ON THE EVOLUTION OF MICROBES:
THE EVOLUTION OF GENOMES WITH RESPECT TO RNA FOLDING

by

Rachel Brower-Sinning

Biology BS, Virginia Tech, 2004

Physics BA, Virginia Tech, 2004

Bioinformatics MS, George Mason University, 2006

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Rachel Brower-Sinning

It was defended on

October 31, 2011

and approved by

Roni Rosenfeld, Ph.D., Professor, Department of Computer Science

Ted M. Ross, Ph.D., Associate Professor,

Microbiology & Molecular Genetics, Center for Vaccine Research

Daniel M. Zuckerman, Associate Professor,

Department of Computational and Systems Biology

Dissertation Advisor: Panayiotis V. Benos, Associate Professor,

Department of Computational and Systems Biology

ON THE EVOLUTION OF MICROBES:

THE EVOLUTION OF GENOMES WITH RESPECT TO RNA FOLDING

Rachel Brower-Sinning, MS

University of Pittsburgh, 2011

We hypothesized that the stringency by which RNA folds (summarized in our analysis by the predicted folding free energy (FFE)) may be under selective pressure, presumably due to its role in (reverse) transcription and translation, and its potential effect on the RNA degradation rate. For bacteria, the RNA folding will depend on the physical properties of their environment. For viruses, this balance needs to be reached for every host the virus is successfully replicated in, and may play a critical role in adapting to new hosts.

In the influenza A virus, we have shown that the FFE of its polymerase genes is evolving through time from lower to higher values, every time an avian segment jumps into humans. We postulated that this may be related to the difference in body temperature between humans and birds, as generally the genes isolated from avian sources have significantly lower FFE than the human isolates.

Furthermore, we can use the FFE and amino acid sequence of the influenza A virus, to classify whether a given virus is similar to others that can jump to and successfully infect human hosts.

In bacteria, we have shown that, consistent with previous studies of GC content, tRNA FFE is linearly correlated with growth temperature; while mRNA FFE is not. Regardless, we showed that the growth conditions are related to mRNA FFE distributions and function.

Furthermore, there is a relation between mRNA FFE and half-life. Finally, we showed that gene expression can be predicted from RNA structure and sequence properties.

In studying RNA folding in both viruses and bacteria, we were able to view the possible association between FFE and environment in two ways: the number of bacterial genomes sequenced allows us to get a sense of what RNA structures and folding energies are required for the bacteria to inhabit a wide variety of environments- everything from the human body to colonizing black smokers on the ocean floor; while the number of influenza A genomes sequenced allows us to determine how the RNA structures change over time. By using both sets of information, we can get a clearer picture of both the importance of RNA structure, and how RNA structure and folding energy evolve as the host environment changes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This research was supported by National Institutes of Health Training grant T32 EB009403 as part of the Howard Hughes Medical Institute- National Institutes of Biomedical Imaging and Bioengineering Interfaces Initiative in support of Rachel Brower-Sinning as a NRSA Institutional Training Grant recipient; and by the Institute for Clinical Research Education and Clinical and Translational Science Institute, grant 5TL1RR24155-4 in support of Rachel Brower-Sinning as a Predoctoral Fellow in Clinical and Translational Research.

This work started by asking of the question “does RNA structure play a role in influenza?” in an early meeting with my primary advisor, Dr. Takis Benos; later evolving into this dissertation. I would like to thank Dr. Benos, and my thesis committee, for making this dissertation possible.

I would also like to thank my family and friends; specifically my parents, James and Michele Brower, and my in-laws Bill and Karen Sinning for their love and support; Matt and Jenn Legler for their support; and my spouse, Craig Sinning, for his love, support and IT expertise. This would have been much more difficult without you.

1.0 INTRODUCTION

According to the central dogma, messenger RNA (mRNA) functions as an intermediary, the means for shuttling information from the DNA genome to the protein workhorses of the cell. To decode the mRNA and synthesize protein, ribosomal RNA (rRNA) and transfer RNA (tRNA) are the key players, with rRNA binding to the mRNA, tRNA translating the mRNA codons, and rRNA forming the peptide bonds between the amino acids. While the primary function of mRNA is to encode the protein sequence, tRNA and rRNA are examples of non-coding RNA (ncRNA), which are functional RNA (RNA not translated into protein). Other classes of non-coding RNA (ncRNA), which display a variety of functions and properties, include: ribozymes, which catalyze chemical reactions; small nuclear RNA (snRNA), which function in RNA splicing, transcription factor regulation and telomere maintenance; small nucleolar RNA (snoRNA), which guide chemical modification of other RNAs; small interfering RNA (siRNA) and microRNA (miRNA), which bind to mRNA and subsequently decrease the concentration of its protein product *via* the RNA interference (RNAi) mechanism.

RNA structure is critical to the function of these ncRNAs. [1] In tRNA, folding is critical for both synthesis and function. [2] For ribozymes, the RNA must be folded into a specific conformation required for the catalytic activity; and this fold is highly associated with its function. [3] For the RNAi pathway, RNA structure is important at two steps- in the synthesis of

the miRNA or siRNA, the structure of the small RNA to its target, and the structure of the target duplex. [4]

As RNA structure is vital to the function of the above classes of RNA, we are asking the question – does mRNA structure matter? Does mRNA structure impact function/expression? Does evolution influence mRNA structure?

1.1 RNA FOLDING

RNA structure is composed of the canonical Watson-Crick base pairs of Guanine (G) pairing with cytosine (C) through three hydrogen bonds, and adenine (A) with uracil (U) with two. There is also the 'wobble' base pair of guanine pairing with uracil with comparable binding to an AU pair. Non-canonical base pairing is also allowed (defined as non-WC, non-wobble pairing), often with lower binding energies.[5] These paired and unpaired nucleic acids then fold to form the characteristic structures found in the secondary structure RNA- such as stems with hair-pin loops, bulge loops, internal loops, and multibranch loops. [6] (Pseudo-knots can also form, but the computational tools that we will be using in this study, RNAfold of the Vienna RNA package 1.6.5 [7-14] and Sfold [15, 16], do not consider this structure.) The RNA secondary helical structure is stable, and additional hydrogen-bonding interactions can occur, giving rise to tertiary structure interactions. [1]

Programs that computationally fold RNA provide two kinds of information: structure and folding free energy (FFE). When we computationally fold the RNA sequence, we are comparing the stabilizing effect of base-pairing (which lowers the free energy) with that of the loops and bulges (which increase the overall free energy). The net effect is that small RNA hairpins can

have FFE on the magnitude of folded proteins, while larger mRNA have free energies orders of magnitude lower. [1] Of the two folding programs used in this study, RNAfold calculates the minimum-free energy structure. [7-14] As RNA can sometimes fold in sub-optimal structures, Sfold calculates the ensemble structures. [15-17]

1.1.1 Can mRNA structure affect gene expression?

While the primary function of mRNA is to encode the genetic information that produces proteins, it also contains the signals that modulate both the translation efficacy and gene expression. [18], [1] These modulators of gene expression are not minor. Large-scale measurements of mRNA and protein abundance in mammalian systems show that mRNA abundance can explain only ~30% of protein abundance; but that including mRNA sequence and structure features can potentially explain ~66% of protein abundance. [19]

The first way may be by directly affecting the ability of the ribosome to translate the mRNA. In a recent study, Kudla *et al*. reported that in *Escherichia coli*, the folding of the region around the ribosome binding sequence (RBS) was associated with protein levels by – presumably– affecting the translation rate [20]. Additionally, Voges *et al.* has found that translation efficiency for mRNA in *E. coli* was dependent on both base pairing probability and GC content of the sequence directly downstream of the start codon, indicating that mRNA secondary structure in this region could hamper translation [21]. Similarly, in 2010 Tuller *et al.* showed that, in general, mRNA folding may function to slow down ribosomes, impeding translation [22].

The second way may be by the interaction of mRNA and RNA binding proteins and RNAi. RNA binding proteins can recognize a wide range of diverse structural motifs, with some

proteins recognizing pseudoknots and others recognizing simple RNA hairpins or single-stranded regions. In some instances, the binding site can obscure the ribosomal binding site (RBS) on the mRNA, while in others the RNA binding protein will alter the structure of the RNA, causing a conformational change that will make the RBS accessible. [23] For RNAi, the mRNA may be targeted by a miRNA or siRNA, resulting in a down-regulation of protein expression. [24]

1.1.2 What features are known to affect mRNA stability?

Rates of mRNA degradation can vary widely within a cell. However, they generally do not exceed the cell doubling time. The balance between mRNA synthesis and degradation, can also vary in response to conditions and environmental signals. [25] This was seen in Bernstein *et al*., 2002, where *E. coli* grown up on minimal media and nutrient rich media showed no correlation in mRNA half-life. [26]

Most of the work involving RNA stability has been focused on the role of the 3' and 5' ends. For some RNA molecules, RNases can degrade the RNA systematically from the 3' to 5' end; however, 5' end secondary structure may impede this process. Additionally, the RNA can be cleaved, with the secondary structures which act as blocks being removed, allowing for the degradation of structured RNA.[25] However, structurally adding secondary structure will not function to increase gene expression. [20, 21, 27]

To better understand the role mRNA structure may play in the cell, we investigated it within two paradigms- how RNA structure may be vital to an RNA virus; and how RNA structure may affect gene expression of prokaryotes. The reason for investigating RNA structure within these two diverse frameworks largely comes down to numbers- the number of viruses and prokaryotic organisms present in our environment is truly staggering, and we know only a small

portion of them.[28-31] Viruses infect all cellular organisms, and are completely dependent on the host cell for energy, amino acids and nucleosides for producing new viral particles, and the cellular machinery for protein production. As such, viruses have the capacity to exist in very diverse environments. Prokaryotic species have been found to inhabit all environments yet discovered: everything from the sea floor, to hot springs and volcanoes, to deep within the Earth, the human body, and everything in between.

1.1.3 RNA Folding and Secondary Structure Prediction

The idea of RNA secondary structure began with Doty et al. in 1959 and the observation that RNA behaved as a set of "irregularly coiled, relatively compact, single polymeric chains", defined by hydrogen bonds between the bases and sequence dependent. [32] The earliest methods developed to predict RNA secondary structure involved the use of comparative sequence analysis to infer pairing between the bases by looking for compensatory base pair changes (i.e., locations where an AU pair in one sequence was replaced by a GC pair in another). [33] In 1975, Fox and Woese used this approach to determine the structure for 5S rRNA; a structure that was later verified. [34-36]

Currently, however, the most popular methods to predict secondary structure is through the use of free energy minimization. [33, 37] The goal of this approach is to predict the RNA secondary structure of a given sequence which has the lowest free energy; the rationale for this being that, at equilibrium, a given RNA sequence with structure S should fluctuate between a folded state *S* and an unfolded state *US*. [33]

$$US \rightleftharpoons S$$

With an equilibrium constant *K*

$$K = {[S]}/{[US]}$$

Given free energy change for the structure, using the Gibbs Free energy, *K* becomes

$$K = e^{-\Delta G(S)/RT}$$

In this paradigm, if the RNA forms multiple structures, the difference between the folding energy of the optimal, minimum FFE structure (opt) and another suboptimal structure (subopt) will quantify the difference in concentration of the structures

$${K_{opt}}/{K_{subopt}} = {[S_{opt}]}/{[S_{subopt}]} = e^{(\Delta G(S_{subopt}) - \Delta G(S_{opt}))/RT}$$

So the minimum FFE structure will be the most abundant conformation at equilibrium. [33] In this study, we use the program RNAfold of the Vienna RNA Package, which calculates the MFE structure using the classic algorithm from Zuker and Stiegler outlined below. [38]

The structure with the minimum FFE is identified by using nearest neighbor free energy parameters in combination with dynamic programming algorithms. [37] An early example of this, mfold, was developed by Zuker in the 1980s. [14, 37] These algorithms were later expanded to identify sub-optimal and near-optimal folding in the late 1980s [37], but suffered from the inability to guarantee an unbiased representation of the secondary structure landscape. [39]

The algorithms which identify the optimal RNA structure through dynamic programming are highly similar to the sequence alignment algorithms using the same approach. These methods are used to identify the optimal structure for a subsequence. The relative stability of these structures is evaluated using published thermodynamic data- and these data sets are being

continually updated. The optimal structure of the entire sequence at each pair is identified through a recursive search. [40] The published thermodynamic data that is used is experimentally acquired- typically microcalorimetry or optical melting (absorbance melting curves) is used to measure the energy invested into the structure of a small nucleotide sequence in solution. [41]

In addition to not being able to generate a weighted view of the ensemble of structures that the RNA can fold into, the accuracy of the MFE methods is limited by several other factors: the free energy nearest neighbor models are incomplete, and not all the structural factors/conformations are recognized by the dynamic programming (DP) algorithms not all the RNA structures are either in equilibrium conditions or in the MFE conformation; some RNA sequences have more than one structural conformation (e.g. riboswitches), and an inability to make statistical representations of the structure. [33, 39, 42]

To circumvent and solve some of these issues of the free energy minimization approach, McCaskill in 1990 devised an algorithm that used dynamic programming to calculate the complete equilibrium partition function; and the probabilities of the various substructures. [12] The partition function is a normalizing metric that uses the energies of all the structures that a RNA primary sequence can possibly form. From this, it is possible to determine just how favorable, or more likely, a predicted structure is. [43] Where the partition function is defined as [12, 33]

$$Q = \sum_{all\ structures\ S} e^{-\Delta G(S)/RT}$$

The probability of any given structure is

$$P(structure\ S) = \frac{e^{-\Delta G(S)/RT}}{Q}$$

From this, we can see that, as the number of possible structures increases, the probability of any particular structure, included the structure with the minimum free energy, is small. [12] This is in contrast to what we saw in the justification of understanding just the MFE structure: we can see that while the MFE may be the most probable structure, its absolute probability may be tiny. One huge advantage of this approach was that that probability of any two nucleotides being paired could be calculated.[12] This revealed that quarter of the predicted base pairs in the MFE structures have a base pairing probability of greater than 99%. [33] The temperature dependence of the partition function also gives a view of the relationship between RNA secondary structure and temperature. [12]

This approach by McCaskill also allowed for the reconstruction of alternative structures and the differences in folding kinetics between them. [12] This allowed for the study of the different energy landscape of RNA folding. [44]

In continuing in this quest for suboptimally folded structures, Wuchty *et. al* in 1999 presented an algorithm that exhaustively generates all the suboptimal structures between the minimum FFE and an arbitrarily user defined cut-off. [45] From this work, one major conclusion was that these sequences whose ground state structure was thermodynamically well defined show a tendency to be buffered against single point mutations and that the number of strictures expands exponentially with increasing ΔG. [33, 45]

Using a Bayesian approach to predict RNA secondary structure, Ding and Lawrence developed an algorithm to statistically sample structures from the Boltzmann ensemble. In this algorithm, they compute the equilibrium partition function using the most recent thermodynamic parameters, but then, as opposed to folding the base pairs deterministically and finding the probability of that outcome, as in previous algorithms, the base pairs are determined

probabilistically. This generates a set of secondary strictures, which is a statistical sample of the complete ensemble of structures. [16, 39, 42]

In addition to generating a picture of the structural landscape, this allows sampling estimates of the probability of structural motifs; which is incredibly useful in determining the accessibility of a given sequence motif. Also, this allows for estimates of the free energy distribution of a RNA sequence. [39]

This method was implemented as Sfold. [46] Additionally, observations of the sampled ensemble of structures revealed distinct structural clusters; suggesting that the Boltzmann ensemble can be represented by clusters, and that the best representative structure can be chosen by using centroids. Compared to MFE techniques, centroids of the ensemble make 30% fewer prediction errors. [15] In an application of folding 100 human mRNA sequences from 425 to 8458 nucleotides in length, Ding et al. found that increasing the length of the sequence did not result in a statistically significant increase in the number of structure clusters, despite the exponential increase in clusters. [47]

1.2 VIRUSES

For this dissertation, we will be focusing on single stranded RNA (ssRNA) viruses, either with sense (positive) or anti-sense (negative) genomes. This genome is then enclosed within a protein

coat, which functions to both protect the RNA genome, and to enable the virus to infect new host cells.

1.2.1 Influenza A Virus

1.2.1.1 Genome and Infection

The influenza A virus belongs to the *Orthomyxoviridae* family. It is a segmented negative ssRNA virus that has been isolated from a number of vertebrate organisms. Its genome consists of eight ssRNA segments, which encode for 10-11 proteins. These proteins are PB2 polymerase, PB1 polymerase, PB1-F2, PA polymerase, hemagglutinin (HA), nucleoprotein (NP), neurominidase (NA), M1 and M2 proteins, and non-structural proteins NS1 and NS2 (NEP). (**Figure 1.1**) [48, 49] The structure of the influenza virus is spherical body, surrounded by a host-derived lipid-bilayer, with viral proteins HA and NA displayed on the surface.

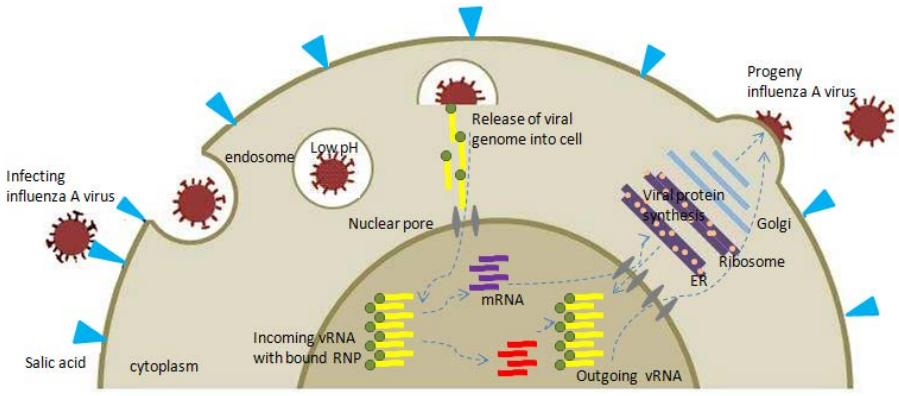

Figure 1.1 A cartoon of how influenza A infects its host cell. This figure was adapted from reactome.org. [50]

To infect the host cell, an activated HA protein binds to a terminal sialic acid on the glycoproteins/glycolipids on the host cells' surface. Immediately following binding, the virus particle is endocytosed, where the low internal pH of the endosome facilities the release of the

genome. Once inside the cell, the loosely NP encapsulated genome segments move into the nucleus, where the RNA- dependent RNA polymerase (RNP) complex of PB1, PB2 and PA, in conjunction with NP, begin to transcribe the viral mRNA (vRNA); the first transcripts are translated to viral proteins, and host mRNA translation is blocked. Other positive stranded transcripts are used as a template for vRNA. [51] (**Figure 1.2**)

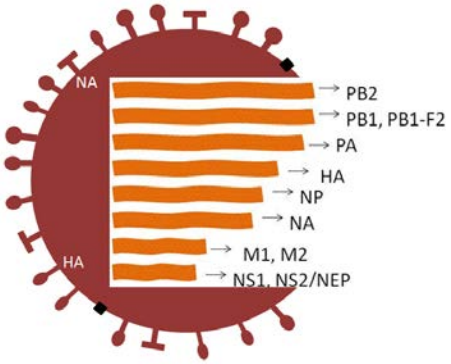

Figure 1.2 A cartoon illustrating the genome and structure of the influenza A virus.

The HA protein is an integral membrane protein, whose primary function is to bind the virus to the host cell receptor. Its second function occurs once the virus has been engulfed, when the low pH environment of the lysosome dramatically alters the structure of the HA molecule, enabling the membranes of the endosome and the virus to fuse, allowing for the un-coating and release of the genome. The HA protein is also the major determinate recognized by the host immune response.[51] The NA protein is also an integral membrane protein, and functions to cleave terminal sialic acid from the glycoproteins/glycolipids, allowing progeny to escape the cell from which they came. As the HA and NA genes are seen on the surface of the viral particle, they are used to determine the serotype of the virus. There are currently fifteen subtypes of the HA gene, and nine of the NA, and all circulate in the aquatic bird population. M1 is the protein matrix protein which forms the protein structure underneath the lipid-bilayer, while M2 functions as a cell-surface transport protein, which is more abundant on the infected host cell

than on the virion. [51] NS1 binds and sequesters RNA, and prevents cellular apoptosis, while NS2 is involved in vRNA nuclear export and interacting with host immune factors. [51, 52]

Within the viruses coat reside the eight genome segments, which are loosely encapsulated by NP molecules. NP not only functions to facilitate the movement of the virus into the nucleus, but also functions in both RNA binding and synthesis, in conjunction the rest of the RNP. [51] Complexes of PB1, PB2 and PA, forming the core of the RNP, are located at the ends of the encapsulated genome segments, ready to facilitate RNA replication. [51, 52]

1.2.1.2 Viral Genomic Changes: Genetic Drift and Antigenic Shift

Transcription by the RNA polymerase complex is extremely error prone, resulting in a high mutation rate of approximately $2*10^{-3}$ substitutions per base position per virus generation. An increased concentration of NP within the host nucleus triggers a shift in transcription production from mRNA to complementary RNA (cRNA) and viral RNA (vRNA) of the viral genome. The vRNA is encapsulated by the NP proteins, moves into the cytoplasm, and together with the viral proteins, buds off as a virus from the host cell.[51]

Due to the modular nature of the genome and an error-prone transcription and replication system, the influenza A virus has several primary modes of mutation. The first is simply mutation due to the inaccuracy of the polymerase complex. The ratio of synonymous to non-synonymous mutations introduced into each gene segment varies: for example, for the human H3N2 HA gene, 57% of all the mutations are synonymous, while for human PB2 that number rises to 90%. [51] Reassortment of whole genome segments between different viruses will allow for the switching of whole genes, potentially giving a virus an advantage it hadn't had before (this can occur when more than one virus infects a single host cell). [51] Antigenic drift can occur gradually, but the slow accumulation of these point mutations, or drastically by genetic

reassortment. Antigenic drift is driven by the host immune pressure on the HA and NA genes, enables the virus to evade the host immune response. Antigenic shift occurs when the virus acquires a completely new surface protein(s) from genome reassortment. [53] Homologous recombination is largely absent as a source of mutation for the human influenza A virus.[54]

1.2.1.3 Viral Ecology and Host Adaptation

The natural reservoir of influenza is in populations of wild aquatic birds, where it causes no outward signs of disease, replicates in the intestinal track, and follows the fecal-oral transmission route. In this natural population, the virus can be best described as being in evolutionary stasis, with limited amino acid changes, and limited antigenic drift. In humans, however, influenza is an infection of the respiratory tract, and is spread by air borne transmission.[51, 52]

Influenza A virus has shown the ability to jump to new hosts. An example of one such jump is a movement from avian to human host. Many times this jump simply causes a dead-end infection in the human without spreading to other individuals. The relatively large number of dead-end infections suggests that host jumping is inefficient, with numerous barriers acting to prevent an expansion of host range.[52] Once successful transmission to a new host occurs, however, the virus undergoes rapid evolution as it adapts [55], with characteristic amino acid changes in the HA, NA, PB2, PB1, M and NS genes. The result of this adaptation is that, when experimentally tested, avian viruses will not replicate well in primates, and vice-versa.[52, 53] Known examples of host-jumping of influenza is the jump from canines from equines [56], and the, still controversial theory that, 1918 pandemic's virus jumped to humans from an avian source. [57, 58]

While the exact viral factors which influence host range is largely unknown, HA plays a central role, and hence a large amount of research has been devoted to this gene. As HA

influences the binding of the cell, its binding of specific sialic acid receptors are known to be a factor in determining host range. Human influenza strains show preferential binding to sialic acid resides linked to galactose by a α2,6 linkage (the linkage which predominates in the human upper respiratory tract), while avian strains preferentially recognize a α2,3 linkage (which is typically seen in the avian host cell and the lower human respiratory tract). As respiratory tracts of pigs contain both α2,6 and α2,3 linkages, this animal is susceptible to both, and may serve as a host to both strains, potentially resulting in novel virus reassortments. [53, 59-63] In humans, avian influenza viruses will replicate in the lower respiratory tract. This infection location, along with the limited replication ability of avian influenza viruses in mammalian cells, functions to limit the spread of the viruses in the human population. [51]

As the NA gene is responsible for cleaving off the progeny virus, it, in the same manner as HA, shows preference for specific sialic acid-galactose linkages. [51, 64]

In PB2, it has been shown that the presence of a glutamic acid in position 627, found in many avian viruses, prevents efficient replication in mouse cells. In human influenza viruses, this amino acid is a lysine. [64, 65] Notably, both H5N1 viruses isolated from human patients, and a virus isolated in a fatal H7N7 infection in the Netherlands had a lysine at this position.[64]

NS1 functions as an interferon (IFN) antagonist, to allow virus replication in IFN competent hosts. This action is necessary and host specific, as introduction of double-stranded RNA (dsRNA) into the cell triggers the activation of numerous immune response transcription factors which work to trigger IFN-β production. NS1 interferes with this pathway, allowing the virus to replicate in the host cell. In addition to being a factor to determine host range of the virus, NS1 activity also influences the pathogenicity of the virus. [51]

Recently, the only serotypes known to have circulated extensively in the human population were H1N1, H2N2 and H3N2. In 1918, H1N1 caused the “Spanish flu” pandemic, and circulated in the human population as the seasonal flu strain, until 1957 when H2N2 “Asian flu” pandemic occurred. The 1918 H1N1 influenza A virus caused the most devastating influenza epidemic of modern recorded history, killing up to 100 million people worldwide.[53] This influenza strain is thought to have possibly been an adaptation of an avian strain, as was discovered by Taubenberger et al. after the recovery and reconstruction of the viral genome from archived specimens and tissue of Alaskan influenza victims buried in the permafrost.[52, 53, 58]

The 1957 “Asian flu” pandemic caused an antigenic shift in the circulating influenza strains, with the H2N2 strain becoming the dominant seasonal strain. While much less severe than the 1918 pandemic, it still caused approximately 1 million deaths worldwide.[51] This pandemic was caused by genetic reassortments of human and avian strains.[53] The 1957 pandemic, was caused by a reassortment of an avian H2N2 virus with the circulating human H1N1 virus, with five segments (PB2, PA, NP, M and NS) coming from the human virus, and three (PB1, HA and NA) coming from the avian.[52]

In 1968 another antigenic shift and pandemic occurred, and “Hong Kong” H3N2 became the dominant strain. In this cause, the circulating H2N2 virus again underwent reassortment with and avian H3 virus, resulting in the H3N2 pandemic. For this virus, PB1 and HA originated from the avian virus, with the rest coming from the human H2N2 virus.[51, 52]

In 1977, however, an H1N1 influenza strain, identical to strains that previously circulated in the 1950s, was re-introduced to the human population.[51, 52]

In 2009, a new pandemic strain of H1N1 emerged, resulting in at least 18,449 deaths worldwide in 214 countries. [66] This strain is capable of human-to-human transmission, and

was derived from several circulating swine viruses.[67] This virus lacks specific genetic markers that were thought to be necessary in adapting to the human population, suggesting a novel molecular determinants responsible for the adaptation to the human population. [68] It is likely that this H1N1 virus will continue to circulate in the human population, becoming another seasonal influenza strain. [69]

To date, both H1N1 and H3N2 strains circulate and cause the seasonal influenza cases.[51, 52]

While the natural reservoir of the virus is in birds, a range of symptoms in response to the infection may be seen. The range of symptoms may be anywhere from asymptomatic, to a mild upper respiratory infection, to rapidly fatal systemic disease. Depending on the virus pathogenicity in chickens and turkeys, the virus will be classified as lowly pathogenic (LPAI), or highly pathogenic (HPAI). The HAs of HPAI viruses may cleave in a broad range of host cells, and may extend to humans. The HA subtype is typically H5 or H7.[61, 64]

Recently, H5N1 influenza has increasingly become a cause of disease and mortality in the human and domestic fowl populations. Prior to 1997, the transmission of avian influenza viruses to humans was not considered to be of serious concern, as only three cases had been reported (two HPAI, one LPAI).[51] That thinking began to change in 1996, the first highly pathogenic H5N1 virus was isolated from a farm goose in China, and in 1997, 18 human H5N1 infections occurred in Hong Kong, resulting in 6 fatalities.[70] Since 2003, there have been 387 human cases, with 245 fatalities, across fifteen countries.[71] The evolution of this virus remains controversial, with some claiming that the current H5N1 strains are no more adapted to humans now than they were three decades ago[72], and others maintaining that the virus is adapting to humans[73] and evolving to a highly pathogenic state.[55]

In 2009, the emergence of a new pandemic influenza strain, an H1N1 virus, occurred. This virus had, at the end of that year, infected greater than 296,000 people worldwide, with 3486 deaths being reported. This virus was the result of multiple reassortments, with PB2 and PA originating from an avian virus of North American lineage that was introduced into swine populations around 1998. PB1 originated from an H3N2 human seasonal virus, which also entered swine populations around the same time. The HA, NA, NP and NS directly descended from the classic swine influenza A North American lineage. NA and M were introduced from birds to the swine populations around 1979.[74, 75]

With the online *Influenza Research Database*, influenza genome sequences are available, from a variety of hosts, making large-scale population research on the virus and its population dynamics possible.[76] It began in 2005 with a few hundred sequences, up to a few thousand today. [49, 76]

1.2.2 West Nile Virus

1.2.2.1 Genome and Infection

Unlike the influenza A virus, the West Nile virus (WNV), is a non-segmented positive ssRNA virus, whose genome encodes a single long open reading frame (ORF). This ORF is translated into a single polypeptide, which is co- and post- translationally cleaved into ten proteins, which are, in order: capsid (C), premembrane (prM), envelope (E), and the non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5).[30, 53]

Replication starts with the synthesis of a genome-length RNA of negative polarity, which then serves as a template for the synthesis of additional positive stranded RNA. Eventually, the positive stranded RNA is up to ten times more abundant than the negative strand. Unlike the

influenza A virus, this does not occur in the nucleus. While studies have indicated that the virus does replicate in association with host cell membrane, budding intermediates of progeny virus particles has not been observed. [30]

1.2.2.2 WNV Ecology

Until 1999, WNV was geographically limited to the, Africa, West Asia, the Middle East, and parts of Europe. In 1999, however, the virus was introduced into Western hemisphere (New York, NY), where it rapidly spread. [30, 53]

The natural transmission of the virus is between a variety of *Culex* mosquito species and passerine birds, with humans and other animals considered incidental, non-amplifying hosts.[30, 53]

While this virus has more genomic RNA sequences than the other ssRNA viruses excepting influenza, it numbers less than 100 complete genome RNA sequences in GenBank.

For these viruses, we studied the mRNA transcript and the vRNA. The motivation behind this was to determine if RNA folding could be another mechanism of host adaptation. As mRNA structure has been hypothesized to influence translation, and these viruses have a need to be able to transcribe and translate their genome immediately upon entering the host cell, it could be hypothesized that RNA folding may be a critical component in viral ecology with regards to host adaptation and range.

