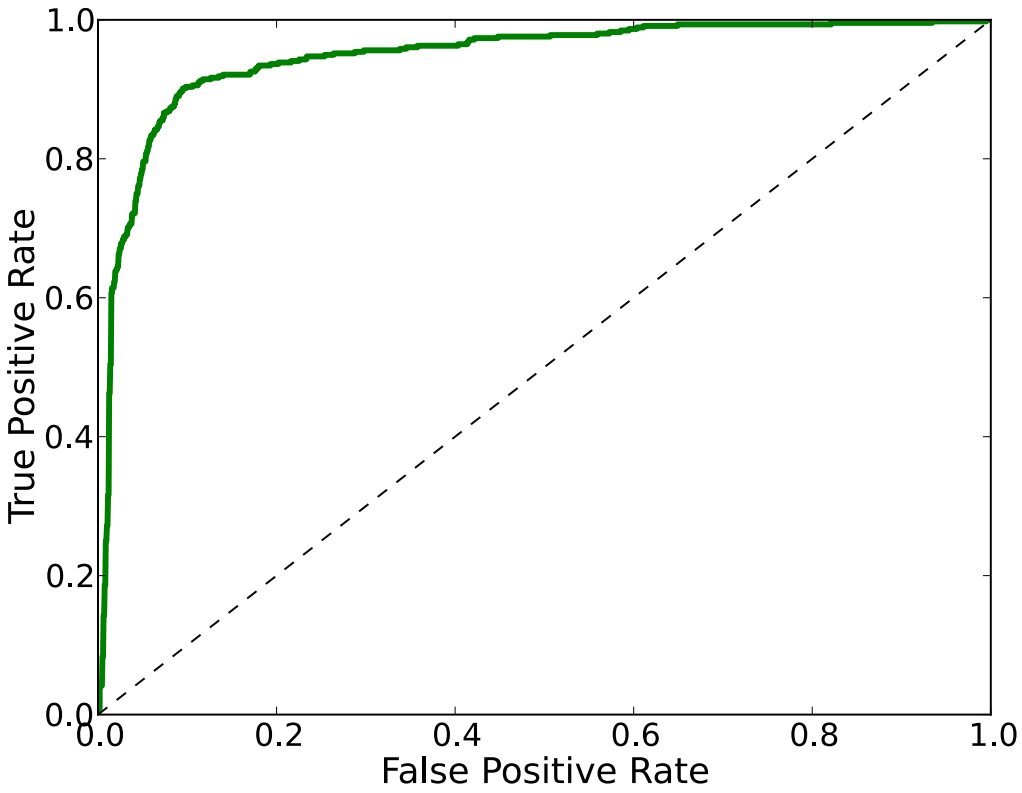

Figure 3.1: Receiver Operator Characteristic curve (ROC) showing true-positive rates for test jelly-roll possessing capsid proteins (n=456) and false-positive rates for test human proteins (n=1400) and test viral polymerases and reverse transcriptases (n=199). Area under ROC curve is 0.9463. The maximum area under the ROC curve ranges from 1 for perfect classifier to 0.5 for a random classifier.

for this classifier for one random realization of the training set is shown in Fig. 3.1. The area under this ROC curve was 0.9463.

The performance of SVM-Caps was studied $\sim$ 2 million proteins from archea, bacteria, fungi, plants, protozoa, invertebrates and non-mammal vertebrates (Methods). In spite of the proteins from these organisms not being present in the training dataset, surprisingly, a good prediction accuracy was found with false-positive rate $< 9\%$ (Table 3.3). The ROC curves for these different classes are shown in Fig. 3.2. The areas under ROC curves were 0.92-0.93 (Table 3.3). These results suggest that SVM-Caps can classify capsid sequences from proteins from non-viral proteins and

viral polymerases and reverse transcriptases with a low false-positive rate.

Table 3.3: Performance of SVM on proteins from other organisms. The training set for SVM comprised of jelly-roll capsids, viral polymerases/reverse transcriptases, and human proteins.

Group	**N**	**False Positives (Rate in %)**	**Area under ROC**
Fungi	734575	63021 (8.57)	0.9224
Protozoa	430365	34428 (8.00)	0.9276
Plants	566219	45517 (8.03)	0.9277
Non-mammalian Vertebrates	258301	16522 (6.40)	0.9376
Invertebrates	631386	49113 (7.78)	0.9278
Archea/Bacteria	97056	8676 (8.94)	0.9232

3.3.2 SVM-Caps can outperform BLASTP in detection of Novel Viral Capsid Sequences from Novel Viral Families

BLAST [38] is a popular local alignment tool which has been used extensively in metagenomic analyses to identify novel species [33] (Methods). In such analyses, a variant of BLAST, called tBLASTx, is used to compare the similarity of translated nucleotide sequences to a target protein database. Since SVM-Caps works with protein sequences, we used the program BLASTP, which is similar to tBLASTx except that it uses proteins sequences as input queries. We performed a comparative study of the prediction powers of BLASTP and SVM under two simulated scenarios. First, we assessed the performance of both approaches to detect novel viruses of known families exhibiting jelly-roll fold in their capsid proteins, and second, to detect novel viruses from novel families potentially possessing jelly-roll fold.

For the first scenario, randomly assigned capsid sequences from the dataset were placed in "known" and "unknown" sequences (using the above-mentioned scheme for ensuring balanced representation of each viral family). BLASTP was then used to detect sequence similarity of the "unknown" capsid sequences to the "known" capsid

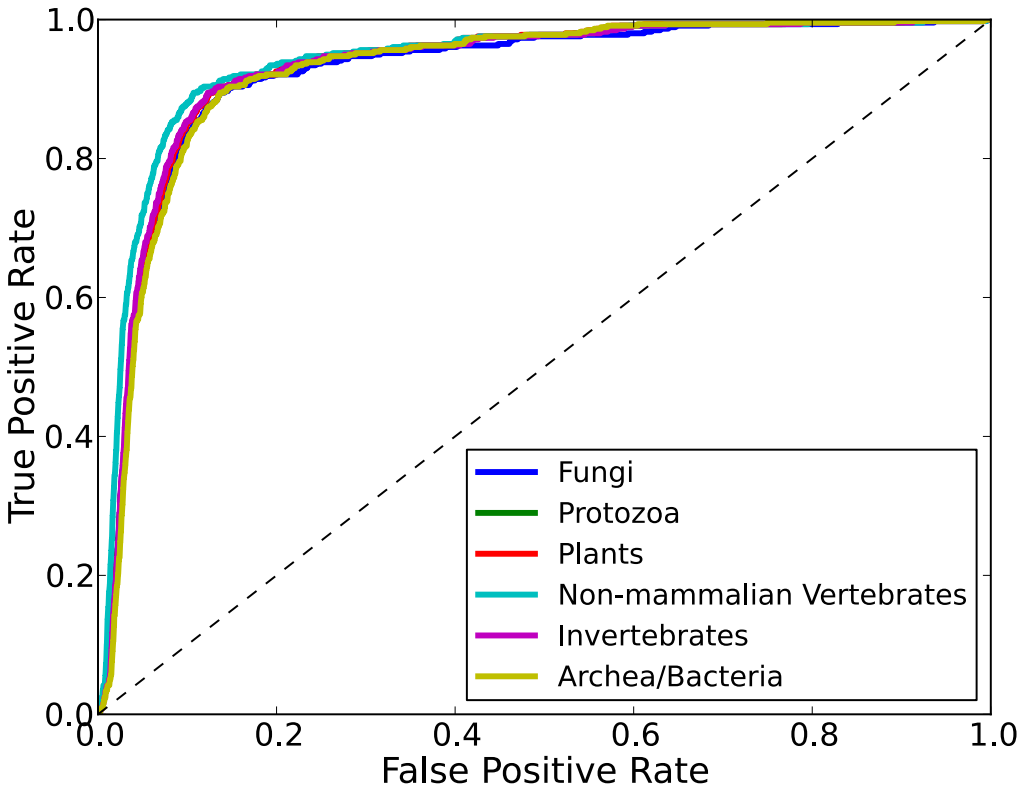

Figure 3.2: ROC curve showing true-positive rates for test jelly-roll possessing capsid proteins (n=456), and proteins from different organismal groups (n=97056-734575) from RefSeq database. Area under each ROC curve was 0.92-0.94 (Table 3.3).

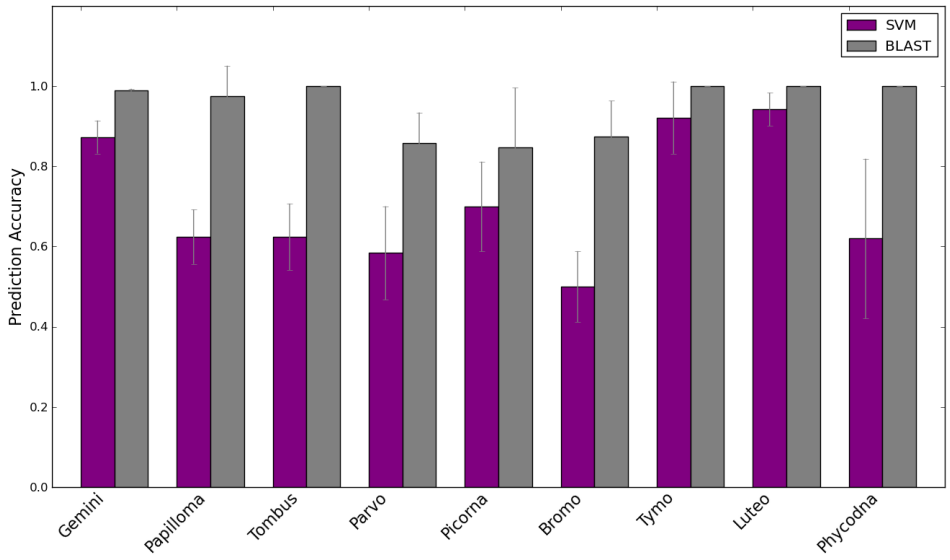

Figure 3.3: Family-wise comparison of performance of BLAST and SVM classifier to detect test jelly-roll capsid sequences when some members of the family are used in training set. The mean prediction accuracies using BLAST and SVM for 20 trials are shown, with grey bars indicating standard deviation. Only families with > 5 "unknown"/test sequences are shown, with results for all jelly-roll possessing families shown in Table 3.4. The names of the viral families are truncated to remove "-viridae" from them.

sequences. A capsid sequence from the "unknown" dataset is "detected" if the e-value obtained by BLASTP is lesser than 5% after correcting for multiple testing (Methods). To assess the performance of SVM-Caps, an SVM was first trained to classify the "known" dataset of capsid sequences against human proteins and viral polymerases and reverse transcriptases as mentioned above. The predictions of this SVM classifier were used to "detect" capsid sequences from the "unknown" dataset. The family-wise prediction powers for BLASTP and SVM, averaged over 20 random realizations of "known" and "unknown" datasets, are shown in Fig. 3.3 and Table 3.4. We found that BLASTP has a higher prediction accuracy (average accuracy = 96%) than SVM (average accuracy = 76%) for all viral families, though for some families the performance is comparable (Table 3.4).

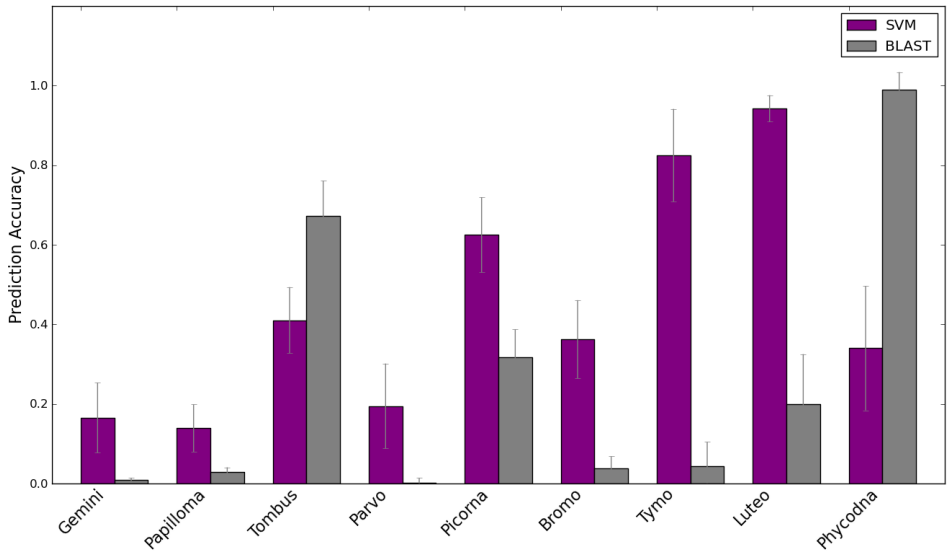

Figure 3.4: Family-wise comparison of performance of BLAST and SVM classifier to detect test jelly-roll capsid sequences when members of the family are *not* used in training set. The mean prediction accuracies using BLAST and SVM for 20 trials are shown, with grey bars indicating standard deviation. Only families with > 5 "unknown"/test sequences are shown, with results for all jelly-roll possessing families shown in Table 3.4. The names of the viral families are truncated to remove "-viridae" from them.

For the second scenario, all capsid sequences from a given viral family were assigned to the "unknown" dataset, and the capsid sequences from other families were retained in the "known" dataset (using the above mentioned scheme to ensure equivalent weightage for each family). The subsequent analysis as described above for the first scenario was repeated, using such "known" and "unknown" datasets for each viral family having the jelly-roll fold. As expected, both SVM-Caps and BLASTP do not perform well in this scenario, with average prediction accuracies of 26% and 11% respectively. But, SVM-Caps was found to have a significantly higher prediction accuracy than BLASTP ($p < 0.0001$, see Fig. 3.4, Table 3.4). Thus, these results together suggest that while BLAST can be more useful for detection of novel viral sequences from extant families, the SVM based approach can be more successful than BLAST for detection of novel viral families.

Table 3.4: Comparison of BLASTP and SVM-Caps performance in situations mimicking detection of novel capsids from extant jelly-roll fold possessing families, and novel capsids from novel jelly-roll possessing families. The accuracies listed are averaged over 20 trials (maximum accuracy = 1). Bold entries indicate significantly better performance for BLAST/SVM ($p < 0.05$).

Viral Family	**Number of test samples**	**Family used in training**		**Family not used in training**	
		SVM-Caps Accuracy	**BLASTP Accuracy**	**SVM-Caps Accuracy**	**BLASTP Accuracy**
Adenoviridae	3	0.38 ± 0.24	0.68 ± 0.2	0 ± 0	0 ± 0
Ascoviridae	1	1 ± 0	1 ± 0	0.3 ± 0.46	1 ± 0
Bromoviridae	16	0.5 ± 0.09	0.88 ± 0.09	0.36 ± 0.1	0.04 ± 0.03
Caliciviridae	3	0.55 ± 0.34	1 ± 0	0.28 ± 0.26	0.7 ± 0.28
Corticoviridae	1	0 ± 0	0 ± 0	0 ± 0	0 ± 0
Dicistroviridae	1	0.65 ± 0.48	1 ± 0	0.65 ± 0.48	0.8 ± 0.4
Geminiviridae	212	0.87 ± 0.04	$\mathbf{0.99 \pm 0}$	$\mathbf{0.17 \pm 0.09}$	0.01 ± 0
Iridoviridae	2	1 ± 0	1 ± 0	0.43 ± 0.29	1 ± 0
Luteoviridae	13	0.94 ± 0.04	1 ± 0	$\mathbf{0.94 \pm 0.03}$	0.2 ± 0.12
Microviridae	4	0.64 ± 0.23	0.9 ± 0.12	0.46 ± 0.29	0 ± 0
Mimiviridae	1	0.7 ± 0.46	1 ± 0	0.7 ± 0.46	1 ± 0
Nodaviridae	4	0.8 ± 0.2	0.88 ± 0.22	0.5 ± 0.25	0.59 ± 0.25
Papillomaviridae	102	0.62 ± 0.07	$\mathbf{0.97 \pm 0.08}$	$\mathbf{0.14 \pm 0.06}$	0.03 ± 0.01
Parvoviridae	19	0.58 ± 0.12	0.86 ± 0.08	0.19 ± 0.11	0 ± 0.01

Phycodnaviridae	5	0.62 ± 0.2	1 ± 0	0.34 ± 0.16	0.99 ± 0.04
Picornaviridae	18	0.7 ± 0.11	0.85 ± 0.15	0.63 ± 0.09	0.32 ± 0.07
Polyomaviridae	4	0.69 ± 0.21	0.88 ± 0.13	0.2 ± 0.17	0 ± 0
Tectiviridae	2	0.78 ± 0.25	0 ± 0	0.78 ± 0.25	0 ± 0
Tetraviridae	2	0.58 ± 0.29	0.98 ± 0.11	0.3 ± 0.33	0 ± 0
Tombusviridae	29	0.62 ± 0.08	$\mathbf{1 \pm 0}$	0.41 ± 0.08	0.67 ± 0.09
Tymoviridae	14	0.92 ± 0.09	1 ± 0	$\mathbf{0.83 \pm 0.12}$	0.04 ± 0.06

3.3.3 Detection of Novel Capsid Sequences in Metagenomic Data on the French Lake Bourget

As an application of the SVM based approach for detection novel jelly-roll capsid sequences, sequences from the viral metagenomic data obtained for water samples from the French freshwater lake, Lake Bourget [103] were analyzed. The metagenomic data collection and sequencing techniques for this study were designed to extract viral sequences, and the authors were able to show no significant contamination due to bacteria. Sequence reads were mapped to known genomes using tBLASTx (thresholds at e-value < 0.001 and BLAST bit-score > 50). They further showed that of the reads which could be mapped to known genomes, 70% mapped to viral sequences, thus indicating a high concentration of viruses in their samples (estimated at 10^7 virus-like-particles/ml of water sample). The sequence data for unknown assembled contigs (n=11,038) and unknown open reading frames (ORFs) (n=28,872) for this project were downloaded from the Metavir web-server [104].

Using SVM-Caps, we found that 1019 unknown ORFs (on 999 unknown contigs), longer than 170 amino acids, were predicted by SVM to be jelly-roll capsid protein

sequences. Because some of the identified novel capsid sequences should have intermediate sequence similarity to known capsid sequences, we extracted a subset of predicted capsid ORFs which were similar, but just below the thresholds of significance, to the known jelly-roll capsids from the dataset. We used BLASTP to extract predicted capsid ORFs similar to known jelly-roll containing capsids (pairwise BLASTP e-value < 0.0001, query size = 1019 proteins, target size = 606 proteins). This procedure resulted in 38 ORFs. Since the smallest genome of a jelly-roll exhibiting virus is 1.76kb (Porcine Circovirus [105]), we focussed on the 6 ORFs (out of the 38 in the previous step) which were on contigs longer than 1.7kb to identify possibly complete genomes (Table 3.5). Surprisingly, using DELTA BLAST [99], it was found that 2 of these 6 ORFs mapped significantly to capsid proteins from Geminiviridae family (e-value $< 10^{-20}$, using the non-redundant database), and 1 ORF mapped to coat protein from Plasmopora Halsteidii Virus A (e-value$= 3 \times 10^{-9}$, using non-redundant database) (Table 3.6). Although these ORFs were classified as "unknown" using tBLASTx (i.e. not meeting the thresholds e-value < 0.001 and bit score > 50 using non-redundant database), a more sensitive search algorithm, DELTA BLAST, was able to detect significant similarity to capsid proteins. Due to the high computation time of DELTA BLAST, its use in large scale metagenomic annotation studies is not suitable. But using SVM-Caps to filter putative capsid proteins can significantly reduce the number of ORFs. Thus, using SVM-Caps predictions to screen for putative capsid proteins followed by more elaborate downstream analysis of these putative capsid ORFs can prove to be a useful strategy.