1.3 PROKARYOTES

Prokaryotes are single cells, typically 0.2μm to 10.0μm in size. They are defined by cellular, not organismal, properties. Typically, they have no nucleus. Because of this translation and transcription are not discrete events. Respiratory and photosynthetic functions are associated with the membrane. Nutrients are required in molecular form. They can adapt and thrive in extremely diverse habitats- everything from the human body to hot springs to the deepest reaches of the ocean.[77]

The organisms we refer to as prokaryotes consist of two dramatically different domains: the *Archaea* and the *Bacteria*. Molecular analysis has shown that the *Archaea*, despite being prokaryotes, are more closely related to eukaryotes than to the *Bacteria*.[77]

The bacteria group may be further subdivided into three groups- those which are gram-negative and have a cell wall; those which are gram-positive and have a cell wall; and those which lack a cell wall altogether. The gram-negative bacteria which have a cell wall typically have a thin peptidoglycan layer with an outer membrane (and stain gram-negative). Examples of this type of bacteria are *Escherichia coli*, *Haemophilis influenzae* and *Thermus aquaticus*. Those which are gram-positive with a cell wall have no outer membrane overtop the peptidoglycan layer, and the peptidoglycan layer is relatively thick (allowing them to stain gram positive). Examples of this are *Staphylococcus aureus* and *Streptococcus pneumoniae*.[77]

The *Archaea* predominantly occur terrestrial and aquatic environments- including those which are hypersaline, hydrothermally or geothermally heated, although a few may be found symbiotic relationships with animals. [77]

Like the virus mentioned above, bacteria are haploid- meaning they have one allele of each gene. They also have a short generation time- with *E. coli* being able to replicate itself in 20

minutes. Bacteria multiply by cell division- creating progeny that are genetically identical to the parent. Even though they divide asexually, they may uptake new DNA in by three mechanisms- transformation, conjugation, and transduction. In transformation, the cell will directly uptake DNA from its environment. In conjugation, the bacteria can pass genetic information directly from one cell to another. In transduction, the viruses that infect bacteria move DNA from one cell to another.[78]

In the prokaryotic cell, the DNA is transcribed into RNA, which may then be translated into a protein. To be transcribed, the promoter attracts the RNA polymerase- and the same polymerase is used to synthesize tRNA, rRNA and mRNA. The promoter region of the DNA has two important regions that enable it to attract and bind the polymerase- an AT-rich sequence about 10bp upstream of transcription (consensus sequence TATAAT), and another sequence approximately 35bp upstream (consensus sequence TTGACA). Once the RNA polymerase has bound to the DNA and transcription has been initiated at a promoter, it will continue until it encounters a transcription termination site. In bacteria, there are two types of termination signals- factor independent and factor dependent. In factor dependent transcription termination, one of three termination factors, either rho (ρ) tau (τ) or NusA, functions to stop transcription (with ρ being the most studied). To terminate transcription, ρ will bind to a specific RNA sequences (characterized by a lot of C's and not much secondary structure) before the polymerase reaches the ρ-dependent termination site. The ρ will then chase the polymerase toward the termination site. When ρ "catches" the polymerase, typically due to stalling at a stop site, the protein will unwind the RNA-DNA hybrid, releasing the transcript and terminating transcription. If the gene is being actively translated, ρ cannot bind to the RNA. In factor independent transcription termination, the termination region consists of two sequences, the first being an inverted repeat,

and the second being a string of A's. In this termination method, the inverted repeat binds to itself, creating a hairpin and destroying the RNA-DNA transcription bubble. The RNA-DNA hybrid is then destabilized when the polymerase hits the adjoining AT-rich region (as AU pairing is less stable than DNA AT base pairing), causing the transcript to fall off and terminate transcription.[78]

In viruses, when we were analyzing the RNA species, we are looking at the viral genome and its transcripts. When analyzing the prokaryotes, however, we can study tRNA, rRNA and mRNA. Both tRNA and rRNA are functional, but as most of the rRNA genes in the bacterial genomes are not yet annotated. As tRNA is functional RNA, being the unit which brings the amino acids to the ribosomes, we can study how changes in the environment the physical environmental conditions interact with tRNA structure. With mRNA we analyze the molecules which get transcribed, the same as for the viruses.

RNA structure is very important for tRNAs and rRNAs [79]. Indeed, previous work has detailed how the GC content of tRNAs and rRNAs strongly correlates with growth temperature/ This same work has also showed that the genomic or mRNA GC content does not [79-81]. This lack of correlation between mRNA GC content and temperature is somewhat surprising, as it was previously believed that melting temperatures of nucleic acids should affect genome evolution globally [81].

While not correlating with growth temperature, genomic GC content has been shown to play a strong role in the codon usage across different species. [27, 82, 83] There are several mutational biases that may affect the patterns seen in both genomic and mRNA GC content. There can be natural or selective pressure on the innate bias in point mutations [27, 82, 84, 85]; there can be selective pressure to prefer certain synonymous mutations over others. [21, 27, 82,

83, 86-88] Even among bacteria, there is a split in the selective pressure exerted on the patterns of synonymous mutations: some species show a strong selective pressure, while others do not [86]. In Sharp *el al.* 2004, it was noted that the species exhibiting strong selective pressure on their codon bias also had faster growth rates and shorter generation times. [86] In Kudla *et al.*, a relation between the strength of codon bias of the GFP construct and growth rate was also observed, with higher codon adaption of the exogenous gene correlated with a faster growth rate. [20] The rate of synonymous substitution and degree of codon bias of a gene may be related, and reflect the genes' translational landscape; and this may differ between genes within a given genome. [87]

While seeking to understand if there is a dependence of RNA structure on physical environmental conditions, we also sought to investigate why this dependence was occurring. As detailed above, previous research has found some correlation between mRNA structure and expression, [1, 18-25] but none to mRNA structure and environmental adaptation. In researching mRNA structure in prokaryotes, we want to uncover just how these properties are linked.

Our hypothesis is that the RNA structure plays a critical role in (reverse) transcription and translation rate, and RNA degradation rate. If the RNA molecule has a lower folding energy and is tightly bound, the translation rate of the molecule may be severely reduced; on the other hand, if the RNA molecule has too high a folding energy, the structure may be more loosely bound, making the molecule more susceptible to degradation. This balance between translation and degradation rates would be particularly important to microbes. For viruses, this balance needs to be reached for every host it wishes to successfully replicate in, and may play a critical

role in viruses emerging and adapting to human hosts. For bacteria, this would indicate an internal balance dependent on the given environment.

The importance of understanding all of these organisms, from both viruses to bacteria, is vital. As Dobzhansky stated it best “nothing in biology makes sense except in the light of evolution”.[89] This is the lens through which I will seek to understand how viruses and bacteria interact with their various hosts and their environment. This knowledge would be invaluable for the development of efficient disease control methods, and for preventing the emergence of new human-adapted viruses. The specific aims for this thesis are detailed below; with the work done for each detailed in the subsequent chapters.

The intent of this dissertation is to study how the genomes of both viruses and bacteria evolve with respect to the folding energy and structure of their RNA. The stability of the genome and its mRNA may play an important role in the ability of a given influenza A virus to spread to both new hosts and new locations due to the dependence of the genome folding energy on temperature. If an influenza virus has a very stable mRNA and vRNA at a given temperature, the ability of its genome to be translated and replicated at a different temperature may be severely impaired. On the other hand, these that have a lower stability may be more susceptible to environmental conditions, and may be degraded before they get a chance to infect a new host. In chapter 2, we will seek to determine the folding energy constraints on the influenza viruses’ genome ability to expand from avian to human populations, and how these constraints affect viral evolution once the virus has established itself within the human population. In chapter 3, we will use both existing knowledge of how the virus emerges and adapts and combine this with the knowledge garnered in chapter 2 in an attempt to be able to predict specifically what virus can propagate in the human population. In chapter 4, we will extrapolate the relation of FFE and

the influenza A virus to other ssRNA viruses. Finally, in chapter 5, we will determine if RNA stability plays a role in prokaryotic evolution as well, as prokaryotes also rely on environmental conditions to set their growth temperature. We will also explore exactly how the FFE may impact gene expression, and understand its importance the cell.

New diseases, caused by both bacteria and viruses are continually emerging; and it is becoming imperative that a greater understanding of the mechanisms of disease emergence and of host-pathogen interactions is achieved. The long term goal of this research project is to do just that- to determine evolutionary constrains of the influenza virus, and to predict when an emerging virus is an epidemic threat; and to explore the significance of these identified constraints in other ssRNA viruses and prokaryotes.

.

2.0 THE EFFECT OF FFE IN THE EVOLUTION OF THE INFLUENZA A VIRUS

In the following work, we addressed how folding energy may be a factor in the ability of the influenza A virus to jump to and establish itself in the human population. This work was published by R Brower-Sinning, DM Carter, CJ Crevar, E Ghedin, TM Ross and PV Benos in *Genome Biology*, February 2009 as "The role of RNA folding free energy in the evolution of the polymerase genes of the influenza A virus".

In this study, I aided in its conceptualization and design, performed the computational analysis and analyzed the data.

2.1 ABSTRACT

Background

The influenza A virus genome is composed of eight single-stranded RNA segments of negative polarity. Although the hemagglutinin and neuraminidase genes are known to play a key role in host adaptation, the polymerase genes (which encode the polymerase segments PB2, PB1, PA) and the nucleoprotein gene are also important for the efficient propagation of the virus in the host and for its adaptation to new hosts. Current efforts to understand the host-specificity of the virus have largely focused on the amino acid differences between avian and human isolates.

Results

Here we show that the folding free energy of the RNA segments may play an equally important role in the evolution and host adaptation of the influenza virus. Folding free energy may affect the stability of the viral RNA and influence the rate of viral protein translation. We found that there is a clear distinction between the avian and human folding free energy distributions for the polymerase and the nucleoprotein genes, with human viruses having substantially higher folding free energy values. This difference is independent of the amino acid composition and the codon bias. Furthermore, the folding free energy values of the commonly circulating human viruses tend to shift towards higher values over the years, after they entered the human population. Finally, our results indicate that the temperature in which the cells grow affects infection efficiency.

Conclusions

Our data suggest for the first time that RNA structure stability may play an important role in the emergence and host shift of influenza A virus. The fact that cell temperature affects virus propagation in mammalian cells could help identify those avian strains that pose a higher threat to humans.

2.2 ARTICLE

The influenza A virus, a member of the *Orthomyxoviridae* family, is an enveloped negative single-stranded RNA virus with a genome consisting of eight individual RNA segments, each packaged into ribonucleoproteins (RNPs) [51]. RNPs are composed of four proteins, each of which is coded by a single segment. Segments 1-3 code for the three subunits of the heterotrimeric RNA-dependent RNA polymerase (PB2, PB1, and PA, respectively) and

segment 5 codes for the nucleoprotein (NP), a protein that binds single-stranded RNA [90]. RNPs are sufficient for replication of the viral RNA, which leads to synthesis of positive strand complementary RNA and transcription to viral mRNA [91]. The proteins that comprise the RNPs play an important role in the adaptation of the avian viruses to humans [92], but the precise mechanism is still unclear. Recently, it was found that the three polymerase genes affect replication of avian influenza viruses [93]. Current efforts to investigate this adaptation mechanism are mainly focused on characteristic amino acid differences between avian and human genes [58]. In some cases, critical amino acid substitutions have been identified that affect species-specific virulence [94-96].

Influenza A viruses are subdivided by antigenic characterization of the hemagglutinin (HA) and neuraminidase (NA) surface glycoproteins (segments 4 and 6, respectively). HA has 16 and NA has 9 different subtypes. The most commonly circulating subtypes in the human population are A/H1N1, A/H2N2, and A/H3N2. The 1918 pandemic was caused by an A/H1N1 strain, whose polymerase genes were probably of avian origin [58]. Since then, there have been two major influenza pandemics (1957 and 1968) caused by A/H2N2 and A/H3N2 subtypes, respectively. Both strains were subject to reassortment. The human virus seems to have acquired three avian segments (HA, NA, and PB1) in the case of the 1957 pandemic, and two avian segments (HA, PB1) in the case of the 1968 pandemic [97]. The other segments are believed to have been circulating in humans since the 1918 pandemic. Currently, A/H3N2 and A/H1N1 (re-introduced into the population in 1977) are circulating in the human population [98].

Predicting the emergence of new circulating influenza strains for annual vaccine development is critical [99]. Recently, the emergence of highly pathogenic avian influenza has been of widespread concern. The majority of these outbreaks involve the direct transmission of

isolates from the A/H5N1 subtype from birds to humans [100, 101]. Since 2004, 385 people have been infected with H5N1 viruses, with 243 fatalities (63%). Other highly pathogenic subtypes associated with disease include A/H9N2, A/H7N7, and A/H7N3.

In this study, we investigate the role of the RNP member proteins in the propagation of the virus in birds and humans. We propose that RNA structure stability, reflected in the folding free energy, plays a critical role in overall influenza virus fitness, having an effect on replication, transmission, and spread to humans. RNA molecules with low folding energies will generally form longer stems that could potentially reduce the translation rate. Also, long stems may trigger the RNA interference mechanism, thus increasing the RNA degradation rate [17, 102], which may also restrict protein production and reduce the overall number of released virions. We note, however, that long imperfect stems, especially in the 3' untranslated regions (UTRs) of the genes, can increase stability.

The discovery of differences between avian and human RNA folding energies represents a novel angle in our understanding of molecular evolutionary adaptation of influenza A virus to various hosts.

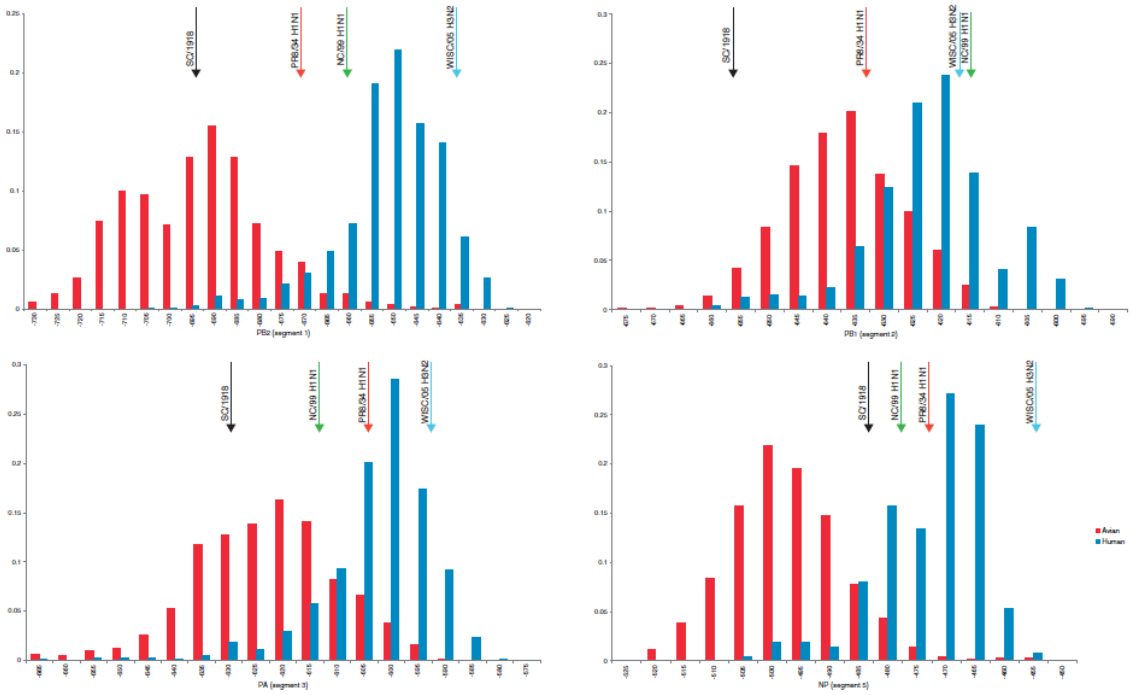

Figure 2.1 Folding free energy distributions for human and avian influenza A polymerase gene segments (in kcal/mol). The black arrows indicate the folding energies for the corresponding 1918 virus segment. Red, A/Puerto Rico/8/1934 (H1N1) (PR8/34); green, A/New Caledonia/20/1999 (H1N1) (NC/99); blue, A/Wisconsin/67/2005 (H3N2) (Wisc/05). The x-axis is the folding energy calculated by the program RNAfold [7-14], and the y-axis is the relative frequency of this folding energy in the viral population.

Results

Influenza A virus genes coding for RNP components exhibit species-specific mRNA folding energies

To investigate whether differences exist in the preferred folding energies of human and avian viruses, the mRNA of genes coding for PB2, PB1, PA (polymerase complex segments 1-3), and NP (segment 5) were folded as described in Materials and methods. Avian and human frequency distributions are found to be distinct in all these genes ($p << 0.01$, Wilcoxon Rank Sum test), with segments 1 (PB2) and 5 (NP) having the most distinct distributions (**Figure 2.1**). A similar discrimination exists between the energy distributions of the avian-derived A/H5N1 strains

isolated from humans and the currently circulating A/H1N1, A/H2N2 and A/H3N2 human strains (**Figure 2.5**; $p << 0.01$ for all segments, Wilcoxon Rank Sum test). This separation coincides with the fact that A/H1N1 and A/H3N2 strains circulate in the human population, whereas human transmission of A/H5N1 isolates is still inefficient. Avian influenza strains from other subtypes, such as A/H7N3 and A/H9N2, also exhibit folding energy preferences at the lower end of the human spectrum (data not shown).

The 1918 outbreak was the worst pandemic in recorded history. It caused severe disease with high mortality in the United States (675,000 total deaths) [97] and worldwide (50 million people)[103]. It was previously suggested that the polymerase genes of the 1918 virus were of avian origin [58]. In agreement with this hypothesis, we found that the folding energies of the polymerase genes (segments 1-3) of the 1918 strain are in the lower 1.5-4% of the human energy distributions and 6.5-67% of the avian distributions. Similarly, Kawaoka *et al.* [98] have suggested that the PB1 segment was of avian origin in the 1957 and 1968 pandemics (caused by A/H2N2 and A/H3N2 strains, respectively). We found the folding energies of the PB1 segments for all 1968 A/H3N2 isolates to be smaller than the average avian values (-655 to -635) and at the very low end of the human range, which supports the hypothesis of the avian origin of this segment. However, all the 1957 A/H2N2 isolates have folding energies in the region between the two distributions (-633 to -623), so we are not able to draw any conclusions in this case (**Figure 2.1**).

Next, we examined whether the observed differences in RNA folding energy distribution between human and avian strains are a by-product of the selection performed at the protein level. Certain amino acids are known to play an important role in host-specificity. For example, Subbarao *et al.*[96] showed that a Glu to Lys substitution at position 627 of the PB2 gene is

sufficient for restoring the virus's ability to replicate in Madin-Darby canine kidney (MDCK) cells. In an attempt to distinguish between the folding energy constraints and the amino acid constraints, we examined whether degenerate codon positions favored an increase or decrease in the *hydrogen bonding potential* between the viruses of the two species. Hydrogen bonding potential is defined as the number of hydrogen bonds a particular base would form if it was paired in the RNA secondary structure (see Materials and methods). While the hydrogen bond potential cannot offer definite proof of whether evolution operates at the folding energy level or not, it is nevertheless indicative of the trend. If amino acid substitutions constitute the only dominant force that drives the evolution of the polymerase genes, then it would be expected that no differences would exist in the number of potential hydrogen bonds in the degenerate positions between the avian and human species. In other words, there would be no increase in the number of A or U bases in human strains compared to the avian strains at these positions. Instead, we found that degenerate positions in the avian strains contained bases with higher bonding potential than in the human strains (**Figure 2.2**). In fact, the differences between the potential hydrogen bond distributions in segments 1, 3, and 5 are similar to the distributions of the folding energies (**Figure 2.1**); and in segment 2 the differences in hydrogen bonding potential are even more profound. In all cases, the observed differences are statistically significant ($p << 0.01$, Wilcoxon Rank Sum test). These results are in agreement with other studies that have found host-specific nucleotide bias for the influenza virus, which was attributed to host mutational bias [104, 105].

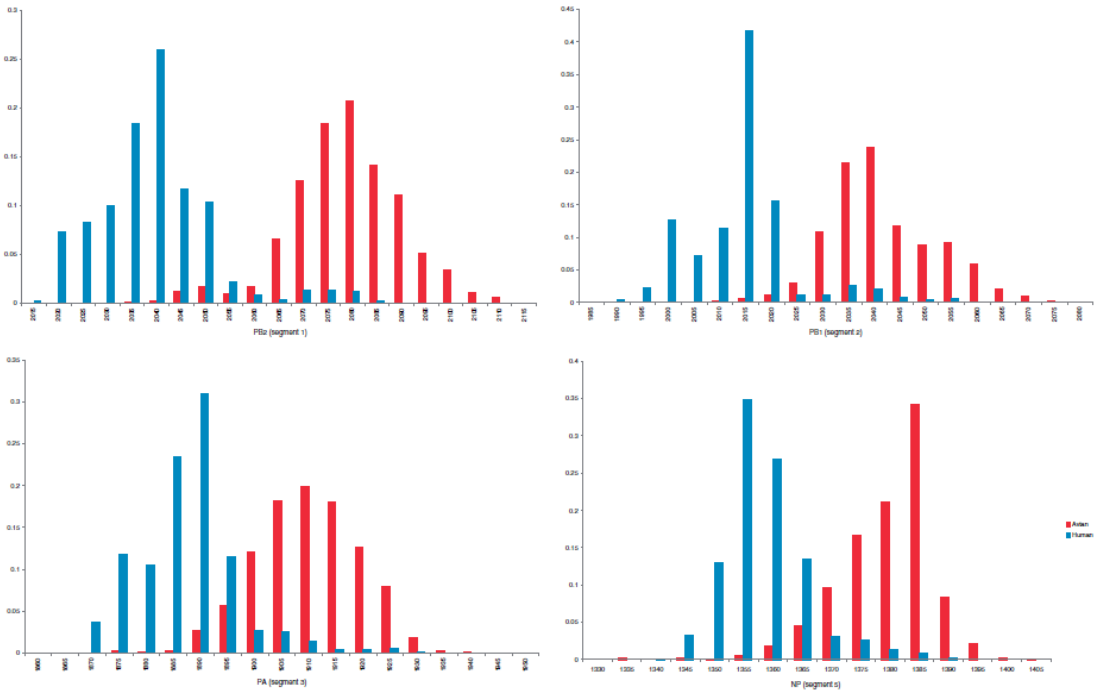

Figure 2.2 Potential hydrogen bond distribution (per segment) at all degenerate codon positions in human and avian influenza A strains. The x-axis is the number of potential hydrogen bonds per segment, while the y-axis represents the relative frequency.

Another factor that might affect the evolution of the nucleotide sequence is the codon usage bias. Each organism uses more frequently a specific set of codons for coding certain amino acid residues. In polioviruses, selection of strongly unfavorable codons can lead to reduced protein translation [106]. Could it be that this is also the case in influenza viruses and that the trend we observe in the degenerate codon positions is the result of a shift towards the host-specific codon bias? We examined this by comparing the codon frequencies of the avian and human influenza A viruses (A/H1N1, A/H3N2 and A/H5N1) to the codon frequencies of avian genes (chicken was used as representative of avian species) and human genes [107]. We found that codon frequencies are similar between the human and chicken genes (R = 0.98), and between human and avian influenza A virus genes (R > 0.97), but not between the virus genes

and the animal species ($R < 0.66$). This suggests that the influenza polymerase genes are not under strong selection to shift towards their host codon usage preferences. In fact, this agrees with the proposed theory that, for species with small population sizes (like humans or birds), the codon usage changes are effectively neutral [108].

Based on these observations, we postulate that the folding free energy of the polymerase and NP gene segments is an important biophysical property of the segments and plays a significant role in the evolution of the virus both within the human population and in the ability of the virus to adapt to the human host when introduced from an avian source.

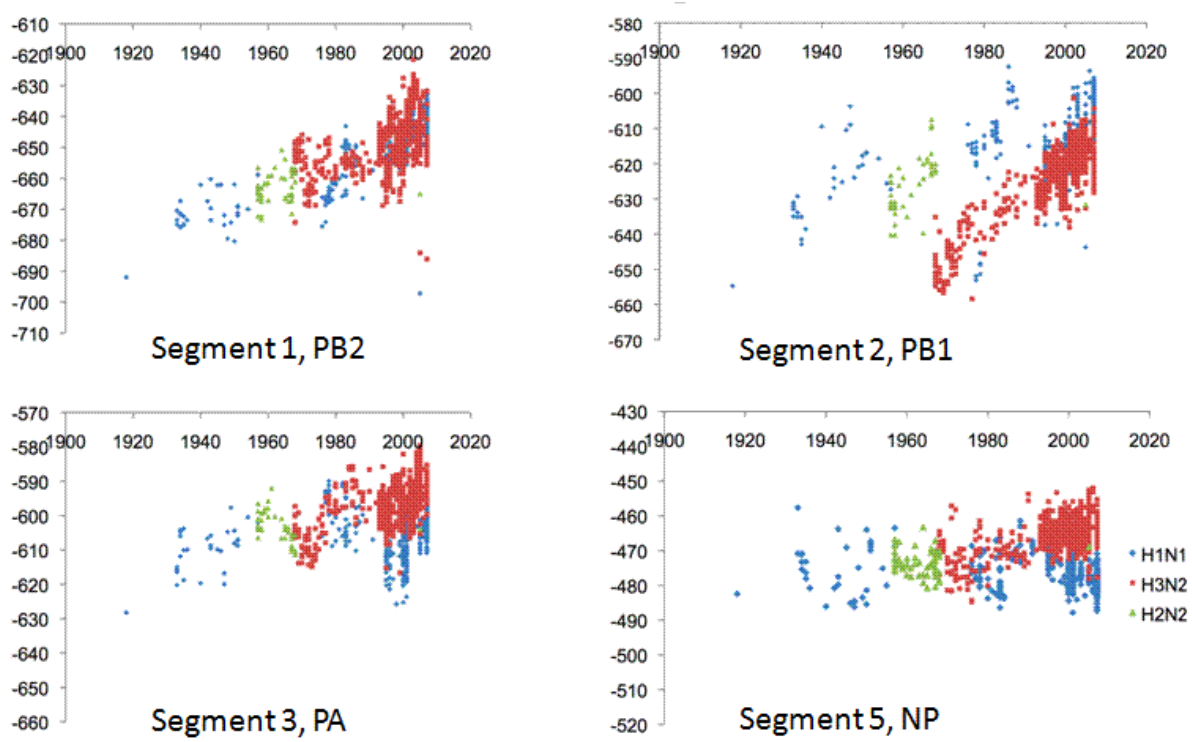

Figure 2.3 Predicted folding free energy of the human influenza A strains (polymerase genes) versus year isolated.

Evolution of folding energies of the polymerase and NP genes

If there is an 'ideal range' of folding free energies for each of the polymerase and NP genes, then strains from subtypes that entered the human population at some point and circulated for many years will tend to progressively shift their folding energies towards this 'ideal' range for humans. To test this evolutionary stasis hypothesis, three of the most recently circulating human influenza A subtypes (A/H1N1, A/H3N2 and A/H2N2) were examined. We found that there was an evolutionary trend towards higher folding energies as strains from these subtypes circulated in the human population (**Figure 2.3**). Although there is no reason to expect that the changes in the folding energy will correlate linearly with the year, we observe in fact such correlation for parts of the evolutionary trend. For example, segment 1 (PB2) of the A/H1N1 strains isolated since 1918 shows a shift towards higher folding energies, which continues after the strain's re-emergence in 1977 ($R = 0.80$, $p << 0.01$). Segment 2 (PB1) also shows some linear correlation for the years 1918-1956 ($R = 0.77$, $p = 10^{-6}$), when the strain was replaced by A/H2N2. During the years that the A/H2N2 strain was in circulation (1957-1967), we observe a weak linear correlation of the folding energies with the year ($R = 0.69$, $p = 10^{-6}$). In 1968 the A/H2N2 strain was replaced by an A/H3N2 strain. The newly introduced segment 2 (from bird viruses) continued having strong correlation of the folding energies with the year until 1998 ($R = 0.89$, $p << 0.01$). Finally, for segment 3 (PA) of the A/H3N2 strain, we observe linear correlation in the years 1968-1985 ($R = 0.75$, $p << 0.01$). Notably, none of the avian strains shows such a pattern over the same time period (**Figure 2.8**).

RNA folding energy and cell temperature

One of the factors that determine RNA folding energy is temperature. If viral RNA and mRNA folding energy affects the efficiency of viral infection and replication, then one would expect that virulence will vary according to the temperature that cells are incubated at and the folding energy of the viral segments. To further investigate this hypothesis, MDCK cells were slowly adapted for growth at two temperatures higher than 37°C (39°C and 40°C) as described in Materials and methods. The slow adaptation allowed cells to adjust to higher temperatures, thus minimizing the risk of injury due to heat shock. The adapted cells showed no difference in their growth rate. Further support for the regular growth of the cells comes from the fact that one of the mammalian influenza viruses, A/Puerto Rico/8/1934 (H1N1) (PR8/34), was able to replicate equally well in MDCK cells incubated at all temperatures in the 37-40°C range (**Table 2.1**).

Table 2.1 Viral titer (PFU/ml) for A/Puerto Rico/8/1934 (PR8/34) and A/New Caledonia/20/1999 (NC/99) H1N1 strains, and for A/Wisconsin/67/2005 (Wisc/05) H3N2 strain. The folding energies for segments 1-3, and 5 are: PR8/34, [-671.33, -633.85, -604.73, -473.22]; NC/99, [-658.78, -615.39, -611.74, -477.67]; Wisc/05,[-637.74, -617.08, -593.41, -455.20].