Table 3.5: Pairwise BLASTP results for putative novel capsid ORFs with known Jelly-roll containing Capsid Sequences

ORF	Target Capsid (GI[6])	Percent identity	Alignment length	E-value	Bit Score
contig18897 gene 3	Maize streak virus (9625667)	23.53	136	2.00×10^{-6}	33.5
contig18897 gene 3	Eragrostis curvula streak virus (229605060)	28.33	60	6.00×10^{-6}	32
contig37564 gene 2	Hibiscus chlorotic ringspot virus (20153394)	26.09	115	2.00×10^{-5}	31.6
contig18897 gene 3	Tobacco yellow dwarf virus (20564137)	26.92	104	2.00×10^{-5}	30
contig18897 gene 3	Wheat dwarf virus (18071200)	25	84	2.00×10^{-5}	30.4
contig20303 gene 2	Canine papillomavirus 4 (164429764)	28.99	69	2.00×10^{-5}	31.2
contig37577 gene 3-partial	Equus caballus papillomavirus 1 (20428635)	26.4	125	2.00×10^{-5}	31.2
contig37537 gene 1-partial	Sweet potato leaf curl South Carolina virus (327409463)	25.53	94	3.00×10^{-5}	28.9
contig18897 gene 3	Sweet potato leaf curl virus (29294540)	24.77	109	4.00×10^{-5}	29.6

[6]RefSeq identifier

contig21296 gene 1-partial	Pseudoalteromonas phage PM2 (9632869)	30.19	53	7.00×10^{-5}	28.5

Table 3.6: DELTA BLAST results for putative capsid ORFs with significant similarity to capsid proteins using non-redundant database

Target Protein	**DELTA BLAST e-value**	**Percent Identity**	**Alignment length**
Contig 18897 Gene 3			
coat protein [Tomato yellow leaf curl virus - Il]	2.00×10^{-28}	17	165
coat protein [Tomato yellow leaf curl Mali virus]	5.00×10^{-28}	16	165
coat protein [Tomato leaf curl virus]	5.00×10^{-28}	19	258
coat protein [Honeysuckle yellow vein mosaic virus]	3.00×10^{-27}	18	146
coat protein [Tomato yellow leaf curl virus-[Minab:Iran]]	4.00×10^{-27}	16	165

Contig 37537 Gene 1 (partial)

Coat protein [Tomato yellow leaf curl virus - Il]	2.00×10^{-22}	17	107
AV1 [Premna leaf curl virus]	3.00×10^{-22}	17	107
coat protein [Tomato yellow leaf curl Mali virus]	4.00×10^{-22}	18	107
coat protein, partial [Tomato yellow leaf curl virus]	7.00×10^{-22}	17	106
coat protein, partial [Watermelon chlorotic stunt virus]	1.00×10^{-21}	18	107
Contig 37564 Gene 2			
putative coat protein, partial [Plasmopara halstedii virus A]	3.00×10^{-9}	30	123
coat protein [Sclerophthora macrospora virus A]	0.014	26	117

Using tertiary structure modelling, we found that the 3 ORFs mentioned above showed characteristic topology similar to the jelly-roll fold. We first used CPHmodels3.2 web-server [102] to identify suitable template protein structures used to model the tertiary structures of the unknown putative capsid ORFs. The best structural templates predicted were the capsid protein of Satellite Tobacco Necrosis Virus (PDB id: 2BUK [106]) for two ORFs, and for one was capsid protein of Tomato Bushy Stunt Virus (PDB id: 2TBV [107]) for one ORF. Both these protein structures are known to possess a single copy of the jelly-roll motif. Using these as optional user-submitted

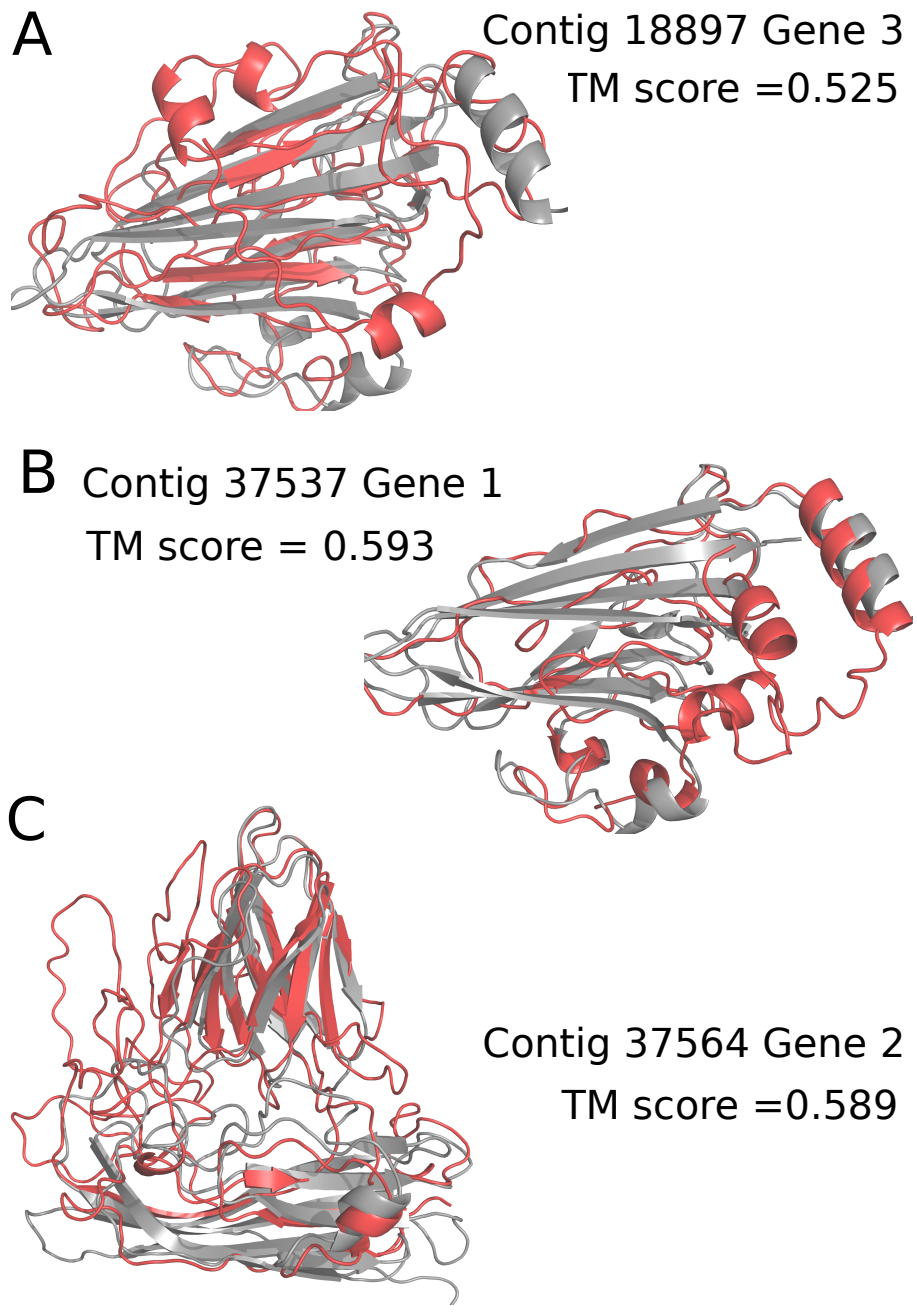

Figure 3.5: Alignment of predicted structure for putative novel jelly-roll containing capsids with structurally most similar known jelly-roll containing capsids. Contig 18897 Gene 3 (A) and Contig 37537 Gene 1 (B) were found to be structurally closest to Satellite Tobacco Necrosis Virus capsid (PDB id: 2BUK/2STV), and Contig 37564 Gene 2 to capsid protein of Tomato Bushy Stunt virus (PDB id: 2TBV). Grey corresponds to the known templates, and red corresponds to the predicted structures for putative capsids. TM-score, a measure of structural similarity, of greater than 0.5 indicates same topology between the template and predicted model [10].

templates for modelling, the web-server I-TASSER [10] was used to perform an in-depth structural prediction for these ORFs. We found that although the structures predicted were not significant (estimated TM-scores: 0.49 ± 0.15 for contig 18897 gene 3, 0.49 ± 0.15 for contig 37537 gene 1, and 0.36 ± 0.12 for contig 37564 gene 2), some of them could have the correct fold (TM-score > 0.5 correlates with the same fold). The structural alignment between predicted models and the given templates indicates that they have similar topology (TM scores of 0.52-0.59) (Fig. 3.5).

None of the predicted structures exhibited the complete eight beta-stranded jelly-roll motif. The topology of the jelly-roll fold (but not the secondary structure) was exhibited by gene 3 on contig 18897 (abbreviated as 18897-3) with 6 of the 8 beta strands of the motif, as well as by gene 2 on contig 37564 (abbreviated as 37564-2) with overlap of 5 out of 7 beta strands, but not by gene 1 on contig 37537. This discrepancy in the number of beta strands in the predicted structures could arise due to limitations of the secondary structure prediction algorithm PSIPRED used by I-TASSER, which has an accuracy of 78% [108]. The predicted structure for 37564-2 also showed striking similarity to the other parts of the capsid protein structure of Tomato Bushy Stunt Virus with several overlapping secondary structures and folds (Fig. 3.5 C). Thus, taken together, the above results suggest that the ORFs 18897-3 and 37564-2 could be jelly-rold containing capsid sequences. Moreover, the low sequence identity to known capsid proteins (Table 3.6) indicates that they could be from novel viral families.

3.4 Discussion & Future Work

In this chapter, a novel method for detection novel jelly-roll containing capsid sequences was presented. The low sequence similarity between viruses, coupled with the poor representation of the enormous viral diversity in characterized databases, limits the effectiveness of sequence similarity based methods for detecting novel viral

sequences. This limitation is borne out in most of the viral metagenomic studies where a large fraction of sequence reads cannot be confidently mapped to any known viruses or other organisms. Several studies have successfully explored the use of alignment-free features in such scenarios to identify novel organisms in spite of low sequence similarity [42, 43]. Most of such studies have focused on microbial metagenomics, and to our knowledge, there exists only one method (MGTAXA) which can be used for identification of viruses using alignment-free features [44]. MGTAXA is a novel Hidden Markov Model based method which can predict the taxonomy of viral sequences using 12 nucleotide frequencies. Although such information is useful, for the problem of detecting whether a sequence is viral or not, this method is much more computationally expensive than our method. Thus, the work developed in this chapter is a crucial addition to this field.

All of the above mentioned methods relying on alignment-free features have two commonalities: a) they use frequencies of short sequence motifs, and b) they use machine learning algorithms to solve the classification problems underlying detection of novel sequences. Our method uses both of these aspects, but deviates in the former by using motifs comprised of a reduced alphabet of amino-acids, rather than of nucleotides. This choice was natural since the aim was to exploit the remarkable structural conservation of the jelly-roll motif exhibited by capsid proteins from diverse viral families. Although methods based on alignment-free features do extract useful signals from genomes, contrary to our method, they are not explicitly based on either sequence/structural conservation or other obvious biological properties (except perhaps the codon biases used by viruses). By using motifs that are based on physical properties of amino-acids and that can be robust to sequence evolution, we were able to construct a high-accuracy classifier for jelly-roll capsid proteins.

The most popular tool used in metagenomic studies to annotate novel genomic sequences is BLAST [38]. Using controlled comparative studies of simulated scenarios, I found that although BLAST performs better than SVM-Caps in detection of novel

members of the same family, it has a significantly lower performance than SVM in detecting members of novel viral families. Because, sequence similarity within viral families is expected to be higher than across families, the higher success of BLAST in the former scenario is expected given its high specificity. In spite of the absence of sequence similarity, alignment-free features can still show similarity [39], and, as in the case of the latter scenario, can lead to a higher performance for the methods using them.

Using SVM-Caps, the viral metagenomic data obtained from water samples from a freshwater lake, Lake Bourget, in France [103] were analyzed to predict the existence of several novel jelly-roll containing capsid sequences. Using tBLASTx, Roux et al. found that in this dataset only 26.4% and 18.4% reads mapped to known organisms and viruses respectively. Furthermore, 91% of the reads that mapped to viruses mapped to Microviridae viruses. This family of small single stranded DNA viruses possesses jelly-roll containing capsid proteins. It is possible that some of the large fraction of unknown sequences could belong to novel families similar to Microviridae, which possess jelly-roll capsid proteins. Thus, our finding of 999 unknown contigs with putative jelly-roll capsid sequences, is plausible. In the absence of methods to validate these unknown contigs en-masse, we used the following way of validating some of these contigs. We reasoned that some of these predicted novel capsid proteins might show intermediate sequence-similarity to known capsid proteins. Using this approach, three candidate sequences were isolated, which showed significant sequence similarity to known capsid proteins using the more sensitive sequence similarity based algorithm DELTA-BLAST [99]. These results are expected as it was shown that DELTA-BLAST is more powerful at detecting sequence similarity between divergent sequences as compared to BLAST [99]. Two out these three novel putative capsids were then showed to exhibit partial structural similarity to jelly-roll containing capsids using protein structure modeling, further validating these candidate ORFs as jelly-roll containing capsid proteins. These results confirm the potential of SVM classifier

to identify putative novel jelly-roll capsid sequences from metagenomic data.

The limitations of this work suggest several future avenues of research. The requirement of intermediate sequence similarity restricted the focus to only few of the ORFs predicted to be capsids. The maximum utility of SVM-Caps lies in detecting divergent capsid sequences. Thus, it would be desirable to be able to use an independent alignment-free method to validate these predicted ORFs as jelly-roll capsids. One possible way could be to use similarity in feature space as a metric. Although intuitive, this approach would need more testing especially since such a metric has the potential to give taxonomic information. Another area of concern which demands future research is to study the sensitivity of SVM-Caps to artefacts of metagenomic assembly. Since metagenomics involves assembly of reads obtained from sequencing environmental DNA/RNA, often hybrid assembled contigs can arise which comprise of genomic sequences from multiple species. It is not clear how sensitive the SVM performance will be on such genomic chimeras, and this issue needs further study.

Chapter 4
Future Work

The research presented in this dissertation suggests several fruitful lines of research. In chapter 2, I discussed the use of a novel strategy to understand transmission bias in H5N1 infections from birds to humans. H5N1 viruses cannot transmit between humans yet, and all human infections are transmitted from birds. Many important human infecting viruses such as Ebola, Rabies, SARS Coronavirus, Dengue, etc. have been a result of such "spillover" events, i.e. infections which jumped from other animals into humans. In most of these viruses, after an initial spillover event from animal to human, human-to-human transmission is also observed. The methods developed in this thesis can be used to understand not only the transmission bias in the initial spillover events, but can also be used to detect selection on short time scales in the latter stages of human-to-human transmission. On the theoretical side, the method introduced by Pan and Deem [79], which enables the calculation of the probability of neutral evolution of amino-acid frequencies on short time scales, assumes the scenario of infinite effective population size. To be applicable to the study of situations such as intra-host micro-evolution of viruses, this method needs to be modified for finite, possibly small, effective population sizes.

In Chapter 3, a novel method was developed to detect unknown capsid sequences using machine learning algorithm on alignment-free features. This approach exploited the conservation of a particular structural fold, the "jelly-roll" fold, in the capsid proteins of viruses from several diverse families. Given the high accuracy of this method exhibits to classify such capsid sequences against virtually all other proteins, it is reasonable to expect that certain essential features underlying the jelly-roll fold

are being used in the classification. An in-depth study of extracting such features and understanding their prevalence with respect the jelly-roll folds of the diverse families could provide clues into the sequence determinants of this conserved fold. Furthermore, Abrescia et al. [6] have discovered other conserved structural motifs in capsid proteins. The methods developed in this chapter can be employed to detect these other type of conserved capsid sequences. Bacteriophages are probably the most numerous organisms in the biosphere, and a large class of them have capsid proteins with a conserved structural motif (distinct from the jelly-roll fold). Thus, an alignment-free method to detect novel capsid sequences from such bacteriophage families would be highly relevant for the viral metagenomic community.

Appendix A

Closely clustering Avian and Human H5N1 Isolates

Table A.1: Closely clustering avian and human isolates identified using a distance cutoff in the principal component space

Accession Number	Host	Date	Virus Name
Egypt			
ABM92273	Human	2007/01	A/Egypt/0636-NAMRU3/2007(H5N1)
ACI06187	Human	12/24/07	A/Egypt/10211-NAMRU3/2007(H5N1)
ACI06188	Human	12/26/07	A/Egypt/10215-NAMRU3/2007(H5N1)
ACI06189	Human	12/29/07	A/Egypt/10216-NAMRU3/2007(H5N1)
ADG21405	Human	12/29/07	A/Egypt/10217-NAMRU3/2007(H5N1)
ABJ90343	Human	2006/10	A/Egypt/12374-NAMRU3/2006(H5N1)
ABP96845	Human	2007	A/Egypt/1394-NAMRU3/2007(H5N1)
ABM54179	Human	2006	A/Egypt/14724-NAMRU3/2006(H5N1)
ABP96847	Human	2007	A/Egypt/1731-NAMRU3/2007(H5N1)
ABP96848	Human	2007	A/Egypt/1902-NAMRU3/2007(H5N1)
ACI06180	Human	02/26/08	A/Egypt/1980-NAMRU3/2008(H5N1)
ABP96849	Human	2007	A/Egypt/2256-NAMRU3/2007(H5N1)
ACI06181	Human	03/02/08	A/Egypt/2289-NAMRU3/2008(H5N1)

ABP96850	Human	03/13/07	A/Egypt/2321-NAMRU3/2007(H5N1)
ACI06182	Human	03/04/08	A/Egypt/2514-NAMRU3/2008(H5N1)
ACI06183	Human	03/08/08	A/Egypt/2546-NAMRU3/2008(H5N1)
ABP96852	Human	2007	A/Egypt/2616-NAMRU3/2007(H5N1)
ABP96853	Human	2007	A/Egypt/2620-NAMRU3/2007(H5N1)
ABU53968	Human	2007	A/Egypt/2629-NAMRU3/2007(H5N1)
ABU53971	Human	2007	A/Egypt/2750-NAMRU3/2007(H5N1)
ABU53972	Human	2007	A/Egypt/2751-NAMRU3/2007(H5N1)
ABK32775	Human	2006	A/Egypt/2763-NAMRU3/2006(H5N1)
ABK32778	Human	2006	A/Egypt/2947-NAMRU3/2006(H5N1)
ABU53966	Human	2006	A/Egypt/2991-NAMRU3/2006(H5N1)
ACI06184	Human	04/05/08	A/Egypt/3158-NAMRU3/2008(H5N1)
ACI06186	Human	04/16/08	A/Egypt/3401-NAMRU3/2008(H5N1)
ABU53973	Human	2007	A/Egypt/4081-NAMRU3/2007(H5N1)
ABU53974	Human	2007	A/Egypt/4082-NAMRU3/2007(H5N1)
ABU53975	Human	2007	A/Egypt/4226-NAMRU3/2007(H5N1)
ABK32782	Human	2006	A/Egypt/5614-NAMRU3/2006(H5N1)
ABU53976	Human	2007	A/Egypt/6251-NAMRU3/2007(H5N1)
ADG21402	Human	05/16/06	A/Egypt/7021-NAMRU3/2006(H5N1)
ACT15310	Human	01/11/09	A/Egypt/N00001/2009(H5N1)
ADG21427	Human	01/11/10	A/Egypt/N00269/2010(H5N1)
ADG21429	Human	01/12/10	A/Egypt/N00270/2010(H5N1)
ACT15312	Human	01/23/09	A/Egypt/N00585/2009(H5N1)
ACT15314	Human	02/03/09	A/Egypt/N00605/2009(H5N1)
ACT15316	Human	02/07/09	A/Egypt/N00606/2009(H5N1)
ADG21431	Human	02/02/10	A/Egypt/N01360/2010(H5N1)
ADG21435	Human	02/12/10	A/Egypt/N01982/2010(H5N1)
ADG21437	Human	02/15/10	A/Egypt/N02038/2010(H5N1)

ACT15320	Human	03/03/09	A/Egypt/N02039/2009(H5N1)
ADG21439	Human	02/13/10	A/Egypt/N02127/2010(H5N1)
ACT15322	Human	03/09/09	A/Egypt/N02407/2009(H5N1)
ADG21441	Human	02/23/10	A/Egypt/N02554/2010(H5N1)
ACT15324	Human	03/14/09	A/Egypt/N02563/2009(H5N1)
ACT15326	Human	03/24/09	A/Egypt/N02752/2009(H5N1)
ADG21443	Human	02/27/10	A/Egypt/N02770/2010(H5N1)
ADG21445	Human	03/03/10	A/Egypt/N03071/2010(H5N1)
ADG21447	Human	03/07/10	A/Egypt/N03072/2010(H5N1)
ACT15328	Human	03/30/09	A/Egypt/N03228/2009(H5N1)
ACT15330	Human	04/01/09	A/Egypt/N03272/2009(H5N1)
ACT15332	Human	04/15/09	A/Egypt/N03434/2009(H5N1)
ACT15334	Human	04/16/09	A/Egypt/N03438/2009(H5N1)
ACT15338	Human	04/19/09	A/Egypt/N03450/2009(H5N1)
ACT15342	Human	05/11/09	A/Egypt/N04394/2009(H5N1)
ACT15345	Human	05/17/09	A/Egypt/N04396/2009(H5N1)
ADG21449	Human	03/31/10	A/Egypt/N04434/2010(H5N1)
ACT15347	Human	05/18/09	A/Egypt/N04526/2009(H5N1)
ACT15349	Human	05/18/09	A/Egypt/N04527/2009(H5N1)
ACT15351	Human	05/25/09	A/Egypt/N04822/2009(H5N1)
ACT15353	Human	05/25/09	A/Egypt/N04823/2009(H5N1)
ADG21407	Human	05/29/09	A/Egypt/N04830/2009(H5N1)
ACT15357	Human	06/06/09	A/Egypt/N05056/2009(H5N1)
ADG21410	Human	06/16/09	A/Egypt/N05912/2009(H5N1)
ADG21412	Human	07/25/09	A/Egypt/N07392/2009(H5N1)
ADG21416	Human	08/01/09	A/Egypt/N08835/2009(H5N1)
ADG21418	Human	08/26/09	A/Egypt/N09174/2009(H5N1)
ADG21422	Human	11/22/09	A/Egypt/N11981/2009(H5N1)

ADG21425	Human	2009/12	A/Egypt/N15262/2009(H5N1)
ACR56220	Avian	05/24/07	A/chicken/Egypt/07175-NLQP/2007(H5N1)
ACA29670	Avian	06/19/07	A/chicken/Egypt/07181-NLQP/2007(H5N1)
ACA29679	Avian	12/24/07	A/chicken/Egypt/07665S-NLQP/2007(H5N1)
ACR56224	Avian	01/04/08	A/chicken/Egypt/0811-NLQP/2008(H5N1)
ACR56251	Avian	01/10/08	A/chicken/Egypt/08124S-NLQP/2008(H5N1)
ADD21350	Avian	2008/01	A/chicken/Egypt/0814S-NLQP/2008(H5N1)
ACR56227	Avian	01/07/08	A/chicken/Egypt/0823-NLQP/2008(H5N1)
ACA29683	Avian	01/03/08	A/chicken/Egypt/0836S-NLQP/2008(H5N1)
ACR56254	Avian	02/09/08	A/chicken/Egypt/08371S-NLQP/2008(H5N1)
ACR56230	Avian	01/16/08	A/chicken/Egypt/0838-NLQP/2008(H5N1)
ACR56233	Avian	01/18/08	A/chicken/Egypt/0847-NLQP/2008(H5N1)
ACR56223	Avian	01/03/08	A/chicken/Egypt/085-NLQP/2008(H5N1)
ACR56234	Avian	01/22/08	A/chicken/Egypt/0850-NLQP/2008(H5N1)
ACR56239	Avian	02/17/08	A/chicken/Egypt/0870-NLQP/2008(H5N1)
ADD21349	Avian	2008/06	A/chicken/Egypt/0883-NLQP/2008(H5N1)
ACR56247	Avian	12/24/08	A/chicken/Egypt/0896-NLQP/2008(H5N1)
ADM85845	Avian	2009/12	A/chicken/Egypt/091317s/2009(H5N1)
ADD21355	Avian	2009/02	A/chicken/Egypt/0915-NLQP/2009(H5N1)
ACX31965	Avian	2009/01	A/chicken/Egypt/092-NLQP/2009(H5N1)
ACX31993	Avian	2009/02	A/chicken/Egypt/0920-NLQP/2009(H5N1)

ADD21378	Avian	2009/01	A/chicken/Egypt/093smg-NLQP/2009(H5N1)
ADD21365	Avian	2009/05	A/chicken/Egypt/09485s-NLQP/2009(H5N1)
ACX31973	Avian	2009/03	A/chicken/Egypt/0960-NLQP/2009(H5N1)
ADB85109	Avian	05/12/09	A/chicken/Egypt/0987-NLQP/2009(H5N1)
AEQ72831	Avian	09/09/10	A/chicken/Egypt/10117/2010(H5N1)
AEQ72839	Avian	10/21/10	A/chicken/Egypt/10132/2010(H5N1)
ADM85868	Avian	2010/02	A/chicken/Egypt/101604v/2010(H5N1)
ADM85880	Avian	2010/03	A/chicken/Egypt/10189s/2010(H5N1)
ADM85881	Avian	2010/03	A/chicken/Egypt/1020d/2010(H5N1)
AEQ72813	Avian	04/13/10	A/chicken/Egypt/1021AD/2010(H5N1)
ADM85883	Avian	2010/03	A/chicken/Egypt/1021L/2010(H5N1)
ADM85888	Avian	2010/05	A/chicken/Egypt/1022L/2010(H5N1)
AEQ72825	Avian	08/01/10	A/chicken/Egypt/10249SF/2010(H5N1)
AEQ72830	Avian	08/08/10	A/chicken/Egypt/10259SF/2010(H5N1)
AEQ72827	Avian	08/04/10	A/chicken/Egypt/10264AG/2010(H5N1)
ADM85852	Avian	2010/01	A/chicken/Egypt/1029/2010(H5N1)
ADM85854	Avian	2010/02	A/chicken/Egypt/1034/2010(H5N1)
ADM85871	Avian	2010/02	A/chicken/Egypt/1034qd/2010(H5N1)
AEQ72828	Avian	08/05/10	A/chicken/Egypt/1038AL/2010(H5N1)
AEQ72842	Avian	12/10/10	A/chicken/Egypt/10413SF/2010(H5N1)
AEQ72835	Avian	09/29/10	A/chicken/Egypt/10513S/2010(H5N1)
ADM85875	Avian	2010/02	A/chicken/Egypt/1052g/2010(H5N1)
ADM85885	Avian	2010/03	A/chicken/Egypt/1058sf/2010(H5N1)
ABN70706	Avian	2006	A/chicken/Egypt/1078-NAMRU3/2006(H5N1)

ABN70707	Avian	2007	A/chicken/Egypt/1079-NAMRU3/2007(H5N1)
ABN70708	Avian	2006	A/chicken/Egypt/1080-NAMRU3/2006(H5N1)
ABN70709	Avian	2006	A/chicken/Egypt/1081-NAMRU3/2006(H5N1)
AEQ72816	Avian	06/08/10	A/chicken/Egypt/1090/2010(H5N1)
AEQ72869	Avian	01/27/11	A/chicken/Egypt/111127V/2011(H5N1)
AEQ72850	Avian	01/10/11	A/chicken/Egypt/1112/2011(H5N1)
AEQ72877	Avian	02/23/11	A/chicken/Egypt/111640V/2011(H5N1)
AEQ72896	Avian	03/21/11	A/chicken/Egypt/1117AF/2011(H5N1)
AEQ72879	Avian	02/27/11	A/chicken/Egypt/11184S/2011(H5N1)
AEQ72890	Avian	03/13/11	A/chicken/Egypt/111945V/2011(H5N1)
AFI44347	Avian	10/05/11	A/chicken/Egypt/1119AF/2011(H5N1)
AEQ72903	Avian	04/26/11	A/chicken/Egypt/1123AL/2011(H5N1)
AEQ72845	Avian	01/04/11	A/chicken/Egypt/112SG/2011(H5N1)
AEQ72884	Avian	03/03/11	A/chicken/Egypt/1134SD/2011(H5N1)
AEQ72861	Avian	01/20/11	A/chicken/Egypt/113AF/2011(H5N1)
AFI44339	Avian	07/08/11	A/chicken/Egypt/11506SF/2011(H5N1)
AEQ72905	Avian	06/13/11	A/chicken/Egypt/11529S/2011(H5N1)
AEQ72894	Avian	03/17/11	A/chicken/Egypt/1155/2011(H5N1)
AEQ72856	Avian	01/12/11	A/chicken/Egypt/1156S/2011(H5N1)
AEQ72868	Avian	01/27/11	A/chicken/Egypt/1158SF/2011(H5N1)
AFI44345	Avian	07/27/11	A/chicken/Egypt/11667s/2011(H5N1)
AFI44343	Avian	07/26/11	A/chicken/Egypt/11672s/2011(H5N1)
AEQ72870	Avian	02/07/11	A/chicken/Egypt/116AF/2011(H5N1)
AEQ72857	Avian	01/14/11	A/chicken/Egypt/1174S/2011(H5N1)
AEQ72872	Avian	02/14/11	A/chicken/Egypt/117AF/2011(H5N1)

AEQ72874	Avian	02/19/11	A/chicken/Egypt/1197AG/2011(H5N1)
AEQ72844	Avian	01/04/11	A/chicken/Egypt/119S/2011(H5N1)
AFI44358	Avian	2012/02	A/chicken/Egypt/12186F-12/2012(H5N1)
AFI44357	Avian	2012/02	A/chicken/Egypt/12186F-9/2012(H5N1)
AFI44356	Avian	01/12/12	A/chicken/Egypt/1219s/2012(H5N1)
ABO64688	Avian	2006	A/chicken/Egypt/12378N3-CLEVB/2006(H5N1)
ABO64689	Avian	2006	A/chicken/Egypt/12379N3-CLEVB/2006(H5N1)
AFI44355	Avian	01/09/12	A/chicken/Egypt/128s/2012(H5N1)
ABN70710	Avian	2007	A/chicken/Egypt/1300-NAMRU3/2007(H5N1)
ACM68979	Avian	02/23/08	A/chicken/Egypt/15NLQP-CLEVB244/2008(H5N1)
ACD64996	Avian	02/25/07	A/chicken/Egypt/1709-1VIR08/2007(H5N1)
ABO64697	Avian	2007	A/chicken/Egypt/1892N3-HK49/2007(H5N1)
ACM68984	Avian	01/25/08	A/chicken/Egypt/22NLQP-CLEVB232/2008(H5N1)
ABY79009	Avian	2007	A/chicken/Egypt/3044NAMRU3-CLEVB59/2007(H5N1)
ABY79010	Avian	2007	A/chicken/Egypt/3045NAMRU3-CLEVB60/2007(H5N1)
ABY79011	Avian	2007	A/chicken/Egypt/3046NAMRU3-CLEVB62/2007(H5N1)
ABY79014	Avian	2007	A/chicken/Egypt/3049NAMRU3-CLEVB75/2007(H5N1)

ABG67712	Avian	2006	A/chicken/Egypt/5611NAMRU3-AN/2006(H5N1)
ABY79019	Avian	2007	A/chicken/Egypt/9383NAMRU3-CLEVB112/2007(H5N1)
ABY79033	Avian	2007	A/chicken/Egypt/9400NAMRU3-CLEVB211/2007(H5N1)
ABD85144	Avian	2006	A/chicken/Egypt/960N3-004/2006(H5N1)
ABW37430	Avian	03/06/07	A/chicken/Egypt/F6/2007(H5N1)
ADY16731	Avian	03/24/09	A/chicken/Egypt/F8/2009(H5N1)
AEP37319	Avian	01/19/11	A/chicken/Egypt/M2773A/2011(H5N1)
ABW37431	Avian	12/25/06	A/chicken/Egypt/R1/2006(H5N1)
ABW37432	Avian	01/01/07	A/chicken/Egypt/R2/2007(H5N1)
ABW37436	Avian	02/26/07	A/chicken/Egypt/R6/2007(H5N1)
BAK23400	Avian	2008/11	A/chicken/Egypt/RIMD12-3/2008(H5N1)
BAK23402	Avian	2008/06	A/chicken/Egypt/RIMD5-3/2008(H5N1)
AEP84526	Avian	02/28/11	A/chicken/Egypt/S2938A/2011(H5N1)
AEP37324	Avian	05/12/11	A/chicken/Egypt/S3280B/2011(H5N1)
ACJ61696	Avian	02/22/06	A/chicken/Qalubia/1/2006(H5N1)
ACA29675	Avian	03/20/07	A/duck/Egypt/07322S-NLQP/2007(H5N1)
ACR56249	Avian	01/03/08	A/duck/Egypt/0845S-NLQP/2008(H5N1)
ADD21352	Avian	2008/04	A/duck/Egypt/08561S-NLQP/2008(H5N1)
ACU16727	Avian	02/20/08	A/duck/Egypt/0871/2008(H5N1)
ACR56240	Avian	04/06/08	A/duck/Egypt/0875-NLQP/2008(H5N1)
ACR56245	Avian	09/26/08	A/duck/Egypt/0891-NLQP/2008(H5N1)
AEA92628	Avian	2008/12	A/duck/Egypt/0897-NLQP/2008(H5N1)
ADD21369	Avian	2009/02	A/duck/Egypt/09118sm-NLQP/2009(H5N1)
ACX31969	Avian	2009/02	A/duck/Egypt/0926-NLQP/2009(H5N1)

ADD21372	Avian	2009/04	A/duck/Egypt/09274sm-NLQP/2009(H5N1)
ACX31992	Avian	2009/01	A/duck/Egypt/093-NLQP/2009(H5N1)
ADD21380	Avian	2009/02	A/duck/Egypt/0930smL-NLQP/2009(H5N1)
ADD21364	Avian	2009/04	A/duck/Egypt/09315s-NLQP/2009(H5N1)
ADD21374	Avian	2009/05	A/duck/Egypt/09332sm-NLQP/2009(H5N1)
ACX31997	Avian	2009/04	A/duck/Egypt/09339S-NLQP/2009(H5N1)
ACX31984	Avian	2009/04	A/duck/Egypt/09349S-NLQP/2009(H5N1)
ADD21368	Avian	2009/01	A/duck/Egypt/0934sm-NLQP/2009(H5N1)
ADD21375	Avian	2009/02	A/duck/Egypt/0945smf-NLQP/2009(H5N1)
ADD21376	Avian	2009/02	A/duck/Egypt/0967smf-NLQP/2009(H5N1)
ADD21359	Avian	2009/04	A/duck/Egypt/0970-NLQP/2009(H5N1)
ACX31975	Avian	2009/04	A/duck/Egypt/0972-NLQP/2009(H5N1)
ACX31966	Avian	2009/01	A/duck/Egypt/099-NLQP/2009(H5N1)
ACX31989	Avian	2009/02	A/duck/Egypt/0990SM-NLQP/2009(H5N1)
AEQ72832	Avian	09/14/10	A/duck/Egypt/10118/2010(H5N1)
AEQ72838	Avian	10/20/10	A/duck/Egypt/10131/2010(H5N1)
ADM85863	Avian	2010/02	A/duck/Egypt/101565v/2010(H5N1)
ADM85882	Avian	2010/03	A/duck/Egypt/10157s/2010(H5N1)
AEQ72818	Avian	06/10/10	A/duck/Egypt/10185SS/2010(H5N1)
ADM85851	Avian	2010/01	A/duck/Egypt/1022/2010(H5N1)
AEQ72821	Avian	06/28/10	A/duck/Egypt/10228SF/2010(H5N1)
AEQ72824	Avian	07/20/10	A/duck/Egypt/10255AG/2010(H5N1)

AEQ72834	Avian	09/22/10	A/duck/Egypt/10290SF/2010(H5N1)
AEQ72836	Avian	10/12/10	A/duck/Egypt/10331SF/2010(H5N1)
AEQ72837	Avian	10/14/10	A/duck/Egypt/10336SF/2010(H5N1)
ADM85847	Avian	2010/01	A/duck/Egypt/103swf/2010(H5N1)
AEQ72819	Avian	06/15/10	A/duck/Egypt/10403S/2010(H5N1)
AEQ72812	Avian	02/24/10	A/duck/Egypt/1046SF/2010(H5N1)
ADM85858	Avian	2010/02	A/duck/Egypt/1053/2010(H5N1)
ADM85884	Avian	2010/02	A/duck/Egypt/1063s/2010(H5N1)
ADM85876	Avian	2010/02	A/duck/Egypt/1068s/2010(H5N1)
ADM85873	Avian	2010/02	A/duck/Egypt/1097s/2010(H5N1)
AEQ72888	Avian	03/08/11	A/duck/Egypt/11106SF/2011(H5N1)
AEQ72880	Avian	02/28/11	A/duck/Egypt/1110AF/2011(H5N1)
AEQ72866	Avian	01/23/11	A/duck/Egypt/11117S/2011(H5N1)
AEQ72852	Avian	01/11/11	A/duck/Egypt/1113SD/2011(H5N1)
AEQ72892	Avian	03/15/11	A/duck/Egypt/1116AF/2011(H5N1)
AEQ72898	Avian	04/04/11	A/duck/Egypt/11175SF/2011(H5N1)
AEQ72889	Avian	03/09/11	A/duck/Egypt/11221S/2011(H5N1)
AEQ72854	Avian	01/11/11	A/duck/Egypt/1123SF/2011(H5N1)
AEQ72891	Avian	03/15/11	A/duck/Egypt/11246S/2011(H5N1)
AEQ72848	Avian	01/06/11	A/duck/Egypt/1125S/2011(H5N1)
AEQ72849	Avian	01/09/11	A/duck/Egypt/1130AG/2011(H5N1)
AEQ72871	Avian	02/13/11	A/duck/Egypt/1174SF/2011(H5N1)
AFI44350	Avian	11/29/11	A/duck/Egypt/11762s/2011(H5N1)
AFI44348	Avian	10/22/11	A/duck/Egypt/1187/2011(H5N1)
AFI44344	Avian	07/27/11	A/duck/Egypt/1198AS/2011(H5N1)
ABO64692	Avian	2006	A/duck/Egypt/12380N3-CLEVB/2006(H5N1)

ABO64693	Avian	2006	A/duck/Egypt/13010N3-CLEVB/2006(H5N1)
ACD64997	Avian	03/04/07	A/duck/Egypt/1709-3VIR08/2007(H5N1)
ABG81040	Avian	2006/06	A/duck/Egypt/2253-3/2006(H5N1)
ABY79008	Avian	2007	A/duck/Egypt/3043NAMRU3-CLEVB56/2007(H5N1)
ABY79012	Avian	2007	A/duck/Egypt/3047NAMRU3-CLEVB63/2007(H5N1)
ABY79833	Avian	2007	A/duck/Egypt/5169-1/2007(H5N1)
ABY79032	Avian	2007	A/duck/Egypt/9399NAMRU3-CLEVB202/2007(H5N1)
BAJ07733	Avian	2007/01	A/duck/Egypt/D2Br210/2007(H5N1)
BAJ07734	Avian	2007/01	A/duck/Egypt/D2Li234/2007(H5N1)
BAJ07736	Avian	2007/01	A/duck/Egypt/D3Li12/2007(H5N1)
ABW37429	Avian	12/25/06	A/duck/Egypt/F5/2006(H5N1)
AEP27003	Avian	12/22/10	A/duck/Egypt/M2583A/2010(H5N1)
AEP37317	Avian	12/22/10	A/duck/Egypt/M2583D/2010(H5N1)
AEP37323	Avian	02/20/11	A/duck/Egypt/M3075B/2011(H5N1)
AEP37318	Avian	12/26/10	A/duck/Egypt/Q2645C/2010(H5N1)
ABW37435	Avian	02/20/07	A/duck/Egypt/R5/2007(H5N1)
ADD21377	Avian	2009/03	A/goose/Egypt/09102smf-NLQP/2009(H5N1)
ADD21379	Avian	2009/02	A/goose/Egypt/0912smg-NLQP/2009(H5N1)
ADM85844	Avian	2009/09	A/goose/Egypt/09134sml/2009(H5N1)
ADM85874	Avian	2010/02	A/goose/Egypt/1057/2010(H5N1)
AEQ72876	Avian	02/22/11	A/goose/Egypt/11162S/2011(H5N1)
AEQ72900	Avian	04/12/11	A/goose/Egypt/11350S/2011(H5N1)

ABW37434	Avian	02/01/07	A/goose/Egypt/R4/2007(H5N1)
AEQ72906	Avian	04/03/11	A/ostrich/Egypt/11139F/2011(H5N1)
ACR56222	Avian	12/25/07	A/peacock/Egypt/07667S-NLQP/2007(H5N1)
AEQ72904	Avian	05/08/11	A/quail/Egypt/1171SG/2011(H5N1)
ABY79015	Avian	2007	A/quail/Egypt/3050NAMRU3-CLEVB77/2007(H5N1)
ABK34513	Avian	2005/12	A/teal/Egypt/14051-NAMRU3/2005(H5N1)
ACA29677	Avian	05/24/07	A/turkey/Egypt/07444S-NLQP/2007(H5N1)
ADD21370	Avian	2009/03	A/turkey/Egypt/09206sm-NLQP/2009(H5N1)
ACX31972	Avian	2009/02	A/turkey/Egypt/0959-NLQP/2009(H5N1)
AEQ72902	Avian	04/20/11	A/turkey/Egypt/112694V/2011(H5N1)
ABG67714	Avian	2006	A/turkey/Egypt/5613NAMRU3-T/2006(H5N1)
ADD13576	Avian	2007	A/turkey/Egypt/7/2007(H5N1)
ABY79031	Avian	2007	A/turkey/Egypt/9398NAMRU3-CLEVB195/2007(H5N1)
ABW37425	Avian	02/25/06	A/turkey/Egypt/F1/2006(H5N1)
ABW37426	Avian	03/01/06	A/turkey/Egypt/F2/2006(H5N1)