	PR8/34 A/H1N1		NC/99 A/H1N1		WISC/05 A/H3N2	
	48 H	96 H	48 H	96 H	48 H	72 H
$37^{O}C$	$2.5 * 10^{8}$	$4.2 * 10^{9}$	$1.0 * 10^{5}$	$1.1 * 10^{9}$	$1.0 * 10^{5}$	$> 10^{6}$
$39^{O}C$	$1.7 * 10^{8}$	$7.4 * 10^{9}$	<100	$< 10^{4}$	$3.0 * 10^{3}$	$3.2 * 10^{8}$
$40^{O}C$	$1.0 * 10^{8}$	$2.0 * 10^{8}$	<100	$< 10^{4}$	<100	<100

MDCK cells, incubated at 37°C, 39°C and 40°C, were infected with one of two A/H1N1 human strains - A/New Caledonia/20/1999 (H1N1) (NC/99), and A/Puerto Rico/8/1934 (H1N1) (PR8/34) - or one A/H3N2 human strain - A/Wisconsin/67/2005 (H3N2) (Wisc/05). Viral replication was measured by plaque assay at various time points post-infection. What becomes apparent from the results in **Table 2.1** is that the viral titer generally decreases with increased temperature, and the rate of decrease depends on the virus. Both NC/99 and Wisc/05 produced

no viral plaques at 40°C, but Wisc/05 produced plaques at 39°C, whereas NC/99 did not. Finally, PR8/34 was found to replicate efficiently at all three temperatures. Notably, all four PR8/34 segments (segments 1-3, and 5) have folding energy values in the range between the human and avian average values (**Figure 2.1**). Compared to PR8/34, NC/99 has higher folding energies for segments 1 and 2 and similar or slightly lower energies for segments 3 and 5. However, the folding energies of segments 1 and 2 of NC/99 are at the extreme end of the avian distribution, which might explain its inability to replicate efficiently at higher temperatures, as indicated by the viral titer values (**Table 2.1**). All four segments of Wisc/05 have RNA folding free energy values higher than the average for human influenza A viruses (**Figure 2.1**). So, based on the hypothesis that cell temperature affects viral replication through the folding energy of the polymerase genes, Wisc/05 is expected to replicate more efficiently at 37°C than at higher temperatures. Consistent with that hypothesis, no plaques were observed when MDCK cells, infected with Wisc/05, were incubated at 40°C, and there were fewer plaques on MDCK cells incubated at 39°C compared to MDCK cells incubated at 37°C (**Table 2.1**).

Ability of the H5N1 influenza A virus to become established in the human population

The ability of an avian virus to jump from the bird population directly to the human population has been recorded for the A/H5N1, A/H7N3, A/H7N2, and A/H9N2 subtypes [109, 110]. Most of these human outbreaks have been limited to a single round of infection from birds to humans with little or no human-to-human transmission. Nevertheless, the A/H5N1 human outbreaks have occurred in at least 16 countries across 3 continents since 1997 [70], and strains of the avian A/H5N1 subtype are considered to be a threat to humans because of their pandemic potential [111]. For this reason, we decided to further examine the folding energies for avian

A/H5N1 isolates. Box plots of the folding energies of segments 3 and 5 were calculated for all observations from the same region when data existed for two or more consecutive years (**Figure 2.4**). Differences in the yearly plots are not statistically significant for all but one of them (Indonesia population, segment 5, $p = 0.04$). This is expected for changes occurring over short periods of time. Nevertheless, these plots show a clear trend towards higher energies from year to year, which would favor adaptation to human hosts according to our hypothesis. For segments 1 and 2 no such trend was observed, but we note that the vast majority of segment 1 and 2 sequences from these regions have folding energies already in the human spectrum (data not shown).

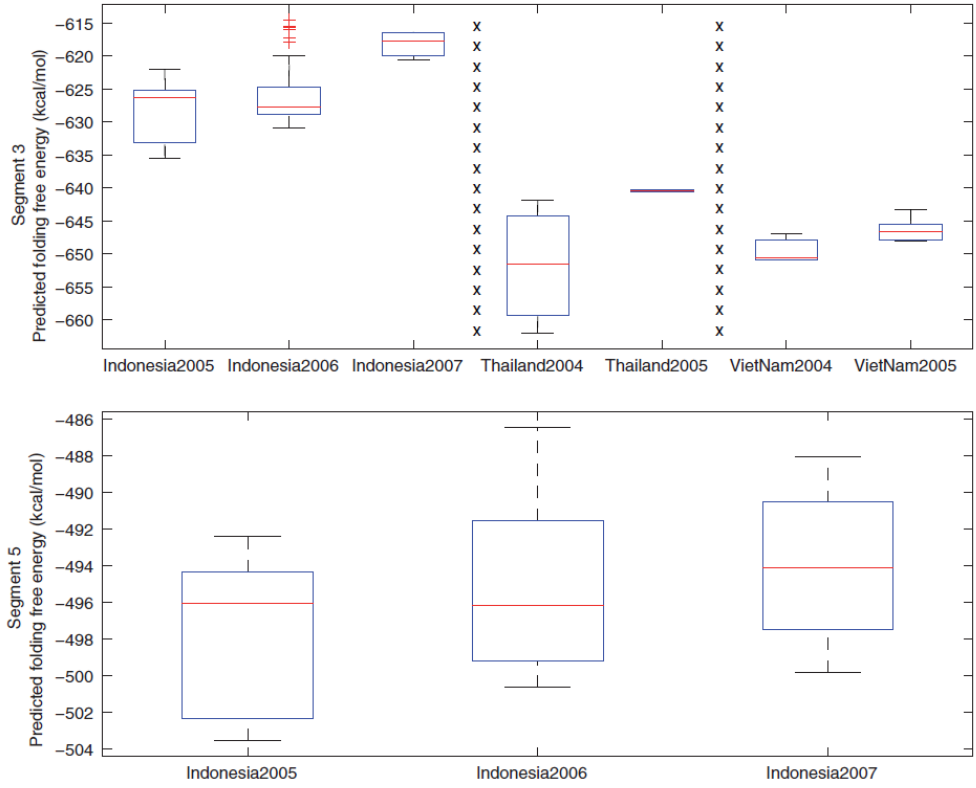

Figure 2.4 Predicted folding free energy of human A/H5N1 cases (polymerase gene segments 3 and 5) arranged by location and year of outbreak.

We also analyzed the folding energies for five A/H5N1 strains that are currently recommended by the World Health Organization for the production of vaccines against potential pandemic A/H5N1 influenza. The 1918 virus was used in this analysis as a low energy limit for the virus to be able to efficiently propagate in humans. The folding energy values of the 1918 virus are among the smallest observed in human viruses, and the virus caused one of the worst pandemics. In all but one case, segments 1-3 of the A/H5N1 viruses had higher folding energies than the corresponding segments of the 1918 strain (**Table 2.2**). The exception is segment 3 of the A/Vietnam/1203/2004 (VN/04) H5N1 strain, with a predicted folding free energy of -651 kcal/mol compared to the 1918 value of -628 kcal/mol. These data suggest that, as far as segments 1-3 are concerned, all but one A/H5N1 strain analyzed (VN/04) have the potential to contribute to efficient transmission from human-to-human and, hence, the establishment of the virus in the human population.

Hatta *et al.* [94] studied the virulence of two H5N1 influenza A strains with respect to residue 627 of the PB2 protein. They found that strain A/Vietnam/1203/2004 with Lys at position 627 of PB2 was three times more efficient in infecting mice cells than A/Vietnam/1204/2004, which has Glu at this position (MLD_{50} of 0.7 compared to 2.1). We folded the two PB2 segments and found them to differ by approximately 2 kcal/mol, with A/Vietnam/1203/2004 having higher energy (-682 versus -684). Although the difference is small, we note that both strains have PB2 folding energies at the extreme low end of the human distribution (**Figure 2.1**). It is possible that at distribution extremes, even small differences can give the virus an evolutionary advantage. In addition, Hatta *et al.* [94] performed site-directed mutagenesis and replaced the amino acid at position PB2-627 in each of the strains with the amino acid of the other strain. The new strains, VN1203PB2-627E and VN1204PB2-627K, had

measured MLD_{50} values of 67.6 and 0.6, respectively. Interestingly, the corresponding folding energies of these mutants are -684.2 (VN1203PB2-627E) and -681.7 (VN1204PB2-627K). It is easy to see that for all four proteins (initial isolates and mutants), the order of the MLD_{50} values coincides with the order of the negative folding energy values (rank correlation coefficient R = -1). In fact, if we exclude mutant VN1203PB2-627E from the analysis (because, practically, it does not infect the cells), the remaining three segments exhibit a strong anti-correlation between MLD_{50} and folding energy values (R = -0.97). In other words, in this case, the virulence of the virus with respect to PB2 seems to be associated with how close its folding energy is to the human average (**Figure 2.1**), with the segments closer to the average being more virulent.

Discussion

In this study, we have analyzed a biophysical property of the RNA segments of the influenza A virus: the folding free energy. We show that folding free energies of the RNP complex genes (PB2, PB1, PA and NP) differ between avian and human viruses and between seasonal human viruses and A/H5N1 viruses isolated from humans. The fact that the other segments do not show such drastic folding energy preferences (data not shown) may reflect the importance of the polymerase genes in escaping the host's cellular response [112].

The choice of focusing on the coding regions (or open reading frames (ORFs)) rather than on the complete segments was dictated by the fact that a large percentage of the sequences in the database (20-48%, depending on the segment and the host species) lack information about the 5' UTR, the 3' UTR, or both. Thus, analyzing the coding regions provided the largest common dataset. Given the small length of the non-coding regions (compared to the ORFs), their effect on the analysis of the folding energies is expected to be small. In other words, it is reasonable to believe that the trends observed in the analysis of the coding regions are

representative of the phenomenon seen for the whole segments. However, non-coding regions can be important for viral RNA replication [113], hence affecting virulence. For example, certain 5' UTRs may enhance the translation efficiency or some 3' UTRs may contain targets for microRNA genes from the host. But these phenomena are independent of the folding energies, so their contribution to virulence is similar to the contribution of HA, NA or the other non-RNP genes, and hence not a subject of our analysis.

Based on the folding energy distributions of the human and avian strains, we postulated that the avian virus segments may fold into a more 'rigid' structure in human cells than in avian cells. Such structure is expected to have long stems. Long stems with no mismatches can result in slower translation rates or increased degradation rates of the mRNA molecules [17, 102]. Either case can result in a reduction in viral fitness. We showed that, in the case of MDCK cells, human strains NC/99 (A/H1N1) and Wisc/05 (A/H3N2), with folding energies of the polymerase genes and NP segment largely in the human range, propagated efficiently at 37°C, but their propagation was diminished at higher temperatures. In contrast, strain PR8/34 (A/H1N1), with folding energies in the region between human and avian average values, propagated equally well at all temperatures. This shows that the cells that were slowly adapted in higher temperatures have no difficulty in propagating human influenza A viruses. It also shows that viruses with high folding energies (in the human range) may have difficulties propagating in birds. Whether avian viruses with very low energies have difficulties propagating in human cells remains to be seen. We note, however, that if this is true, then the host's body temperature may impose an additional barrier to cross-species transmission. Finally, we found that the RNA folding free energy of the A/Vietnam/1203/2004 and A/Vietnam/1204/2004 H5N1 viruses and the mutant VN1204PB2-627K show a nearly perfect inverse correlation with the measured MLD_{50} values ($R = -0.97$).

The effect of the folding energy on the evolution of the virus appears to be independent of the concurrent amino acid changes in the polymerase and NP genes, and independent of the codon usage bias. In addition, human influenza A strains have increasingly higher folding energies over time (within a certain range), especially when their folding energy starting points are close to the avian range.

Taken together, these results suggest that the folding free energy of the RNA molecules of the polymerase segments is an important factor in the evolution of the influenza A virus. Previous research in this area was focused on amino acid changes, especially in the HA, NA, and PB2 genes [94-96], where a number of mutations were found to be critical for host adaptation of the virus. The fact that the 1918 A/H1N1 has segments 1-3 with RNA folding free energies in the lowest part of the human spectrum (**Figure 2.1**) is indicative of the importance of the NA and HA genes in the success of replication and host adaptation [114].

In agreement with previous studies [58], our data support the idea that the polymerase genes (PB2, PB1, PA) of the 1918 A/H1N1 virus were of avian origin, since they are outside of the spectrum of the A/H1N1 folding energies and in the lower spectrum of folding energies of all human viruses. Also, our results support the hypothesis that the PB1 segment in the 1968 pandemic (but not necessarily in the 1957 pandemic) was of avian origin. The possibility of an avian influenza A virus strain crossing the host barrier and successfully propagating in humans has been controversial [111, 115]. So far, cases of avian-to-human transmission are limited, both in number and virulence. From the folding free energy perspective and in light of the results above, we can postulate that avian viruses whose RNP complex genes have folding energies in the corresponding human spectra will have increased chances to establish themselves in the human population. So far, no avian virus has been found with all its RNP segments in the human

range, although this might reflect gaps in the sequence data. Nevertheless, should a re-assortment and the necessary amino acid changes occur in HA segments coding for glycoproteins with specificity for human receptors (sialic acid alpha-2,6-galactose), it is possible that an avian A/H5N1 strain may cause a pandemic in humans.

To our knowledge, this is the first time that RNA folding was identified as a factor in the evolution and adaptation of the influenza A virus. Taken together, our results are consistent with the hypothesis that the host's body temperature may play an important role in the host adaptation of a virus, although clearly more experimentation is required. Interestingly, the folding free energy distribution of the swine viruses is intermediate between the avian and human distributions (**Figure 2.7**) and the swine is known as an intermediate host (possibly as a 'mixing vessel') for avian viruses jumping into humans. The swine's mean body temperature range is 37.8-38.6°C [116], which is also intermediate between avian and human body temperature ranges. Also, the folding free energy distributions of the avian viral genes become indistinguishable from the human distributions if the avian genes are folded at 38°C (**Figure 2.6**) in Additional data file 1). Having said that, the evolution of the influenza A virus is complicated and the folding free energy hypothesis cannot explain all observations. The RNP complex genes of the 1918 virus, for example, have very small folding free energies compared to the rest of the human viral genes and still caused one of the most devastating pandemics in history. Waterfowl birds present another interesting case. Influenza viruses isolated from chickens can seamlessly circulate in waterfowl birds, although the latter generally have higher average body temperatures [117]. On the other hand, the body temperature of waterfowl birds varies substantially between different organs, as well as the bird's activity during the day [118], which adds to the complexity of the evolutionary forces shaping the propagation of the virus.

Conclusions

This study is mainly based on computational analysis of the available influenza data. The results support the intriguing hypothesis that the RNA folding free energy of the polymerase genes plays an important role in the evolution and host specificity of the influenza A virus. We hope these results will stimulate further biochemical research on the subject. For example, isogenic chimeric viruses with different polymerase genes, but the same HA and NA segments, can be used to further test the hypothesis of viral replication dependence on temperature in human and avian cells. One of the challenges will be to combine amino acid composition, mRNA folding energy and other factors in a single evolutionary analysis framework. To that extent, work on animal models is necessary to help understand the mechanism by which RNA folding free energies shape the adaptation of the influenza virus from birds to humans.

Materials and methods

Sequences and codon usage tables

Influenza A sequences, isolated from human, and avian species, were downloaded from NCBI's Influenza Virus Resource Database [119] in March 2008. For the calculation of the folding energy distributions, we used all available human and avian strains with at least one complete ORF sequence (human: A/H1N1, A/H1N2, A/H2N2, A/H3N2, A/H5N1, A/H7N3, A/H9N2; avian: A/H1N1, A/H1N2, A/H1N3, A/H1N5, A/H1N6, A/H1N9, A/H2N1, A/H2N2, A/H2N3, A/H2N4, A/H2N5, A/H2N7, A/H2N8, A/H2N9, A/H3N1, A/H3N2, A/H3N3, A/H3N4, A/H3N5, A/H3N6, A/H3N8, A/H4N1, A/H4N2, A/H4N3, A/H4N4, A/H4N5, A/H4N6, A/H4N8, A/H4N9, A/H5N1, A/H5N2, A/H5N3, A/H5N6, A/H5N7, A/H5N8,

A/H5N9, A/H6N1, A/H6N2, A/H6N3, A/H6N4, A/H6N5, A/H6N6, A/H6N8, A/H6N9, A/H7N1, A/H7N2, A/H7N3, A/H7N4, A/H7N5, A/H7N7, A/H7N8, A/H7N9, A/H8N2, A/H8N4, A/H9N1, A/H9N2, A/H9N4, A/H9N5, A/H9N6, A/H10N1, A/H10N2, A/H10N3, A/H10N4, A/H10N5, A/H10N6, A/H10N7, A/H10N8, A/H10N9, A/H11N1, A/H11N2, A/H11N3, A/H11N6, A/H11N8, A/H11N9, A/H12N1, A/H12N4, A/H12N5, A/H12N9, A/H13N2, A/H13N3, A/H13N6, A/H13N9, A/H14N5, A/H14N6, A/H15N2, A/H15N8, A/H15N9, A/H16N3). The vast majority of the bird strains were isolated from chicken and duck (about equal number of sequences from each species). For the analysis of the folding free energies versus time, we used the more commonly circulating human strains (A/H1N1, A/H2N2, and A/H3N2). Only sequences corresponding to the complete ORF of each segment were considered for reasons we describe in the text. A complete ORF was defined as having both a start and a stop codon. The position of the start codon was determined by a multiple protein sequence alignment of each segment in each species, for a total of eight multiple alignments (four genes, two species). There are no length differences between the corresponding human and avian segments, although the four segments vary between them in terms of protein length (340-759 amino acids) and GC content (42.7-47% for human and 43-47.5% for avian mRNAs). If two or more segment sequences were identical at the nucleotide level, only one of them was used in the analysis. As we explained above, the choice of focusing on the ORF was dictated by the fact that the majority of the sequences in the database contain partial or no non-coding sequence. Thus, analyzing only the ORFs provided the largest possible dataset. Codon usage tables for human and chicken were obtained from the current version (September 2007) of the Codon Usage Tabulated from the GenBank (CUTG) database [107].

RNA folding

The folding free energy of each segment was computed using the Vienna RNA (version 1.6.5) package's RNAfold program [7-14], with the default parameters, save temperature, which was varied as we describe in the text.

Hydrogen bonding potential

The hydrogen bonding potential on the degenerate codon positions was calculated by assigning two hydrogen bonds to an A or U, and three to a C or G in every degenerate codon position. G•U pairs were not considered in this analysis, since it would have made it difficult to assign a number of hydrogen bonds to Gs and Us if the structure was unknown (or differed depending on the molecule). The bond assignment is based on the primary sequence, not the predicted secondary structure.

MDCK cell adaptation and plaque assays

MDCK cells were adapted for efficient growth at temperatures higher than 37°C (namely, 39°C, and 40°C). To minimize cell injury due to heat-shock and to ensure that cells are responsive to the viruses, we passaged them at higher temperatures gradually over a period of 21 days. MDCK cells were propagated in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal calf serum in 5% CO_2 and the temperature was increased by 0.2°C every three days. Aliquots of cells adapted for efficient growth at 39°C and 40°C were frozen at -80°C. Viruses were propagated and harvested from supernatants in cells grown at 37°C. MDCK cells plated in 6-well tissue culture plates were inoculated with 0.1 ml of virus serially diluted in DMEM. Virus was adsorbed to cells for 1 h, with shaking every 15 minutes. Wells were overlaid

with 1.6% w/v Bacto agar (DIFCO, BD Diagnostic Systems, Palo Alto, CA, USA) mixed 1:1 with L-15 media (Cambrex, East Rutherford, NJ, USA) containing antibiotics and fungizone, with 0.6 μg/ml trypsin (Sigma, St Louis, MO, USA). Plates were inverted and incubated for 2-3 days. Wells were then overlaid with 1.8% w/v Bacto agar mixed 1:1 with 2× Medium 199 containing 0.05 mg/ml neutral red, and plates were incubated for two additional days to visualize plaques. Plaques were counted and compared to uninfected cells. The ability of the PR8/34 (A/H1N1) virus to infect cells equally efficiently at all temperatures further suggests that any potential heat-shock effect is negligible.

Abbreviations

DMEM: Dulbecco's modified Eagle's medium; HA: hemagglutinin; MDCK: Madin-Darby canine kidney cells; NA: neuraminidase; NP: nucleoprotein; ORF: open reading frame; RNP: ribonucleoprotein; UTR: untranslated region.

Authors' contributions

PVB and RB-S conceived and designed the study, performed the computational analyses, and analyzed the data. DMC and CJC infected cells and collected viral titer data under the direction of TMR. PVB, RB-S, TMR and EG wrote the paper.

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 contains four figures showing various plots of folding energies (referenced in the main text) and one table listing the folding energies of vaccine strains WHO and CDC use against H5 influenza.

Acknowledgements

We thank David Lipman, Cassandra Miller-Butterworth, Roni Rosenfeld, and Paul Samollow for useful discussions and suggestions. We also thank the three anonymous reviewers for their constructive criticism. This work was supported by NIH-NIAID contract N01AI50018 and by NIH awards 1R01LM009657-01 (PVB), U01AI077771 (TMR) and R01GM083602 (TMR).

2.3 SUPPLEMENTARY MATERIALS

Table 2.2 Predicted folding energies of the five A/H5N1 strains that WHO and CDC use as vaccine strains against H5 influenza. For comparison purposes, the values for the 1918 strain were included. Bold letters indicate the smallest value of the segment; red letters indicate a smaller folding energy than the corresponding 1918 segment.

	Segment			
	1	**2**	**3**	**5**
A/VietNam/1203/2004 (VN/04)	-682.38	-627.92	**-650.94**	-495.72
A/Indonesia/05/2005 (Indo/05)	-680.61	-618.52	-625.42	-494.32
A/Hong Kong/156/1997	-652.82	-640.43	-612.30	-487.12
A/Hong Kong/483/1997	-660.72	-642.36	-622.63	-489.30
A/Hong Kong/486/1997	-657.71	-637.76	-609.70	**-497.60**
A/South Carolina/1/1918	**-691.92**	**-654.75**	-628.20	-482.42

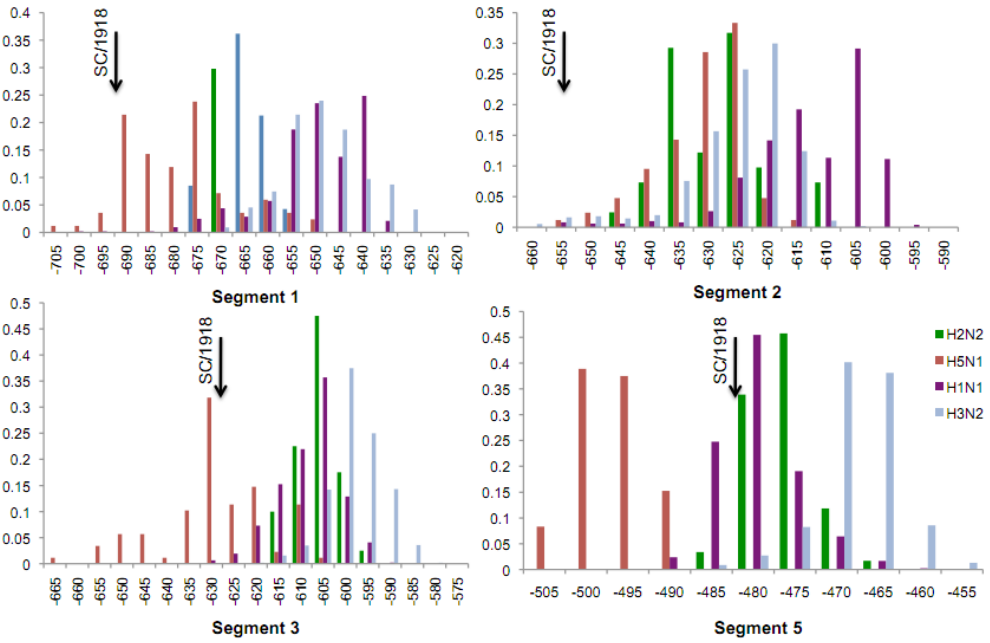

Figure 2.5 Folding free energies for all human influenza A polymerase gene segments (in kcal/mol). The arrows indicate the folding energies for the corresponding A/South Carolina/1/1918 (H1N1) virus segments.

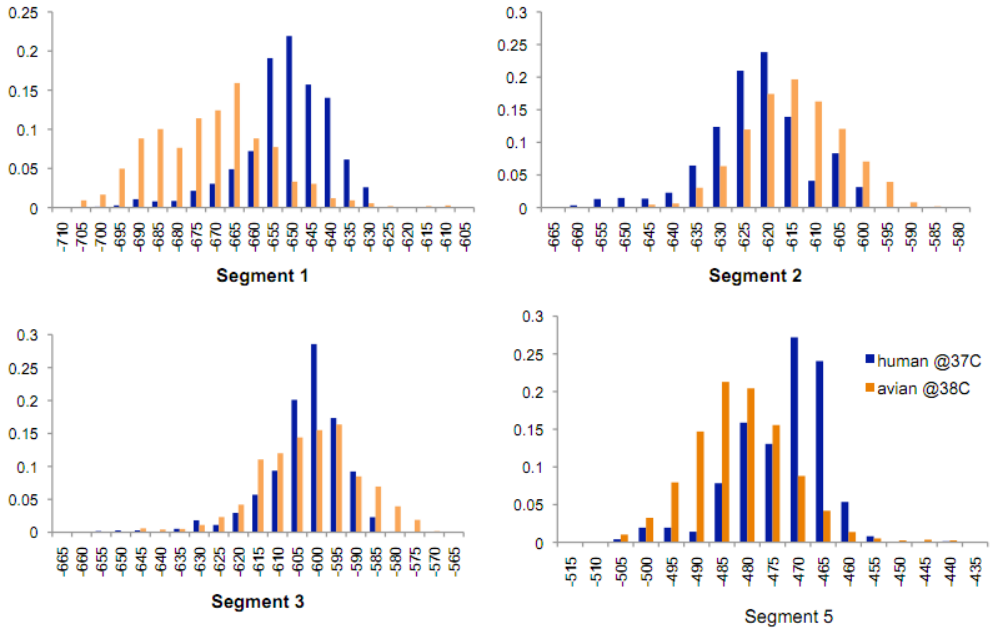

Figure 2.6 Distribution of folding free energy (in kcal/mol) for human, and avian influenza A strains folded at the body temperature of the host.

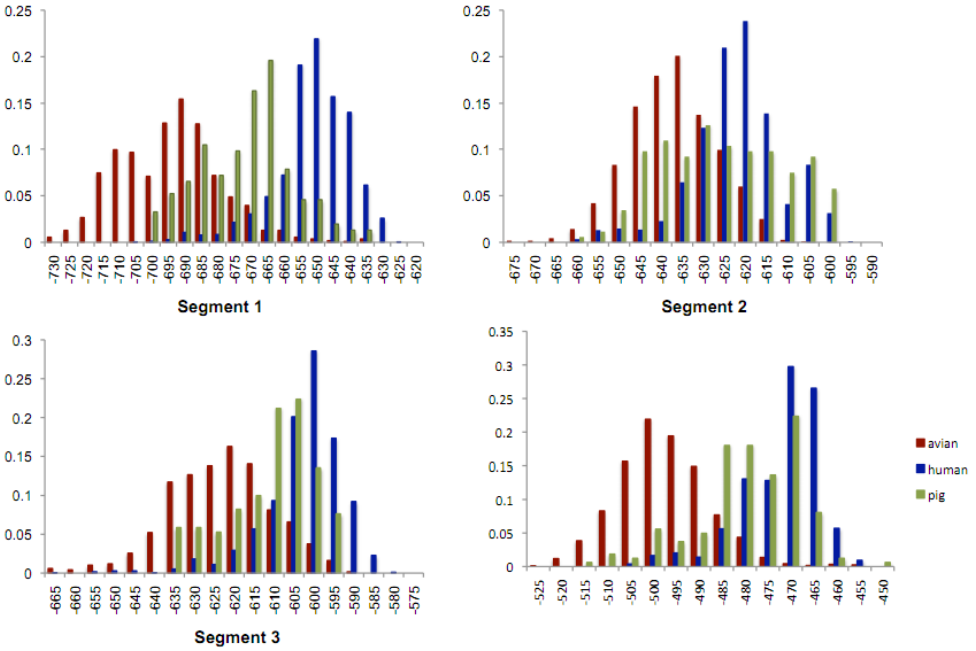

Figure 2.7 Folding free energy distributions for human, swine and avian influenza A polymerase gene segments (in kcal/mol).

Figure 2.8 Predicted folding free energy of the avian influenza A strains (polymerase genes) vs. year isolated.

3.0 USING FFE AND AMINO ACID FEATURES TO CLASSIFY INFLUENZA A IN THE HUMAN POPULATION

This work was done with the goal of being able to classify which influenza viruses had the capacity to infect and successfully propagate within the human population; a project made more relevant by the pandemic of 2009 H1N1 influenza that jumped to humans from swine. [73, 75] This work is in preparation for submission.

In the work detailed below, I conceived and designed this study with a good deal of advice and help from my committee, performed the computational analysis and analyzed the data,

3.1 ABSTRACT

While the pandemics of 1957, 1968 and 2009 were nowhere near as costly as the pandemic of 1918, they still resulted in an increase in lives lost compared to the typical seasonal strains of influenza. [51] The threat of H5N1 turning into a pandemic is also real. The ability to predict if a new strain of influenza has the potential of causing a pandemic is critical in terms of preventing or limit one, and thus is an important public health issue.

In chapter 2, we showed that influenza A polymerase genes fold more tightly in birds compared to humans. Here, we sought to determine whether any features of the influenza A polymerase genes and NP are associated with a successful entering and propagation of the virus

in the human population, independently of the rest of the genes in the influenza A genome. In this work we investigate two things. First, if we could pick key amino acid changes in a high throughput manner that could differentiate the human strains from those infecting other hosts. Such mutations may important polymorphisms for successful infection of the human host. Second, we will assess whether these changes (in addition to the folding of the mRNA) can be useful in classifying whether a given strain of the influenza A virus is similar to other strains that have the ability to enter and replicate efficiently in the human population.

Our analysis shows that, for these 4 gene segments, the identification of few amino acid positions in the protein, in conjunction with the folding energy of the mRNA, are able to accurately distinguish the human seasonal viruses' polymerase gene segments from strains from other hosts or from H5N1 human strains. Using our model, we were also able to accurately classify that the polymerase gene segments and NP of the strains of pandemic 2009 H1N1 as being similar to other influenza viruses that have circulated in the human population. We also further illustrate that swine harbor avian, human, and swine identified isolates, possibly fulfilling their role as a 'mixing vessel'. [59, 60, 62, 63, 120]

3.2 ATRICLE

INTRODUCTION

With the ability of influenza A to jump into the human population and cause pandemics, there is a need to understand what genomic features enable this. In 1918, an influenza pandemic, thought to have resulted from an avian virus entering the human population, [58, 98, 121] resulted in the death of an estimated 50 million people worldwide. [103] In 1957 and 1968, two

more influenza pandemics occurred through reassortments of a seasonal virus with an avian virus. [51, 52] In 2009, a new pandemic strain emerged, resulting in at least 18,449 deaths worldwide in 214 countries. [66] There is also the threat of a highly pathogenic H5N1 strain emerging into a pandemic strain. [100, 101, 122]

While computational modeling has appeared to focus on predicting when/how an influenza pandemic is occurring and how to mitigate it [111, 123-136], it might also be important to determine if a given virus might have the ability to cause one. In this study, we sought to determine amino acid mutations in the polymerase genes that might be associated with efficient replication of the virus in the human population; and to build a model to classify whether or not a given influenza virus is similar to those influenza viruses which have successfully circulated in the human population in the past. We focused on the polymerase genes PB2, PB1 and PA, which function together as the viruses' polymerase complex, and NP, which binds to the RNA to facilitate polymerase activity. [51] While the HA and NA genes are vital for the virus to recognize, bind to and escape from the host cell, a large amount of research has gone into determining which mutations in these genes are necessary for the virus to spread within the human population. [51-53, 55] Additionally, recombination experiments between avian and human viruses have yielded viruses with identical HA and NA genes but with vastly different replication and infection capabilities [137-139]; underscoring the importance of the internal genes, which include PB2, PB1, PA and NP. Furthermore, an important determinant of host range is an amino acid in PB2 in position 627- a lysine in human viruses. [64, 65]

By focusing on genes PB2, PB1, PA and NP, we are attempting to build a tool that will correctly classify which combinations of these genes can successfully replicate and propagate in humans; a step toward predicting which influenza A viruses could do the same.