China

ABD28180	Human	11/01/05	A/Anhui/1/2005(H5N1)
ADG59080	Human	2005	A/Anhui/1/2005(H5N1)
AEO89065	Human	12/10/06	A/Anhui/1/2006(H5N1)
AEO89082	Human	03/17/07	A/Anhui/1/2007(H5N1)

ABD28181	Human	11/11/05	A/Anhui/2/2005(H5N1)
ADG59048	Human	2005	A/Anhui/2/2005(H5N1)
AEO89118	Human	12/04/08	A/Beijing/1/2009(H5N1)
ABR10842	Human	2006	A/China/2006(H5N1)
ABE68932	Avian	2005	A/Duck/Fujian/1734/05(H5N1)
ACJ68610	Human	12/06/05	A/Fujian/1/2005(H5N1)
ACJ68612	Human	03/04/06	A/Guangdong/01/2006(H5N1)
AEO89109	Human	02/16/08	A/Guangdong/1/2008(H5N1)
AEO89048	Human	06/03/06	A/Guangdong/2/2006(H5N1)
ABD28182	Human	11/23/05	A/Guangxi/1/2005(H5N1)
ADG59086	Human	2005	A/Guangxi/1/2005(H5N1)
AEO89100	Human	02/12/08	A/Guangxi/1/2008(H5N1)
ABI34142	Human	2006	A/Guangzhou/1/2006(H5N1)
AEO89154	Human	01/15/09	A/Guizhou/1/2009(H5N1)
AEO89030	Human	04/01/06	A/Hubei/1/2006(H5N1)
ACJ68607	Human	01/27/06	A/Hunan/1/2006(H5N1)
AEO89091	Human	01/16/08	A/Hunan/1/2008(H5N1)
AEO89136	Human	01/08/09	A/Hunan/1/2009(H5N1)
AEO89172	Human	01/23/09	A/Hunan/2/2009(H5N1)
ACB87563	Human	2007	A/Jiangsu/2/2007(H5N1)
ACJ68613	Human	12/04/05	A/Jiangxi/1/2005(H5N1)
AEO89127	Human	01/05/09	A/Shandong/1/2009(H5N1)
AEO89021	Human	03/13/06	A/Shanghai/1/2006(H5N1)
ACJ68609	Human	01/03/06	A/Sichuan/1/2006(H5N1)
AEO89039	Human	04/16/06	A/Sichuan/3/2006(H5N1)
AEO89145	Human	01/10/09	A/Xinjiang/1/2009(H5N1)
ABG23657	Human	2006	A/Zhejiang/16/2006(H5N1)
ACN39410	Avian	01/28/07	A/chicken/Anhui/1089/2007(H5N1)

ABJ96761	Avian	2005	A/chicken/Fujian/11933/2005(H5N1)
ABJ96763	Avian	2005	A/chicken/Fujian/12239/2005(H5N1)
ABJ97047	Avian	2006	A/chicken/Guangxi/683/2006(H5N1)
ADG59055	Avian	2008	A/chicken/Guizhou/7/2008(H5N1)
ACN39419	Avian	01/29/07	A/chicken/Hubei/2856/2007(H5N1)
ACN39421	Avian	01/29/07	A/chicken/Hubei/3002/2007(H5N1)
ADG59050	Avian	2009	A/chicken/Hunan/1/2009(H5N1)
ACN39415	Avian	01/23/07	A/chicken/Hunan/1793/2007(H5N1)
ADG59088	Avian	2005	A/chicken/Hunan/21/2005(H5N1)
ACN39422	Avian	11/30/06	A/chicken/Hunan/3157/2006(H5N1)
ADG59069	Avian	2009	A/chicken/Shandong/A-1/2009(H5N1)
ABJ96712	Avian	2006	A/chicken/Shantou/1233/2006(H5N1)
ABJ96718	Avian	2006	A/chicken/Shantou/3840/2006(H5N1)
ADG59062	Avian	2008	A/chicken/Tibet/6/2008(H5N1)
ADG59083	Avian	2005	A/duck/Anhui/56/2005(H5N1)
ADD10580	Avian	12/15/08	A/duck/Eastern China/008/2008(H5N5)
ADD10569	Avian	01/15/09	A/duck/Eastern China/031/2009(H5N5)
ADD10558	Avian	12/15/08	A/duck/Eastern China/108/2008(H5N1)
ADD10547	Avian	01/15/09	A/duck/Eastern China/909/2009(H5N1)
ABJ96762	Avian	2005	A/duck/Fujian/12032/2005(H5N1)
ABJ96765	Avian	2006	A/duck/Fujian/668/2006(H5N1)
ABJ96963	Avian	2006	A/duck/Guangxi/1258/2006(H5N1)
ABJ96961	Avian	2006	A/duck/Guangxi/1436/2006(H5N1)
ABJ96675	Avian	2006	A/duck/Guangxi/150/2006(H5N1)
ABJ96956	Avian	2006	A/duck/Guangxi/1830/2006(H5N1)
ABJ96955	Avian	2006	A/duck/Guangxi/2143/2006(H5N1)
ABJ96679	Avian	2006	A/duck/Guangxi/392/2006(H5N1)
ABJ96668	Avian	2005	A/duck/Guangxi/5075/2005(H5N1)

ABJ96672	Avian	2005	A/duck/Guangxi/5457/2005(H5N1)
ABJ96682	Avian	2006	A/duck/Guangxi/744/2006(H5N1)
ABJ96706	Avian	2006	A/duck/Guiyang/1260/2006(H5N1)
ACN39412	Avian	12/30/06	A/duck/Henan/1647/2006(H5N1)
ACN39414	Avian	12/30/06	A/duck/Henan/1652/2006(H5N1)
ACN39420	Avian	01/29/07	A/duck/Hubei/2911/2007(H5N1)
ADG59047	Avian	2005	A/duck/Hubei/49/2005(H5N1)
ACF16400	Avian	2006	A/duck/Hubei/Hangmei01/2006(H5N1)
ACN39416	Avian	01/23/07	A/duck/Hunan/1930/2007(H5N1)
ACN39417	Avian	01/23/07	A/duck/Hunan/1964/2007(H5N1)
ACN39418	Avian	01/23/07	A/duck/Hunan/1994/2007(H5N1)
ACN39423	Avian	11/30/06	A/duck/Hunan/3315/2006(H5N1)
ACN39424	Avian	11/30/06	A/duck/Hunan/3340/2006(H5N1)
ACN39409	Avian	12/15/06	A/duck/Hunan/689/2006(H5N1)
ADG59076	Avian	2005	A/duck/Jiangxi/80/2005(H5N1)
ADC97015	Avian	12/15/08	A/duck/Shandong/009/2008(H5N1)
ABJ96709	Avian	2005	A/duck/Shantou/13323/2005(H5N1)
ACH85377	Avian	2006	A/duck/Yunnan/4873/2006(H5N1)
ACH85399	Avian	2006	A/duck/Yunnan/6490/2006(H5N1)
ABJ96662	Avian	2005	A/goose/Guangxi/4289/2005(H5N1)
ABJ96673	Avian	2006	A/goose/Guangxi/52/2006(H5N1)
ABJ96671	Avian	2005	A/goose/Guangxi/5414/2005(H5N1)
ADG59046	Avian	2005	A/goose/Hubei/65/2005(H5N1)
ABJ96716	Avian	2006	A/goose/Shantou/3265/2006(H5N1)
ABJ96717	Avian	2006	A/goose/Shantou/3295/2006(H5N1)
ABJ96748	Avian	2006	A/goose/Yunnan/1143/2006(H5N1)
ACH85443	Avian	2006	A/goose/Yunnan/4985/2006(H5N1)
ABJ96742	Avian	2005	A/goose/Yunnan/6169/2005(H5N1)

ACH85509	Avian	2006	A/goose/Yunnan/6193/2006(H5N1)
ACZ54018	Avian	01/10/07	A/lesser kestrel/Heilongjiang/194/2007(H5N1)
ABW21647	Avian	2005	A/mallard/Huadong/S/2005(H5N1)
ABW21657	Avian	2005	A/mallard/Huadong/lk/2005(H5N1)
ADG59077	Avian	2006	A/shrike/Tibet/13/2006(H5N1)
ACR48937	Avian	2008	A/tree sparrow/Jiangsu/1/2008(H5N1)
ABX83938	Avian	2005	A/wild duck/Hunan/021/2005(H5N1)
ABX83951	Avian	2005	A/wild duck/Hunan/211/2005(H5N1)

Indonesia

ABU99134	Avian	2006	A/Chicken/Indonesia/Siak1631-2/2006(H5N1)
ABU99093	Avian	05/01/06	A/Chicken/West Java/SMI-ENDRI2/2006(H5N1)
ABU99086	Avian	07/01/06	A/Chicken/West Java/TASIK2/2006(H5N1)
ABU99083	Avian	08/01/06	A/Chicken/West Java/TASIKSOL/2006(H5N1)
ABW06336	Human	11/24/05	A/Indonesia/195H/2005(H5N1)
ABW06315	Human	2005	A/Indonesia/245H/2005(H5N1)
ABW06287	Human	02/03/06	A/Indonesia/298H/2006(H5N1)
ABW06244	Human	02/21/06	A/Indonesia/341H/2006(H5N1)
ABP51969	Human	07/08/05	A/Indonesia/5/2005(H5N1)
ABW06169	Human	05/10/06	A/Indonesia/542H/2006(H5N1)
ABW06222	Human	05/29/06	A/Indonesia/567H/2006(H5N1)
ABW06117	Human	06/13/06	A/Indonesia/583H/2006(H5N1)
ABW06139	Human	07/06/06	A/Indonesia/604H/2006(H5N1)

ABW06367	Human	2005	A/Indonesia/7/2005(H5N1)
ABM90434	Human	01/05/07	A/Indonesia/CDC1031/2007(H5N1)
ABM90478	Human	01/06/07	A/Indonesia/CDC1032/2007(H5N1)
ABM90489	Human	01/07/07	A/Indonesia/CDC1032N/2007(H5N1)
ABM90511	Human	01/11/07	A/Indonesia/CDC1046/2007(H5N1)
ABM90522	Human	01/11/07	A/Indonesia/CDC1046T/2007(H5N1)
ABI36040	Human	11/08/05	A/Indonesia/CDC184/2005(H5N1)
ABI36041	Human	11/12/05	A/Indonesia/CDC194P/2005(H5N1)
ABI36042	Human	12/13/05	A/Indonesia/CDC287E/2005(H5N1)
ABI36044	Human	12/15/05	A/Indonesia/CDC292N/2005(H5N1)
ABI36050	Human	01/14/06	A/Indonesia/CDC329/2006(H5N1)
ABI36051	Human	01/30/06	A/Indonesia/CDC357/2006(H5N1)
ABI36057	Human	02/20/06	A/Indonesia/CDC390/2006(H5N1)
ABI36198	Human	03/23/06	A/Indonesia/CDC523/2006(H5N1)
ABI36295	Human	04/26/06	A/Indonesia/CDC582/2006(H5N1)
ABI36318	Human	05/18/06	A/Indonesia/CDC623/2006(H5N1)
ABI36384	Human	05/23/06	A/Indonesia/CDC634P/2006(H5N1)
ABI36469	Human	05/29/06	A/Indonesia/CDC644T/2006(H5N1)
ABI36450	Human	07/06/06	A/Indonesia/CDC699/2006(H5N1)
ABI49396	Human	08/05/06	A/Indonesia/CDC739/2006(H5N1)
ABI49407	Human	08/07/06	A/Indonesia/CDC742/2006(H5N1)
ABI49415	Human	2006	A/Indonesia/CDC759/2006(H5N1)
ABL31755	Human	09/24/06	A/Indonesia/CDC836/2006(H5N1)
ABL31780	Human	10/14/06	A/Indonesia/CDC887/2006(H5N1)
ABL07008	Human	11/10/06	A/Indonesia/CDC938/2006(H5N1)
ABW74701	Human	2006	A/Indonesia/TLL001/2006(H5N1)
ABW74702	Human	2006	A/Indonesia/TLL002/2006(H5N1)
ABW74703	Human	2006	A/Indonesia/TLL003/2006(H5N1)

ABW74704	Human	2006	A/Indonesia/TLL004/2006(H5N1)
ABW74707	Human	2006	A/Indonesia/TLL007/2006(H5N1)
ABW74708	Human	2006	A/Indonesia/TLL008/2006(H5N1)
ABW74709	Human	2006	A/Indonesia/TLL009/2006(H5N1)
ABW74714	Human	2006	A/Indonesia/TLL014/2006(H5N1)
ABU99081	Avian	09/01/06	A/Muscovy Duck/Jakarta/HABWIN/2006(H5N1)
ADB07927	Avian	2007/01	A/Muscovy duck/West Java/Bks3/2007(H5N1)
ABU99084	Avian	08/01/06	A/Quail/Jakarta/JU1/2006(H5N1)
AEH42723	Avian	05/02/07	A/chicken/Badung/BBVD-175/2007(H5N1)
AEH42724	Avian	05/15/07	A/chicken/Badung/BBVD-205/2007(H5N1)
AEH42734	Avian	09/03/07	A/chicken/Badung/BBVD-532/2007(H5N1)
AEH42711	Avian	07/26/07	A/chicken/Bangli/BBVD-387ab/2007(H5N1)
AEH42712	Avian	09/13/07	A/chicken/Bangli/BBVD-555ab/2007(H5N1)
AEH42714	Avian	09/18/07	A/chicken/Bangli/BBVD-563/2007(H5N1)
AEH59134	Avian	07/06/07	A/chicken/Bantul/BBVW-446-24454/2007(H5N1)
AEH59146	Avian	07/10/07	A/chicken/Bantul/BBVW-482-22234/2007(H5N1)
AEH42686	Avian	03/27/07	A/chicken/Denpasar/BBVD-145/2007(H5N1)
AEH42688	Avian	06/14/07	A/chicken/Denpasar/BBVD-291/2007(H5N1)
AEH42690	Avian	08/15/07	A/chicken/Denpasar/BBVD-430/2007(H5N1)

AEH42694	Avian	08/27/07	A/chicken/Denpasar/BBVD-494/2007(H5N1)
BAL61222	Avian	2010/05	A/chicken/EastJava/UT551/2010(H5N1)
BAL61218	Avian	2010/04	A/chicken/EastKalimantan/UT498/2010(H5N1)
AEH42700	Avian	06/04/07	A/chicken/Flores Timur/BBVD-256/2007(H5N1)
AEH42697	Avian	08/21/07	A/chicken/Gianyar/BBVD-458/2007(H5N1)
BAK42591	Avian	2010/06	A/chicken/Indonesia/D10015/2010(H5N1)
AEH42721	Avian	08/21/07	A/chicken/Klungkung/BBVD-455/2007(H5N1)
AEH42722	Avian	08/23/07	A/chicken/Klungkung/BBVD-484/2007(H5N1)
AEH59202	Avian	11/04/07	A/chicken/Kulon Progo/BBVW-822-545/2007(H5N1)
AEH59204	Avian	11/26/07	A/chicken/Kulon Progo/BBVW-922-511/2007(H5N1)
AEH59165	Avian	09/14/07	A/chicken/Magelang/BBVW-662-762/2007(H5N1)
AEH59166	Avian	09/14/07	A/chicken/Magelang/BBVW-662-762A/2007(H5N1)
AEH59167	Avian	09/14/07	A/chicken/Magelang/BBVW-662-763/2007(H5N1)
AEH59175	Avian	08/25/07	A/chicken/Magelang/BBVW-667-944/2007(H5N1)
ADB07926	Avian	03/07/07	A/chicken/Pessel/BPPVRII/2007(H5N1)
AEH59149	Avian	07/24/07	A/chicken/Sleman/BBVW-493-214/2007(H5N1)

AEH59226	Avian	01/26/08	A/chicken/Sleman/BBVW-82-65/2008(H5N1)
AEH42705	Avian	07/10/07	A/chicken/Tabanan/BBVD-339/2007(H5N1)
AEH42706	Avian	08/21/07	A/chicken/Tabanan/BBVD-461/2007(H5N1)
ADB07921	Avian	01/18/08	A/chicken/West Java/Smi-Acul/2008(H5N1)
ADB07930	Avian	2005/02	A/chicken/West Java/Smi-Hay/2005(H5N1)
AEH59119	Avian	05/05/07	A/duck/Bantul/BBVW-224-24466/2007(H5N1)
AEH59120	Avian	06/19/07	A/duck/Bantul/BBVW-358-24381/2007(H5N1)
AEH59181	Avian	09/21/07	A/duck/Bantul/BBVW-678-2D403/2007(H5N1)
AEH59206	Avian	12/02/07	A/duck/Bantul/BBVW-949-2D362/2007(H5N1)
AEH59157	Avian	08/19/07	A/duck/Kulon Progo/BBVW-618-11001/2007(H5N1)
AEH59213	Avian	01/15/08	A/duck/Magelang/BBVW-24-44380/2008(H5N1)
AEH59127	Avian	05/28/07	A/duck/Magelang/BBVW-604-44401/2007(H5N1)
AEH59122	Avian	07/24/07	A/duck/Sleman/BBVW-379-34423/2007(H5N1)
AEH59125	Avian	05/15/07	A/duck/Sleman/BBVW-598-32237/2007(H5N1)

ADB07929	Avian	2006/09	A/muscovy duck/Jakarta/Sum106/2006(H5N1)
AEH42693	Avian	08/23/07	A/peaceful dove/Denpasar/BBVD-480/2007(H5N1)

Appendix B

Jelly Roll Capsid Proteins in Dataset

Table B.1: Capsid protein sequences in dataset from families known to possess jelly-roll fold

RefSeq GI	Protein Name	Virus	Viral Family
9626565	capsid protein IX	Human adenovirus F	Adenoviridae
224531380	L1 gene product	Human adenovirus A	Adenoviridae
9626630	L1 gene product	Human adenovirus A	Adenoviridae
9626631	L1 gene product	Human adenovirus A	Adenoviridae
9626166	capsid protein IX	Human adenovirus C	Adenoviridae
9628847	L1 52K	Fowl adenovirus A	Adenoviridae
190340978	capsid protein IX	Human adenovirus D	Adenoviridae
197944770	capsid protein IX	Human adenovirus B	Adenoviridae
51527268	capsid protein IX	Human adenovirus E	Adenoviridae
388570686	capsid protein III	Goose adenovirus 4	Adenoviridae
197944730	capsid protein IX	Human adenovirus B	Adenoviridae
115298543	50.9 kDa Major capsid protein	Spodoptera frugiperda ascovirus 1a	Ascoviridae
116326831	major capsid protein	Trichoplusia ni ascovirus 2c	Ascoviridae

116326851	major capsid protein	Trichoplusia ni ascovirus 2c	Ascoviridae
134287212	major capsid protein	Heliothis virescens ascovirus 3e	Ascoviridae
400355027	coat protein	Amazon lily mild mottle virus	Bromoviridae
9626930	coat protein	Alfalfa mosaic virus	Bromoviridae
19744916	coat protein	Apple mosaic virus	Bromoviridae
20087043	coat protein	Cowpea chlorotic mottle virus	Bromoviridae
20087061	coat protein	Citrus leaf rugose virus	Bromoviridae
20143448	coat protein	Elm mottle virus	Bromoviridae
20177490	coat protein	Pelargonium zonate spot virus	Bromoviridae
20178605	capsid protein	Olive latent virus 2	Bromoviridae
20564220	coat protein	Tulare apple mosaic virus	Bromoviridae
20564159	capsid protein	Tomato aspermy virus	Bromoviridae
21426911	coat protein	Broad bean mottle virus	Bromoviridae
24817636	coat protein	Prunus necrotic ringspot virus	Bromoviridae
46393303	coat protein	Parietaria mottle virus	Bromoviridae
50428572	coat protein	Humulus japonicus latent virus	Bromoviridae

22550383	coat protein	Spring beauty latent virus	Bromoviridae
56692635	coat protein	Fragaria chiloensis latent virus	Bromoviridae
66090971	coat protein	Cassia yellow blotch virus	Bromoviridae
9626474	capsid protein	Cucumber mosaic virus	Bromoviridae
9632352	coat protein	Peanut stunt virus	Bromoviridae
20564171	coat protein	Tobacco streak virus	Bromoviridae
98960848	coat protein	Prune dwarf virus	Bromoviridae
119943075	coat protein	Strawberry necrotic shock virus	Bromoviridae
148717841	coat protein	Citrus variegation virus	Bromoviridae
212525351	coat protein	Blackberry chlorotic ringspot virus	Bromoviridae
224808904	coat protein	Gayfeather mild mottle virus	Bromoviridae
261041623	coat protein	Melandrium yellow fleck virus	Bromoviridae
28392818	capsid protein	Vesicular exanthema of swine virus	Caliciviridae
21699778	VP2 minor capsid protein	Calicivirus strain NB	Caliciviridae
28392841	capsid protein	Walrus calicivirus	Caliciviridae
28392850	capsid protein	Canine calicivirus	Caliciviridae
28268489	capsid protein	Feline calicivirus	Caliciviridae

50284530	capsid protein	Sapovirus Mc10	Caliciviridae
60677689	VP2 minor capsid protein	Calicivirus isolate TCG	Caliciviridae
9790294	minor capsid protein	Rabbit hemorrhagic disease virus	Caliciviridae
106060736	58 kd capsid protein	Norwalk virus	Caliciviridae
113478396	capsid protein VP1	Murine norovirus 1	Caliciviridae
194268062	capsid	Steller sea lion vesivirus	Caliciviridae
9632869	major capsid protein P2	Pseudoalteromonas phage PM2	Corticoviridae
10314011	capsid protein	Acute bee paralysis virus	Dicistroviridae
9629652	capsid polyprotein	partial	Drosophila C virus
395406757	coat protein	Wheat yellow dwarf virus	Geminiviridae
311788803	coat protein	Ageratum leaf curl Cameroon virus	Geminiviridae
401817561	coat protein	French bean severe leaf curl virus	Geminiviridae
401829594	coat protein	Soybean chlorotic spot virus	Geminiviridae
9632378	coat protein	Potato yellow mosaic Panama virus	Geminiviridae
9630668	coat protein	Bean dwarf mosaic virus	Geminiviridae
9632990	AV1 coat protein	Pepper leaf curl virus	Geminiviridae

9632991	AV2 precoat protein	Pepper leaf curl virus	Geminiviridae
9626668	coat protein	Chloris striate mosaic virus	Geminiviridae
9626689	coat protein	Digitaria streak virus	Geminiviridae
9626987	coat protein	Tomato golden mosaic virus	Geminiviridae
9627992	coat protein	Panicum streak virus	Geminiviridae
9630701	coat protein	Tomato mottle virus	Geminiviridae
9629637	coat protein	Tomato mottle Taino virus	Geminiviridae
9632369	coat protein	Sida golden mosaic virus	Geminiviridae
9845215	coat protein	Cotton leaf curl Gezira virus	Geminiviridae
10257476	capsid protein	Horseradish curly top virus	Geminiviridae
10518471	capsid protein	Tomato rugose mosaic virus	Geminiviridae
14647160	coat protein	Ageratum yellow vein Sri Lanka virus	Geminiviridae
14717138	coat protein	Cucurbit leaf crumple virus	Geminiviridae
16507264	coat protein	Cotton leaf curl Rajasthan virus	Geminiviridae
18249858	V1 coat protein	Soybean crinkle leaf virus	Geminiviridae
18450239	coat protein	Miscanthus streak virus	Geminiviridae

19073914	coat protein AV1	Bhendi yellow vein mosaic virus	Geminiviridae
19919897	coat protein	Bean calico mosaic virus	Geminiviridae
20143471	coat protein	Eupatorium yellow vein virus	Geminiviridae
20153402	coat protein	Honeysuckle yellow vein mosaic virus	Geminiviridae
20178611	capsid protein	Tomato chlorotic mottle virus	Geminiviridae
20279532	coat protein	Watermelon chlorotic stunt virus	Geminiviridae
20428541	coat protein	Sugarcane streak virus	Geminiviridae
20522147	coat protein	South African cassava mosaic virus	Geminiviridae
20564137	coat protein	Tobacco yellow dwarf virus	Geminiviridae
108518228	coat protein	Tomato golden mottle virus	Geminiviridae
20806030	coat protein AV1	Dicliptera yellow mottle virus	Geminiviridae
20806025	coat protein	Squash yellow mild mottle virus	Geminiviridae
20806016	coat protein	Cabbage leaf curl virus	Geminiviridae
20806521	coat protein	Tomato mosaic Havana virus	Geminiviridae

21165976	V1 (coat protein)	Tomato leaf curl Bangalore virus	Geminiviridae
21218465	coat protein	Tomato leaf curl virus	Geminiviridae
21218472	coat protein	Tomato leaf curl Karnataka virus	Geminiviridae
21218479	coat protein	Tomato leaf curl Taiwan virus	Geminiviridae
21493008	coat protein	Bean golden mosaic virus	Geminiviridae
9626467	coat protein	Bean golden yellow mosaic virus	Geminiviridae
21911443	coat protein	Hollyhock leaf crumple virus	Geminiviridae
22128598	coat protein	Pepper golden mosaic virus	Geminiviridae
22726210	Coat protein	Papaya leaf curl virus	Geminiviridae
22788706	coat protein	Tomato leaf curl Vietnam virus	Geminiviridae
23096166	coat protein	Pepper leaf curl Bangladesh virus	Geminiviridae
28209381	coat protein	Tomato leaf curl Gujarat virus	Geminiviridae
28380573	coat protein V2	Tomato yellow leaf curl Malaga virus	Geminiviridae
28872847	coat protein	Cotton leaf curl Alabad virus	Geminiviridae
28872854	coat protein	Cotton leaf curl Kokhran virus	Geminiviridae

28913322	coat protein	Cotton leaf curl Multan virus	Geminiviridae
28976186	coat protein	Tomato leaf curl Laos virus	Geminiviridae
28976179	coat protein	Tomato leaf curl Bangladesh virus	Geminiviridae
29135252	coat protein	Ageratum yellow vein Taiwan virus	Geminiviridae
29135245	coat protein	Chilli leaf curl virus	Geminiviridae
29171764	coat protein	Malvastrum yellow vein virus	Geminiviridae
29243868	coat protein	Sida golden mosaic Florida virus	Geminiviridae
29243847	coat protein CP	Sida mottle virus	Geminiviridae
29243861	coat protein	Potato yellow mosaic Trinidad virus	Geminiviridae
29243890	coat protein	Sweet potato leaf curl Georgia virus	Geminiviridae
166162031	capsid protein	Rhynchosia golden mosaic virus	Geminiviridae
29294568	coat protein	Tomato leaf curl Sri Lanka virus	Geminiviridae
29294561	coat protein	Tomato leaf curl Malaysia virus	Geminiviridae
29294599	coat protein	Tobacco leaf curl Japan virus	Geminiviridae
29337257	coat protein	Sida golden mosaic Costa Rica virus	Geminiviridae

29337262	coat protein	Sida golden mosaic Honduras virus	Geminiviridae
29337271	coat protein	Sida yellow vein virus	Geminiviridae
29501757	coat protein	Okra yellow vein mosaic virus	Geminiviridae
29502196	coat protein V1	Tomato curly stunt virus	Geminiviridae
30146804	capsid protein	Beet mild curly top virus	Geminiviridae
31442399	coat protein	Luffa yellow mosaic virus	Geminiviridae
32493266	coat protein	Tomato leaf curl Java virus	Geminiviridae
32493273	coat protein	Tomato leaf curl Philippines virus	Geminiviridae
30146793	coat protein	Sugarcane streak Reunion virus	Geminiviridae
41057578	coat protein	Sida micrantha mosaic virus	Geminiviridae
41057588	coat protein	Dolichos yellow mosaic virus	Geminiviridae
41057726	coat protein	Pepper yellow vein Mali virus	Geminiviridae
41057733	coat protein	Tomato leaf curl Mali virus	Geminiviridae
45445709	coat protein	Tomato yellow leaf curl Kanchanaburi virus	Geminiviridae

9629911	coat protein	Sugarcane streak Egypt virus	Geminiviridae
9626216	coat protein	Beet curly top virus	Geminiviridae
46359748	coat protein	Tomato chino La Paz virus	Geminiviridae
46359762	coat protein	Squash leaf curl Philippines virus	Geminiviridae
46395058	coat protein	Tomato leaf curl Sudan virus	Geminiviridae
46402154	capsid protein	Spinach curly top virus	Geminiviridae
22128603	coat protein	Macroptilium mosaic Puerto Rico virus	Geminiviridae
22128612	coat protein	Macroptilium yellow mosaic Florida virus	Geminiviridae
29243883	coat protein	Tobacco leaf curl Kochi virus	Geminiviridae
29251561	AV1 coat protein	Squash mild leaf curl virus	Geminiviridae
22128022	coat protein	Stachytarpheta leaf curl virus	Geminiviridae
29243838	capsid protein	Tomato severe leaf curl virus	Geminiviridae
20564206	capsid protein	Tomato yellow leaf curl Sardinia virus	Geminiviridae
22128015	coat protein	Ageratum yellow vein China virus	Geminiviridae

29337248	coat protein	East African cassava mosaic Zanzibar virus	Geminiviridae
28867232	coat protein AV1	Cotton leaf crumple virus	Geminiviridae
9625667	26.8 kD virion capsid protein	Maize streak virus	Geminiviridae
20564195	coat protein	Tomato pseudo-curly top virus	Geminiviridae
187476505	coat protein	Macroptilium yellow mosaic virus	Geminiviridae
19881402	coat protein	Bean yellow dwarf virus	Geminiviridae
20806050	coat protein	Sri Lankan cassava mosaic virus	Geminiviridae
9632882	coat protein	Tomato yellow leaf curl Thailand virus	Geminiviridae
29126597	coat protein	East African cassava mosaic Cameroon virus	Geminiviridae
21426902	coat protein	Tomato yellow leaf curl virus	Geminiviridae
19352428	coat protein	Ageratum enation virus	Geminiviridae
29028718	coat protein	Chayote yellow mosaic virus	Geminiviridae
23395821	coat protein	Croton yellow vein mosaic virus	Geminiviridae

60683948	coat protein	Tomato leaf curl Mayotte virus	Geminiviridae
45390218	coat protein	Honeysuckle yellow vein virus	Geminiviridae
57790518	coat protein	Malvastrum yellow vein Yunnan virus	Geminiviridae
20340274	coat protein	Tobacco curly shoot virus	Geminiviridae
24432117	coat protein	Tobacco leaf curl Yunnan virus	Geminiviridae
28916653	coat protein	Mungbean yellow mosaic India virus	Geminiviridae
21594414	coat protein	Tomato yellow leaf curl China virus	Geminiviridae
29294540	coat protein AV1	Sweet potato leaf curl virus	Geminiviridae
18071200	Coat protein	Wheat dwarf virus	Geminiviridae
29243854	coat protein CP	Sida yellow mosaic virus	Geminiviridae
71849680	coat protein	Cotton leaf curl Bangalore virus	Geminiviridae
83571709	coat protein	Sida leaf curl virus	Geminiviridae
386361877	coat protein	Wheat dwarf India virus	Geminiviridae
85667887	coat protein	Tomato leaf curl Joydebpur virus	Geminiviridae
85667900	coat protein	Tomato yellow spot virus	Geminiviridae

85718608	coat protein	Vernonia yellow vein virus	Geminiviridae
387600897	capsid protein	Maize streak Reunion virus	Geminiviridae
387600873	coat protein	Tomato yellow leaf distortion virus	Geminiviridae
92918809	coat protein	Merremia mosaic virus	Geminiviridae
108519231	coat protein	Sida mosaic Sinaloa virus	Geminiviridae
393186609	capsid protein	French bean leaf curl virus-Kanpur	Geminiviridae
110084002	coat protein	Siegesbeckia yellow vein virus	Geminiviridae
112180296	coat protein	Pepper yellow leaf curl Indonesia virus	Geminiviridae
113200772	coat protein	Euphorbia mosaic virus	Geminiviridae
113972273	coat protein	Malvastrum leaf curl Guangdong virus	Geminiviridae
113972280	coat protein	Siegesbeckia yellow vein Guangxi virus	Geminiviridae
114067527	coat protein	Tomato leaf curl Guangxi virus	Geminiviridae
115350046	capsid protein	Tomato yellow leaf curl Guangdong virus	Geminiviridae
115353272	capsid protein	Tomato leaf curl Guangdong virus	Geminiviridae

115353279	coat protein	Okra yellow crinkle virus	Geminiviridae
116294318	coat protein	Desmodium leaf distortion virus	Geminiviridae
116294326	coat protein	Corchorus yellow spot virus	Geminiviridae
116536744	coat protein	Tomato leaf curl Pune virus	Geminiviridae
117530757	coat protein	Malvastrum yellow mosaic virus	Geminiviridae
121614280	coat protein	Sida yellow mosaic Yucatan virus	Geminiviridae
122809022	coat protein	Crassocephalum yellow vein virus - Jinghong	Geminiviridae
126030107	coat protein	Tomato leaf curl Arusha virus	Geminiviridae
126031756	coat protein	Tomato leaf curl Seychelles virus	Geminiviridae
126640092	coat protein	Mesta yellow vein mosaic virus	Geminiviridae
146411801	coat protein	Clerodendron yellow mosaic virus	Geminiviridae
148659001	coat protein	Pepper curly top virus	Geminiviridae
149980612	coat protein	Tomato leaf curl Sinaloa virus	Geminiviridae

149980618	coat protein	Tomato severe rugose virus	Geminiviridae
164564314	coat protein	Radish leaf curl virus	Geminiviridae
166153477	capsid protein	Chickpea chlorotic dwarf Sudan virus	Geminiviridae
167046007	coat protein	Tomato leaf curl Ghana virus	Geminiviridae
168164405	coat protein	Eragrostis streak virus	Geminiviridae
169303567	coat protein	Beet curly top Iran virus	Geminiviridae
169822863	coat protein	Tomato leaf curl Cebu virus	Geminiviridae
169822877	coat protein	Tomato leaf curl Cotabato virus	Geminiviridae
169822870	coat protein	Tomato leaf curl Mindanao virus	Geminiviridae
189303449	coat protein	Urochloa streak virus	Geminiviridae
189311145	Coat protein	Barley dwarf virus	Geminiviridae
189311150	Coat protein	Oat dwarf virus	Geminiviridae
189475231	coat protein	Mesta yellow vein mosaic Bahraich virus	Geminiviridae
190149264	coat protein	Pumpkin yellow mosaic Malaysia virus	Geminiviridae
190336466	coat protein	Wissadula golden mosaic virus	Geminiviridae
190336475	coat protein	Macroptilium golden mosaic virus	Geminiviridae

190336514	capsid protein	Allamanda leaf curl virus	Geminiviridae
190336524	coat protein	Tomato leaf curl Palampur virus	Geminiviridae
190336531	coat protein	Blainvillea yellow spot virus	Geminiviridae
190336540	coat protein	Tomato common mosaic virus	Geminiviridae
190336549	coat protein	Tomato mild mosaic virus	Geminiviridae
194322733	coat protein	Chickpea chlorotic dwarf virus	Geminiviridae
195535991	coat protein	Tomato leaf curl Kumasi virus	Geminiviridae
196049397	coat protein	Tomato leaf curl Kerala virus	Geminiviridae
197322511	coat protein	Okra mottle virus	Geminiviridae
197914889	capsid protein	Pepper yellow dwarf virus - New Mexico	Geminiviridae
203449527	coat protein	Jatropha leaf curl virus	Geminiviridae
219552922	coat protein	Gossypium darwinii symptomless virus	Geminiviridae
222822600	coat protein	Bhendi yellow vein Bhubhaneswar virus	Geminiviridae
224581830	coat protein	Pedilanthus leaf curl virus	Geminiviridae

224591440	coat protein	Cotton leaf curl Burewala virus	Geminiviridae
225847630	coat protein	Rhynchosia golden mosaic Yucatan virus	Geminiviridae
226202301	coat protein	Tomato leaf curl Patna virus	Geminiviridae
229605060	coat protein	Eragrostis curvula streak virus	Geminiviridae
239740601	coat protein	Passionfruit severe leaf distortion virus	Geminiviridae
254729453	coat protein	Okra leaf curl virus	Geminiviridae
255983866	coat protein	Tomato leaf curl Hainan virus	Geminiviridae
262396921	coat protein	Saccharum streak virus	Geminiviridae
281372527	coat protein	Tomato leaf curl Cameroon virus	Geminiviridae
281372529	coat protein	Tomato leaf curl Cameroon virus	Geminiviridae
296006067	coat protein	Okra yellow mosaic Mexico virus	Geminiviridae
296006054	coat protein	Sida golden mottle virus	Geminiviridae
296005649	coat protein	Abutilon Brazil virus	Geminiviridae
296011098	coat protein	Soybean mild mottle virus	Geminiviridae
296040241	coat protein	Soybean chlorotic blotch virus	Geminiviridae

306478724	coat protein	Turnip curly top virus	Geminiviridae
304360753	coat protein	Sida golden mosaic Florida virus-Malvastrum	Geminiviridae
308125299	coat protein	Digitaria didactyla striate mosaic virus	Geminiviridae
308814739	coat protein	Tobacco leaf curl Pusa virus	Geminiviridae
310288307	capsid protein	Spinach severe curly top virus	Geminiviridae
311788794	coat protein	Malvastrum yellow vein Changa Manga virus	Geminiviridae
313493546	coat protein	Okra leaf curl Cameroon virus	Geminiviridae
315570368	coat protein	Bromus catharticus striate mosaic virus	Geminiviridae
317453631	coat protein	Rhynchosia yellow mosaic India virus	Geminiviridae
321961794	coat protein	Sida mosaic Bolivia virus 2	Geminiviridae
321961749	coat protein	Sida mosaic Bolivia virus 1	Geminiviridae
321961703	coat protein	Abutilon mosaic Bolivia virus	Geminiviridae
322840371	coat protein	Spinach curly top Arizona virus	Geminiviridae

323361059	capsid protein	Tomato mottle leaf curl Zulia virus	Geminiviridae
327409463	coat protein	Sweet potato leaf curl South Carolina virus	Geminiviridae
330370662	coat protein	Cleome golden mosaic virus	Geminiviridae
332290614	coat protein	Bean yellow mosaic Mexico virus	Geminiviridae
332290619	coat protein	Rhynchosai mild mosaic virus	Geminiviridae
332290627	coat protein	Merremia mosaic Puerto Rico virus	Geminiviridae
333595895	coat protein	Eragrostis minor streak virus	Geminiviridae
334847636	coat protein	Tobacco yellow crinkle virus	Geminiviridae
372204031	coat protein	Tomato dwarf leaf virus	Geminiviridae
403516100	coat protein	Paspalum striate mosaic virus	Geminiviridae
404184257	capsid protein	Paspalum dilatatum striate mosaic virus	Geminiviridae
404184274	capsid protein	Sporobolus striate mosaic virus 1	Geminiviridae
404184291	capsid protein	Sporobolus striate mosaic virus 2	Geminiviridae
404184316	capsid protein	Digitaria ciliaris striate mosaic virus	Geminiviridae

9695414	major capsid protein	Lymphocystis disease virus 1	Iridoviridae
45686022	major capsid protein	Ambystoma tigrinum virus	Iridoviridae
48843722	major capsid protein	Lymphocystis disease virus - isolate China	Iridoviridae
56692709	major capsid protein	Singapore grouper iridovirus	Iridoviridae
49237388	major capsid protein	Frog virus 3	Iridoviridae
388260085	major capsid protein	European catfish virus	Iridoviridae
9634965	coat protein P3	Cereal yellow dwarf virus - RPS	Luteoviridae
9634104	coat protein P3	Barley yellow dwarf virus-PAS	Luteoviridae
15150437	coat protein	Soybean dwarf virus	Luteoviridae
18314287	coat protein P3	Bean leafroll virus	Luteoviridae
20178352	coat protein	Pea enation mosaic virus-1	Luteoviridae
20219028	viral coat protein	Barley yellow dwarf virus-MAV	Luteoviridae
30146801	coat protein P3	Cereal yellow dwarf virus - RPV	Luteoviridae
30248021	coat protein	Beet western yellows virus	Luteoviridae
29366748	coat protein P3	Barley yellow dwarf virus-GAV	Luteoviridae
51980899	coat protein	Carrot red leaf virus	Luteoviridae

9629164	coat protein	Potato leafroll virus	Luteoviridae
20428575	coat protein	Turnip yellows virus	Luteoviridae
20260790	coat protein	Cucurbit aphid-borne yellows virus	Luteoviridae
19881392	coat protein	Beet mild yellowing virus	Luteoviridae
30146775	coat protein P3	Barley yellow dwarf virus-PAV	Luteoviridae
9632982	capsid protein	Sugarcane yellow leaf virus	Luteoviridae
110645396	coat protein	Chickpea chlorotic stunt virus	Luteoviridae
189418881	coat protein	Melon aphid-borne yellows virus	Luteoviridae
189418888	coat protein P3	Rose spring dwarf-associated virus	Luteoviridae
253729533	coat protein readthrough	Wheat yellow dwarf virus-GPV	Luteoviridae
253729534	coat protein	Wheat yellow dwarf virus-GPV	Luteoviridae
308125289	P3 coat protein	Cotton leafroll dwarf virus	Luteoviridae
348020289	coat protein	Brassica yellows virus	Luteoviridae
17402852	capsid protein VP2-related protein	Chlamydia phage phiCPG1	Microviridae
17402853	capsid protein VP3	Chlamydia phage phiCPG1	Microviridae