When trying to identify if a new strain of influenza can enter the human population by jumping from another host species, the first response is too look at whether similar events have already occurred. Here, we run into a problem. Prior to 2009, we had only one instance where an entire influenza strain is thought to have jumped into humans from another host, the 1918 pandemic strain. For PB2, PA and NP, we have just this one instance, and one sequence. For PB1, which is thought to have entered the human population *via* reassortment in 1957 and 1968 as well, we have three instances. **Figure 3.1** is a small graphic depicting these events. As such, we are trying to classify a when a novel influenza virus is similar to other viruses that have successfully propagated in the human population from a data set of one. We can to identify which mutations are relevant to the virus host jumping into humans, but we also must separate out these mutations from other mutations the virus may have picked up along the way.

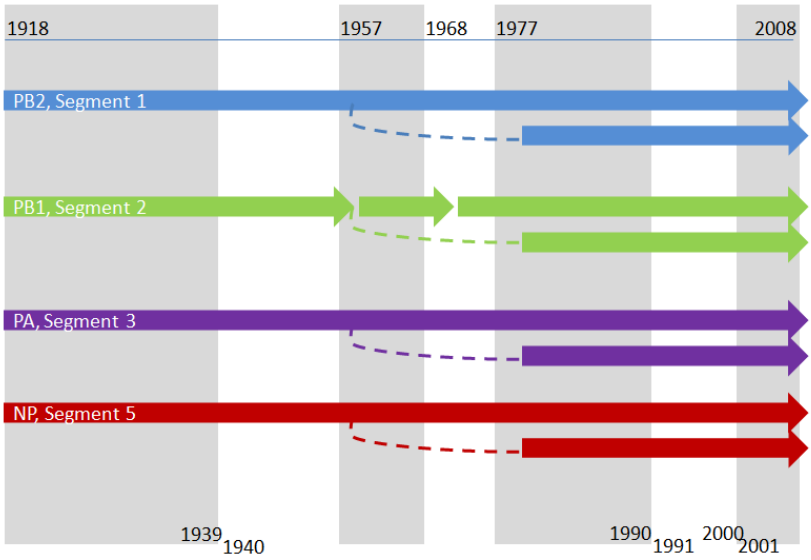

Figure 3.1 A time line depicting the entry of each segment into the human population prior to the pandemic influenza H1N1 in 2009. The dashes indicate that the H1N1 viruses that starting circulating in 1977 is the same as the virus that stopped circulation 1957 as a result of that years' pandemic. The grey and while background blocks denote the time intervals used later in this study.

This problem is further compounded by the fact that the influenza virus is still currently circulating in the human host, and acquiring new mutations as it does so. These new mutations may not relate to the initial mutations that allowed the virus to enter the human population at the earlier time point, but may relate to other evolutionary pressures on the host-virus relationship. In addition, this relatedness of the data violates the idea that the data points are independent from each other, a problem that we will refer to as the *lineage problem*.

In addition to the fact that the viruses are related, the sequence database does not have its influenza sequences uniformly distributed. The database is predominately dominated by strains isolated in the last few decades.

With only one or a few instances of these segments jumping into the human population, compounded by the relatedness of the data, and the non-uniform nature of its distribution throughout the past century, trying to predict which features of the polymerase genes are important to the virus entering the human population is rather difficult.

In an attempt to mitigate these issues, we divided that data up into seven discrete time intervals. Due to the sparse nature of the data, and the fact that we tried to take into account the antigenic drifts and shifts of the virus, some time intervals are a good deal larger than others. By doing this, we are trying to both limit the effect of the uneven data distribution; limit the affect of the sequence relatedness and mutations acquired after new host entry; and to possibly identify other mutations that are hallmarks of the viruses' evolution within the human host.

METHODS

Using the coding regions of the PB2, PB1, PA and NP genes, we divided the sequence data into four groups- **(1)** avian, **(2)** swine, **(3)** human seasonal, and **(4)** human isolated strains

that were not seasonal. The distribution of the sequences across the populations is detailed in **Table 3.10**. What information we had to use as input is the sequence and multiple sequence alignment for each gene, and the sequences' folding energy.

As influenza in humans has been steadily evolving, we divided human seasonal influenza A up into seven time intervals: **(1)** 1918-1939, **(2)** 1940-1957, **(3)** 1958-1968, **(4)** 1969-1976, **(5)** 1977-1990, **(6)** 1991-2000, and **(7)** 2001-2008. Each division of the human seasonal influenza A was used to build a model (of the amino acid sequence and folding energy characteristics) with the non-human seasonal strains. In effect, there are 7 individual models from which we make our classifications from, one for each time interval (above). This was done to mitigate the problem of having limited instances of these influenza A segments entering the human population. While dividing the data into the intervals does not solve this problem, we can use this method as a means to sort out the mutations that occurred early on in the viruses' adaptation to humans from those that occurred later.

We call a strain "classified as human" if any of the seven models classifies it as "human". 10-fold cross validation was used to build the models to score the amino acid and folding energy of the pre-2008 influenza strains. The entire datasets of pre-2008 strains were used to train the model for the pH1N1 data set. Pandemic H1N1 was not part of the training dataset, and we used it as an independent test set.

Part 1: Using a Naïve Bayes Algorithm on the entire sequence. We first tested the ability to classify the host of the viral segment by using the Naïve Bayes algorithm on the entire sequence each of the four genome segments' sequences. In this, we calculated the probability of a given sequence belonging to a given population label, using the entire amino acid sequence. [140]

$$P(label = l|sequence = s) = \frac{P(sequence = s|label = l) \cdot P(label = l)}{P(sequence = s)}$$

Because the data are scarce, we can use the Bayes rule together with the assumption that each position is independent

$$P(label = l|sequence = s) = \frac{P(label = l)\prod_i P(a.a.\,at\ pos\ i = a|label = l)}{\sum_j P(label = j)\prod_i P(a.a.\,at\ pos\ i = a|label = j)}$$

In the above equation *P(label=l)* is estimated by the percentage of sequences that have *label=l* and *P(a.a. at pos i = a*) is also calculated over all sequences.

Because we are interested in the label, the above equation can be evaluated for the argmax label of a given sequence

$$Label = \text{argmax}_l \frac{P(label = l)\prod_i P(a.a.\,at\ pos\ i = a|label = l)}{\sum_j P(label = j)\prod_i P(a.a.\,at\ pos\ i = a|label = j)}$$

Because the denominator doesn't depend on label=*l*, and is constant for the given sequence, it can be simplified to:

$$Label = \text{argmax}_l\ P(label = l)\prod_i P(a.a.\,at\ pos\ i = a|label = l)$$

By using logs:

$$LabelScore = \text{argmax}_l\ [\log(P(label = l)) + \sum_i \log\ (P(a.a.\,at\ pos\ i = a|label = l))]$$

The output is then the score of the segment given a label; or the population with the maximum score. When outputting the score, this is a numeric value. From this, we can get the percent likely-hood that the sequence belongs to a given population; with the maximum value corresponding to the more likely population. As we are evaluating a given sequence over all the

time intervals, we get a score on what population the sequence belongs in for each time intervals (seven scores). From these 7 scores, if the sequence is identified as human seasonal for any time interval, the sequence is labeled as human.

Part 2: Identifying a few Informative Positions. Several studies have revealed that the change of a subset of amino acids may be related to host range. [58, 64, 65] With this knowledge, we looked for a small subset of positions that we could use to identify the host. Chen *et al.* 2009 has been working to catalog the mutations that separate the human circulating and infecting influenza A segments from the avian influenza A segments (they call these changes in amino acid (AA) sequence that result in host change signatures).[74] However, of the identified 47 non-glycoprotein signatures that separate the pre-2009 influenza sequences from birds, only 8 were identified in the pH1N1 influenza. To identify the most informative mutations computationally, we used mutual information (MI) to rank each position relatively to the human seasonal label using the multiple sequence alignment of the amino acid sequence for each gene segment. The MI is calculated using the formula below where the MI_i and $p_i(a)$ refer to a position-specific quantities.

$$MI_i(aa, label) = \sum_{l \in \{label\}} \sum_{a \in \{a.a.\}} p_i(l,a) \log \left(\frac{p_i(l,a)}{p(l) \cdot p_i(a)} \right)$$

We then ranked the positions by the MI score relative to the human seasonal label. The top twenty informative positions were used as features, starting at two and iteratively adding one position. In this manner, we were able to use the positions to generate a probability, or score, that related each sequence to its most likely label.

The model is built using the counts of the amino acids in each group for each position used a feature for each sequence in the training set. Pseudo counts were used, so no probability

of a given amino acid is zero. From this, the scores are calculated as detailed above, with the highest score indicating which group the sequence is in. A flow chart of the process is presented in **Figure 3.2**.

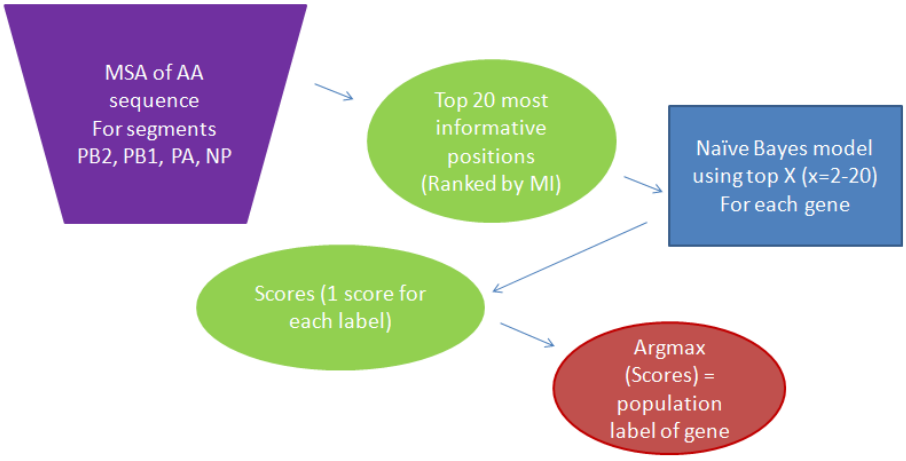

Figure 3.2 A flow diagram of the process outlined above. Using the multiple sequence alignment of the amino acid sequence, we can get the most informative positions. We can then use Naïve Bayes to generate a score (and this population label) for each sequence. The final output is a call as to what the host group is (avian, human seasonal, human bird, or swine). In this example, the purple trapezoid represents the input, green ovals intermediate outputs, blue rectangles the model, and red ovals the final output. MSA: multiple sequence alignment; AA sequence: amino acid sequence; MI: mutual information.

We ran this in 3 sets of training and testing data: a 10-fold cross validation set (in the time interval specified, 10-fold cross validation was done on the human seasonal and non-human seasonal sequences in this interval, and the other human seasonal sequences were used as a test set); and then training only on the human seasonal sequences of a given time interval; and then testing with the other human seasonal sequences.

Part 3: Accounting for Evolution and Adaptation. When using the human sequences, we run into the issue that the segments have entered the human population only one time (in the

case of PB1, three times). This is an issue, as we have a difficult time differentiating between which mutations are necessary for the virus to circulate in the human population, and which mutations arose by happenstance and were simply carried down the evolutionary line but do not relate to host specificity.

In our previous work, we demonstrated that the folding energy of these influenza A genes may be an important feature in the host range of the virus. Furthermore, the folding energy of these genes was continually evolving to a higher folding energy (less thermodynamically stable) as continual passage through the human population occurred. [141] As folding energy of these genes is related to the evolution of the genes in the human population, we can screen against the subsequent amino acid changes by accounting for folding energy.

$$P(amino\ acid\ i | segment\ FFE) = P(amino\ acid\ i)$$

To account for this, we filtered the ranked positions by checking their independence to the folding energy of the segment in the human seasonal population. We allowed plus/minus 10%. The scores were then calculated in the same manner as above. This was used as filter on the previously identified positions: if the probability of an amino acid was found to be independent of folding energy in the human population, this position was kept: otherwise, it was skipped over.

The process is summarized in **Figure 3.3**.

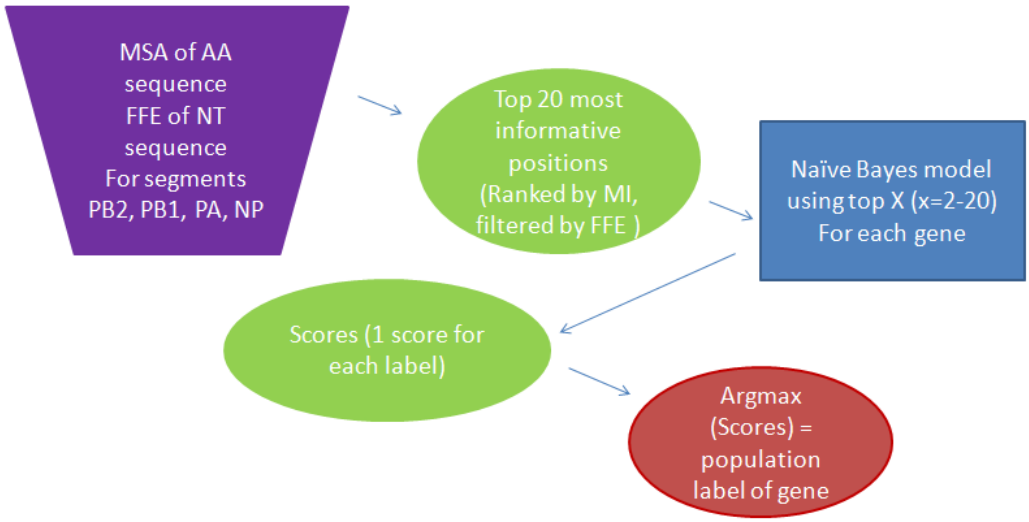

Figure 3.3 A flow diagram of the process outlined above. Using the multiple sequence alignment of the amino acid and FFE of the sequence, we can get the most informative positions, and screen them by the folding energy distributions. We can then use Naïve Bayes to generate a score (and this population label) for each sequence. The final output is a call as to what the host group is (avian, human seasonal, human bird, or swine). In this example, the purple trapezoid represents the input, green ovals intermediate output, blue rectangles the model, and red ovals the final output. MSA: multiple sequence alignment; AA/NT sequence: amino acid/nucleotide sequence; MI: mutual information.

The same training sets and scoring were used as in the above example.

Part 4: Computationally classifying if a specific polymerase gene set is similar to those which propagate in the human population. Work from several lab groups, which are building reassorted influenza viruses from pH1N1, human seasonal, and avian viruses, suggest that swapping a few segments with a human virus could make an non-human virus able to replicate in mammals. [137-139] We then grouped segments together by genomes to look at the combined effect of PB2, PB1, PA and NP.

We can also gage how well the folding energy of each segment relates to the folding energy of the segment in different populations. In this, we assumed that for each population, the

folding energy distribution is normal around the sample mean. If the folding energy of a specific segment was within ±1 standard deviation of the mean, it was given a 68% chance of being from that population; ±2 standard deviation of the mean, it was given a 27% chance of being from that population; more than that was given a 5% chance of being from that population.

To classify whether the polymerase gene set could propagate in the human population, we used the glmnet [142, 143] package in MatLab, which performs regression with a lasso penalty, to build a model using the amino acid and folding energy scores of each segment in each population as input. Each strain has a vector of 32 inputs (there are 16 inputs per gene segment, 4 per population. The FFE score and the amino acid score is present for each of the 4 populations). The model is run 7 times, with each human subpopulation being used once (or once for each time interval). This yields 7 calls as to what population a given sequence belongs in. If, at any point, the virus identifies as human in any human subpopulation, it is marked as human seasonal.

We compiled this model twice for each data set; once as a binomial model for classifying human/non-human hosts of the virus, and once as a multinomial model with weights (the weight of each data point was inversely proportional to the number of data points in the model with the same label- this model was able to detect origin of virus).

This process is summarized in **Figure 3.4**.

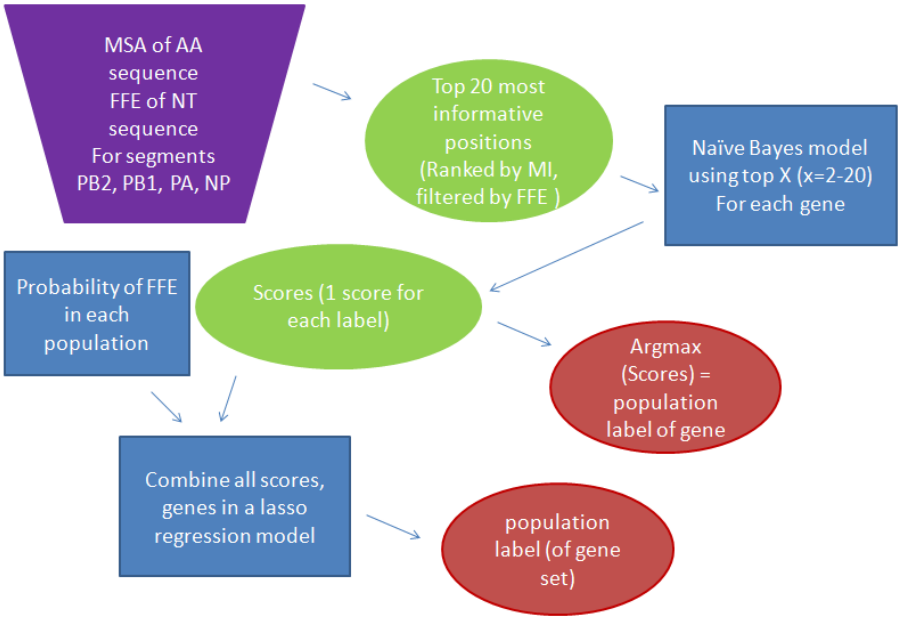

Figure 3.4 A flow diagram of the process outlined above. Using the process outlined in Figure 3.3, we added a regression model that combines the scores and folding energy probability of an influenza viruses 4 gene segments. The final output is a call as to what the host group is (avian/human seasonal/human bird/ swine if the model is multinomial or human/non-human if binomial). From this, we have a population label classifying the host of the entire gene set together. In this example, the purple trapezoid represents the input, green ovals intermediate output, blue rectangles the model, and red ovals the final output. MSA: multiple sequence alignment; AA/NT sequence: amino acid/nucleotide sequence; MI: mutual information**.**

For the data, we ran the models once on 10-fold cross validated data using the LabelScores generated by each time interval model, and then again using the data that was trained only on the human sequences of the specific time interval and all the non-human sequences.

RESULTS

Part 1: Using a Naïve Bayes Algorithm on the entire sequence. With this approach, calculating the label using the entire amino acid sequence, did not result in a good classification tool for identifying influenza A sequences host group. While it does a decent job of classifying the influenza strains from pre-2008 (the strains it was tested on using 10-fold cross validation), it failed to correctly classify the pandemic 2009 H1N1 sequences human for all but segment 2.

Table 3.1 Classification efficiency of the Naïve Bayes classifier on the entire amino acid sequence of each segment using 10-fold cross validation; the classifier classifies a sequence as "human" or "non-human" as we describe in the text.

Segment	Sensitivity (pre 2008) (%)	Specificity(pre 2008) (%)	pH1N1 classified as human seasonal (%)
1	98.5	99.7	0
2	96.1	94.9	99.2
3	99.0	98.8	0
5	98.35	98.3	0

In this model, for the most part, the pre-2008 human influenza segments are classified as human; this is also seen for the classification of avian segments. For the human bird and swine segments, however, this is not the case. The human bird segments are largely classified as avian segments. The swine segments are classified as either swine, avian or human seasonal, perhaps underscoring the ability of pigs to act as a 'mixing vessel'. [59, 60, 62, 63, 120]

Table 3.2 Sensitivity of the Naïve Bayes classifier only on swine sequences. In this case, the classifier classifies a sequence as "swine", "human", or "avian", showing that the swine sequences classify to the other population labels

Segment	Swine (%)	Human (%)	Avian (%)
1	42.8	9.4	47.8
2	41.3	31.5	27.2
3	43.1	8.8	48.1
4	63.1	10.8	26.1

Part 2: Identifying a few Informative Positions. The positions with the highest mutual information content are summarized in **Table 3.3**. Of the 10 amino acid positions in the polymerases (PB2, PB1, PA) noted by Taubenberger *et al*. 2005 [58] that differentiate human and avian influenza strains, all 10 were identified in the top 20 amino acids positions, with all but 3 being in the top 10.

Table 3.3 A list of informative amino acid positions, as identified by mutual information on the amino acid sequence alignments. Numbers 1-10 designate the top 10 features (amino acid positions) for each model for each segment. These positions were then used to generate a Naïve Bayes model that was used to classify the sequences.

Segment	Model	1	2	3	4	5	6	7	8	9	10
1	1918 to 1939	474	198	660	183	367	194	65	472	8	466
	1940 to 1957	474	198	660	367	8	291	587	452	183	270
	1958 to 1968	660	474	198	452	367	587	8	104	270	291
	1969 to 1976	474	660	198	675	452	367	337	8	587	381
	1977 to 1990	660	474	198	367	587	63	8	104	270	291
	1991 to 2000	587	660	104	63	566	80	474	612	673	367
	2001 to 2008	587	104	612	80	63	566	660	673	270	43
2	1918 to 1939	360	580	429	374	653	472	575	338	399	434
	1940 to 1957	360	580	429	653	374	472	575	338	399	434
	1958 to 1968	360	580	374	429	653	338	472	399	434	751

	1969 to 1976	360	580	374	429	211	653	472	338	399	326
	1977 to 1990	360	580	740	335	429	653	472	374	56	575
	1991 to 2000	580	335	360	326	583	740	429	211	620	215
	2001 to 2008	326	580	335	360	583	429	211	215	740	485
3	1918 to 1939	399	64	408	347	320	54	381	271	240	331
	1940 to 1957	399	64	320	408	54	81	240	347	355	71
	1958 to 1968	399	320	64	54	408	355	381	141	240	347
	1969 to 1976	399	320	64	141	54	408	355	381	276	347
	1977 to 1990	399	320	64	54	355	381	408	336	65	551
	1991 to 2000	320	399	64	336	276	551	27	267	65	420
	2001 to 2008	399	336	551	267	420	320	342	64	65	27
5	1918 to 1939	99	356	429	135	304	350	349	288	424	188
	1940 to 1957	99	35	304	35	432	350	429	349	288	424
	1958 to 1968	99	356	304	455	135	350	449	422	429	349
	1969 to 1976	99	356	304	350	135	449	429	349	455	32
	1977 to 1990	99	356	304	135	350	312	32	282	60	15
	1991 to 2000	99	372	356	304	312	454	282	60	352	15
	2001 to 2008	99	374	312	441	282	454	60	15	421	356

Using these positions, we scored each sequence, and identified the population with the highest score; the results are summarized in **Figure 3.5**. While this method has improved sensitivity and specificity (see **Table 3.4)**, it still fails to classify the pandemic 2009 H1N1 sequences as being human for all but segment 2.

Table 3.4 Maxim**um** classification efficiency of the Naïve Bayes classifier on the top 20 acid sequence of each segment; the classifier classifies a sequence as "human" or "non-human" as we describe in the text, using 10-fold cross validation

Segment	Sensitivity (pre 2008) (%)	Specificity(pre 2008) (%)	pH1N1 classified as human seasonal (%)
1	98.6	98.8	0
2	96.7	95.6	99.15
3	99.0	98.9	0
5	98.3	98.2	4.9

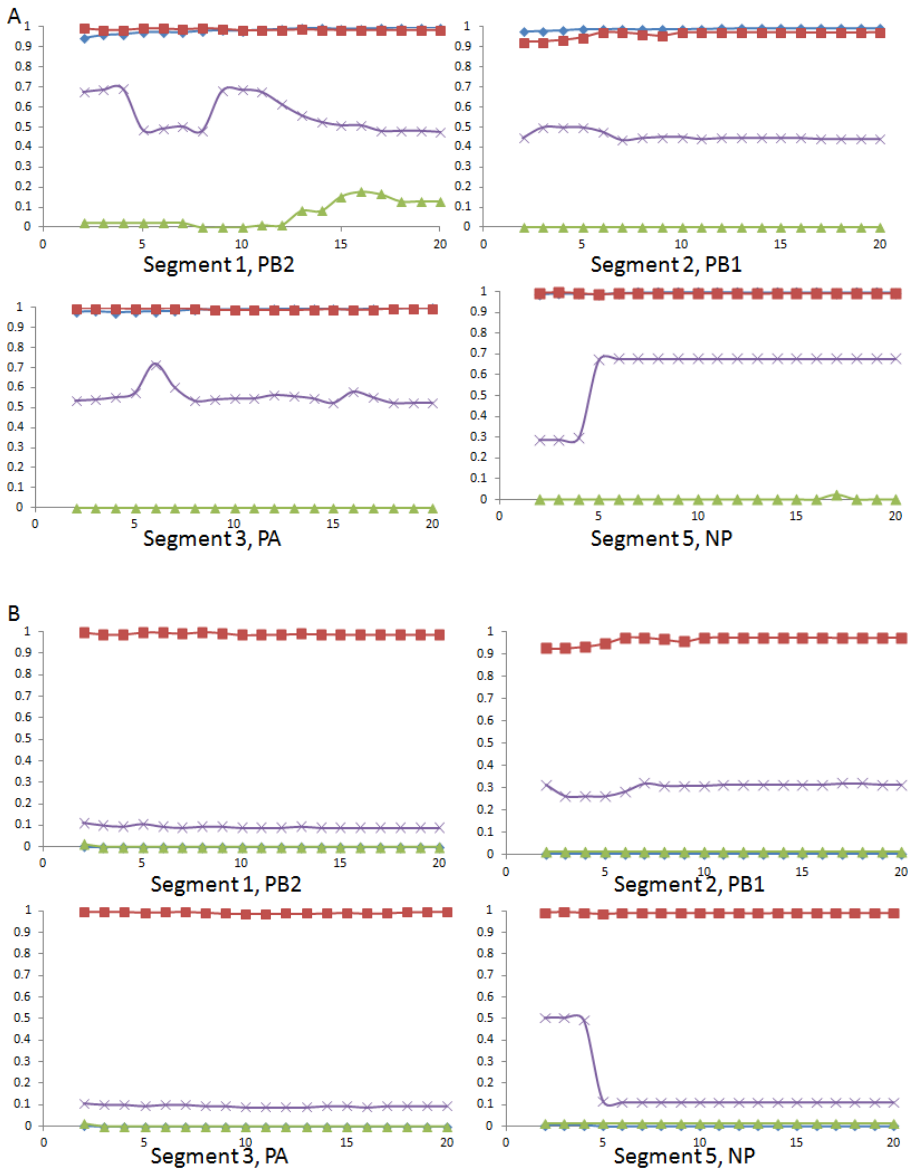

Figure 3.5 A. The ability of the classifier to classify each influenza A segment as being in its own group, using the positions identified by mutual information and 10-fold cross-validation. **B.** The ability of the classifier to classify each sequence as being human seasonal, using the positions identified by mutual information and 10-fold cross-validation. Blue is avian sequences classified as avian; red shows human seasonal classified as human seasonal, purple shows swine classified as swine; and green is human bird classified as human bird.

When we analyze the ability of the model to human seasonal sequences, we can see that at 10-fold cross validation that it does a good job. When we change the training and testing data, and look at the ability of a single time interval to predict both itself and the other time intervals, we see that a slightly different picture emerges, with some time intervals have better classification power than others.

For segment 1, in the first time interval of 1918 to 1939, there is a limited number of sequences, and the population is diverse in its amino acid composition across the informative positions. Because of this, the predictive model build with sequences from this interval cannot accurately predict either the sequences in its own interval or in any other time interval. Starting with the second time interval, however, we see that the models trained with these sequences are able to actually predict pretty well the sequences from their own time intervals and most of the other time intervals. Segments 3 and 5 show a similar pattern to segment 1, (Although for segment 3, the first time interval is useful).

For segment 2, we see a different story, with the later time intervals (after the last introduction in 1968) being able to correctly classify each other's segments, and the earlier time intervals not having this ability. This may be related to the repeated instances of this gene segment entering the human population.

These results are summarized in **Table 3.5**.

Table 3.5 Maximum classification efficiency of the Naïve Bayes classifier on the top 20 amino acid sequence features of each segment; in this case, we are testing the ability of each time interval to classify sequences that belong to both itself, and the other intervals. In this case, the self-classification is summed in one number, as accuracy of the self classifications across all time intervals. The classifications across the other time intervals are separated according the interval that was trained on, as the classification sensitivity varied widely.

Segment	Interval Classifying Self (%)	Interval Classifying Other Intervals' Sequences (%)						
		1918 - 1939	1940 – 1957	1958 - 1968	1969 - 1976	1977 – 1990	1991 – 2000	2001 - 2008
1	98.8	0	98.8	98.6	98.6	98.9	99.5	99.5
2	95.0	0.02	0.27	0.01	66.6	94.1	85.4	72.5
3	99.4	74.4	74.4	82.3	82.4	98.6	99.4	99.6
5	98.6	0.01	59.1	54.0	73.5	99.4	97.9	97.1

Part 3: Accounting for Evolution and Adaptation. The positions with the highest mutual information content are summarized in **Table 3.6**. Of the 10 amino acid positions in the polymerases (PB2, PB1, PA) noted by Taubenberger *et al*. 2005 [58] that differentiate human and avian influenza strains, all 10 were identified in the top 20 amino acid positions, and improved to all but 2 being in the top 10.