12085136	major capsid protein	Bdellovibrio phage phiMH2K	Microviridae
12085140	minor capsid protein	Bdellovibrio phage phiMH2K	Microviridae
12085142	minor capsid protein	Bdellovibrio phage phiMH2K	Microviridae
19387569	capsid protein	Spiroplasma phage 4	Microviridae
9626346	major coat protein	Enterobacteria phage G4 sensu lato	Microviridae
9625357	capsid morphogenesis protein	Enterobacteria phage alpha3	Microviridae
9625360	capsid morphogenesis protein	Enterobacteria phage alpha3	Microviridae
9625363	major coat protein	Enterobacteria phage alpha3	Microviridae
393707864	viral coat protein VP1	Microvirus CA82	Microviridae
9791177	minor capsid protein	Chlamydia phage CPAR39	Microviridae
9791178	major capsid protein	Chlamydia phage CPAR39	Microviridae
9791182	minor capsid protein	Chlamydia phage CPAR39	Microviridae
311977809	capsid protein 1	Acanthamoeba polyphaga mimivirus	Mimiviridae
38018422	capsid protein beta	Nodamura virus	Nodaviridae
38018423	capsid protein gamma	Nodamura virus	Nodaviridae

19551003	coat protein	Striped Jack nervous necrosis virus	Nodaviridae
25121783	mature capsid protein gamma	Pariacato virus	Nodaviridae
25121784	mature capsid protein beta	Pariacato virus	Nodaviridae
22681118	coat protein	Epinephelus tauvina nervous necrosis virus	Nodaviridae
34610125	capsid protein	Macrobrachium rosenbergii nodavirus	Nodaviridae
98960855	coat protein	Redspotted grouper nervous necrosis virus	Nodaviridae
194351533	capsid protein	Barfin flounder virus BF93Hok	Nodaviridae
262396907	coat protein	Barfin flounder nervous necrosis virus	Nodaviridae
262396912	coat protein	Tiger puffer nervous necrosis virus	Nodaviridae
320339431	capsid protein	Penaeus vannamei nodavirus	Nodaviridae
322867243	capsid protein	Santeuil nodavirus	Nodaviridae
9628556	minor capsid protein L2	Human papillomavirus type 50	Papillomaviridae
9628557	major capsid protein L1	Human papillomavirus type 50	Papillomaviridae

9628572	minor capsid protein L2	Human papillomavirus type 60	Papillomaviridae
9628573	major capsid protein L1	Human papillomavirus type 60	Papillomaviridae
9628548	minor capsid protein L2	Human papillomavirus type 48	Papillomaviridae
9628549	major capsid protein L1	Human papillomavirus type 48	Papillomaviridae
9627492	L1 protein	Iotapapillomavirus 1	Papillomaviridae
9628453	L1 protein	Alphapapillomavirus 12	Papillomaviridae
9635140	minor capsid protein L2	Kappapapillomavirus 1	Papillomaviridae
9635141	major capsid protein L1	Kappapapillomavirus 1	Papillomaviridae
9627071	minor capsid protein	Deltapapillomavirus 2	Papillomaviridae
9627073	major capsid protein	Deltapapillomavirus 2	Papillomaviridae
13186230	L1	Kappapapillomavirus 2	Papillomaviridae
21326234	minor capsid protein L2	Thetapapillomavirus 1	Papillomaviridae
21326235	major capsid protein L1	Thetapapillomavirus 1	Papillomaviridae

9629728	L1	Common chimpanzee papillomavirus 1	Papillomaviridae
27531793	L1	Human papillomavirus type 92	Papillomaviridae
30315801	L1 protein	Felis domesticus papillomavirus type 1	Papillomaviridae
56698754	minor capsid protein	Trichechus manatus papillomavirus 1	Papillomaviridae
56698755	major capsid protein	Trichechus manatus papillomavirus 1	Papillomaviridae
56693043	L1	Canis familiaris papillomavirus 2	Papillomaviridae
62362152	minor capsid protein	Erethizon dorsatum papillomavirus 1	Papillomaviridae
62362153	major capsid protein	Erethizon dorsatum papillomavirus 1	Papillomaviridae
18138524	L1 protein	Omikronpapillomavirus 1	Papillomaviridae
9628580	minor capsid protein L2	Alphapapillomavirus 3	Papillomaviridae
9628581	major capsid protein L1	Alphapapillomavirus 3	Papillomaviridae
9627086	minor capsid protein	Deltapapillomavirus 1	Papillomaviridae
9627087	major capsid protein	Deltapapillomavirus 1	Papillomaviridae
9626061	minor capsid protein L2	Human papillomavirus type 6b	Papillomaviridae

9626062	major capsid protein L1	Human papillomavirus type 6b	Papillomaviridae
9627369	L1 protein	Human papillomavirus type 49	Papillomaviridae
9628443	minor capsid protein	Alphapapillomavirus 13	Papillomaviridae
9628444	major capsid protein	Alphapapillomavirus 13	Papillomaviridae
68304295	L1	Procyon lotor papillomavirus 1	Papillomaviridae
20428634	minor capsid protein	Equus caballus papillomavirus 1	Papillomaviridae
20428635	major capsid protein	Equus caballus papillomavirus 1	Papillomaviridae
37595916	L1	Human papillomavirus type 96	Papillomaviridae
9626067	minor capsid protein L2	Mupapillomavirus 1	Papillomaviridae
9626068	major capsid protein L1	Mupapillomavirus 1	Papillomaviridae
9627107	minor capsid protein	Human papillomavirus type 16	Papillomaviridae

9627108	major capsid L1 protein	Human papillomavirus type 16	Papillomaviridae
9626077	L1 protein	Alphapapillomavirus 7	Papillomaviridae
9627152	minor capsid protein	Human papillomavirus type 5	Papillomaviridae
9627153	major capsid protein	Human papillomavirus type 5	Papillomaviridae
13186282	late protein L1	Alphapapillomavirus 4	Papillomaviridae
9627064	capsid protein L1	Deltapapillomavirus 4	Papillomaviridae
386576358	L1 gene product	Papio hamadryas papillomavirus type 1	Papillomaviridae
82547789	L1 protein	Bovine papillomavirus 7	Papillomaviridae
389656406	minor capsid protein	Human papillomavirus type 135	Papillomaviridae
389656407	major capsid protein	Human papillomavirus type 135	Papillomaviridae
389656414	minor capsid protein	Human papillomavirus type 136	Papillomaviridae
389656415	major capsid protein	Human papillomavirus type 136	Papillomaviridae

389656422	minor capsid protein	Human papillomavirus type 137	Papillomaviridae
389656423	major capsid protein	Human papillomavirus type 137	Papillomaviridae
389656430	minor capsid protein	Human papillomavirus type 140	Papillomaviridae
389656431	major capsid protein	Human papillomavirus type 140	Papillomaviridae
389656438	minor capsid protein	Human papillomavirus type 144	Papillomaviridae
389656439	major capsid protein	Human papillomavirus type 144	Papillomaviridae
97331433	L1	Capra hircus papillomavirus 1	Papillomaviridae
392283764	L1 protein	Phocoena phocoena papillomavirus 1	Papillomaviridae
392283736	L1 protein	Phocoena phocoena papillomavirus 2	Papillomaviridae
392283743	L1 protein	Phocoena phocoena papillomavirus 4	Papillomaviridae
109390375	L1	Tursiops truncatus papillomavirus 2	Papillomaviridae

109390388	L1 protein	Human papillomavirus type 103	Papillomaviridae
109390395	L1 protein	Human papillomavirus 101	Papillomaviridae
113200770	L1	Rousettus aegyptiacus papillomavirus 1	Papillomaviridae
116536734	L1	Mastomys coucha papillomavirus 2	Papillomaviridae
118129787	L1	Micromys minutus papillomavirus 1	Papillomaviridae
156522778	L1	Bovine papillomavirus 8	Papillomaviridae
164429764	minor capsid protein	Canine papillomavirus 4	Papillomaviridae
164429769	major capsid protein	Canine papillomavirus 4	Papillomaviridae
167600372	L1 protein	Human papillomavirus type 88	Papillomaviridae
189043083	L1	Ursus maritimus papillomavirus 1	Papillomaviridae
189475226	minor capsid protein	Bandicoot papillomatosis carcinomatosis virus type 2	Papillomaviridae

189475227	major capsid protein	Bandicoot papillomatosis carcinomatosis virus type 2	Papillomaviridae
194268072	L1	Capreolus capreolus papillomavirus 1	Papillomaviridae
195661191	L1	Tursiops truncatus papillomavirus 1	Papillomaviridae
206599542	L1	Sus scrofa papillomavirus 1	Papillomaviridae
212286054	L2 capsid protein	Caretta caretta papillomavirus 1	Papillomaviridae
212286055	L1 capsid protein	Caretta caretta papillomavirus 1	Papillomaviridae
212286062	L2 capsid protein	Chelonia mydas papillomavirus 1	Papillomaviridae
212286063	L1 capsid protein	Chelonia mydas papillomavirus 1	Papillomaviridae
218685643	minor capsid protein	Erinaceus europaeus papillomavirus 1	Papillomaviridae
218685644	major capsid protein	Erinaceus europaeus papillomavirus 1	Papillomaviridae
224983327	early protein L1	Human papillomavirus type 108	Papillomaviridae
225927575	L1 protein	Human papillomavirus 112	Papillomaviridae

225927567	L1 protein	Human papillomavirus 109	Papillomaviridae
254810670	L1 protein	Human papillomavirus 116	Papillomaviridae
256352182	L1	Francolinus leucoscepus papillomavirus 1	Papillomaviridae
257136432	L1	Rattus norvegicus papillomavirus 1 EES-2009	Papillomaviridae
258611057	L1	Canine papillomavirus 5	Papillomaviridae
258611065	L1	Lambdapapillomavirus 3	Papillomaviridae
296040258	L1	Bettongia penicillata papillomavirus 1	Papillomaviridae
297342362	minor capsid protein L2	Human papillomavirus 121	Papillomaviridae
297342363	major capsid protein L1	Human papillomavirus 121	Papillomaviridae
301173450	L1	Mus musculus papillomavirus type 1	Papillomaviridae
304522121	L1	Gammapapillomavirus HPV127	Papillomaviridae
319976690	L1	Human papillomavirus type 128	Papillomaviridae

319976682	L1	Human papillomavirus type 129	Papillomaviridae
319976674	L1	Human papillomavirus type 131	Papillomaviridae
319962674	L1	Human papillomavirus type 132	Papillomaviridae
319962666	L1	Human papillomavirus type 134	Papillomaviridae
326631683	L1	Camelus dromedarius papillomavirus type 1	Papillomaviridae
326631691	L1	Camelus dromedarius papillomavirus type 2	Papillomaviridae
332288084	L1	Zalophus californianus papillomavirus 1	Papillomaviridae
338209364	L1	Macaca fascicularis papillomavirus 2	Papillomaviridae
338209371	L1	Colobus guereza papillomavirus type 2	Papillomaviridae
347750419	L1	Morelia spilota papillomavirus 1	Papillomaviridae
347750427	L1	Canine papillomavirus 8	Papillomaviridae

363540895	L1	Canine papillomavirus 9	Papillomaviridae
379059608	L1 gene product	Trichechus manatus latirostris papillomavirus 2	Papillomaviridae
404184239	L1 protein	Crocuta crocuta papillomavirus 1	Papillomaviridae
26335933	capsid protein 2	Porcine parvovirus	Parvoviridae
26335936	capsid protein 1	Porcine parvovirus	Parvoviridae
356457874	capsid protein 1	Human parvovirus B19	Parvoviridae
356457875	capsid protein 2	Human parvovirus B19	Parvoviridae
29823072	Coat protein VP1	LuIII virus	Parvoviridae
33235705	capsid protein	Blattella germanica densovirus	Parvoviridae
9627949	capsid protein	Mouse parvovirus 1	Parvoviridae
51593843	capsid protein VP	Muscovy duck parvovirus	Parvoviridae
51593844	capsid protein VP	Muscovy duck parvovirus	Parvoviridae
51593845	capsid protein VP	Muscovy duck parvovirus	Parvoviridae
294441963	37 kDa coat protein	Infectious hypodermal and hematopoietic necrosis virus	Parvoviridae
51555746	capsid protein	Snake parvovirus 1	Parvoviridae

9626079	Non-capsid protein	H-1 parvovirus	Parvoviridae
9626080	coat protein	H-1 parvovirus	Parvoviridae
19263355	Coat protein VP1	H-1 parvovirus	Parvoviridae
312460625	coat protein VP1/VP2	Aedes albopictus densovirus	Parvoviridae
40795677	capsid protein VP1	Canine parvovirus	Parvoviridae
40795678	capsid protein VP2	Canine parvovirus	Parvoviridae
23334615	capsid protein	Bombyx mori densovirus 5	Parvoviridae
40804367	major capsid protein VP3	Bombyx mori densovirus 5	Parvoviridae
40804368	major capsid protein VP1	Bombyx mori densovirus 5	Parvoviridae
208429854	viral capsid protein	Anopheles gambiae densonucleosis virus	Parvoviridae
229342014	capsid protein	Aedes aegypti densovirus	Parvoviridae
238621219	capsid protein VP1	Human bocavirus 4	Parvoviridae
238621220	capsid protein VP2	Human bocavirus 4	Parvoviridae
302317869	capsid protein	Bocavirus gorilla/GBoV1/2009	Parvoviridae
302317870	capsid protein	Bocavirus gorilla/GBoV1/2009	Parvoviridae
312126260	capsid	Porcine parvovirus 4	Parvoviridae
322688189	capsid protein	Mosquito densovirus BR/07	Parvoviridae
340025681	Capsid protein	Paramecium bursaria Chlorella virus 1	Phycodnaviridae

9631580	Capsid protein	Paramecium bursaria Chlorella virus 1	Phycodnaviridae
340025827	Capsid protein	Paramecium bursaria Chlorella virus 1	Phycodnaviridae
340025832	Capsid protein	Paramecium bursaria Chlorella virus 1	Phycodnaviridae
9631998	Major capsid protein	Paramecium bursaria Chlorella virus 1	Phycodnaviridae
9632117	Capsid protein	Paramecium bursaria Chlorella virus 1	Phycodnaviridae
9632153	Capsid protein	Paramecium bursaria Chlorella virus 1	Phycodnaviridae
314055152	major capsid protein	Ostreococcus tauri virus 2	Phycodnaviridae
314055157	major capsid protein	Ostreococcus tauri virus 2	Phycodnaviridae
314055161	major capsid protein	Ostreococcus tauri virus 2	Phycodnaviridae
314055180	major capsid protein	Ostreococcus tauri virus 2	Phycodnaviridae
314055181	major capsid protein	Ostreococcus tauri virus 2	Phycodnaviridae
314055249	major capsid protein	Ostreococcus tauri virus 2	Phycodnaviridae
314055301	major capsid protein	Ostreococcus tauri virus 2	Phycodnaviridae
314055305	major capsid protein	Ostreococcus tauri virus 2	Phycodnaviridae

25121840	coat protein VP4	Human enterovirus C	Picornaviridae
25121841	coat protein VP2	Human enterovirus C	Picornaviridae
25121842	coat protein VP3	Human enterovirus C	Picornaviridae
25121843	coat protein VP1	Human enterovirus C	Picornaviridae
25121875	capsid protein VP4	Theilovirus	Picornaviridae
25121876	capsid protein VP2	Theilovirus	Picornaviridae
25121877	capsid protein VP3	Theilovirus	Picornaviridae
25121878	capsid protein VP1	Theilovirus	Picornaviridae
25121919	capsid protein 1A	Human enterovirus D	Picornaviridae
25121920	capsid protein 1B	Human enterovirus D	Picornaviridae
25121921	capsid protein 1C version 3	Human enterovirus D	Picornaviridae
25121922	capsid protein 1C version 1	Human enterovirus D	Picornaviridae
25121923	capsid protein 1C version 2	Human enterovirus D	Picornaviridae
25121924	capsid protein 1D version 2	Human enterovirus D	Picornaviridae
25121925	capsid protein 1D version 1	Human enterovirus D	Picornaviridae
25121926	capsid protein 1D version 3	Human enterovirus D	Picornaviridae
25121796	capsid protein 1A	Porcine sapelovirus 1	Picornaviridae
25121797	capsid protein 1B	Porcine sapelovirus 1	Picornaviridae
25121798	capsid protein 1C	Porcine sapelovirus 1	Picornaviridae
25121800	capsid protein 1D	Porcine sapelovirus 1	Picornaviridae
182406747	capsid protein VP4	Saffold virus	Picornaviridae
182406748	capsid protein VP2	Saffold virus	Picornaviridae

182406749	capsid protein VP3	Saffold virus	Picornaviridae
182406750	capsid protein VP1	Saffold virus	Picornaviridae
189418896	capsid protein VP4	Human TMEV-like cardiovirus	Picornaviridae
189418897	capsid protein VP2	Human TMEV-like cardiovirus	Picornaviridae
189418898	capsid protein VP3	Human TMEV-like cardiovirus	Picornaviridae
189418899	capsid protein VP1	Human TMEV-like cardiovirus	Picornaviridae
9626486	major capsid protein VP1	Bovine polyomavirus	Polyomaviridae
297591902	Major capsid protein VP1	Simian virus 40	Polyomaviridae
9626981	capsid protein VP1	Murine pneumotropic virus	Polyomaviridae
28376616	capsid protein VP2	Murine pneumotropic virus	Polyomaviridae
28376617	capsid protein VP3	Murine pneumotropic virus	Polyomaviridae
30315612	VP2 capsid protein	African green monkey polyomavirus	Polyomaviridae
30315613	VP3 capsid protein	African green monkey polyomavirus	Polyomaviridae
30315614	VP1 capsid protein	African green monkey polyomavirus	Polyomaviridae
9627024	VP1 (major capsid protein)	Murine polyomavirus	Polyomaviridae

9627026	VP2 (capsid protein)	Murine polyomavirus	Polyomaviridae
9627025	VP3 (capsid protein)	Murine polyomavirus	Polyomaviridae
393738581	major capsid protein VP1	MW polyomavirus	Polyomaviridae
393738582	capsid protein VP2	MW polyomavirus	Polyomaviridae
393738583	capsid protein VP3	MW polyomavirus	Polyomaviridae
9626357	minor capsid protein	Enterobacteria phage PRD1	Tectiviridae
9626361	major capsid protein	Enterobacteria phage PRD1	Tectiviridae
9626366	minor capsid protein	Enterobacteria phage PRD1	Tectiviridae
211956457	capsid protein	Bacillus phage AP50	Tectiviridae
211956463	major capsid protein	Bacillus phage AP50	Tectiviridae
211956469	minor capsid protein	Bacillus phage AP50	Tectiviridae
9632274	coat protein	Helicoverpa armigera stunt virus	Tetraviridae
38018443	large capsid protein	Euprosterna elaeasa virus	Tetraviridae
38018444	small capsid protein	Euprosterna elaeasa virus	Tetraviridae
9631281	capsid protein	Nudaurelia capensis beta virus	Tetraviridae
48697171	capsid protein p71	Dendrolimus punctatus tetravirus	Tetraviridae
394743614	capsid protein	Lettuce necrotic stunt virus	Tombusviridae

9626674	coat protein	Cucumber necrosis virus	Tombusviridae
9627424	coat protein	Cardamine chlorotic fleck virus	Tombusviridae
19774247	29 kDa coat protein	Tobacco necrosis virus D	Tombusviridae
20087051	capsid	Cowpea mottle virus	Tombusviridae
20428582	capsid protein	Red clover necrotic mosaic virus	Tombusviridae
20522129	capsid protein	Sweet clover necrotic mosaic virus	Tombusviridae
20564144	coat protein	Turnip crinkle virus	Tombusviridae
30018249	coat protein CP	Pear latent virus	Tombusviridae
30018255	coat protein	Cucumber Bulgarian latent virus	Tombusviridae
32469482	coat protein	Pea stem necrosis virus	Tombusviridae
39163651	coat protein	Johnsongrass chlorotic stripe mosaic virus	Tombusviridae
39163636	coat protein	Pelargonium flower break virus	Tombusviridae
50080148	coat protein	Beet black scorch virus	Tombusviridae
11072115	26 kDa capsid protein	Panicum mosaic virus	Tombusviridae
9629190	capsid protein	Saguaro cactus virus	Tombusviridae
9633807	coat protein	Carnation mottle virus	Tombusviridae

20087027	coat protein	Cymbidium ringspot virus	Tombusviridae
189042977	capsid protein	Carnation Italian ringspot virus	Tombusviridae
20177494	coat protein (p48)	Oat chlorotic stunt virus	Tombusviridae
9629523	capsid protein	Leek white stripe virus	Tombusviridae
9628879	capsid protein	Olive latent virus 1	Tombusviridae
9634678	capsid protein (p38)	Japanese iris necrotic ring virus	Tombusviridae
20162542	coat protein	Maize chlorotic mottle virus	Tombusviridae
9629187	coat protein	Tobacco necrosis virus A	Tombusviridae
9790331	p41 capsid protein	Tomato bushy stunt virus	Tombusviridae
62327386	coat protein	Olive mild mosaic virus	Tombusviridae
20153394	coat protein	Hibiscus chlorotic ringspot virus	Tombusviridae
9629513	capsid protein	Galinsoga mosaic virus	Tombusviridae
38707977	coat protein	Pelargonium necrotic spot virus	Tombusviridae
66478135	coat protein	Pelargonium line pattern virus	Tombusviridae

9626977	capsid protein	Melon necrotic spot virus	Tombusviridae
9625554	coat protein of 41 kDa	Artichoke mottled crinkle virus	Tombusviridae
85718598	coat protein	Angelonia flower break virus	Tombusviridae
89888604	coat protein	Cucumber leaf spot virus	Tombusviridae
94536600	coat protein	Lisianthus necrosis virus	Tombusviridae
126010923	coat protein	Nootka lupine vein-clearing virus	Tombusviridae
212498617	capsid protein	Grapevine Algerian latent virus	Tombusviridae
216905812	coat protein	Soybean yellow mottle mosaic virus	Tombusviridae
9626697	coat protein	Eggplant mosaic virus	Tymoviridae
9629159	coat protein	Kennedya yellow mosaic virus	Tymoviridae
9627013	coat protein	Ononis yellow mosaic virus	Tymoviridae
20177482	coat protein	Physalis mottle virus	Tymoviridae
62327637	21 kDa capsid protein	Citrus sudden death-associated virus	Tymoviridae
62327641	capsid protein	Citrus sudden death-associated virus	Tymoviridae
18138527	coat protein	Grapevine fleck virus	Tymoviridae

21686953	coat protein	Turnip yellow mosaic virus	Tymoviridae
9629257	21 kDa capsid protein	Oat blue dwarf virus	Tymoviridae
9634118	coat protein	Poinsettia mosaic virus	Tymoviridae
25013992	capsid protein	Poinsettia mosaic virus	Tymoviridae
11067740	coat protein	Chayote mosaic virus	Tymoviridae
82524284	coat protein	Dulcamara mottle virus	Tymoviridae
148724440	coat protein	Okra mosaic virus	Tymoviridae
194473851	coat protein CP	Diascia yellow mottle virus	Tymoviridae
212498836	coat protein	Scrophularia mottle virus	Tymoviridae
212498907	coat protein	Nemesia ring necrosis virus	Tymoviridae
212498988	coat protein	Plantago mottle virus	Tymoviridae
212525936	coat protein	Anagyris vein yellowing virus	Tymoviridae
226201771	capsid protein	Grapevine Syrah Virus-1	Tymoviridae
289522105	coat protein	Olive latent virus 3	Tymoviridae
296006065	coat protein	Chiltepin yellow mosaic virus	Tymoviridae
326537272	coat protein	Fig fleck-associated virus	Tymoviridae

339276126	coat protein	Asclepias asymptomatic virus	Tymoviridae

Appendix C

Mathematical Formulation of Support Vector Machine Algorithm

In Chapter 3, we used Support Vector Machine (SVM) to learn classification between jelly-roll containing capsid proteins and other proteins. SVM is a maximum margin supervised classifier, which works by finding a separating hypersurfaces between the two or more classes of training data [97]. Once this separating hypersurface is found, it can be used for predicting the classification of novel samples. In this appendix, I discuss the mathematical formulation of this algorithm. I have closely followed the discussion in the text by Bishop [96].

The input for the SVM comprises of representations of training samples in a feature space, and their actual classes. Let the training samples be represented by, $\{\mathbf{x_i}\}$, where $i = 1, .., N$ and each $\mathbf{x_i}$ is a feature vector with dimensionality D. For simplicity, I will work with two class classification. Let the class labels for training samples be $\{y_i\}$, where $i = 1, ..., N$ and each $t_i \in \{-1, 1\}$. The SVM algorithm then seeks to construct a function which will ideally be positive for training samples of class 1 and negative for training samples of class -1. In the case of the linear classification, the ansatz for this function, known as "decision function", is

$$y(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \tag{C.1}$$

where b is a constant, and $\mathbf{w}$ is the normal to the separating hyperplane $y(\mathbf{x}) = 0$. In principle, there can be many separating hyperplanes for the training data. The separating hyperplane chosen by SVM is the one which maximizes the distance to the

closest points of each class (maximizing "margin"). Using straigthforward geometry, it can be shown that the distance of a point $\mathbf{x}$ to a plane specified by $y(\mathbf{x}) = 0$ is given by $abs(y(\mathbf{x}))/|\mathbf{w}|$. Assuming correct classification for all training samples implies $t_i y_i(\mathbf{x_i}) = |y_i(\mathbf{x_i})|$. Thus, the problem of finding the maximum margin separating hyperplane can be stated as

$$\max_{\mathbf{w},b} \left\{ \min_i \left[\frac{1}{|\mathbf{w}|} t_i \left(\mathbf{w} \cdot \mathbf{x_i} + b \right) \right] \right\} \tag{C.2}$$

subject to the constraints

$$t_i \left(\mathbf{w} \cdot \mathbf{x_i} + b \right) > 0 \ , \ \ i = 1, \ldots, N \tag{C.3}$$

Because the sign of the function $y(\mathbf{x})$ predicts the class of a sample, one can arbitrarily scale the function by a positive constant to still retain this property. Using this freedom, the distance of the closest training samples to the separating plane can be set to 1. Thus, Eq. (C.2) becomes the maximization of $1/|w|$, which is equivalent to *minimization* of $|\mathbf{w}|^2$, a problem in quadratic programming. The constraints now change – since the closest samples to the hyperplane are set to be at distance 1, the training samples thus obey the constraints

$$t_i \left(\mathbf{w} \cdot \mathbf{x_i} + b \right) \geq 1 \ , \ \ i = 1, \ldots, N \tag{C.4}$$

The maximization problem in the presence of constraints can be solved by introducing Lagrange multipliers a_i.

$$L \left(\mathbf{w}, b, \{a_i\} \right) = \frac{1}{2} \mathbf{w}^2 - \sum_i a_i \left[t_i \left(\mathbf{w} \cdot \mathbf{x_i} + b \right) - 1 \right] \tag{C.5}$$

Stationary solutions for $\mathbf{w}, b, \{a_i\}$ then obey

$$\mathbf{w} = \sum_i a_i t_i \mathbf{x} \tag{C.6}$$

$$\sum_i a_i t_i = 0 \tag{C.7}$$

In addition to these, the Karush-Kuhn-Tucker relations for inequality constraints imply

$$a_i \geq 0 \tag{C.8}$$

$$a_i \left[t_i \left(\mathbf{w} \cdot \mathbf{x_i} + b \right) - 1 \right] = 0 \tag{C.9}$$

By definition, all the samples except the samples closest to the hyperplane obey the strict inequality $t_i \left(\mathbf{w} \cdot \mathbf{x_i} + b \right) > 1$. Thus, Eq. (C.9) implies that $a_i \neq 0$ for only the samples (strictly) closest to the separating hyperplane. Using Eq. (C.6), it is evident that the normal to the plane $\mathbf{w}$ is a superposition of only these samples. Therefore, these samples are called "support vectors". The numerical solution then proceeds using the methods of quadratic programming to find a_i.

There are two important generalizations of this basic formulation. First, training samples can be separable using a non-linear hypersurface and not a hyperplane. This extension can be made by resolving the constraints in Eqs. (C.6, C.7) to get

$$\tilde{L}(\{a_i\}) = \sum_i a_i - \frac{1}{2} \sum_{i,j} a_i a_j t_i t_j k(\mathbf{x_i}, \mathbf{x_j}) \tag{C.10}$$

where I have introduced the kernel function $k(\mathbf{x_i}, \mathbf{x_j}) \cdot x_i x_j$. Using different forms for $k(\mathbf{x_i}, \mathbf{x_j})$ can result in non-linear separating hypersurfaces. For example some of the popular choices for kernel functions are

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\gamma |\mathbf{x_i} - \mathbf{x_j}|\right)^2 \text{ where } \gamma > 0 \qquad \text{(radial basis function)} \tag{C.11}$$

$$k(\mathbf{x_i}, \mathbf{x_j}) = \tanh\left(\mathbf{x_i} \cdot \mathbf{x_j} + \gamma\right) \qquad \text{(sigmoidal)} \tag{C.12}$$

$$k(\mathbf{x_i}, \mathbf{x_j}) = \left(\mathbf{x_i} \cdot \mathbf{x_j} + \gamma\right)^d \qquad \text{(polynomial)} \tag{C.13}$$

In each case, the decision function is given by

$$y(\mathbf{x}) = \sum_i a_i t_i k(\mathbf{x_i}, \mathbf{x}) + b \tag{C.14}$$

The second generalization to the original SVM algorithm addresses the problem of "soft" classification. Often training data are not completely separable but we still wish

to classify the data within some error margins. For this, an error function e_i is assigned to each point. For training samples which are well classified, i.e. $t_i y(\mathbf{x_i}) \geq 1$, the error function is $e_i = 0$. But for points which are not correctly classified $e_i = |t_i - y(\mathbf{x_i})|$. To introduce "softness" in classification, we can then make the following addition to the minimization function

$$L(\mathbf{w}, \mathbf{e_i}) = \frac{1}{2} w^2 + C \sum_i e_i \tag{C.15}$$

where C controls the amount of "softness" and is inversely proportional to the mean error rate. In the limit of $C \to \infty$, we recover the "hard" classification of above, and e_i are unconstrained in the limit of $C \to 0$. Using a pre-determined value for C, one can then repeat the procedure above to reduce this problem to a quadratic programming problem, which can be solved using established numerical techniques.

Bibliography

[1] Chantal Hulo, Edouard de Castro, Patrick Masson, Lydie Bougueleret, Amos Bairoch, Ioannis Xenarios, and Philippe Le Mercier. Viralzone: a knowledge resource to understand virus diversity. *Nucleic Acids Research*, 39(suppl 1):D576–D582, 2011.

[2] D Baltimore. Expression of animal virus genomes. *Bacteriological Reviews*, 35(3):235–241, 1971.

[3] Thomas D. Goddard, Conrad C. Huang, and Thomas E. Ferrin. Software extensions to {UCSF} chimera for interactive visualization of large molecular assemblies. *Structure*, 13(3):473 – 482, 2005.

[4] Shanshan Cheng and Charles L. Brooks, III. Viral capsid proteins are segregated in structural fold space. *PLoS Comput Biol*, 9(2):e1002905, 02 2013.

[5] Reza Khayat, Liang Tang, Eric T. Larson, C. Martin Lawrence, Mark Young, and John E. Johnson. Structure of an archaeal virus capsid protein reveals a common ancestry to eukaryotic and bacterial viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 102(52):18944–18949, 2005.

[6] Nicola GA Abrescia, Dennis H Bamford, Jonathan M Grimes, and David I Stuart. Structure unifies the viral universe. *Annual review of biochemistry*, 81:795–822, 2012.

[7] Shinya Yamada, Yasuo Suzuki, Takashi Suzuki, Mai Q Le, Chairul A Nidom,

Yuko Sakai-Tagawa, Yukiko Muramoto, Mutsumi Ito, Maki Kiso, Taisuke Horimoto, Kyoko Shinya, Toshihiko Sawada, Makoto Kiso, Taiichi Usui, Takeomi Murata, Yipu Lin, Alan Hay, Lesley F Haire, David J Stevens, Rupert J Russell, Steven J Gamblin, John J Skehel, and Yoshihiro Kawaoka. Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors. *Nature*, 444(7117):378–82, November 2006.

[8] Warren L DeLano. The pymol molecular graphics system. 2002.

[9] Giovanni Cattoli, Adelaide Milani, Nigel Temperton, Bianca Zecchin, Alessandra Buratin, Eleonora Molesti, Mona Meherez Aly, Abdel Arafa, and Ilaria Capua. Antigenic drift in H5N1 avian influenza virus in poultry is driven by mutations in major antigenic sites of the hemagglutinin molecule analogous to those for human influenza virus. *Journal of virology*, 85(17):8718–24, September 2011.

[10] Ambrish Roy, Alper Kucukural, and Yang Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols*, 5(4):725–38, April 2010.

[11] James H. Strauss and Ellen G. Strauss. *Viruses and Human Disease, Second Edition*. Academic Press, 2007.

[12] Eugene V Koonin, Tatiana G Senkevich, and Valerian V Dolja. The ancient Virus World and evolution of cells. *Biology direct*, 1(1):29, January 2006.

[13] RC Gallo, PS Sarin, EP Gelmann, M Robert-Guroff, E Richardson, VS Kalyanaraman, D Mann, GD Sidhu, RE Stahl, S Zolla-Pazner, J Leibowitch, and M Popovic. Isolation of human t-cell leukemia virus in acquired immune deficiency syndrome (aids). *Science*, 220(4599):865–867, 1983.

[14] F Barre-Sinoussi, JC Chermann, F Rey, MT Nugeyre, S Chamaret, J Gruest, C Dauguet, C Axler-Blin, F Vezinet-Brun, C Rouzioux, W Rozenbaum, and

L Montagnier. Isolation of a t-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (aids). *Science*, 220(4599):868–871, 1983.

[15] Peter Palese. Influenza: old and new threats. *Nature Medicine*, 10(12 Suppl):S82–S87, 2004.

[16] Patrick S Moore and Yuan Chang. Why do viruses cause cancer? highlights of the first century of human tumour virology. *Nature Reviews Cancer*, 10(12):878–889, 2010.

[17] Yu Li, Darin S Carroll, Shea N Gardner, Matthew C Walsh, Elizabeth A Vitalis, and Inger K Damon. On the origin of smallpox: correlating variola phylogenics with historical smallpox records. *Proceedings of the National Academy of Sciences of the United States of America*, 104(40):15787–92, October 2007.

[18] Steffen Mueller, Eckard Wimmer, and Jeronimo Cello. Poliovirus and poliomyelitis: a tale of guts, brains, and an accidental event. *Virus Research*, 111(2):175–193, 2005.

[19] WHO Fact Sheet: Rabies. *http://www.who.int/mediacentre/factsheets/fs099/en/*, accessed August 8, 2013.

[20] Barker LF, Shulman N, Murray R, and et al. Transmission of serum hepatitis. *JAMA*, 211(9):1509–1512, 1970.

[21] AMQ King, MJ Adams, EB Carstens, and EJ Lefcowits. Ninth report of the international committee on taxonomy of viruses, 2011.

[22] Nadège Philippe, Matthieu Legendre, Gabriel Doutre, Yohann Couté, Olivier Poirot, Magali Lescot, Defne Arslan, Virginie Seltzer, Lionel Bertaux, Christophe Bruley, et al. Pandoraviruses: Amoeba viruses with genomes up

to 2.5 mb reaching that of parasitic eukaryotes. *Science*, 341(6143):281–286, 2013.

[23] Carl R Woese. Bacterial evolution. *Microbiological reviews*, 51(2):221, 1987.

[24] Natalya Yutin, Pere Puigb, Eugene V. Koonin, and Yuri I. Wolf. Phylogenomics of prokaryotic ribosomal proteins. *PLoS ONE*, 7(5):e36972, 05 2012.

[25] Valerian V Dolja and Eugene V Koonin. *Capsid-Less RNA Viruses*. John Wiley and Sons, Ltd, 2012.

[26] David Prangishvili, Patrick Forterre, and Roger A Garrett. Viruses of the archaea: a unifying view. *Nature Reviews Microbiology*, 4(11):837–848, 2006.

[27] D. L. Caspar and A Klug. Physical principles in the construction of regular viruses. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 27, pages 1–24. Cold Spring Harbor Laboratory Press, 1962.

[28] Mart Krupovic and Dennis H Bamford. Virus evolution: how far does the double β-barrel viral lineage extend? *Nature Reviews Microbiology*, 6(12):941–948, 2008.

[29] Mart Krupovic and Dennis H Bamford. *Protein Conservation in Virus Evolution*. John Wiley & Sons, Ltd, 2011.

[30] Jo Handelsman, Michelle R. Rondon, Sean F. Brady, Jon Clardy, and Robert M. Goodman. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, 5(10):R245 – R249, 1998.

[31] Mya Breitbart, Peter Salamon, Bjarne Andresen, Joseph M. Mahaffy, Anca M. Segall, David Mead, Farooq Azam, and Forest Rohwer. Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences*, 99(22):14250–14255, 2002.

[32] Robert A Edwards and Forest Rohwer. Viral metagenomics. *Nature Reviews Microbiology*, 3(6):504–510, 2005.

[33] John L Mokili, Forest Rohwer, and Bas E Dutilh. Metagenomics and future perspectives in virus discovery. *Current opinion in virology*, 2(1):63–77, February 2012.

[34] L Fancello, Didier Raoult, and C Desnues. Computational tools for viral metagenomics and their application in clinical research. *Virology*, 2012.

[35] Gustavo Palacios, Julian Druce, Lei Du, Thomas Tran, Chris Birch, Thomas Briese, Sean Conlan, Phenix-Lan Quan, Jeffrey Hui, John Marshall, et al. A new arenavirus in a cluster of fatal transplant-associated diseases. *New England journal of medicine*, 358(10):991–998, 2008.

[36] Joseph G Victoria, Amit Kapoor, Linlin Li, Olga Blinkova, Beth Slikas, Chunlin Wang, Asif Naeem, Sohail Zaidi, and Eric Delwart. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *Journal of virology*, 83(9):4642–4651, 2009.

[37] Dana Willner, Mike Furlan, Matthew Haynes, Robert Schmieder, Florent E Angly, Joas Silva, Sassan Tammadoni, Bahador Nosrat, Douglas Conrad, and Forest Rohwer. Metagenomic analysis of respiratory tract dna viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One*, 4(10):e7370, 2009.

[38] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[39] Susana Vinga and Jonas Almeida. Alignment-free sequence comparisona review. *Bioinformatics*, 19(4):513–523, 2003.

[40] Alice C McHardy and Isidore Rigoutsos. What's in the mix: phylogenetic classification of metagenome sequence samples. *Current opinion in microbiology*, 10(5):499–503, 2007.

[41] Hanno Teeling, Jost Waldmann, Thierry Lombardot, Margarete Bauer, and Frank O Glöckner. Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC bioinformatics*, 5(1):163, 2004.

[42] Alice Carolyn McHardy, Héctor García Martín, Aristotelis Tsirigos, Philip Hugenholtz, and Isidore Rigoutsos. Accurate phylogenetic classification of variable-length DNA fragments. *Nature methods*, 4(1):63–72, January 2007.

[43] Arthur Brady and Steven L Salzberg. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature methods*, 6(9):673–6, September 2009.

[44] Shannon J Williamson, Lisa Zeigler Allen, Hernan A Lorenzi, Douglas W Fadrosh, Daniel Brami, Mathangi Thiagarajan, John P McCrow, Andrey Tovchigrechko, Shibu Yooseph, and J Craig Venter. Metagenomic exploration of viruses throughout the indian ocean. *PloS one*, 7(10):e42047, 2012.

[45] Robert G Webster, William J Bean, Owen T Gorman, Thomas M Chambers, and Yoshihiro Kawaoka. Evolution and ecology of influenza a viruses. *Microbiological reviews*, 56(1):152–179, 1992.