Table 3.6 Below is the list of the top 10 features for each model for each segment**.** These positions were identified by mutual information on the amino acid sequence alignment, and then filtered by using the folding energy distributions (detailed in part 3). These positions were then used to generate a Naïve Bayes model that was used to classify the sequences

Segment	Model	1	2	3	4	5	6	7	8	9	10
1	1918 to 1939	474	198	183	367	65	472	8	466	675	237
	1940 to 1957	474	198	8	291	587	452	183	270	63	65
	1958 to 1968	660	474	198	452	367	587	8	104	270	291
	1969 to 1976	474	660	198	675	452	367	337	8	587	381
	1977 to 1990	660	474	198	367	587	63	8	104	270	291

	1991 to 2000	587	660	104	63	566	474	673	367	198	270
	2001 to 2008	587	104	612	63	566	660	673	270	43	474
2	1918 to 1939	360	580	429	374	653	472	575	338	399	434
	1940 to 1957	360	580	338	399	434	751	151	79	128	647
	1958 to 1968	360	580	374	653	338	472	399	434	751	151
	1969 to 1976	580	374	211	653	472	338	399	326	434	751
	1977 to 1990	360	740	429	178	338	399	434	751	151	181
	1991 to 2000	580	335	112	148	13	399	338	151	181	434
	2001 to 2008	326	580	335	112	13	148	383	399	151	339
3	1918 to 1939	399	64	408	347	320	381	240	331	351	141
	1940 to 1957	399	64	408	54	381	240	347	355	271	331
	1958 to 1968	399	320	64	54	408	355	381	240	347	271
	1969 to 1976	54	408	355	381	276	390	27	711	519	93
	1977 to 1990	399	320	54	355	381	408	336	551	267	27
	1991 to 2000	399	336	551	27	267	420	54	224	355	56
	2001 to 2008	399	336	551	267	27	224	54	355	99	56
5	1918 to 1939	99	356	135	304	350	349	288	188	20	455
	1940 to 1957	99	304	432	350	429	349	288	424	188	20
	1958 to 1968	99	356	304	135	350	449	422	429	349	33
	1969 to 1976	99	356	304	350	135	449	429	349	455	32
	1977 to 1990	356	304	350	312	32	282	60	15	454	441
	1991 to 2000	99	356	304	312	454	282	60	15	441	421
	2001 to 2008	99	374	312	441	282	454	60	15	421	356

Using these positions, we scored each sequence, and identified the population with the highest score; the results are summarized in **Figure 3.6 and Table 3.7**. While we improve on the algorithm above, we only classify the pandemic 2009 H1N1 sequences as being human for segments 2 and 5. The pandemic 2009 H1N1 influenza A virus was a triple reassortment of viruses circulating in the swine population: PB2 and PA are thought to be from an H1N1 avian

virus that entered the swine population; PB1 is thought to have originated from a human seasonal H3N2 virus that entered the swine population; and NP is thought to have come from a classical swine virus of the N. American lineage.[73] In this example, we are correctly classifying PB2 (which is of human influenza ancestry), and NP (ancestry is thought to be similar to that of human 1918 H1N1 pandemic[144]).

Table 3.7 Maxim**um** classification efficiency of the Naïve Bayes classifier on the top 20 amino acid sequence features of each segment; the classifier classifies a sequence as "human" or "non-human" as we describe in the text.

Segment	Sensitivity (pre 2008) (%)	Specificity(pre 2008)(%)	pH1N1 classified as human seasonal (%)
1	99.0	98.8	0
2	98.4	98.0	99.2
3	99.0	98.8	0
5	98.4	98.3	100

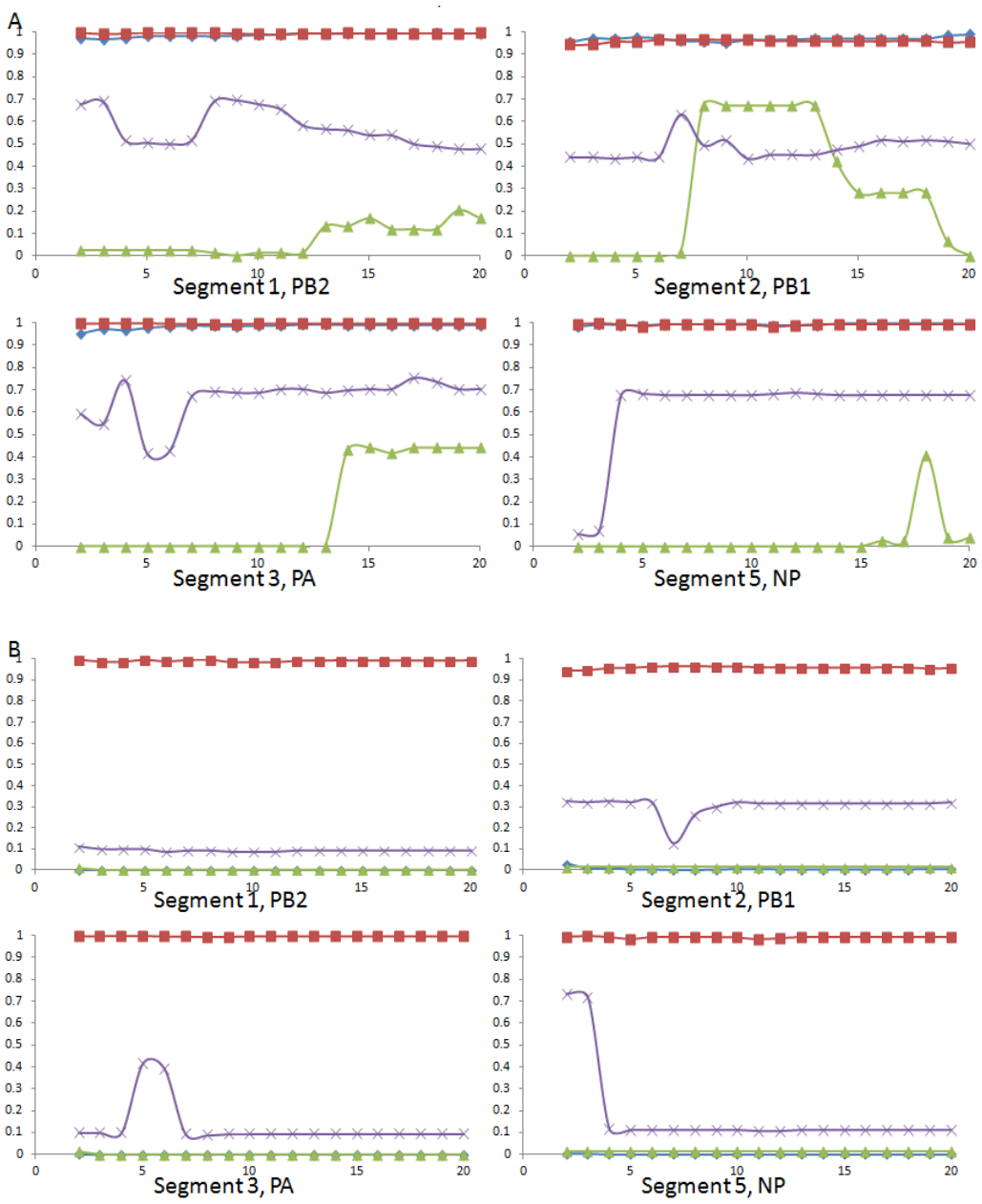

Figure 3.6 A. The ability of the classifier to classify each influenza A segment as being in its own group, using the positions identified by mutual information and screened by segment folding energy and 10-fold cross-validation. **B.** The ability of the classifier to each sequence as being human seasonal, using the positions identified by mutual information and screened by segment folding energy and 10-fold cross-validation. Blue is avian sequences classified as avian; red shows human seasonal classified as human seasonal, purple shows swine classified as swine; and green is human bird classified as human bird.

When we analyze the ability of the above model, after we filter the MI positions using folding energy, to predict human seasonal sequences, we can see that at 10-fold cross validation it does a great good job. When we change the training and testing data, and look at the ability of a single time interval to classify both itself and the other time intervals, we see a slightly different picture emerge.

We see a similar result as before, with for segment 1, the first time interval of 1918 to 1939, the number of sequences is too small, and the population to divided in amino acid composition, for it to be able to accurately classify the sequences in its own or any other time interval. Starting with the second time interval, however, we see that the time intervals are able to actually correctly classify sequences from its own and most of the other time intervals as well. Segments 3 and 5 show a similar pattern to segment 1, although for segment 3, the first time interval is informative).

For segment 2, we see a different story, with the later time intervals (after the last introduction in 1968) being able to correctly classify each other's segments, and the earlier time intervals not having this ability.

The results are summarized in **Table 3.8**.

Table 3.8 Maximum classification efficiency of the Naïve Bayes classifier on the top 20 features of each segment; in this case, we are testing the ability of each time interval model to correctly classify sequences in both itself, and the other intervals' sequences. Self-classification is summed as one number, as accuracy of the self classifications across all time intervals. The classifications across the other time intervals are separated according to the interval that the classifier was trained on, as the classification sensitivity varies widely.

Segment	Interval Classifying	Interval Classifying Other Intervals' Sequences						
		1918 -	1940 –	1958 -	1969 -	1977 –	1991 –	2001 -

	Self	1939	1957	1968	1976	1990	2000	2008
1	99.0	0	98.8	98.6	98.6	89.9	99.5	99.5
2	94.2	0.02	0	0.01	65.6	70.7	83.4	83.0
3	98.7	74.4	74.2	82.3	86.6	98.6	99.5	99.6
5	98.2	0.02	33.5	54.3	73.5	99.4	99.1	97.1

Part 4: Computationally classifying a specific polymerase gene set as similar to those propagating in the human population. The results of the binomial model, using the folding energy distributions and the part 3, are summarized in **Figure 3.7**, and the multinomial model in **Figure 3.8**. The binomial model has a drop in sensitivity to about 90%, its specificity can go up to 100%, and this model accurately classifies the pandemic 2009 H1N1 viruses' gene set as being similar to other human influenza A viruses.

Of the 10% of sequences that are miss-classified as human influenza A segment sets, most are swine viruses. This may indicate a pool of viruses that may have the ability to jump into a new host. Of the one virus labeled as human seasonal that is sometimes classified as not, identifies highest with the swine viruses using a blast search of the influenza database [76], indicating this may be a virus that isn't a human seasonal virus.

The multinomial model has a better sensitivity (97.9%), and it also classifies the pandemic 2009 H1N1 sequences (but over a smaller range of features when compared to the binomial model). Once again, the segment sets that were miss-classified as human influenza sets are mostly swine viruses.

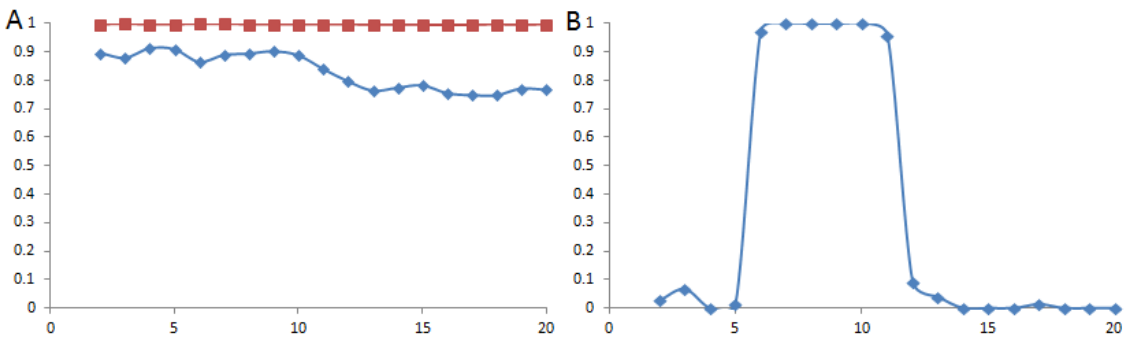

Figure 3.7 Classifying the ability of the virus' 4 gene set to replicate in the human population using a binomial model. A denotes the sequences trained by 10-fold cross validation that are prior to the pandemic 2009 H1N1 viruses. B denotes training on the data in A, but testing on the pandemic 2009 H1N1 gene set. The positions were determined by mutual information screened by folding energy. Blue denotes the non-human seasonal sequences identified as non-human; while red denotes the human sequences.

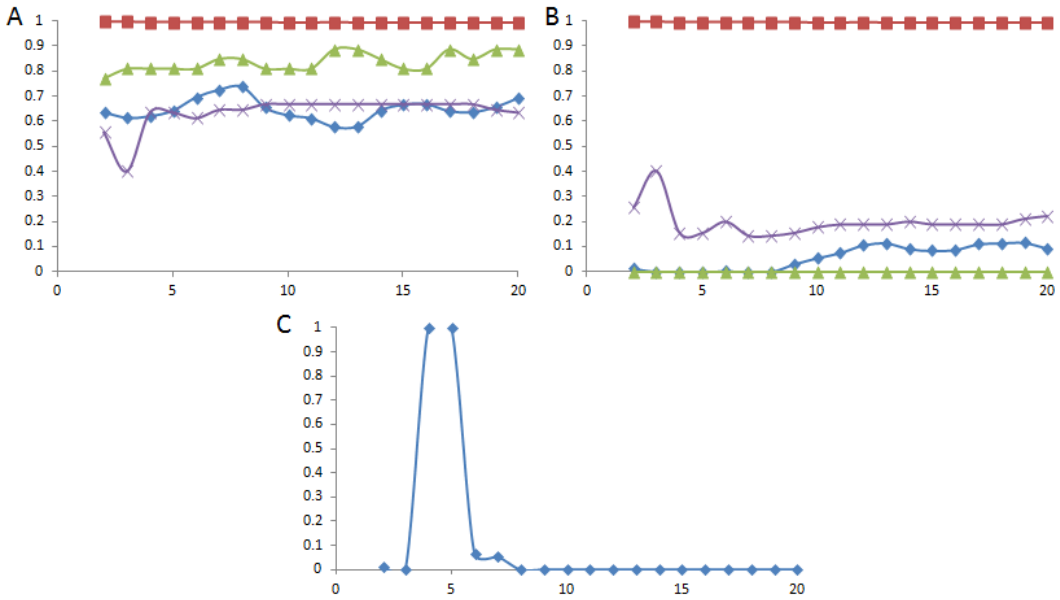

Figure 3.8 Classifying the ability of the virus' 4 gene set to replicate in the human population using a multinomial model. A denotes the sequences trained by 10-fold cross validation that are prior to the pandemic 2009 H1N1 viruses that are identified as the correct group; B denotes the viruses that are denotes the viruses denoted as human seasonal; C denotes the ability of the model to correctly classify the pandemic 2009 H1N1 gene set. The positions were determined by mutual information screened by folding energy. Blue denotes the avian sequences identified as avian; while red denotes the human sequences; purple denotes swine sequences as swine; and green denotes human bird.

Both of the models that took into account the folding energy of the polymerase gene segments were able to classify the pandemic 2009 H1N1 virus as a virus could propagate in the human population. To compare the model performance, we generated the same models for the scores of part 2 (16 inputs). Both of these models fail to classify pandemic 2009 H1N1 as being similar to other human seasonal viruses. A flow diagram of these models is presented in **Figure 3.9**.

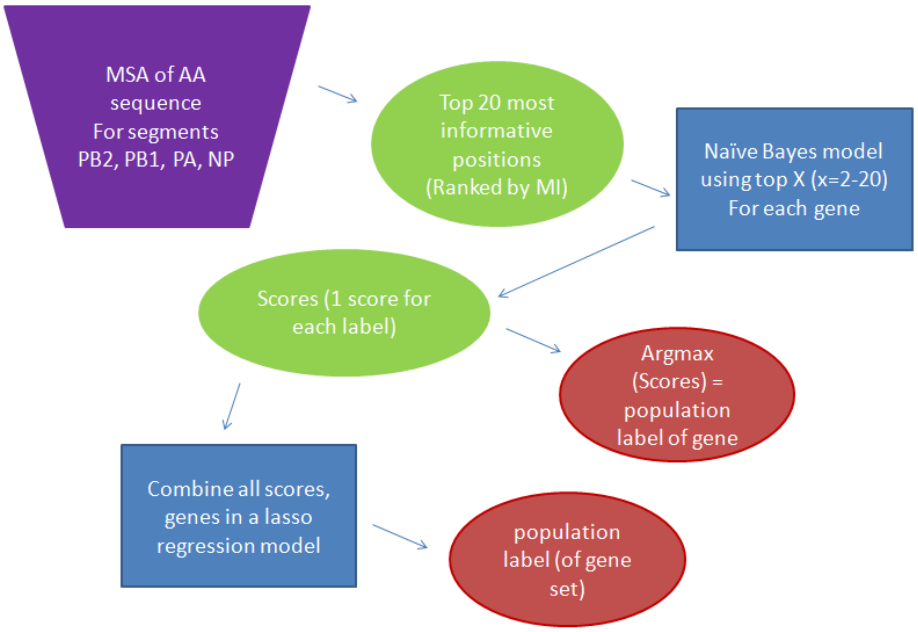

Figure 3.9 A flow diagram of the process outlined above. This differs from the diagram in Figure 3.4 in that we are not using any information about the sequences FFE. In this example, the purple trapezoid represents the input, green ovals intermediate output, blue rectangles the model, and red ovals the final output. MSA: multiple sequence alignment; AA sequence: amino acid sequence; MI: mutual information.

To further evaluate the models, in addition to the 10-fold cross validation, we also looked at each time interval model's ability to classify its own polymerase gene set and the human gene sets of the other time intervals. This data is presented in appendix B. From this data, it is evident that, when the binomial data set is severely unbalanced between human and non-human sequences, as is the case on the earliest time intervals, the classifier does not function well at all.

In looking at the cross-classification power, both of the multinomial weighted models both (the one that took into account the FFE and the one that did not) performed well- the self-classification efficiency is almost 100% in all models, and there is a great deal of overlap in the models' ability to correctly classify sequences in other time intervals. Multiple time interval models have the ability to correctly classify both virtually 100% of their own segment sets, but also almost 100% of most of the other time intervals segment sets as well. The major exception

to that is that the time intervals 1918-1939 and 2001-2008, which are not good classifiers of each others' segment sets being thought to be human seasonal. This may be reflective of the lineage problem detailed previously. (The tables showing this data are in **Appendix A**, and summarized in **Table 3.3**.)

Table 3.9 Maximum classification efficiency of the Regression classifier on the top 20 features of each segment; in this case, we are looking at the ability of each time interval to correctly classify both itself, and the other intervals' sequences. Self-classification efficiency is summed as one number (i.e., the average accuracy of the self classifications across all time intervals). The classifications across the other time intervals are separated according to the interval that the classifier was trained on, as the classification sensitivity varies widely. In the table, BI stands for the binomial un-weighted models and MULTI for the multinomial weighted models. MI stands for the model just using the 16 input vector (just the scores from part 2), MI+FFE for the 32-input vector (scores from part 3, and the FFE scores).

Regression Model	Interval Classifying Self	Interval Classifying Other Intervals' Sequences						
		1918 - 1939	1940 – 1957	1958 - 1968	1969 - 1976	1977 – 1990	1991 – 2000	2001 - 2008
Bi MI+FFE	95.5	0	0	0	64.2	99.2	99.2	99.2
Multi MI_FFE	99.7	79.6	99.5	99.3	98.3	99.6	99.2	98.9
Bi MI	95.5	0	0	0	65.6	99.6	99.2	99.2
Multi MI	99.7	99.5	98.8	99.4	99.4	99.6	99.2	99.2

DISCUSSION

With a multinomial model using the scores from part 3 and the folding energy distributions, we can demonstrate that we may be able correctly classify the polymerase and NP gene sets of the pandemic 2009 H1N1 as similar to those which have circulated in the human population. This is an important result on two fronts- first it demonstrates that folding energy of the polymerase genes is an important feature by which the virus adapts, and second that we can potentially use the properties of all the polymerase genes to identify viruses that are more likely

to replicate in the human population. The first may point to the importance of the folding energy in the viruses' adaptation to the human population. The second may perhaps indicate the need to evaluate the entire virus for its ability to propagate within the entire population, rather than evaluating individual gene segments.

An interesting result from the multinomial weighted model, using both FFE and amino acid scores, is that, while the model can classify the human sequences very well (including the pandemic 2009 H1N1 virus), it struggles a bit on correctly classifying avian, swine, and human bird sequences. The swine influenza sequences overlap with the other three host categories (human bird sequences to a lesser extent); human bird appears split between its own group and the avian group; and avian sequences appear primarily split between avian and human bird, and to a lesser extent swine and human seasonal. This may reflect the ability of the avian and swine viruses to perhaps infect a wider range of host species; and that human bird is not a self-propagating group. Alternatively, this may reflect sequence features that are present in the other segments of the virus, such as segment 4 and 6 (coding for HA or NA genes, respectively).

There doesn't appear to be another study that attempts to classify the polymerase genes (and NP) of the virus with regards to potential hosts; most work is done on determining when an epidemic is occurring or modeling the course of one. There is one paper, Chen *et al*. 2009, that looks into developing a genomic signature of the human viruses by looking at the differences between human and avian and human and swine viruses' by entropy calculations. In this paper, however, they succeed only to a limited extent in that they succeed in finding only 8 out of 10 of the essential amino acid positions in the polymerases (PB2, PB1, PA) that are known to differentiate human and avian influenza strains. [58] A critical point, though, is that they do not attempt to predict when/if a pandemic may occur.

This work is an attempt to partially fill this hole, by identifying sets of PB2, PB1, PA and NP that can work together to jump into and propagate within the human population.

MATERIALS

The data was periodically downloaded from the NIAID IRD through the website www.fludb.org. [76] Only sequences whose CDS length was 760, 758, 717 and 499 amino acids in length, respectively, were used in this study. The genomes were grouped using the virus names: if there were one and only one sequence for each segment was a virus considered complete and used in this portion of the study. Because of this, there were fewer sequences used in this portion of the study.

Table 3.10 The numbers of segments/viruses used.

	Avian	Human Seasonal	Human Bird	Swine	Pandemic H1N1
Segment 1	1039	1174	84	180	121
Segment 2	984	1645	64	184	119
Segment 3	1108	1591	86	181	134
Segment 5	1083	1395	79	195	122
Virus	492	447	26	90	75

3.3 SUPPLEMENTARY MATERIALS

The Genbank identification numbers used in this study are listed in the appendix B.

4.0 THE EFFECT OF FFE ON OTHER SSRNA VIRUSES

My role in the below work was to perform the computational and genome analysis.

We hypothesized that the relationship of RNA folding energy to host specificity was not restricted to the influenza A virus- this could be a property of RNA synthesis and structure. If this was the case, then we should be able to find other ssRNA viruses that shared this property. Therefore, we analyzed the genome of the West Nile Virus (WNV), a positive ssRNA virus of the family *Flaviviridae*. [145] This work on analyzing the folding free energy 0f WNV was presented at an NIH NIBIB Training Grantees Meeting, June 2010 as a poster and talk titled "*Potential Role of RNA Folding in West Nile Virus"*.

This virus first entered North America in New York City in 1999 [146], resulting in substantial die-offs of the wild bird population, primarily American crows. [147] The WNV strain circulating at this time, NY99, has been shown to have high virulence in the American crows.[145, 146] The mortality of the WNV has been shown to decreasing in over time [148], with the minimum abundance of the crow populations occurring in 2004-2005.[146]

When we map the folding energy of the US isolated WNV strains from 1999, we see an increase in folding energy of the virus (**Figure 4.1**), which corresponds to the reported reduction in lethality in American crows.

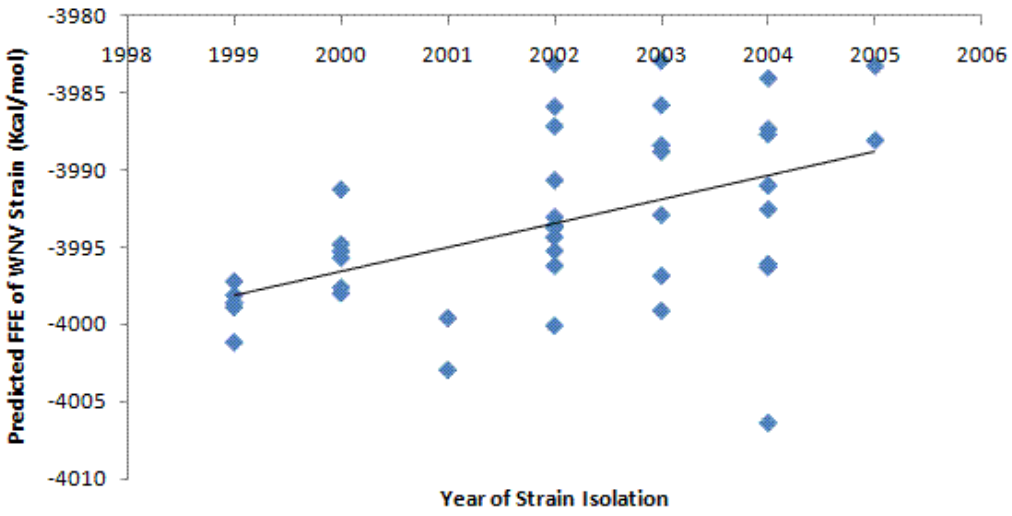

Figure 4.1. Predicted FFE of WNV across US from 1999 to 2005

Interestingly, a similar strain with only 11 amino acid differences to the NY99 strain, KEN3829, has low mortality in American crows. When the replication profiles of these two viruses was compared and analyzed by RM Kinney et al in 2006 [145], it was seen that the KEN3829 had marked decrease in viral titers when compared to NY99 at temperatures above 43^{o}C. Furthermore, they noted that the replication of viral RNA was compromised in KEN3829 at the higher temperatures. In addition, the body temperature of the American crow has been reported to be 40.2 to 41.4^{o}C, with temperatures 43-45^{o}C representing high temperatures. [145]

When we compare ability of these two strains to replicate at high temperatures with the folding free energy of the RNA, we see that the NY99 strain has a lower free energy (-3998.91 kcal/mol), than KEN3829 (-3720.58 kcal/mol), which may explain the ability of NY99 to grow more efficiently at high temperatures when compared to KEN3829, which failed to effectively replicate its RNA genome at high temperatures.[145] We hypothesize that the large difference in folding energy (278 kcal/mol) may be enough to increase the openness of the RNA structure up

to greater access to cellular degradation mechanisms at high temperature such that the replication of the virus is impeded relative to lower temperatures.

These results explain the difference in virulence of these strains- the inability of the KEN32829 virus to replicate effectively at high temperatures allows the birds which show increased temperatures as a result of the infection to survive, while this temperature fails to sufficiently impede NY99 replication.

Both of these pieces of evidence taken together- the evolution of the WNV virus in the US to higher folding energies while the virus reduces lethality and the correlation between RNA folding energy and genome replication at high temperatures- indicate that RNA folding energy may be a factor in the evolution of the WNV virus as well as the influenza virus.

While FFE appears to play a role in influenza A, we sought to address the question of whether this property is applicable to negative ssRNA viruses. To do this, we folded the polymerase genes of the negative ssRNA viruses in the GenBank database. These viruses were then grouped into two groups according to their ability to cause human disease, as was outlined in the study by Lloyd-Smith *et al.* in 2009.[149] We grouped those negative ssRNA viruses which can cause spillover disease only in human together with those viruses which can cause spillover and stuttering chains of human infection. We then compared this group with the combined group of viruses are human diseases and those which have the ability to cause outbreaks in one group. (**Table 4.1**). The differences in the means of the two populations was significant, with the human disease and outbreak causing viruses having a higher FFE on average -0.28 kcal mol^{-1} $base^{-1}$ versus -0.30 kcal mol^{-1} $base^{-1}$ in the spillover only group (p-value = 10^{-5}, two sample t-test assuming unequal variances)(**Figure 4.2**).

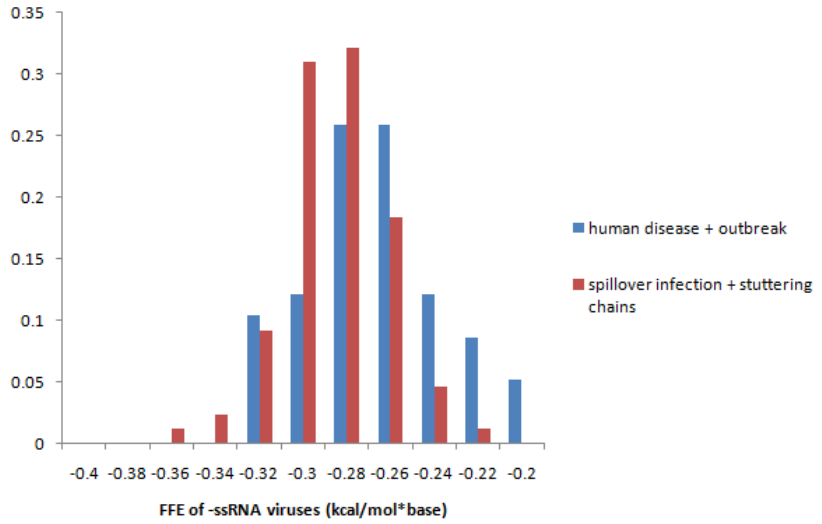

Figure 4.2. Predicted FFE negative ssRNA viruses, sorted by ability to cause sustainable disease in the human population.

While not definitive by any account, this may be indicative of a trend that FFE of the genome is an important host adaption trait for ssRNA viruses.