[46] Rafael Lozano, Mohsen Naghavi, Kyle Foreman, Stephen Lim, Kenji Shibuya, Victor Aboyans, Jerry Abraham, Timothy Adair, Rakesh Aggarwal, Stephanie Y Ahn, Mohammad A AlMazroa, Miriam Alvarado, H Ross Anderson, Laurie M Anderson, Kathryn G Andrews, Charles Atkinson, Larry M Baddour, Suzanne Barker-Collo, David H Bartels, Michelle L Bell, Emelia J Benjamin, Derrick Bennett, Kavi Bhalla, Boris Bikbov, Aref Bin Abdulhak,

Gretchen Birbeck, Fiona Blyth, Ian Bolliger, Soufiane Boufous, Chiara Bucello, Michael Burch, Peter Burney, Jonathan Carapetis, Honglei Chen, David Chou, Sumeet S Chugh, Luc E Coffeng, Steven D Colan, Samantha Colquhoun, K Ellicott Colson, John Condon, Myles D Connor, Leslie T Cooper, Matthew Corriere, Monica Cortinovis, Karen Courville de Vaccaro, William Couser, Benjamin C Cowie, Michael H Criqui, Marita Cross, Kaustubh C Dabhadkar, Nabila Dahodwala, Diego De Leo, Louisa Degenhardt, Allyne Delossantos, Julie Denenberg, Don C Des Jarlais, Samath D Dharmaratne, E Ray Dorsey, Tim Driscoll, Herbert Duber, Beth Ebel, Patricia J Erwin, Patricia Espindola, Majid Ezzati, Valery Feigin, Abraham D Flaxman, Mohammad H Forouzanfar, Francis Gerry R Fowkes, Richard Franklin, Marlene Fransen, Michael K Freeman, Sherine E Gabriel, Emmanuela Gakidou, Flavio Gaspari, Richard F Gillum, Diego Gonzalez-Medina, Yara A Halasa, Diana Haring, James E Harrison, Rasmus Havmoeller, Roderick J Hay, Bruno Hoen, Peter J Hotez, Damian Hoy, Kathryn H Jacobsen, Spencer L James, Rashmi Jasrasaria, Sudha Jayaraman, Nicole Johns, Ganesan Karthikeyan, Nicholas Kassebaum, Andre Keren, Jon-Paul Khoo, Lisa Marie Knowlton, Olive Kobusingye, Adofo Koranteng, Rita Krishnamurthi, Michael Lipnick, Steven E Lipshultz, Summer Lockett Ohno, Jacqueline Mabweijano, Michael F MacIntyre, Leslie Mallinger, Lyn March, Guy B Marks, Robin Marks, Akira Matsumori, Richard Matzopoulos, Bongani M Mayosi, John H McAnulty, Mary M McDermott, John McGrath, Ziad A Memish, George A Mensah, Tony R Merriman, Catherine Michaud, Matthew Miller, Ted R Miller, Charles Mock, Ana Olga Mocumbi, Ali A Mokdad, Andrew Moran, Kim Mulholland, M Nathan Nair, Luigi Naldi, K M Venkat Narayan, Kiumarss Nasseri, Paul Norman, Martin O'Donnell, Saad B Omer, Katrina Ortblad, Richard Osborne, Doruk Ozgediz, Bishnu Pahari, Jeyaraj Durai Pandian, Andrea Panozo Rivero, Rogelio Perez Padilla, Fernando Perez-Ruiz, Norberto Perico, David Phillips, Kelsey Pierce, C Arden Pope III, Esteban

Porrini, Farshad Pourmalek, Murugesan Raju, Dharani Ranganathan, Jrgen T Rehm, David B Rein, Guiseppe Remuzzi, Frederick P Rivara, Thomas Roberts, Felipe Rodriguez De Len, Lisa C Rosenfeld, Lesley Rushton, Ralph L Sacco, Joshua A Salomon, Uchechukwu Sampson, Ella Sanman, David C Schwebel, Maria Segui-Gomez, Donald S Shepard, David Singh, Jessica Singleton, Karen Sliwa, Emma Smith, Andrew Steer, Jennifer A Taylor, Bernadette Thomas, Imad M Tleyjeh, Jeffrey A Towbin, Thomas Truelsen, Eduardo A Undurraga, N Venketasubramanian, Lakshmi Vijayakumar, Theo Vos, Gregory R Wagner, Mengru Wang, Wenzhi Wang, Kerrianne Watt, Martin A Weinstock, Robert Weintraub, James D Wilkinson, Anthony D Woolf, Sarah Wulf, Pon-Hsiu Yeh, Paul Yip, Azadeh Zabetian, Zhi-Jie Zheng, Alan D Lopez, and Christopher JL Murray. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *The Lancet*, 380(9859):2095 – 2128, 2013.

[47] Taisuke Horimoto and Yoshihiro Kawaoka. Influenza: lessons from past pandemics, warnings from current incidents. *Nature Reviews Microbiology*, 3(8):591–600, 2005.

[48] Stephanie Sonnberg, Richard J Webby, and Robert G Webster. Natural history of highly pathogenic avian influenza h5n1. *Virus research*, 2013.

[49] Maurice R Hilleman. Realities and enigmas of human viral influenza: pathogenesis, epidemiology and control. *Vaccine*, 20(25):3068–3087, 2002.

[50] Alan J Hay, Victoria Gregory, Alan R Douglas, and Yi Pu Lin. The evolution of human influenza viruses. *Philos Trans R Soc Lond B Biol Sci*, 356(1416):1861–70, 2001.

[51] Edwin D Kilbourne. Influenza pandemics of the 20th century. *Emerging infectious diseases*, 12(1):9, 2006.

[52] Marc P Girard, John S Tam, Olga M Assossou, and Marie Paule Kieny. The 2009 a (h1n1) influenza virus pandemic: A review. *Vaccine*, 28(31):4895–4902, 2010.

[53] Christopher W Potter. A history of influenza. *Journal of applied microbiology*, 91(4):572–579, 2001.

[54] Jeffery K Taubenberger, Ann H Reid, Raina M Lourens, Ruixue Wang, Guozhong Jin, and Thomas G Fanning. Characterization of the 1918 influenza virus polymerase genes. *Nature*, 437(7060):889–893, 2005.

[55] Gavin JD Smith, Dhanasekaran Vijaykrishna, Justin Bahl, Samantha J Lycett, Michael Worobey, Oliver G Pybus, Siu Kit Ma, Chung Lam Cheung, Jayna Raghwani, Samir Bhatt, et al. Origins and evolutionary genomics of the 2009 swine-origin h1n1 influenza a epidemic. *Nature*, 459(7250):1122–1125, 2009.

[56] Yohei Watanabe, Madiha S Ibrahim, Yasuo Suzuki, and Kazuyoshi Ikuta. The changing nature of avian influenza a virus (h5n1). *Trends in microbiology*, 20(1):11–20, 2012.

[57] Kumnuan Ungchusak, Prasert Auewarakul, Scott F Dowell, Rungrueng Kitphati, Wattana Auwanit, Pilaipan Puthavathana, Mongkol Uiprasertkul, Kobporn Boonnak, Chakrarat Pittayawonganon, Nancy J Cox, et al. Probable person-to-person transmission of avian influenza a (h5n1). *New England Journal of Medicine*, 352(4):333–340, 2005.

[58] Kyoko Shinya, Masahito Ebina, Shinya Yamada, Masao Ono, Noriyuki Kasai, and Yoshihiro Kawaoka. Avian flu: influenza virus receptors in the human airway. *Nature*, 440(7083):435–6, March 2006.

[59] Masaki Imai, Tokiko Watanabe, Masato Hatta, Subash C Das, Makoto Ozawa, Kyoko Shinya, Gongxun Zhong, Anthony Hanson, Hiroaki Katsura, Shinji

Watanabe, Chengjun Li, Eiryo Kawakami, Shinya Yamada, Maki Kiso, Yasuo Suzuki, Eileen A Maher, Gabriele Neumann, and Yoshihiro Kawaoka. Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature*, 486(7403):420–8, June 2012.

[60] Sander Herfst, Eefje J A Schrauwen, Martin Linster, Salin Chutinimitkul, Emmie de Wit, Vincent J Munster, Erin M Sorrell, Theo M Bestebroer, David F Burke, Derek J Smith, Guus F Rimmelzwaan, Albert D M E Osterhaus, and Ron A M Fouchier. Airborne transmission of influenza A/H5N1 virus between ferrets. *Science (New York, N.Y.)*, 336(6088):1534–41, June 2012.

[61] Colin A Russell, Judith M Fonville, André E X Brown, David F Burke, David L Smith, Sarah L James, Sander Herfst, Sander van Boheemen, Martin Linster, Eefje J Schrauwen, Leah Katzelnick, Ana Mosterín, Thijs Kuiken, Eileen Maher, Gabriele Neumann, Albert D M E Osterhaus, Yoshihiro Kawaoka, Ron A M Fouchier, and Derek J Smith. The potential for respiratory droplet-transmissible A/H5N1 influenza virus to evolve in a mammalian host. *Science (New York, N.Y.)*, 336(6088):1541–7, June 2012.

[62] Dhanasekaran Vijaykrishna, Justin Bahl, Steven Riley, Lian Duan, Jin Xia Zhang, Honglin Chen, J S Malik Peiris, Gavin J D Smith, and Yi Guan. Evolutionary dynamics and emergence of panzootic H5N1 influenza viruses. *PLoS pathogens*, 4(9):e1000161, January 2008.

[63] Honglin Chen, GJD Smith, SY Zhang, K Qin, J Wang, KS Li, RG Webster, JSM Peiris, and Y Guan. Avian flu: H5n1 virus outbreak in migratory waterfowl. *Nature*, 436(7048):191–192, 2005.

[64] Magdi D Saad, S Ahmed Luay, Mohamed A Gamal-Eldein, Mohamed K Fouda,

Fouda M Khalil, Samuel L Yingst, Michael A Parker, and Marshall R Montevillel. Possible avian influenza (h5n1) from migratory bird, egypt. *Emerging infectious diseases*, 13(7):1120, 2007.

[65] Yohei Watanabe, Madiha S Ibrahim, Hany F Ellakany, Norihito Kawashita, Rika Mizuike, Hiroaki Hiramatsu, Nogluk Sriwilaijaroen, Tatsuya Takagi, Yasuo Suzuki, and Kazuyoshi Ikuta. Acquisition of human-type receptor binding specificity by new H5N1 influenza virus sublineages during their emergence in birds in Egypt. *PLoS pathogens*, 7(5):e1002068, May 2011.

[66] Guang-Wu Chen, Shih-Cheng Chang, Chee-Keng Mok, Yu-Luan Lo, Yu-Nong Kung, Ji-Hung Huang, Yun-Han Shih, Ji-Yi Wang, Chiayn Chiang, Chi-Jene Chen, et al. Genomic signatures of human versus avian influenza a viruses. *Emerging infectious diseases*, 12(9):1353, 2006.

[67] David B Finkelstein, Suraj Mukatira, Perdeep K Mehta, John C Obenauer, Xiaoping Su, Robert G Webster, and Clayton W Naeve. Persistent host markers in pandemic and h5n1 influenza viruses. *Journal of virology*, 81(19):10292–10299, 2007.

[68] Olivo Miotto, AT Heiny, Randy Albrecht, Adolfo García-Sastre, Tin Wee Tan, J Thomas August, and Vladimir Brusic. Complete-proteome mapping of human influenza a adaptive mutations: implications for human transmissibility of zoonotic strains. *PloS one*, 5(2):e9025, 2010.

[69] Suryaprakash Sambhara and Gregory A Poland. H5N1 Avian influenza: preventive and therapeutic strategies against a pandemic. *Annual review of medicine*, 61:187–98, January 2010.

[70] J S Malik Peiris, Menno D de Jong, and Yi Guan. Avian influenza virus (H5N1): a threat to human health. *Clinical microbiology reviews*, 20(2):243–67, April 2007.

[71] Y Guan, L L M Poon, C Y Cheung, T M Ellis, W Lim, A S Lipatov, K H Chan, K M Sturm-Ramirez, C L Cheung, Y H C Leung, K Y Yuen, R G Webster, and J S M Peiris. H5N1 influenza: a protean pandemic threat. *Proceedings of the National Academy of Sciences of the United States of America*, 101(21):8156–61, May 2004.

[72] Kelvin KW To, Kenneth HL Ng, Tak-Lun Que, Jacky MC Chan, Kay-Yan Tsang, Alan KL Tsang, Honglin Chen, and Kwok-Yung Yuen. Avian influenza A H5N1 virus: a continuous threat to humans. *Emerging Microbes & Infections*, 1(9):e25, September 2012.

[73] G J D Smith, T S P Naipospos, T D Nguyen, M D de Jong, D Vijaykrishna, T B Usman, S S Hassan, T V Nguyen, T V Dao, N A Bui, Y H C Leung, C L Cheung, J M Rayner, J X Zhang, L J Zhang, L L M Poon, K S Li, V C Nguyen, T T Hien, J Farrar, R G Webster, H Chen, J S M Peiris, and Y Guan. Evolution and adaptation of H5N1 influenza virus in avian and human hosts in Indonesia and Vietnam. *Virology*, 350(2):258–68, July 2006.

[74] Venkata R S K Duvvuri, Bhargavi Duvvuri, Wilfred R Cuff, Gillian E Wu, and Jianhong Wu. Role of positive selection pressure on the evolution of H5N1 hemagglutinin. *Genomics, proteomics & bioinformatics*, 7(1-2):47–56, June 2009.

[75] Kaifa Wei, Yanfeng Chen, Juan Chen, Lingjuan Wu, and Daoxin Xie. Evolution and adaptation of hemagglutinin gene of human H5N1 influenza virus. *Virus genes*, 44(3):450–8, June 2012.

[76] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[77] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006.

[78] Eric Jones, Travis Oliphant, and Pearu Peterson. SciPy: Open source scientific tools for Python. *http://www.scipy.org/*, 2001.

[79] Keyao Pan and Michael W Deem. Quantifying selection and diversity in viruses by entropy methods, with application to the haemagglutinin of H3N2 influenza. *Journal of the Royal Society, Interface / the Royal Society*, 8(64):1644–53, November 2011.

[80] Cuong Cao Dang, Quang Si Le, Olivier Gascuel, and Vinh Sy Le. FLU, an amino acid substitution model for influenza proteins. *BMC evolutionary biology*, 10(1):99, January 2010.

[81] Tommy Tsan-Yuk Lam, Chung-Chau Hon, Philippe Lemey, Oliver G Pybus, Mang Shi, Hein Min Tun, Jun Li, Jingwei Jiang, Edward C Holmes, and Frederick Chi-Ching Leung. Phylodynamics of h5n1 avian influenza virus in indonesia. *Molecular Ecology*, 21(12):3062–3077, 2012.

[82] Stéphane Guindon and Olivier Gascuel. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, 52(5):696–704, October 2003.

[83] Ziheng Yang. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *Journal of Molecular evolution*, 39(3):306–314, 1994.

[84] A Arafa, D Suarez, S G Kholosy, M K Hassan, S Nasef, A Selim, G Dauphin, M Kim, J Yilma, D Swayne, and M M Aly. Evolution of highly pathogenic avian influenza H5N1 viruses in Egypt indicating progressive adaptation. *Archives of virology*, 157(10):1931–47, October 2012.

[85] E M Abdelwhab, Abdel-Satar Arafa, Jürgen Stech, Christian Grund, Olga Stech, Marcus Graeber-Gerberding, Martin Beer, Mohamed K Hassan, Mona M

Aly, Timm C Harder, and Hafez M Hafez. Diversifying evolution of highly pathogenic H5N1 avian influenza virus in Egypt from 2006 to 2011. *Virus genes*, 45(1):14–23, August 2012.

[86] Calogero Terregino, Anna Toffan, Filippo Cilloni, Isabella Monne, Elena Bertoli, Lilia Castellanos, Nadim Amarin, Marzia Mancin, and Ilaria Capua. Evaluation of the protection induced by avian influenza vaccines containing a 1994 Mexican H5N2 LPAI seed strain against a 2008 Egyptian H5N1 HPAI virus belonging to clade 2.2.1 by means of serological and in vivo tests. *Avian pathology : journal of the W.V.P.A*, 39(3):215–22, June 2010.

[87] Ding-Kwo Chang, Shu-Fang Cheng, Eric Kantchev, Chi-Hui Lin, and Yu-Tsan Liu. Membrane interaction and structure of the transmembrane domain of influenza hemagglutinin and its fusion peptide complex. *BMC biology*, 6(1):2, 2008.

[88] Mya Breitbart and Forest Rohwer. Here a virus, there a virus, everywhere the same virus? *Trends in microbiology*, 13(6):278–84, June 2005.

[89] Elizabeth A Dinsdale, Robert A Edwards, Dana Hall, Florent Angly, Mya Breitbart, Jennifer M Brulc, Mike Furlan, Christelle Desnues, Matthew Haynes, Linlin Li, et al. Functional metagenomic profiling of nine biomes. *Nature*, 452(7187):629–632, 2008.

[90] Didier Raoult and Patrick Forterre. Redefining viruses: lessons from mimivirus. *Nature Reviews Microbiology*, 6(4):315–319, 2008.

[91] Patrick Forterre. The origin of viruses and their possible roles in major evolutionary transitions. *Virus research*, 117(1):5–16, 2006.

[92] Gareth Chelvanayagam, Jaap Heringa, and Patrick Argos. Anatomy and evolution of proteins displaying the viral capsid jellyroll topology. *Journal of molecular biology*, 228(1):220–242, 1992.

[93] Guohong Albert Wu, Se-Ran Jun, Gregory E Sims, and Sung-Hou Kim. Whole-proteome phylogeny of large dsdna virus families by an alignment-free method. *Proceedings of the National Academy of Sciences*, 106(31):12826–12831, 2009.

[94] Yi Wang, Henry CM Leung, Siu-Ming Yiu, and Francis YL Chin. Metacluster 4.0: a novel binning algorithm for ngs reads and huge number of species. *Journal of Computational Biology*, 19(2):241–249, 2012.

[95] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(suppl 1):D61–D65, 2007.

[96] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[97] Corinna Cortes and Vladimir Vapnik. Support vector machine. *Machine learning*, 20(3):273–297, 1995.

[98] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[99] Grzegorz M Boratyn, AA Schaffer, Richa Agarwala, Stephen F Altschul, David J Lipman, and Thomas L Madden. Domain enhanced lookup time accelerated blast. *Biol Direct*, 7(1):12, 2012.

[100] Aron Marchler-Bauer, Shennan Lu, John B Anderson, Farideh Chitsaz, Myra K Derbyshire, Carol DeWeese-Scott, Jessica H Fong, Lewis Y Geer, Renata C Geer, Noreen R Gonzales, et al. Cdd: a conserved domain database for the functional annotation of proteins. *Nucleic acids research*, 39(suppl 1):D225–D229, 2011.

[101] Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *science*, 338(6110):1042–1046, 2012.

[102] Morten Nielsen, Claus Lundegaard, Ole Lund, and Thomas Nordahl Petersen. Cphmodels-3.0remote homology modeling using structure-guided sequence profiles. *Nucleic acids research*, 38(suppl 2):W576–W581, 2010.

[103] Simon Roux, Francois Enault, Agnès Robin, Viviane Ravet, Sébastien Personnic, Sébastien Theil, Jonathan Colombet, Télesphore Sime-Ngando, and Didier Debroas. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PloS one*, 7(3):e33641, 2012.

[104] Simon Roux, Michaël Faubladier, Antoine Mahul, Nils Paulhe, Aurélien Bernard, Didier Debroas, and François Enault. Metavir: a web server dedicated to virome analysis. *Bioinformatics*, 27(21):3074–3075, 2011.

[105] Brian M Meehan, Julie L Creelan, M Stewart McNulty, and Daniel Todd. Sequence of porcine circovirus dna: affinities with plant circoviruses. *Journal of General Virology*, 78(1):221–227, 1997.

[106] T Alwyn Jones and Lars Liljas. Structure of satellite tobacco necrosis virus after crystallographic refinement at 2.5 å resolution. *Journal of molecular biology*, 177(4):735–767, 1984.

[107] P Hopper, SC Harrison, and RT Sauer. Structure of tomato bushy stunt virus: V. coat protein sequence determination and its structural implications. *Journal of molecular biology*, 177(4):701–713, 1984.

[108] Liam J McGuffin, Kevin Bryson, and David T Jones. The psipred protein structure prediction server. *Bioinformatics*, 16(4):404–405, 2000.

Vita

Education

- B.Tech. Engineering Physics, Indian Institute of Technology Bombay – 2001-2005.
- Ph.D. Physics, Rutgers the State University of New Jersey – 2006-present.

Publications

- *Lactase persistence and lipid pathway selection in the Maasai*
  K. Wagh, A. Bhatia, S. Lukic, G. Alexe, A. Reddy, V. Ravikumar, M. Seiler, M. Boemo, M. Yao, L. Cronk, A. Naqvi, S. Ganesan, G. Bhanot, A. Levine. PLoS ONE **7(9)** e44751 (2012).
- *The Link between Integrability, Level Crossings, and Exact Solution in Quantum Models*
  H. K. Owusu, K. Wagh, E. A. Yuzbashyan, J. Phys. A: Math. Theor. **42** 035206 (2009).
- *Equilibration problem for the generalized Langevin equation*
  Abhishek Dhar, Kshitij Wagh, Europhys. Lett. **79** 60003 (2007).
- *Top production at the Tevatron/LHC and nonstandard, strongly interacting spin one particles*
  Debajyoti Choudhury, Rohini M. Godbole, Ritesh K. Singh, Kshitij Wagh, Phys. Lett. B **657** 69 (2007).