Table 4.1 List of negative ssRNA viruses used for **Figure 4.2**

Virus	**GenBank Genome**	**Group**	**Ref**
Australian Bat Lyssavirus (ABL)	NC_003243	spillover/stuttering	[150]
Andes Virus (ANDV)	NC_003468	spillover/stuttering	[149, 151]
Borna Disease Virus (BDV)	NC_001607	spillover/stuttering	[152]
Bunyamwera Virus (BUNV)	NC_001925	spillover/stuttering	[153]
Crimean-Congo hemorrhagic fever (CCFHV)	NC_005301	spillover/stuttering	[154]
Chapare Virus (CV)	NC_010563	spillover/stuttering	[155, 156]
Dobrava-Belgrade Virus (DOBV)	NC_005235	spillover/stuttering	[157, 158]
European Bat Lyssavirus 1(EBLV1)	NC_009527	spillover/stuttering	[159]
European Bat Lyssavirus 2(EBLV2)	NC_009528	spillover/stuttering	[159]
Flexal Virus (FELV)	NC_010759	spillover/stuttering	[160]

Guanarito Virus (GTOV)	NC_005082	spillover/stuttering	[160]
Hantaan Virus (HTNV)	NC_005222	spillover/stuttering	[158]
Hendra Virus (HV)	NC_001906	spillover/stuttering	[149]
Hantavirus (HVZ)	NC_006435	spillover/stuttering	[158]
Ippy Virus (IppV)	NC_007906	spillover/stuttering	[160, 161]
Junin Virus (JV)	NC_005080	spillover/stuttering	[160]
La Crosse Virus (LACV)	NC_004108	spillover/stuttering	[51]
Lassa Virus (LASV)	NC_004297	spillover/stuttering	[149]
Lymphocytic Choriomeningitis Virus (LCMV)	NC_004291	spillover/stuttering	[160]
Lujo Virus	NC_012777	spillover/stuttering	[155]
Machupo Viurs (MACV)	NC_005079	spillover/stuttering	[149, 160]
Mobala Virus (MblV)	NC_007904	spillover/stuttering	[160, 161]
Menangle Virus (MENV)	NC_007620	spillover/stuttering	[162]
Mokola Virus (MOKV)	NC_006429	spillover/stuttering	[160, 161]
Nipah Virus (NV)	NC_002728	spillover/stuttering	[149]
Puumala Virus (PUUV)	NC_005224	spillover/stuttering	[149, 158]
Rabies Virus (RABV)	NC_001542	spillover/stuttering	[149]
Rift Valley Fever Virus (RVFV)	NC_014397	spillover/stuttering	[51]
Sabia Virus (SabV)	NC_006313	spillover/stuttering	[160]
Seoul Virus (SEOV)	NC_005238	spillover/stuttering	[158]
Toscana Virus (SFNV)	NC_006319	spillover/stuttering	[51, 163]
Sin Nombre Virus (SNV)	NC_005217	spillover/stuttering	[149, 158]
Tacaribe Virus (TCRV)	NC_004292	spillover/stuttering	[160]
Thottapalayam Virus (TPMV)	NC_010707	spillover/stuttering	[51]
Tula Virus (TULV)	NC_005226	spillover/stuttering	[158]
Whitewater Arroyo Virus (WWAV)	NC_010703	spillover/stuttering	[160]
H5N1 Influenza		spillover/stuttering	[149]

Bundibugyo Ebolavirus (BDBV)	NC_014373	disease/outbreak	[149]
Cote d'Iviore Ebolavirus (CIEBOV)	NC_014372	disease/outbreak	[149]
Influenza B Virus (FluBV)	NC_002204, NC_002205, NC_002206, NC_002208	disease/outbreak	[51]
Influenza C Virus (FluCV)	NC_006307, NC_006308, NC_006309, NC_006311	disease/outbreak	[51]
Human Metapneumovirus (HMPV)	NC_004148	disease/outbreak	[51]
Human Parainfluenza Virus 1 (HPV)	NC_003461	disease/outbreak	[51]
Human Parainfluenza Virus 2 (HPIV2)	NC_003443	disease/outbreak	[51]
Human Parainfluenza Virus 3 (HPIV3)	NC_001796	disease/outbreak	[51]
Human Respiratory Syncytial Virus (HRSV)	NC_001781, NC_001803	disease/outbreak	[51]
Lake Victoria Marburgvirus (MARV)	NC_001608	disease/outbreak	[149]
Measles Virus (MeV)	NC_001498	disease/outbreak	[51]
Mumps Virus (MuV)	NC_002200	disease/outbreak	[30]
Oropouche Virus (OROV)	NC_005776	disease/outbreak	[51]
Sudan Ebolavirus (SEV)	NC_006432	disease/outbreak	[149]
Parainfluenza Virus 5 (SPV5)	NC_006430	disease/outbreak	[164]
Zaire Ebolavirus (ZEBO)	NC_002549	disease/outbreak	[149]
H3N2 Influenza human A		disease/outbreak	[51]
H1N1 1918 Influenza A		disease/outbreak	[51]

5.0 THE EFFECT OF FFE ON BACTERIAL EVOLUTION AND ADAPTATION

In the following work, we addressed how folding energy may play a role in the effect that the physical environment has on prokaryotic RNA structure and gene expression, has been submitted for publication as R Brower-Sinning and PV Benos as “The effect of physical environmental properties on prokaryotic RNA structure and gene expression”, and was presented in part at the American Society for Microbiology General Meeting in May 2011 as “Evolution of Prokaryotic RNA structure is influenced by physical environment properties”.

My role in the below study was to conceptualize and design the study, and to perform the computational analyses and to analyze the data.

5.1 ABSTRACT

Background. The idea that the physical environment plays a role in the evolution, structure, and function of prokaryotic genomes is intuitive and exciting. There is a well established relationship between growth temperature and the GC content of tRNA and rRNA, but surprisingly not of the mRNA [79, 81, 165]. Recent studies have revealed that the folding sequence around the ribosome binding site is associated with translated protein levels [20], and that translation efficiency may be influenced by GC content [21, 22]. However, so far, no

systematic study has been performed to associate the environmental factors such as temperature and pressure with the evolution of the microbes at the RNA level.

Methods. We present a computational study, in which, we investigate the ways the physical properties of the environment an organism lives and grows in may affect its evolution at the genomic level. We analyze various RNA types in gram-positive and gram-negative bacteria that live and thrive in very diverse environments and we re-examine the association between RNA folding free energy and the environmental conditions, such as optimal growth temperature and pressure.

Results. We show that tRNA folding is directly related to the environmental conditions. We confirm that mRNA, in general, does not clearly correlate with environmental conditions, but the mRNA half-life may be affected. Additionally, we show clear differences in the organization of mRNA folding based on temperature and pressure. Finally, we show that the folding around the translation start codon is the most important determinant of the translation efficiency, but folding on the whole open reading frame plays also a significant role in it.

Conclusions. Our data are consistent with the hypothesis that the RNA folding is under selection imposed by the environmental conditions. A preference for an mRNA structure, as influenced by growth temperature and other physical conditions, could explain at least partly the innate GC and AT preferences seen bacteria. This is an important but previously overlooked factor in the genomic evolution of prokaryotes.

5.2 ARTICLE

INTRODUCTION

The idea that an organism's function and performance is affected by temperature is both intuitive and well documented. Indeed, biologists have linked temperature with a wide range of biological phenomena, such as growth, survival and reproduction as well as population density and species diversity. At the molecular level, temperature is proven to be an essential component in protein reaction kinetics [166] and affect gene expression (heat-shock and cold-shock genes). However, the role the physical environmental properties play as evolutionary forces that shape an organism's mRNA has been largely overlooked, with few exceptions. Bacteria inhabit diverse environments, everything from volcanoes and hot springs to the human body, and they possess no mechanism to regulate their internal temperature, pressure, and other physical properties. As such they need to adapt to the external environmental conditions for optimal growth. Guanine (G) pairs with cytosine (C) through three hydrogen bonds, and adenine (A) with uracil (U) with two. Thus, it can be hypothesized that if folding is important for the RNA function, then organisms living at higher temperatures will tend to have RNAs with higher GC content [167].

RNA structure is very important for tRNAs and rRNAs [79]. Indeed, previous work has detailed how the GC content of tRNAs and rRNAs strongly correlates with growth temperature, but has also showed that the genomic or mRNA GC content does not [79-81]. This lack of correlation between mRNA GC content and temperature is somewhat surprising, as it was

previously believed that melting temperatures of nucleic acids should affect genome evolution globally [81].

In a recent study, Kudla *et al*. reported that in *Escherichia coli*, the folding of the region around the ribosome binding sequence (RBS) was associated with protein levels by – presumably– affecting the translation rate [20]. Additionally, Voges *et al.* has found that translation efficiency for mRNA in *E. coli* was dependent on both base pairing probability and GC content of the sequence directly downstream of the start codon, indicating that mRNA secondary structure in this region could hamper translation [21]. Similarly, in 2010 Tuller *et al.* showed that, in general, mRNA folding may function to slow down ribosomes, impeding translation [22].

While not correlating with growth temperature, genomic GC content has been shown to play a strong role in the codon usage across different species [27, 82, 83]. There are several mutational biases that may affect the patterns seen in both genomic and mRNA GC content. There can be natural or selective pressure on the innate bias in point mutations [27, 82, 84, 85]; there can be selective pressure to prefer certain synonymous mutations over others [21, 27, 82, 83, 86-88]. Even among bacteria, there is a split in the selective pressure exerted on the patterns of synonymous mutations: some species show a strong selective pressure, while others do not [86]. In Sharp *el al.* 2004, it was noted that the species exhibiting strong selective pressure on their codon bias also had faster growth rates and shorter generation times [86]. In Kudla *et al.*, a relation between the strength of codon bias of the GFP construct and growth rate was also observed, with higher codon adaption of the exogenous gene correlated with a faster growth rate ($R=0.54$, p-value $<9*10^{-13}$) [20]. The rate of synonymous substitution and degree of codon bias

of a gene may be related, and reflect the genes' translational landscape; and this may differ between genes within a given genome [87].

The idea that codon bias may influence or be influenced by both the properties of mRNA or translation efficiency is a powerful one. It has been hypothesized that this codon bias may play an important role by increasing either translational efficiency or accuracy or both [27, 168-172]. Several groups have looked for a link, although with limited success. Kudla *et al*. found little to no correlation between protein abundance and codon bias [20]. However, they do state that local sequence patterns do influence gene expression, again suggesting the importance of mRNA structure in regulating expression levels [20]. Additionally, Voges *et al.* also reported little or no correlation [21]. In a previous study, we showed that the polymerase genes of the human influenza A virus fold at significantly higher energy levels (less tight folding), than the avian isolates [141]. We also found that in the human isolates, the predicted folding free energy (FFE) evolves through time towards higher values since the 1918 pandemic, which is presumed to be a bird-to-human transmission [58, 141]. We hypothesized that this difference in FFE is due to the difference in the corresponding host's body temperature (humans: 37^{o}C; birds: 39-42^{o}C, depending on the type of bird). By infecting temperature adapted MDCK cells with different strains of the influenza A virus, we showed that there is a temperature dependent association between the RNA folding stringency and viral propagation [141].

In this paper, we seek to further investigate whether the physical properties of the environment an organism lives and grows in affect its genomic evolution. For that reason, we analyzed different RNA types in various bacteria species and examined the association between RNA FFE and the environmental conditions, such as optimal growth temperature and pressure.

RESULTS

tRNA folding free energy (FFE) is strongly associated with growth temperature in prokaryotes. It is known that the mRNA GC content, and by extension, the FFE, in prokaryotes does not correlate very well with the preferred growth temperature of the specific organism [81]. This is also depicted in **Figure 5.9**, where, although the thermophilic bacteria tend to have generally lower FFEs, there is a large amount of overlap between their distributions. In addition, the two Thermotoga species (*T. neapolitana*, *T. maritima*) seem to break the rule, since they both have unusually high FFEs (**Figure 5.9**). On the other hand, the tRNA GC content has been found to be correlated to the preferred growth temperature [79] (also see **Figure 5.1A**). As expected, the tRNA FFEs in prokaryotes (bacteria and archaea) correlate very well with the optimal growth temperature (R=0.91; **Figure 5.1B**). This difference raises the question: what causes FFE to be associated with growth temperature in tRNA genes, but not in mRNA? One reasonable hypothesis is that RNA structure plays a functional role in the former, but not so much in the latter. Indeed, in order to function properly, tRNAs need to fold to a particular structure, similarly to many of the proteins in the cell. However, in order to maintain the structure at high temperatures, tRNAs need to have lower FFEs.

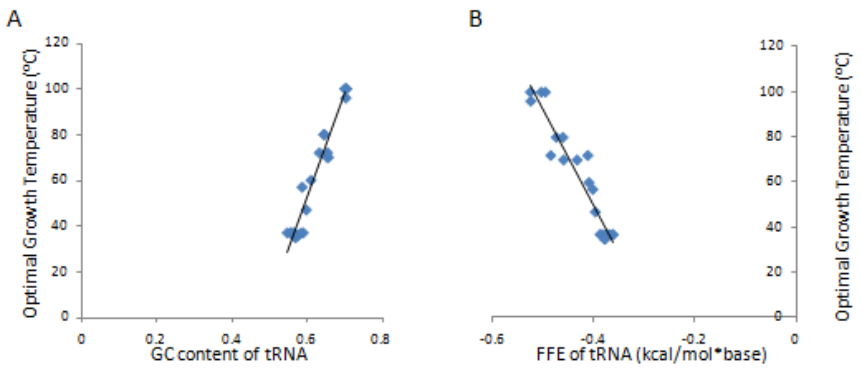

Figure 5.1 A. Correlation of tRNA GC with optimal growth temperature. **B.** Correlation of tRNA GC with optimal growth temperature. (To account for varying length of the RNA species, all values have been normalized according to length, and as such have units of kcal*mol^{-1}*base^{-1}.)

Pair-wise comparisons in different bacteria reveal an extensive agreement between FFE and growth temperature. The correlation coefficient is a very broad measure of association between two variables. Furthermore, the optimal growth temperature does not need to be *linearly* correlated with the FFE across all evolutionary distances and all temperatures. In order to investigate the issue in more detail, we performed a series of pair-wise statistical tests. In order to better explain the methodology of analysis we followed, we present two examples: one of distantly related species and one of closely related species.

Comparing mRNA FFEs in two distant species. Two gram$^-$ bacteria that live in drastically different temperatures were examined: *E. coli* (37^{o}C) and *Thermus aquaticus* (70^{o}C) [77, 173]. Their FFE distributions are characteristically different (**Figure 5.2**) and their average FFE values are -0.241 for *E. coli* and -0.508 for *T. aquaticus*, respectively. In this case a *t*-test shows these values to be statistically significantly different (*p*-value << 0.001).

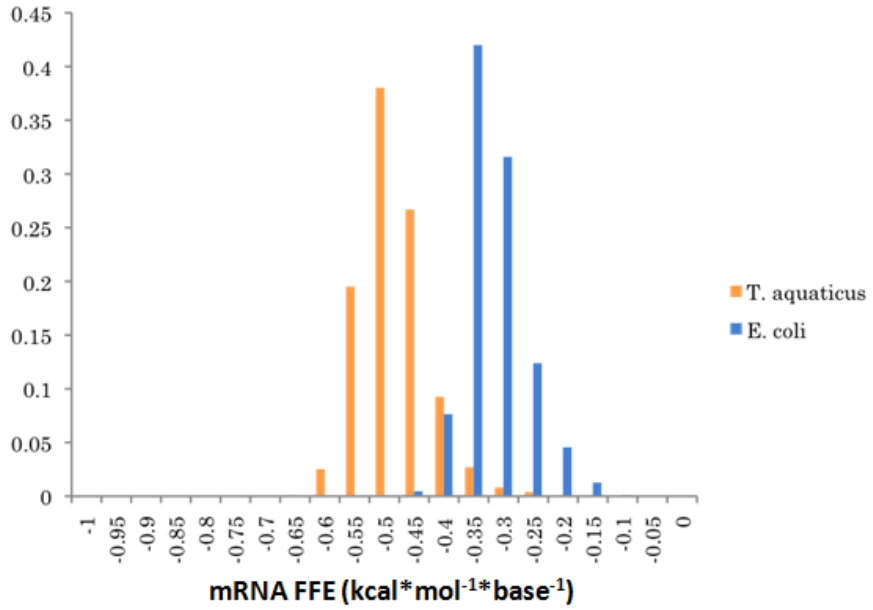

Figure 5.2 mRNA folding energy probability distributions for *T. aquaticus* and *E. coli*. To account for varying length of the RNA species, all values have been normalized according to length.

Comparing mRNA FFEs in two closely related species. *Mycobacterium tuberculosis* and *Mycobacterium leprae* are two closely related human pathogens, but they are agents of vastly differnet diseases (tuberculosis and leprosy, respectively). They have also different optimal growth temperatures (37^{o}C and 33^{o}C, respectively), and *M. leprae* is not viable at 37^{o}C [174-176]. As expected, the two mRNA FFE distributions show a large degree of overlap (**Figure 5.3A**), although *M. tuberculosis* has lower average FFE per base than *M. lepreae* (FFE_{Mt} = -0.462, and FFE_{Ml} = -0.396, respectively). The direction of the difference is consistent with the hypothesis that increased optimal growth temperature may lead to lower mRNA FFEs, but is it significant? In such closely related species, a *t*-test or other statistical tests are not expected to identify significant differences, indicating that the two distributions are overall similar. However, when we compared the orthologous genes in these two species, we found that in 97% of the cases (1276 of 1313 genes), the *M. tuberculosis* gene had a lower FFE than the *M. leprae*

orthologue (**Figure 5.3B**). The difference between the means of the orthologous pairs is statistically significant (paired *t*-test *p*-value << 0.01), also agreeing with the above hypothesis.

We examined more closely the few orthologous genes where the *M. leprae* has the lower FFE value. Of the 37 such genes, 2 have possible ortholog reporting errors, or have multiple *M. tuberculosis* high-score hits. At least 6 of the remaining 35 are highly similar to other punitative genes in the *M. leprae* genome, and another 24 are annotated as "probable", "punitive", or "hypothetical" genes. Finally, the remaining 5 are ribosomal proteins.

One characteristic of the *M. leprae* genome is that contains a relatively large number of pseudogenes (n=1149). We found that the annotated pseudogenes typically have a higher folding energy than the regular genes (**Figure 5.3C**). This may indicate that with a lack of evolutionary pressure to maintain a certain RNA structure the FFE drifts to higher values. This may be indicative of the general tendency of the genome as a whole for lower GC content.

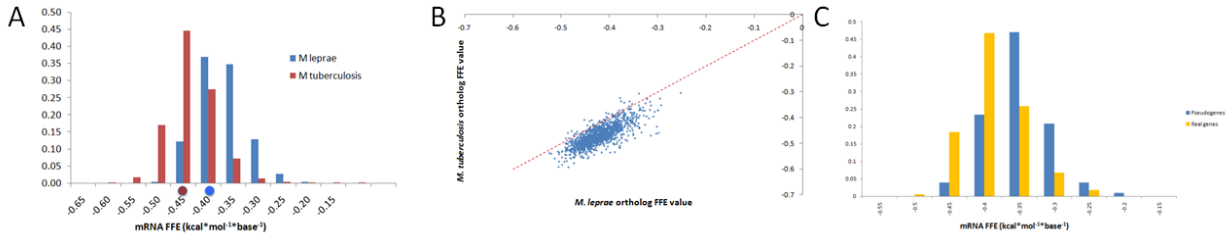

Figure 5.3A. mRNA folding energy probability distributions for *M. tuberculosis* and *M. leprae.* To account for varying length of the RNA species, all values have been normalized according to length. The red and blue dots indicate the average FFE in the two species. **B.** Comparison of the FFE (per base) of *M. leprae* and *M. tuberculosis* orthologous gene pairs. The line is the diagonal, where the two orthologues have equal folding energy. The units in both axes are kcal*mol^{-1}*base^{-1}. **C**. FFE probability distributions for *M. leprae* pseudogenes and non-pseudogenes. To account for varying length of the RNA species, all values have been normalized according to length.

Global pairwise FFE comparison. The above examples indicate that the RNA folding (for both coding and non-coding RNA) may be important for function and expression. To investigate further we extended the analysis to 12 bacterial species that live in a wide range of

temperatures. When we restricted the comparison to orthologous genes for each pair of species, we see that all mesophiles are significantly different from the thermophiles (**Figure 5.4A**). There also appears to be differences within the mesophile group, between the gut-specific bacteria *Salmenolla enterica*, *E. coli* and *Lactobacilus casei*, and the primarily lung infecting bacteria *Klebsiella pneumonia*, *Haemophilus influenzae*, *Streptococcus pyogenes* and the two *Staphylococcus aureus* strains. All comparisons between species across these two groups show statistically significant differences, whereas comparisons within groups are generally not statistically different (**Figure 5.4A**). We also used the *t*-test *p*-values from the pairwise comparisons as distances in a clustogram construction (**Figure 5.4B**). Interestingly, the first statistically significant split occurs between mesophiles and theormophiles. Within the mesophilic group, we see that the one bacterium with a different optimal growth temperature (*H. influenza*) is the first split; and the two gut-specific gram$^-$ bacteria, *E. coli* and *S. enterica*, the second (although this second is not statistical significant based on the bootstrap value). Similar to the global analysis results in the two Mycobacteria above, we found no clear distinction of mesophiles and thermophiles based on comparisons of the mRNA FFE for all genes (**Figure 5.10**).

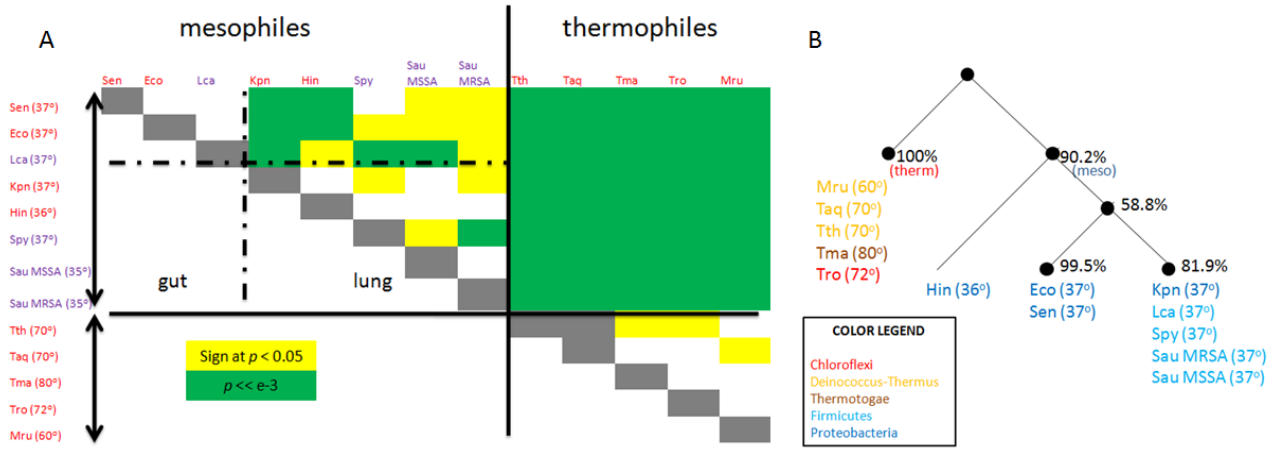

Figure 5.4 A. Comparisons of orthologous genes of different bacterial species. For species abbreviations, please see *Materials and Methods*. Red: gram$^-$; purple: gram$^+$. Significance levels are $p << 10^{-3}$ (green), and $p < 0.05$ (yellow). **B.** The clustering of *p*-values from the paired *t*-test for orthologous genes. The phyla of the bacteria are color-coded. The tree was constructed by Bootstrapping the *p*-values from **Figure 4A**.

Cold-shock genes have higher FFEs in both *E. coli* and *T. aquaticus*. Based on the above data, we hypothesized that the mRNA structure (here summarized by its overall FFE value) is an important component of the mRNA function and it is thus evolutionary constrained. If this is true, then one might expect that the mRNA of the cold-shock response genes will tend to have generally higher than average FFEs. In order to assess the validity of this prediction, we analyzed the FFE of the cold-shock genes annotated as such in the GenBank database. Consistent with our hypothesis, we found that the annotated cold-shock genes in both *T. aquaticus* and *E. coli* have higher than average FFE values (**Table 5.1**; *p*-value for *E. coli* <0.001, binominal test; for *T. aquaticus* there are not enough examples).

Table 5.1 FFE of the annotated cold-shock genes of *E. coli* and *T. aquaticus*. In all cases, the cold-shock genes have FFEs in the higher end of the distribution spectrum and higher than the average FFE values (avg FFE is -0.341 for *E. coli* and -0.508 for *T. aquaticus*).

Organism	Gene /Locus Tag	FFE
E. coli	CspA/3659	-0.295
	CspE/0654	-0.278
	CspD/0895	-0.302
	CspH/0986	-0.279
	CspG/0987	-0.274
	CspC/1766	-0.292
T. aquaticus	4827	-0.435
	4615	-0.398

***E. coli* mRNA half-life corresponds well with the FFE**. The previous result indicates that although there is no global correlation between mRNA FFE and temperature, such correlation exists for certain classes of genes. The question now becomes what the biological mechanism is that drives this correlation. Unlike tRNA, the mRNA structure does not have an obvious functional role. However, mRNA needs to interact with a number of RNA-binding proteins for participating in various biological processes, including translation and degradation. Previous studies have shown that in *E. coli* the folding of the 5'-UTR (but not the whole mRNA) affects the translation rate [20, 177]. Another group showed how the initial translation sequence just downstream of the start codon could influence translation efficiency [21]. Similarly, certain hairpin structures are protective to the RNA molecules [178]. Therefore, may be the variability we observed in the mRNA FFEs previously is related to different degradation rates of these mRNAs. In order to examine whether the mRNA structure affects somehow its half-life, we plotted the FFE range of *E. coli* K12 MG1665 genes with different half-lives using the data presented in Bernstein, *et al.* [26]. We know that all mRNAs, regardless of their half-life time,

have FFEs that are normally distributed around the same mean (*t*-test, p< 0.05; **Figure5.2**). However, in general, the more stable mRNAs have shorter FFE ranges. This is true for *E. coli* growing on M9 (**Figure 5.5A**) and on LB media (**Figure 5.5B**). When we compared the FFEs of the genes with the top 10% *vs.* the lower 10% in terms of half-life, we found them to be significantly different (*p*-values of 10^{-4} and 10^{-14} for *E. coli* growing on M9 and LB media, respectively; two-sample Kolmogorov-Smirnov test). We note that these are very different conditions and there is little correlation of the mRNA half-life between them (R= 0.28). Furthermore, this observation does not hold if we limit the analysis to the 5'-UTR only (**Figure 5.11**).

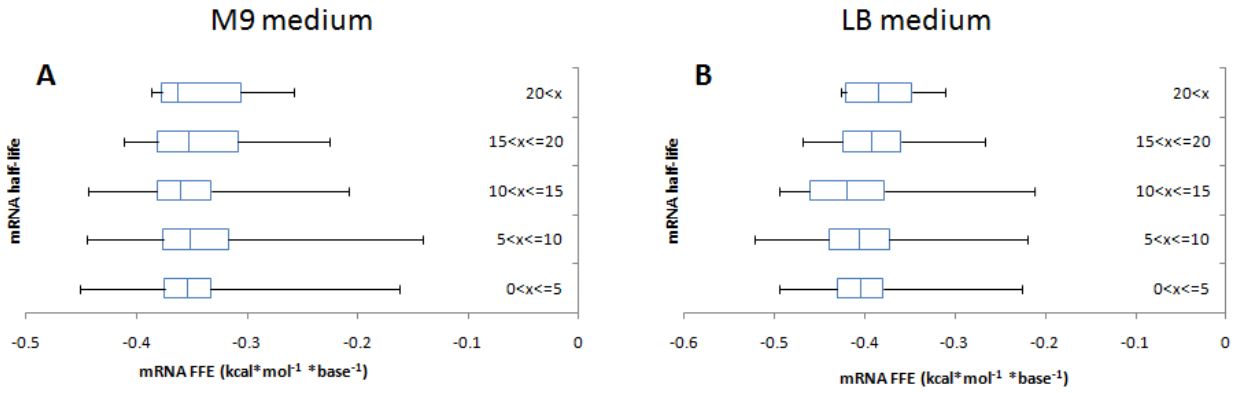

Figure 5.5 Relation between mRNA FFE and half-life in *E. coli* K12 MG1665 grown in **A.** M9 media, and **B.** LB media. Half-life data is from Bernstein, *et al.* [26]

***E. coli* mRNA FFE may relate to correlation between codon usage and tRNA abundance.** Given that codon usage is often associated with tRNA abundance in single cell organisms, and especially in highly expressed genes [179, 180], we were interested to see whether FFE plays a role in this correlation. We tested this by dividing the mRNA FFE distribution into seven bins, based on the number of standard deviations a gene's FFE is away from the mean FFE in *E. coli*. For tRNA abundance in *E. coli*, we used the data from Varenne *et*

al. [181]. We found that the stronger correlation between tRNA abundance and codon usage is for those mRNAs that are within one standard deviation around the mean FFE; and the correlation is reduced progressively and symmetrically as we move more away from the mean (**Table 5.2**). In summary, we showed that long-lived mRNAs have FFEs closer to the center of the FFE distribution, and their codon usage closely matches the tRNA concentrations for those mRNAs with FFEs around the center of the FFE distribution.

Table 5.2 Correlation between codon usage and mRNA FFE in *E. coli*. Each row contains an FFE range as a function of a number of standard deviations (SD) from the mean. The correlation is between FFE and tRNA concentration. Data from Varenne *et al.* [181]. Star (*) indicates significance *p*-value < 10^{-5}.

SD from avg mRNA FFE	R
< -3	0.751
[-3, -2]	0.832*
[-2, -1]	0.890*
[-1, 1]	0.925*
[1. 2]	0.890*
[2, 3]	0.802*
> 3	0.835*

Predicting gene expression from mRNA sequence. Plotkin and colleagues sought to understand how RNA structure impacts gene expression [20]. They used an *E. coli* expression vector to create a library of GFP constructs with identical 5' UTRs and (GFP) amino acid sequences. By introducing synonymous mutations in the coding sequence of the GFP, they varied the mRNA folding. The measured GFP levels are indicative of the stability of the mRNA and the translation efficiency (combined). They analyzed the GFP levels *vs.* the mRNA folding and they found that the folding in the first 1/3 of the sequence (288 bases) is highly correlated with the GFP levels (self-validation R=0.6, *p*-value<10^{-15}), but the folding energy of the entire mRNA was not (R=0.16, *p*-value=0.051) [20]. While we have shown that mRNA structure may

be related to its half-life, one question remains: can we use mRNA sequence and structure features to predict protein output?

We reanalyzed the GFP protein expression and sequence data from the Kudla *et al.* [20] study to address this question. We performed a lasso regression on a variety of sequence and structural features, with 10-fold cross validation, as we describe in *Materials and Methods*. The model containing the first 8 features performs remarkably well in predicting the GFP protein levels (R=0.79, *p*-value = 0; see **Table 5.3** for details). Since, however, four of the features in this model are highly redundant (features #3, 5, 6, 8, **Table 5.3**) we also did the analysis eliminating three of them at a time. We found that even including only the first four, not substantially redundant features we can explain ~60% of the variance (R=0.774, **Table 5.3**), while adding feature #7 ("Bonding strength at 20 nt window around -25") can explain another 1% (R=0.779, **Table 5.3**). A scatterplot of the predicted *versus* the actual gene expression levels can be seen in **Figure 5.6**.

Table 5.3 . The features that best predict GFP expression levels. Data from Kudla *et al.* [20]. The top 8 features can explain ~63% of the variation, while a substantial amount of variation (60%) can be explained by only the top 4, less redundant, features. For description of the features, please see *Methods*.

Feature	Applied on (mRNA part)	$R_{feature}$ (*p*-val)	R_{model}
Avg FFE (Sfold)	First 1/3	0.544 (0.0)	0.544
Min FFE (RNAfold)	ORF	0.055 (0.53)	0.706
Bonding strength	±10bp around ATG	0.434 (0.0)	0.765
Codon-tRNA corr	ORF	0.245 (2e-3)	0.774
Bonding strength	100bp window around ATG	0.488 (0.0)	0.778
Bonding strength	±120bp around of ATG	0.290 (3e-4)	0.786
Bonding strength	20bp window around -25	0.379 (e-6)	0.790
Bonding strength	±15bp around ATG	0.257 (0.0013)	0.792

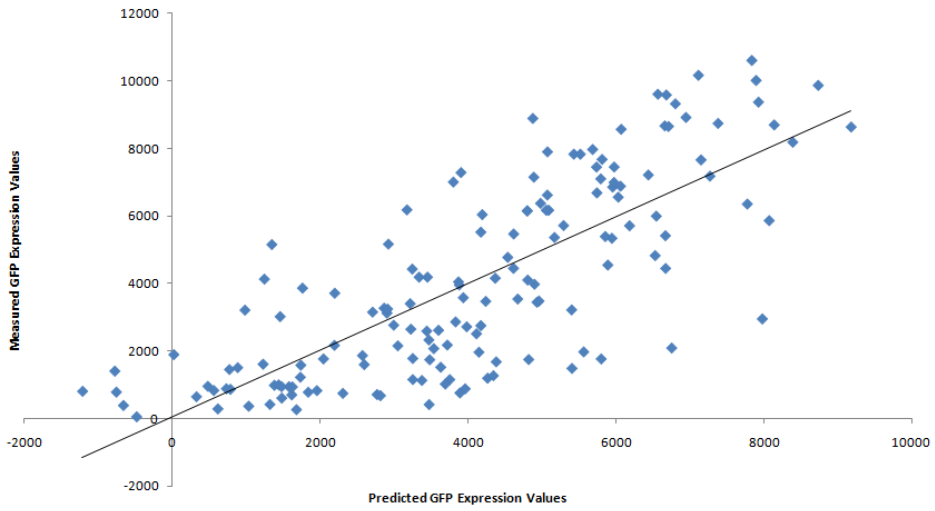

Figure 5.6 Predicting GFP expression using mRNA features. (GFP data is from Kudla et al., 2009 [20]).

Potential role of pressure on tRNA structure. By examining relations between optimal growth temperature and RNA structure, we have found evidence that support the idea that RNA structure is not only important for the function of tRNAs, but it is also related to aspects of mRNA half-life and –possibly– the efficiency of translation. However, temperature is not the only physical property that affects structure. Pressure can also have a similar effect. As has been shown in the literature, increasing pressure stabilizes the hydrogen bonds, similarly to a temperature decrease [182]. Since decreasing RNA FFE is correlated with increasing temperature, one should expect that increasing pressure should exert the same effect. When examining prokaryotes from various oceanic depths and temperatures, we find that neither temperature nor pressure alone correlate well with the tRNA FFE (**Figure 5.7A**). However, there is a strong correlation between tRNA FFE and the conjunction of temperature *and* pressure for a plot resistant to outliers (R=0.94 for all prokaryotes). What is also interesting in this plot is that we recover the effect that the folding energy is stabilized by lower pressure and higher temperature, as seen in the slope of the best-fit plane (**Figure 5.7B**).

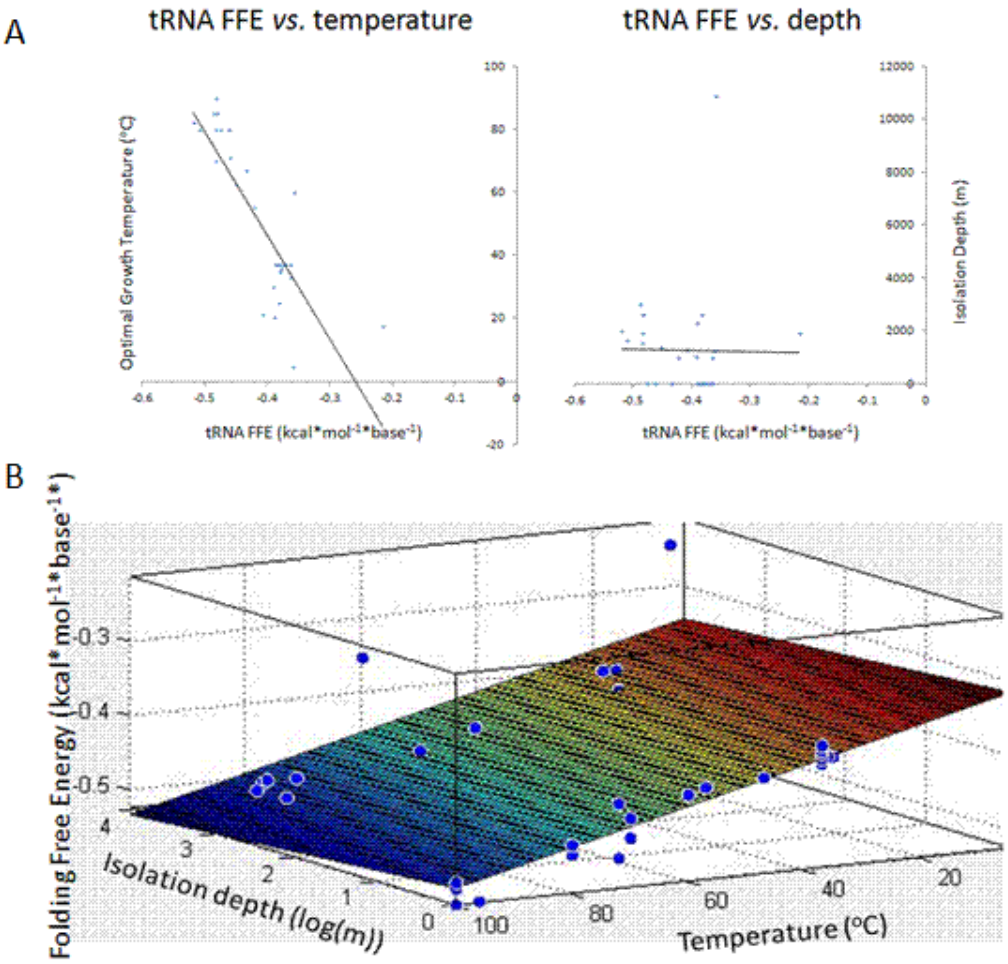

Figure 5.7 A. Scatterplots of tRNA FFE and optimal growth temperature (left, R=-0.83, *p-value*=0.0), or isolation depth (right, R=-0.012, *p-value*=0.94) for bacteria living at different temperature or pressure conditions. To account for varying length of the RNA species, all values have been normalized according to length. **B.** 3D scatterplot of tRNA FFE, temperature and pressure (R=-0.92 for the bacteria only, R=-0.94 for all prokaryotes, *p-value*=0.0).

Comparisons of marine prokaryotes show divisions based on pressure and temperature. We also compared the orthologous genes between the marine-isolated bacteria, in an analysis similar to that of **Figure 5.4**. As expected, we observe similarities between low-temp

moderate pressure prokaryotes and high-temp high-pressure prokaryotes, but not between high-temp, moderate-pressure and high-temp high-pressure organisms (**Figure 5.8**).

Figure 5.8 FFE comparison for orthologous genes from prokaryotes under conditions of varying temperature and pressure

DISCUSSION

In the past we have shown that the host's body temperature can influence the evolution of the polymerase genes of the influenza A virus [141]. The mechanistic explanation we proposed was that an "optimal" mRNA structure is important for rapid transcription and translation of these critical genes during the early stages of infection. We also postulated that mRNA structure may affect its half-life by altering its interactions with RNA-binding proteins. Prokaryotic organisms are more complex than viruses, and a difference in environmental temperature will affect many aspects of the cell's physiology, which will need to adapt to survive and successfully replicate [183]. Also, prokaryotic genomes evolve slower than RNA viruses. In this context, it is not surprising that mRNA folding does not generally correlate well with optimal growth temperature. On the other hand, tRNA FFE correlates strongly with optimal growth temperature,

since correct structure is fundamental to the function of this class of genes. This also shows how certain physical properties affect the evolution of nucleic acids in perhaps unexpected ways.

In this paper, we embarked on the task of investigating in more detail the association between the FFE of prokaryotic RNAs (tRNAs, mRNAs) and two environmental physical properties: optimal growth temperature and pressure. Central hypothesis is that mRNA folding should affect protein translation and RNA degradation. Previous studies have shown this to be partly true. Kudla, *et al*. showed that the FFE of the 5'-UTR (but not the whole mRNA) correlates with the translation efficiency [20]; while others have shown that the region downstream and general folding may also play a role in translation [21, 22].

We hypothesize that the degree of mRNA structure stability will play a role for both how accessible the molecule is for translation, but also in how the molecule is targeted for degradation. In other words, the structure of the mRNA molecule will play an important role in determining the half-life of the mRNA, and may influence the amount of protein produced by a given transcript. For *E. coli*, we found that mRNAs with longer half-life have a more narrow range of FFEs around the organisms average FFE, perhaps to balance the competing needs of translation and degradation. Additionally, these genes in the center of the distribution display a strong correlation of codon usage with the corresponding tRNA concentrations, potentially illustrating the connections between codon usage and mRNA structure and protein abundance. This might give an interesting and previously overlooked explanation for the origin of codon bias in prokaryotes. It is generally believed that codon bias emerges as a result of the abundance of tRNAs [179, 184-187]. But what creates differences in tRNA abundance in different organisms? Our theory predicts that the need for certain "tightness" in mRNA structure affects the GC content of these genes. This, in turn, selects certain codons; and the organism adapts by

providing more abundance for the tRNAs that correspond to these codons. Of course, at this point this is just a theory. We hope that follow-up experiments will be able to test this theory.

We also showed that for both *E. coli* and *T. aquaticus*, the annotated cold-shock genes fall in the higher-folding energy portion of their respective mRNA folding distributions. This is also consistent with the hypothesis that the RNA structure is biologically important, thus when the organism enters cold shock conditions, the activated genes need to have mRNAs with higher FFE values to fold properly in the colder environment. While it is evident that there is no clear linear relationship between the optimal growth temperature and the mRNA FFE, there is a clear difference between mesophilic and thermophilic bacteria, when orthologous gene pairs are compared.

Furthermore, we were able to find a model that explains GFP protein levels from sequence and structural features (data from Kudla *et al.* [20]). This indicates that the mRNA itself may be instrumental in determining protein levels. Combined with preservation of the distributions of the mRNA folding energy within the mesophilic and thermophilic bacteria, this may indicate that the relative protein expression is conserved.

Finally, we showed that for organisms living at various pressure and temperature environments neither of the two physical properties alone is correlated well with the FFE. However, the combination of the two is.

All the above observations are supporting the idea that the structure of the mRNA has a much more active role in regulating its half-life and its translation rate than previously thought. We hope that further experiments will provide more insights to this fascinating phenomenon.

MATERIALS AND METHODS

Data

Sequence data: prokaryotic genomic sequences were downloaded from GenBank. A complete list of organisms, accession numbers, and environmental properties and sources, are listed in **Table 4**.

Table 5.4 List of prokaryotic organisms used, genome number, growth temperature and isolation depth.

Organism name	Accession Number	Version	Temp (^{o}C) / Isolation Depth (m)	Sources
Aciduliprofundum boonei T469	CP001941.1	GI:289533465	70 / 1920	GenBank
Archaeoglobus profundus DSM5631	CP001857.1	GI:284011125	82 / 2000	[77]/ GenBank, depth is Guaymas Basin [188]
Deferribacter desulfuricans SSM1	AP011529.1	GI:290753135	62.5 / 1385	[189]
Deinococcus geothermalis DSM11300	CP000359.1	GI:94554390	47	[77]
Escherichia coli 536	CP000247.1	GI:110341805	37	[173]
Escherichia coli 55989	CU928145.2	GI:218350208	37	[173]
Escherichia coli K-12 MG1655	U00096.2	GI:48994873	37	[173]
Geobacillus kaustophilus HTA426	NC_006510.1	GI:56418535	60 / 10897	[190]
Geobacillus thermodenitrificans NG80-2	CP000557.1	GI:134265192	57	[191]
Haemophilus influenza PittEE	CP000671.1	GI:148715293	36	[192]
Hydrogenobacter thermophilus TK6	AP011112.1	GI:288786720	72	[77]
Idiomarina	AE017340.1	GI:56178122	21 /	Genus

loihiensis L2TR			1300	optimum growth range [77] / [193]
Klebsiella pneumonia 342	CP000964.1	GI:206564770	37	[173]
Lactobacillus casei ATC 334	CP000423	GI:116103724	37	[191]
Lactobacillus johnsonii NCC 533	AE017198.1	GI:41584196	37	[191]
Lactobacillus plantarum JDM 1	CP001617.1	GI:254044096	37	[191]
Meiothermus ruber DSM1279	CP001743.1	GI:290469363	60	[77]
Methanocaldococcus fervens SG86	CP001696.1	GI:256793173	85 / 2600	[77] / GenBank
Methanocaldococcus infernus ME	CP002009.1	GI:295433502	85 / 3000	[77] / GenBank
Methanocaldococcus Sp F3406-22	CP001901.1	GI:288937946	90 / 1552	[194]
Methanocaldococcus vulcanius M7	CP001787.1	GI:261369124	80 / 2600	[77] / GenBank
Nitratiruptor sp SB155-2	AP009178.1	GI:151421614	55 / 1000	[195]
Oceanobacillus iheyensis HTE831	BA000028.3	GI:42632302	30 / 1050	[191] / [196]
Pyrobaculum aerophilum str IM2	AE009441.1	GI:18308975	100	[77]
Pyrobaculum islandicum DSM 4184	CP000504.1	GI:119672928	100	[77]
Pyrococcus abyssi	AL096836.1	GI:30407140	96	[77]
Pyrococcus furiosus DSM 3638	AE009950.1	GI:18980902	100	[77]
Salmonella Enterica serovar Typhi str Ty2	NC_004631.1	GI:29140543	37	[173]
Shewanella piezotolerans WP3	CP000472.1	GI:212554395	17.5 / 1914	[165]
Staphylococcus aureus MRSA 252	BX571856.1	GI:49240382	35	[191]
Staphylococcus aureus MSSA 476	BX571857.1	GI:49243355	35	[191]
Streptococcus pneumoniae ATCC	FM211187.1	GI:220673408	37	[191]

700669				
Streptococcus pyogenes M1GAS	AE004092.1	GI:14286347	37	[191]
Sulfurovum sp NCB37-1	AP009179.1	GI:151423458	33 / 1000	[195]
Thermococcus onnurineus NA1	CP000855.1	GI:212008101	80 / 1650	[197]
Thermomicrobium roseum DSM5159	CP001275.1	GI:221155340	72	[77]
Thermotoga maritima MSB8	AE000512.1	GI:12057205	80	[77]
Thermotoga neapolitana DSM 4359	CP000916.1	GI:221571364	80	[77]
Thermus aquaticus Y51MC23	ABVK02000001.1	GI:218244499	70	[77]
Thermus thermphilus HB27	AE017221.1	GI:46197919	70	[77]
Thiomicrospira crunogena XCL2	CP000109.2	GI:110744159	25 / 2600	[198]
Vibrio Sp EX25	CP001805.1, CP001806.1	GI:262335977,GI:262338893	20.5 / 2300	[192] / Depth of East Pacific Rise[188]
Zunongwangia profunda SMA87	CP001650.1	GI:294979899	4.7 / 1245	[199]

5' UTRs: In this study, we have defined the 5' UTRs as 50bp upstream of a given gene, provided that does not overlap with another gene on the same strand.

mRNA half-life: The mRNA half-life data was obtained from Bernstein *et al.* [26].

Methods

RNA folding: All genes were taken from the above GenBank files, and folded using RNAfold of the Vienna RNA package 1.6.5 [7-14].

Ortholog lists: Orthologs for bacteria in the same family were obtained from the Sanger ortholog list (*S. aureus* strains MSSA and MRSA [200] and *M. tuberculosis* and *M. leprae* comparisons [201]). In the absence of this information, or no orthologous genes were identified for that prokaryotic pair.

For the others, we used the reciprocal best-self hit method [202-204] to identify orthologs: *blastx* ran for every gene of one species against the formatted database of another species. If two genes identified each other as their best hit, with a significance score of less than 10, we consider them orthologs. If a gene was identified as an ortholog to more than one partner, we used the pair with the highest score; if they had the same score, we used the pair with the longest match; if they had the same match length, we randomly selected one of the gene pairs to use.

To determine if the difference in folding free energy between a set of prokaryotes' orthologous genes was significantly different to the null hypothesis (of that the orthologous gene difference is either equal to difference in mRNA folding free energy between the two organisms, or is equal to zero), we performed a paired *t*-test for sample means.

Because of the large gene size of each genome, we only compared a subset of genes. For each of the species, we compared, using an unpaired *t*-test assuming unequal sample size and variance, only the set of genes that were found to have orthologs in every comparison. We then clustered the *p*-values of these comparisons using Matlab to construct the clusters from the resulting agglomerative hierarchal tree. The number of clusters was determined by generating the clusters until the non-statistically significant relationships ($p<0.01$) had been recovered. (This is seen in **Figure 5.4B**.)

GFP Expression Prediction: We used the GFP mutated sequences and expression values from the Kudla *et al.* study [20] to address the problem of predicting protein levels. To do so we generated and tested a number of features that we list below. We define the ***pairing potential***, P_{pair}(b), as the probability that the base *b* will be paired in the secondary structure (predicted by Sfold [15, 16]). We also define the ***bonding strength,*** P_{bond}(b), as the product of the P_{pair}(b) with the number of H-bonds the base *b* might form in the secondary structure. The features we used can be broadly divided into four main groups:

a. FFE-related features. Minimum and average FFE for (1) the entire sequence, (2) the first 1/3, (3) the last 2/3, (4) the ORF of the GFP mRNA. Minimum and average FFEs are calculated by programs RNAfold [7-14] and Sfold [15, 16], respectively. We also included the difference between the FFE of the individual sequence and the median FFE of all *E. coli* sequences, in order to obtain information regarding to where the individual sequence falls with respect to other *E. coli* sequences.

b. Correlation features. The correlation coefficient between codon usage and tRNA concentration.

c. Structural features. The P_{bind}(b) and P_{bond}(b): (1) over a 5, 10, 20, or 100 bp window around the translation start site, (2) over a 20, or 40 bp window around position -25, and (3) in the interval between RBS and start codon.

d. Sequence features. The longest run of Us (the A, G, and C counts varied to a much less extent than the U); the length of longest run of AU nucleotides; and the percent AU of the sequence.

In order to build an efficient predictive model, we used the *glmnet* package in MatLab, which performs regression with a lasso penalty, fitting lambda by coordinate descent. We added

features incrementally using a greedy approach. In each step, the feature that was added, was the one that showed the largest correlation of predicted and measured expression values, conditional on the existing features in the model.

5.3 SUPPLEMENTARY MATERIALS

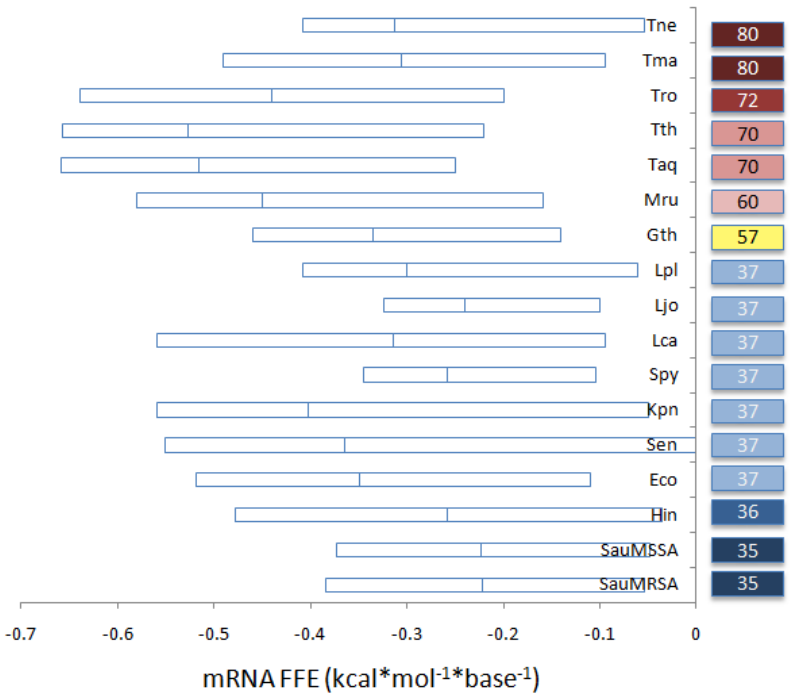

Figure 5.9 Global mRNA folding free energy distributions. For species names, please refer to *Materials and Methods*.

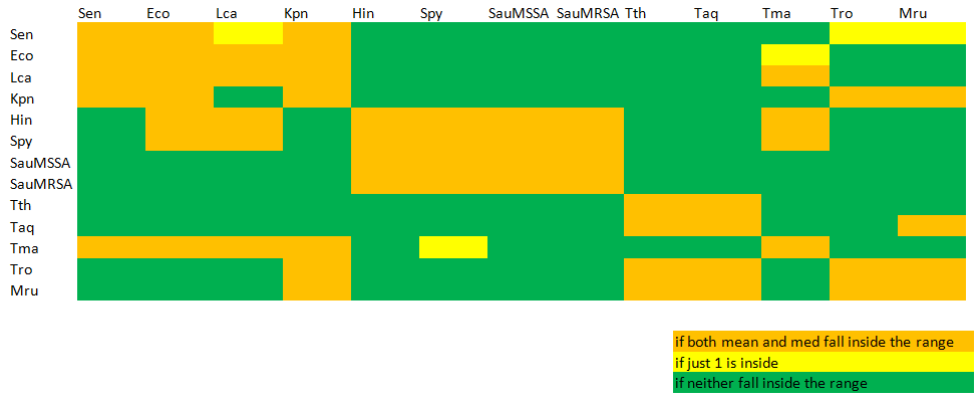

Figure 5.10 Comparison of mRNA FFE in different bacteria. In this chart, the mean/median of mRNA folding of the bacterium on the left is compared with the middle 90% of mRNA FFE of the bacterium on the top.

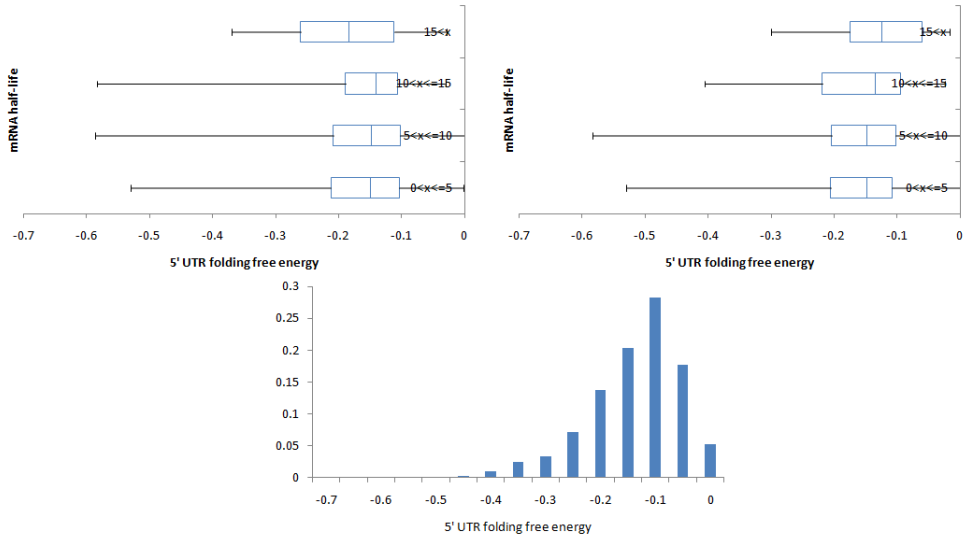

Figure 5.11 Shows the relation of the 5' UTR and mRNA half-life in E. coli K12 MG1665 as grown in **A.** M9 media, and **B.** LB media. The R-squared correlation coefficent between the mRNA half-life when E. coli is grown on these two media is 0.08. **C**. is the histogram of 5' UTR folding free energy of this strain of E. coli, and illustrates that the longest mRNA half-lifes occur in the center of the distribution. The mRNA half-life data is from Bernstein, *et al* 2002 *[26]*.

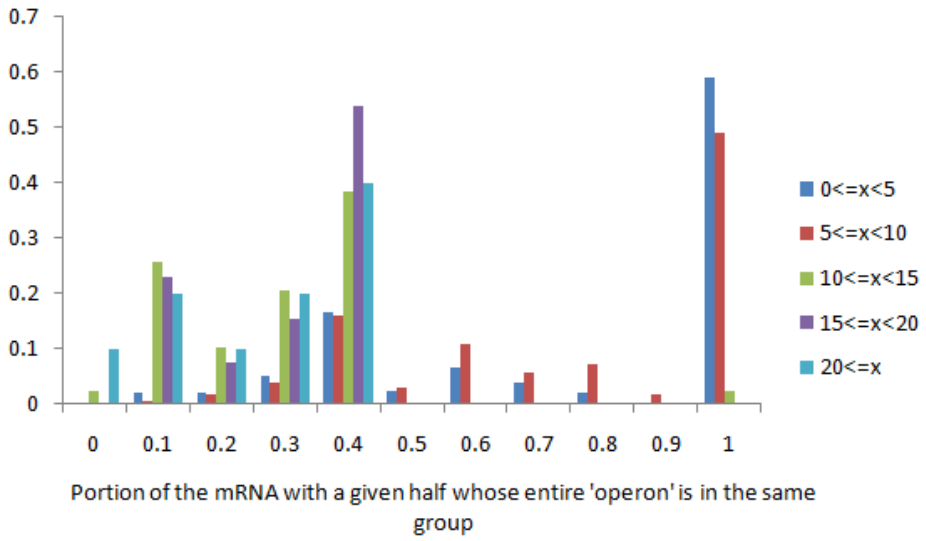

Figure 5.12 Shows the relation of the mRNA half-life and how close the genes are located to each other. In this case, an operon was defined as genes that are in the same DNA strand, and are located closer than 50bp to each other. This histogram is depicting the portion of the mRNA within a given half-life interval whose entire 'operon' is in the same half-life interval. The mRNA half-life data is from Bernstein, *et al* 2002 [26].

6.0 SUMMARY AND CONCLUSIONS

By studying how FFE evolves in influenza A virus, we found that the folding energy between avian and human isolates were dramatically different, with the avian isolates having a substantially lower folding energy. Furthermore, once the virus jumped into the human population, the folding energy of the polymerase gene segments undergoes evolution towards higher values. It was also shown that this difference in folding energy relates to the proficiency of the virus to replicate at specific temperatures: with some human viruses replicating more efficiently at human body temperature, and some avian viruses replicating more efficiently at avian body temperatures. This suggests that FFE of the polymerase genes may be an important factor in the evolution of the virus and its ability to jump between hosts.

We then used the folding energy and amino acid composition of the virus to determine if a set of polymerase genes of the influenza A virus is similar to those which have replicate in the human population. This is a useful tool, as it can correctly classify the host of not only strains that have successfully entered and replicated in the human population such as the seasonal influenza strains, 1918 H1N1 and the pandemic 2009 H1N1 virus, but also several currently circulating strains. The FFE of the virus is a vital feature, in addition to the amino acid sequence, in our ability to predict the virus as being able to replicate in the human population.

It also appears, by way of studying other negative ssRNA viruses, as well as the positive ssRNA West Nile Virus, that folding energy may play an important role in other viruses as well.

We see that an increase in folding energy of WNV corresponds to a decrease of mortality of the virus in American crows; and that, when comparing WNV viruses of two vastly different folding energies, a higher folding energy corresponded to the crow's ability to clear the virus when running a fever. By looking at other negative ssRNA viruses, we see that folding energy of the polymerase genes may be related to the ability of the virus to cause sustained outbreaks in the human population.

But what does all this mean? Why would RNA folding be so important to the RNA viruses? To answer this, we looked to the prokaryotes. What we were able to first identify is that the tRNA FFE does vary directly with temperature- comparable to previous studies of tRNA GC. With the hypothesis that mRNA structure stability may play a role in how accessible the molecule is for translation, balanced against its degradation rate; we looked at the correlation of half-life and FFE in *E. coli*. We found that that the mRNA genes with the longest half-lives had a more narrow range of FFE around the average. Additionally, we found that the genes in the center of the FFE distribution codon usage correlated strongly with tRNA concentrations. This might give an interesting and previously overlooked explanation for the origin of codon bias in prokaryotes.

Furthermore, we were also able to predict the GFP protein levels from the sequence and structural features of the GFP mRNA, illustrating that the structure of the mRNA itself may help dictate protein expression levels. This, in combination with the preservation of orthologous mRNA FFE distributions within the mesophilic and thermophilc bacteria, may indicate how protein expression is conserved. Also, we show that the same trends hold when pressure and temperature is varied (as opposed to just temperature).

Together, this work indicates is that the structure and folding of the mRNA, in addition to its primary sequence, is important. By studying the ssRNA viruses, we can see that the FFE is a feature that is selected for as the virus evolves and enters new host populations. By studying prokaryotes we can see that mRNA FFE is related to environmental adaptation, mRNA half-life and protein expression levels. For both ssRNA viruses and prokaryotes, RNA FFE appears to be another genome feature that adapts to the environment: for ssRNA viruses, RNA FFE must be adapted to the host cell; and for prokaryotes RNA FFE is adapted to the physical environment. The selection pressure behind both adaptations appears to be driven by RNA half-life and protein expression.

APPENDIX A

THE ABILITY OF THE POLYMERSE GENE CLASSIFERS TO CORRECTLY CLASSIFY HUMAN SEQUENCES BOTH WITH-IN AND WITH-OUT THEMSELVES (WHEN TRAINED ON THEMSELVES.

BINOMIAL MODEL MI-IFFE

FEATURES	1918-1939	1940-1957	1958-1968	1969-1976	1977-1990	1991-2000	2001-2008
2	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0.75	1	0.9189	0.9667	0.9574	0.6834
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
3	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0.75	1	0.9189	0.9667	0.9574	0.6834
	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	1	0.9459	0.9833	1	0.9974
4	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0.75	1	0.8919	0.9833	0.9489	0.9525
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
5	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0

	0	0	0.619	0.973	0.5167	0.7191	0.6306
	0	0.75	1	0.973	0.9833	1	0.9894
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
6	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.619	0.973	0.5167	0.7191	0.6306
	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
7	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.619	0.973	0.4833	0.0085	0.0026
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
8	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.619	0.973	0.4833	0.0085	0.0026
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
9	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.619	0.973	0.4833	0.0085	0.0026
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
10	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.619	0.973	0.4833	0.0085	0.0026
	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
11	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.619	0.973	0.4833	0.0085	0.0026

	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
12	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.7619	0.973	0.4	0.6383	0.4697
	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
13	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0.25	1	0.973	0.3833	0.7489	0.5699
	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
14	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0.75	1	0.973	0.45	0.7447	0.5699
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.997
	0	1	1	0.973	0.9833	1	0.994
15	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.7619	0.973	0.5333	0.6979	0.4908
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
16	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0.25	0.9524	0.973	0.5333	0.8128	0.6016
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
17	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0.375	1	0.973	0.4	0.783	0.5989
	0	1	1	0.973	0.9833	1	0.9974

	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
18	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.7619	0.973	0.5333	0.8085	0.628
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
19	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.7619	0.973	0.5333	0.8085	0.628
	0	1	1	0.973	0.9833	1	0.9947
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
20	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.7619	0.973	0.5167	0.8085	0.6359
	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974

BINOMIAL MODEL MI

FEATURES	1918-1939	1940-1957	1958-1968	1969-1976	1977-1990	1991-2000	2001-2008
2	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0.75	1	0.9189	0.7833	0.7745	0.628
	0	0.75	1	0.9189	0.9667	0.9617	0.6887
	0	0.75	1	0.973	0.9833	1	0.9894
	0	1	1	0.973	0.9833	1	0.9974
3	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0.75	0.9524	0.9189	0.65	0.7745	0.628
	0	0.75	0.9524	0.9189	0.8833	0.7787	0.6332
	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	1	0.9459	0.9833	1	0.9974

4	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.7143	0.9189	0.4833	0.7702	0.6253
	0	0.75	0.9524	0.9189	0.9833	0.9957	0.686
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
5	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.7143	0.9189	0.4833	0.7702	0.6253
	0	0.75	1	0.973	0.9833	1	0.9921
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
6	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.619	0.9189	0.4667	0.6638	0.6095
	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
7	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.619	0.973	0.4833	0.0085	0.0026
	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
8	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.619	0.973	0.4833	0.0085	0.0026
	0	0.75	0.9524	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
9	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.619	0.973	0.4833	0.0085	0.0026
	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
10	0	0	0	0	0	0	0

	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.619	0.973	0.4833	0.0085	0.0026
	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
11	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.619	0.973	0.4833	0.0085	0.0026
	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
12	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.7619	0.973	0.4	0.6298	0.467
	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
13	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0.25	1	0.973	0.3833	0.034	0.1821
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
14	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0.75	1	0.973	0.45	0.7362	0.5673
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
15	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.7619	0.973	0.5167	0.0298	0.0053
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
16	0	0	0	0	0	0	0
	0	0	0	0	0	0	0

	0	0	0	0	0	0	0
	0	0.25	0.9524	0.973	0.5333	0.8043	0.6016
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
17	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0.375	0.9524	0.973	0.5	0.7745	0.5937
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
18	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.7619	0.973	0.5333	0.6809	0.3509
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
19	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.7619	0.973	0.5333	0.7745	0.3509
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
20	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0	0	0.7619	0.973	0.5333	0.6681	0.2269
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974

MULTINOMIAL MODEL WITH WEIGHTS MI-IFF

FEATURES	1918-1939	1940-1957	1958-1968	1969-1976	1977-1990	1991-2000	2001-2008
2	1	1	1	0.973	0.8833	0.217	0.3905
	0	1	1	0.973	0.75	0.8128	0.6332
	1	1	1	0.973	0.9667	0.9532	0.9446
	1	1	1	0.973	0.9667	0.9957	0.9789

	1	1	1	0.973	1	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	0.25	0	0.7568	0.9667	0.9957	0.9974
3	1	0.875	1	0.9189	0.85	0.1872	0.3668
	0	1	1	0.973	0.9667	0.9745	0.9736
	0	1	1	0.973	0.8833	0.7745	0.628
	1	1	1	0.973	0.65	0.8085	0.6332
	1	1	1	0.973	1	1	0.9921
	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	0.8571	0.9459	0.9833	1	0.9974
4	1	1	1	0.973	0.9667	0.9574	0.6359
	0	1	1	0.973	0.9667	0.9745	0.9657
	0	1	1	0.973	0.9667	0.9915	0.9763
	0	1	1	0.973	0.9667	0.9447	0.9842
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.9459	0.9833	1	0.9974
5	1	0.25	0	0	0.0333	0.0043	0.0026
	0	1	1	0.973	0.9833	1	0.9657
	0	1	1	0.973	0.9833	0.9872	0.9789
	0	0.75	1	0.973	0.65	0.7702	0.8945
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.9459	0.9833	1	0.9974
6	1	1	1	0.7568	0.7333	0.0511	0.0079
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9947
	0	0.75	1	0.973	0.9667	0.9957	0.9842
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
7	1	0.75	0	0	0.35	0.1617	0.3483
	1	1	1	0.973	0.9833	1	0.9868
	0	1	1	0.973	0.9	0.817	0.6385
	0	0.75	1	0.973	0.65	0.834	0.8918
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
8	1	1	0.9048	0.4595	0.4667	0.1702	0.3509
	0	1	1	0.973	0.9833	1	0.9947
	0	1	1	0.973	0.9	0.8255	0.8865
	0	0.75	1	0.973	0.6667	0.7872	0.8918
	0	1	1	0.973	0.9833	1	0.9974

	0	1	1	0.973	0.9833	1	0.9947
	0	1	1	0.973	0.9833	1	0.9974
9	1	1	0.9048	0.4324	0.4667	0.1787	0.3509
	1	1	1	0.973	0.9833	1	0.8602
	0	1	1	0.973	0.9	0.817	0.6359
	0	0.75	1	0.973	0.9833	0.983	0.9868
	0	1	1	0.973	0.9833	1	0.9894
	0	1	1	0.973	0.9833	1	0.9894
	0	1	1	0.973	0.9833	1	0.9974
10	1	1	0.2381	0.027	0.4667	0.1787	0.3536
	1	1	1	0.973	0.9833	0.4043	0.4749
	0	1	1	0.973	0.9	0.8128	0.6385
	0	0.75	1	0.973	0.9833	0.8426	0.9024
	0	1	1	0.973	0.9833	1	0.9894
	0	1	1	0.973	0.9833	1	0.9894
	0	1	1	0.973	0.9833	1	0.9974
11	1	0.75	0	0.027	0.3667	0.1319	0.3325
	1	1	1	0.973	0.9833	0.4085	0.5567
	0	1	1	0.973	0.9	0.8128	0.6412
	0	0.75	1	0.973	0.9833	0.9319	0.9288
	0	1	1	0.973	0.9833	1	0.9894
	0	1	1	0.973	0.9833	1	0.9894
	0	0.75	1	0.973	0.9833	1	0.9974
12	1	0.75	0	0.1081	0.3667	0.0255	0.2982
	1	1	1	0.8919	0.9667	0.2	0.3641
	0	1	1	0.973	0.9833	1	0.7388
	0	1	1	0.973	0.9833	0.9915	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9894
	0	0.75	1	0.973	0.9833	1	0.9974
13	1	0.75	0	0	0.35	0.0255	0.0132
	1	1	1	0.8919	0.9667	0.2	0.3641
	0	1	1	0.973	0.9	0.8213	0.6385
	0	1	1	0.973	0.9833	0.9872	0.9947
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9894
	0	1	1	0.973	0.9833	1	0.9974
14	1	0.75	0	0	0.45	0.1447	0.3351
	1	1	1	0.9189	0.9667	0.2043	0.3615
	0	1	1	0.973	0.95	0.9106	0.6412
	0	1	1	0.973	0.9833	0.9915	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9894

	0	1	1	0.973	0.9833	1	0.9974
15	1	0.75	0	0	0.45	0.1617	0.3351
	1	1	1	0.9459	0.9667	0.2085	0.3615
	0	1	1	0.973	0.9	0.8213	0.6412
	0	1	1	0.973	0.9833	0.9957	0.9974
	0	1	1	0.973	0.9833	1	0.9947
	0	1	1	0.973	0.9833	1	0.9894
	0	1	1	0.973	0.9833	1	0.9974
16	1	0.75	0	0.027	0.4667	0.183	0.3509
	1	1	1	0.9459	0.9667	0.1957	0.3615
	0	1	1	0.973	0.9	0.8213	0.6412
	0	1	1	0.973	0.9833	0.9915	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9894
	0	1	1	0.973	0.9833	1	0.9974
17	1	0.75	0.2381	0.3784	0.45	0.1277	0.3509
	1	1	1	0.9189	0.9667	0.1957	0.3615
	0	1	1	0.973	0.9833	0.9915	0.9525
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9921
	0	1	1	0.973	0.9833	1	0.9974
18	1	0.75	0.2381	0.3784	0.45	0.0723	0.343
	1	1	1	0.9189	0.9833	0.1957	0.3615
	0	1	1	0.973	0.9833	0.9404	0.6412
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9921
	0	1	1	0.973	0.9833	1	0.9974
19	1	0.75	0.2857	0.4054	0.4333	0.0596	0.3272
	1	1	1	0.973	0.9833	0.9787	0.7071
	0	1	1	0.973	0.9833	0.9447	0.6438
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9921
	0	1	1	0.973	0.9833	1	0.9974
20	1	0.75	0.2857	0.6486	0.5333	0.0596	0.3641
	0	1	1	0.973	0.9833	1	0.9921
	0	1	1	0.973	0.9833	0.9702	0.7071
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9921
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974

MULTINOMIAL MODEL WITH WEIGHTS MI

FEATURES	1918-1939	1940-1957	1958-1968	1969-1976	1977-1990	1991-2000	2001-2008
2	1	1	1	0.973	0.8833	0.217	0.3905
	1	1	1	0.973	0.75	0.8128	0.6332
	1	1	1	0.973	0.9667	0.9532	0.9763
	1	1	1	0.973	0.9833	1	0.9921
	1	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9921
	0	0.25	0.2381	0.973	0.9833	1	0.9974
3	1	0.875	1	0.973	0.85	0.1872	0.3668
	0	1	1	0.973	0.8833	0.7745	0.6306
	0	1	1	0.973	0.8833	0.7745	0.628
	1	1	1	0.973	0.9833	0.9915	0.9815
	0	1	1	0.973	1	1	0.9894
	0	0.75	1	0.973	0.9833	1	0.9974
	0	1	0.8571	0.9459	0.9833	1	0.9974
4	1	1	1	0.973	0.9833	1	0.9947
	0	1	1	0.973	0.8833	0.7702	0.6253
	1	1	1	0.973	0.9833	1	0.9947
	0	1	1	0.973	0.9833	0.9915	0.9947
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
5	1	1	1	0.8378	0.85	0.6553	0.0422
	0	1	1	0.973	0.9667	0.9787	0.9683
	0	1	1	0.973	0.9833	1	0.9921
	1	1	1	0.973	0.9833	0.9957	0.9947
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
6	1	1	1	0.4054	0.4	0.0085	0.0053
	1	1	1	0.973	0.9667	0.9957	0.9868
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9947
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
7	1	0.75	0	0	0.2667	0.0043	0.0132
	1	1	1	0.973	0.9667	0.9447	0.8522
	0	1	1	0.973	0.9833	0.9872	0.9789

	0	0.75	1	0.973	0.9667	0.9319	0.9815
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
8	1	0.75	0	0	0.1167	0.0043	0.0185
	1	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9	0.8085	0.8813
	0	0.75	1	0.973	0.9833	0.966	0.9815
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
9	1	1	0.9048	0.4054	0.4667	0.1745	0.3536
	1	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9	0.8213	0.6227
	0	0.875	1	0.973	0.9833	0.9872	0.9842
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
10	1	1	1	0.7297	0.4833	0.183	0.3562
	1	1	1	0.973	0.9833	1	0.9499
	0	1	1	0.973	0.9	0.8213	0.6359
	0	0.875	1	0.973	0.9833	0.966	0.9657
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
11	1	1	1	0.7568	0.4833	0.183	0.3509
	1	1	1	0.973	0.9667	0.9489	0.5963
	0	1	1	0.973	0.9	0.8128	0.6306
	0	1	1	0.973	0.9833	1	0.9894
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
12	1	1	0.9048	0.4865	0.4833	0.183	0.3588
	1	1	1	0.973	0.9667	0.9532	0.5963
	0	1	1	0.973	0.9	0.8255	0.6385
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
13	1	0.75	0	0	0.35	0.0596	0.3219
	1	1	1	0.9189	0.9333	0.2128	0.3562
	0	1	1	0.973	0.9	0.8255	0.6385
	0	1	1	0.973	0.9833	1	0.9763

	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	0.9957	0.9921
14	1	0.75	0	0	0.35	0.0638	0.0554
	1	1	1	0.9189	0.9333	0.2128	0.3562
	0	1	1	0.973	0.9833	1	0.9789
	0	1	1	0.973	0.9833	1	0.9842
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9947
15	1	0.75	0	0	0.35	0.0681	0.3219
	1	1	1	0.9189	0.9833	0.9149	0.5858
	1	1	1	0.973	0.9833	0.9957	0.9683
	0	1	1	0.973	0.9833	1	0.9894
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9947
16	1	0.75	0	0	0.45	0.1702	0.3351
	1	1	1	0.973	0.9833	0.9447	0.5937
	0	1	1	0.973	0.9833	0.9872	0.6702
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9947
17	1	0.75	0	0.027	0.45	0.1787	0.3562
	1	1	1	0.973	0.9833	0.9447	0.7203
	0	1	1	0.973	0.9833	1	0.942
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
18	1	0.75	0	0	0.45	0.1787	0.3509
	1	1	1	0.973	0.9833	0.9872	0.7256
	0	1	1	0.973	0.9833	1	0.9842
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9947
	0	0.875	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
19	1	0.75	0	0	0.45	0.1745	0.3456
	1	1	1	0.973	0.9833	0.9957	0.7282
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9894

	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974
20	1	0.75	0	0	0.45	0.166	0.3404
	1	1	1	0.973	0.9833	0.9447	0.7203
	1	1	1	0.973	0.9833	1	0.9947
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9947
	0	1	1	0.973	0.9833	1	0.9974
	0	1	1	0.973	0.9833	1	0.9974

APPENDIX B

THE LIST OF GENBANK IDENTIFIEERS USED IN THE CHAPTER 3.

For segment 1: AB166859, AB188813, AB188821, AB189050, AB189058, AB256663, AB256671, AB256679, AB256687, AB256695, AB256703, AB256711, AB256719, AB256727, AB256735, AB256743, AB261850, AF156430, AF255624, AF457672, AF457681, AF457689, AF457697, AF457705, AF457714, AF508640, AF508641, AF508642, AF508645, AF508649, AJ291395, AJ410495, AJ410496, AJ410497, AJ410498, AJ410499, AJ410500, AJ427305, AJ620347, AJ627485, AJ627487, AM262530, AM262531, AM262532, AM262533, AM262534, AM503066, AM503067, AM503068, AM503069, AM503070, AM503071, AM503072, AM503073, AY253750, AY342414, AY609309, AY616766, AY646085, AY648294, AY650276, AY651705, AY651707, AY651713, AY651714, AY651722, AY651723, AY651724, AY651725, AY651726, AY651728, AY653193, AY676023, AY684703, AY724251, AY737286, AY737293, AY770084, AY818128, AY849782, AY849783, AY950279, AY950281, AY950282, AY950283, CY005453, CY005460, CY005483, CY005506, CY005612, CY005800, CY005831, CY005838, CY005844, CY005895, CY005913, CY014182, CY014190, CY014643, CY014670, CY014678, CY014760, CY014771, CY014785, CY014793, CY014799, CY014805, CY014828, CY014836, CY014843, CY014887,

CY014900, CY014916, CY014998, CY015026, CY015039, CY015053, CY015058, CY015072, CY015080, CY015088, CY015096, CY015103, CY015108, CY015114, CY015120, CY015126, CY016283, CY016291, CY016299, CY016307, CY016794, CY016818, CY016842, CY016850, CY016874, CY016882, CY016914, CY016922, CY016930, CY016938, CY016946, CY016954, CY017058, CY017066, CY017186, CY017410, CY018956, CY020236, CY020588, CY020596, CY020604, CY020620, CY020628, CY020636, CY020644, CY020652, CY020660, CY020676, CY020684, CY020692, CY020700, CY021364, CY021372, CY021380, CY021388, CY021396, CY021412, CY021420, CY021428, CY021492, CY021500, CY021516, CY021532, CY021540, CY021548, CY021556, CY022084, CY022268, CY022620, CY022628, CY022636, CY022660, CY022668, CY022676, CY022684, CY022692, CY022700, CY022716, CY022772, CY022820, CY022836, CY024761, CY024769, CY024793, CY024841, CY024849, CY024873, CY024889, CY025084, CY025116, CY025124, CY025132, CY025148, CY025156, CY025172, CY025180, CY025188, CY025196, CY028514, CY028547, CY028579, CY028587, CY028603, CY028635, CY028675, CY028683, CY028691, CY028723, CY029004, CY029088, CY029179, CY029193, CY029235, CY029242, CY029249, CY029431, CY029459, CY029466, CY029473, CY029480, CY029840, CY029920, CY029945, CY029953, CY029961, CY029969, CY029977, DQ064544, DQ064545, DQ064546, DQ064547, DQ064548, DQ064549, DQ064550, DQ064551, DQ064552, DQ064553, DQ064554, DQ064555, DQ064556, DQ064557, DQ064558, DQ064559, DQ064560, DQ064561, DQ064564, DQ064565, DQ064566, DQ064567, DQ064568, DQ064569, DQ138166, DQ138171, DQ138172, DQ138177, DQ138178, DQ323675, DQ334757, DQ334765, DQ334773, DQ335778, DQ351870, DQ351871, DQ351872, DQ366327, DQ376871, DQ376872, DQ376873, DQ376874, DQ376875, DQ376878, DQ376880, DQ376882, DQ376883, DQ376884, DQ376885, DQ376886,

DQ376887, DQ376888, DQ376890, DQ376891, DQ376892, DQ376893, DQ376894, DQ376896, DQ376897, DQ376899, DQ376900, DQ376902, DQ376904, DQ376905, DQ449639, DQ469995, DQ485205, DQ485213, DQ492819, DQ492824, DQ492825, DQ492826, DQ492827, DQ492828, DQ492829, DQ492830, DQ492831, DQ492840, DQ492841, DQ492842, DQ492843, DQ492844, DQ520851, DQ529292, DQ650670, DQ914812, DQ979863, DQ991301, DQ991317, DQ991325, DQ997085, DQ997101, DQ997110, DQ997121, DQ997132, DQ997149, DQ997162, DQ997181, DQ997193, DQ997225, DQ997270, DQ997298, DQ997306, DQ997317, DQ997323, DQ997339, DQ997383, DQ997496, DQ997546, EF015558, EF063554, EF063555, EF063557, EF070740, EF178516, EF205203, EF205204, EF205208, EF205209, EF362425, EF474443, EF551042, EF593099, EF597481, EF605591, EF605604, EF681867, EF681875, EU081871, EU084903, EU084907, EU084909, EU084913, EU084916, EU084920, EU084927, EU084946, EU086233, EU086244, EU086263, EU086299, EU148363, EU148371, EU148379, EU148395, EU148403, EU148411, EU148427, EU148435, EU148443, EU148451, EU163436, EU182258, EU182262, EU182270, EU182293, EU182297, EU182306, EU182319, EU277840, EU277848, EU365368, EU402405, EU414265, EU429983, EU429984, EU429986, EU429987, M21851, M73514, AB049153, AB049154, AB212277, AB239300, AB239307, AB259709, AB262460, AB264769, AB274963, AB284321, AB284985, AB286653, AB300437, AB300438, AB301913, AB304144, AF144300, AF250476, AY038798, AY233387, AY518367, AY585504, AY585505, AY585506, AY585507, AY585508, AY585509, AY585510, AY585511, AY585512, AY585513, AY585514, AY585515, AY585516, AY585517, AY585518, AY585519, AY585520, AY585521, AY585522, AY585523, AY585524, AY633315, AY651706, AY651716, AY651729, AY651748, AY676021, AY676022, AY676024, AY703829, AY724256,

AY737301, AY856861, AY950284, AY950285, CY003854, CY003862, CY003870, CY003878, CY003886, CY003894, CY003901, CY003906, CY003913, CY003921, CY003929, CY003943, CY003951, CY003959, CY003967, CY003975, CY003983, CY003991, CY003999, CY004006, CY004011, CY004018, CY004025, CY004032, CY004042, CY004061, CY004065, CY004071, CY004101, CY004106, CY004113, CY004121, CY004128, CY004153, CY004169, CY004177, CY004185, CY004209, CY004217, CY004225, CY004233, CY004249, CY004257, CY004265, CY004273, CY004281, CY004289, CY004299, CY004309, CY004324, CY004338, CY004345, CY004352, CY004359, CY004366, CY004372, CY004379, CY004386, CY004389, CY004405, CY004419, CY004427, CY004433, CY004439, CY004450, CY004457, CY004465, CY004473, CY004481, CY004489, CY004497, CY004503, CY004514, CY004522, CY004530, CY004538, CY004545, CY004553, CY004561, CY004567, CY004574, CY004587, CY004591, CY004599, CY004606, CY004613, CY004620, CY004625, CY004630, CY004634, CY004641, CY004649, CY004656, CY004661, CY004669, CY004677, CY004680, CY004687, CY004691, CY004699, CY004709, CY004716, CY004721, CY004728, CY004742, CY004749, CY004755, CY004762, CY004769, CY004776, CY004783, CY004790, CY004797, CY004817, CY004824, CY004830, CY004837, CY004842, CY004846, CY004853, CY004860, CY004867, CY004874, CY004880, CY004886, CY004891, CY004896, CY004903, CY004910, CY004918, CY004924, CY004932, CY004946, CY004953, CY004960, CY004976, CY004983, CY004987, CY004994, CY005008, CY005015, CY005034, CY005038, CY005043, CY005050, CY005057, CY005064, CY005071, CY005078, CY005085, CY005091, CY005098, CY005105, CY005120, CY005126, CY005147, CY005153, CY005165, CY005173, CY005187, CY005194, CY005199, CY005206, CY005218, CY005231, CY005238, CY005249, CY005256, CY005263, CY005270, CY005273, CY005280, CY005285, CY005297, CY005304, CY005311, CY005317, CY005330, CY005337,

CY005343, CY005350, CY005357, CY005364, CY005371, CY005405, CY005412, CY005420, CY005427, CY005437, CY005468, CY005475, CY005493, CY005500, CY005512, CY005518, CY005530, CY005537, CY005545, CY005559, CY005562, CY005574, CY005582, CY005589, CY005596, CY005604, CY005618, CY005624, CY005631, CY005638, CY005646, CY005652, CY005659, CY005665, CY005671, CY005678, CY005685, CY005690, CY005698, CY005715, CY005717, CY005745, CY005755, CY005764, CY005776, CY005783, CY005813, CY005819, CY005822, CY005827, CY005851, CY005858, CY005865, CY005873, CY005880, CY005888, CY005902, CY011035, CY011047, CY011055, CY011063, CY011119, CY011255, CY012807, CY012815, CY012823, CY012831, CY012839, CY012847, CY013254, CY013262, CY013270, CY013870, CY014525, CY014555, CY014559, CY014577, CY014585, CY014596, CY014602, CY014614, CY014632, CY014635, CY014686, CY014693, CY014701, CY014709, CY014716, CY014732, CY014738, CY014746, CY014813, CY014820, CY014848, CY014856, CY014864, CY014871, CY014879, CY014895, CY014908, CY014928, CY014936, CY014944, CY014952, CY014960, CY014967, CY014975, CY014991, CY015046, CY015064, CY015134, CY015142, CY015155, CY015450, CY015458, CY015466, CY015474, CY015483, CY015491, CY015499, CY015507, CY016131, CY016139, CY016147, CY016155, CY016163, CY016171, CY016179, CY016187, CY016195, CY016402, CY016418, CY016426, CY016618, CY016626, CY016786, CY016802, CY016810, CY016826, CY016834, CY016890, CY016898, CY016906, CY016962, CY017034, CY017042, CY017050, CY017074, CY017082, CY017282, CY017418, CY017700, CY017708, CY017716, CY017740, CY017748, CY017756, CY017764, CY017772, CY017780, CY017788, CY017796, CY017844, CY017852, CY017860, CY017868, CY018006, CY018014, CY018022, CY018884, CY018892, CY018900, CY018908, CY018916, CY018924, CY019204, CY020356, CY020724, CY020732, CY020740, CY020748, CY020756, CY020764, CY020772,

CY020780, CY020788, CY020796, CY020804, CY020812, CY020820, CY020828, CY020836, CY020844, CY020852, CY020868, CY020876, CY020884, CY020892, CY020908, CY020932, CY020940, CY020948, CY020956, CY020964, CY020972, CY020980, CY020988, CY020996, CY021132, CY021140, CY021148, CY021172, CY021180, CY021188, CY021204, CY021212, CY021220, CY021228, CY021244, CY021252, CY021260, CY021268, CY021276, CY021284, CY021300, CY021308, CY021324, CY021332, CY021340, CY021348, CY021356, CY021404, CY021436, CY021452, CY021460, CY021468, CY021476, CY021484, CY021564, CY021588, CY021596, CY021620, CY021628, CY021644, CY021652, CY021660, CY021668, CY021676, CY021692, CY021868, CY021876, CY021884, CY021900, CY022644, CY022652, CY022724, CY022748, CY022852, CY024753, CY024785, CY024801, CY024905, CY025204, CY028242, CY028250, CY028258, CY028266, CY028290, CY028651, CY028699, CY029025, CY029039, CY029123, CY029228, CY029312, CY029319, CY029403, CY029410, CY029438, CY029445, CY029452, CY029487, CY029501, CY029848, CY029888, CY029896, CY029904, CY029928, CY029936, DQ017495, DQ017501, DQ017509, DQ017510, DQ064562, DQ073399, DQ073400, DQ073401, DQ073402, DQ095756, DQ095761, DQ095762, DQ095763, DQ138167, DQ138179, DQ138180, DQ232607, DQ232608, DQ236087, DQ237956, DQ251441, DQ251449, DQ343506, DQ358750, DQ363916, DQ363921, DQ366303, DQ366311, DQ366335, DQ376870, DQ376877, DQ376879, DQ376881, DQ376895, DQ376898, DQ376901, DQ376903, DQ386305, DQ389161, DQ407243, DQ449647, DQ454106, DQ464357, DQ465397, DQ482668, DQ492834, DQ525418, DQ681211, DQ681215, DQ681220, DQ681225, DQ822187, DQ822195, DQ835799, DQ835800, DQ835804, DQ835805, DQ835806, DQ835807, DQ852607, DQ864507, DQ864713, DQ914807, DQ989958, DQ989966, DQ989974, DQ989984, DQ989992, DQ990000,

DQ997093, DQ997169, DQ997178, DQ997282, DQ997404, DQ997416, DQ997519, DQ997528, DQ997530, EF061121, EF112199, EF112200, EF112201, EF112202, EF112203, EF112205, EF112206, EF112207, EF112208, EF112209, EF112210, EF112211, EF112212, EF112218, EF112221, EF178510, EF205205, EF205206, EF210572, EF392844, EF554800, EF592492, EF597475, EF597484, EF597485, EF597486, EF597498, EF634329, EF634337, EU026013, EU026021, EU026036, EU026044, EU026051, EU026057, EU026089, EU026097, EU026109, EU026117, EU030966, EU030974, EU030982, EU084936, EU084940, EU084949, EU086262, EU094470, EU158170, EU233722, EU257707, EU263342, EU263350, EU296249, EU329174, EU329189, EU430499, EU430505, EU486848, M73516, M73518, M73523, M73525, AF389115, CY009331, CY009451, CY009611, CY010795, CY019962, CY020452, CY020476, DQ208309, EF467818, J02179, V00603, AX399724, AY209934, AY209935, CY008995, CY009283, CY009339, CY009347, CY009459, CY009603, CY009619, CY013278, CY014983, CY019954, CY019978, CY020292, CY020388, CY020468, CY021708, CY021716, CY021812, CY021828, CY021908, CY022020, CY022028, CY022100, DQ508838, EF633442, M73521, M81575, M81581, M81587, X99035, AF348170, AY209936, AY209937, AY209938, AY209939, AY209940, AY209941, AY209942, AY209943, AY209944, AY209945, AY209946, AY209947, AY209948, AY209949, AY209950, AY209951, AY210137, AY210138, AY210139, CY006218, CY008163, CY011127, CY015515, CY019898, CY019914, CY020324, CY020380, CY020396, CY020420, CY020524, CY020532, CY020548, CY020556, CY021020, CY021028, CY021076, CY021116, CY021796, CY021820, CY021852, CY021940, CY022004, CY022092, DQ508878, M23970, M73524, M91713, AX350184, AY210141, AY210142, AY210143, AY210144, AY210145, AY210146, AY210147, AY210148, AY210149, AY210150, CY002103, CY002503, CY002751, CY003503, CY003535, CY003559,

CY003735, CY006051, CY006226, CY006306, CY006314, CY006690, CY006722, CY006730, CY006818, CY006826, CY006834, CY006842, CY006890, CY006914, CY007978, CY008467, CY008683, CY008699, CY008707, CY009011, CY009067, CY009355, CY019906, CY019922, CY021084, CY021092, CY021100, CY021124, CY021604, CY021836, CY021844, CY021948, CY021964, CY022945, CY022953, CY026146, M91712, CY002095, CY002759, CY003071, CY003359, CY003495, CY003511, CY003519, CY003527, CY003543, CY003551, CY003727, CY003743, CY003751, CY006059, CY006106, CY006210, CY006322, CY006330, CY006698, CY006706, CY006714, CY006738, CY006762, CY006770, CY006858, CY006898, CY007626, CY008130, CY008179, CY008459, CY008675, CY008715, CY008731, CY008739, CY008747, CY009059, CY009075, CY009291, CY009299, CY009627, CY010371, CY010379, CY010763, CY010883, CY010891, CY010899, CY010915, CY010923, CY010939, CY010955, CY010963, CY010971, CY010979, CY011295, CY011319, CY011487, CY012887, CY012895, CY013302, CY015531, CY017210, CY017218, CY017250, CY017258, CY017434, CY019044, CY019052, CY019060, CY019068, CY019076, CY019100, CY019108, CY019228, CY019244, CY019746, CY019762, CY019770, CY019778, CY019786, CY019970, CY020172, CY020180, CY020188, CY020196, CY020228, CY020300, CY020332, CY020444, CY020460, CY020484, CY020492, CY020572, CY020580, CY021036, CY021044, CY021108, CY021724, CY021732, CY021740, CY021804, CY021916, CY024932, CY026418, CY028731, DQ508822, DQ508830, DQ508846, DQ508870, DQ508894, M38277, M73517, U62543, X15283, AF037417, AB126635, CY000008, CY000016, CY000024, CY000032, CY000040, CY000048, CY000056, CY000072, CY000080, CY000088, CY000096, CY000104, CY000112, CY000120, CY000128, CY000136, CY000144, CY000152, CY000168, CY000176, CY000192, CY000208, CY000216, CY000224, CY000232, CY000240, CY000248, CY000255, CY000264, CY000280, CY000288,

CY000296, CY000304, CY000312, CY000320, CY000328, CY000336, CY000344, CY000352, CY000360, CY000368, CY000376, CY000384, CY000392, CY000400, CY000408, CY000416, CY000424, CY000432, CY000440, CY000448, CY000480, CY000488, CY000504, CY000512, CY000519, CY000528, CY000544, CY000552, CY000560, CY000568, CY000576, CY000583, CY000592, CY000760, CY000768, CY000784, CY000792, CY000872, CY000880, CY000888, CY000896, CY000908, CY000940, CY000956, CY000964, CY000972, CY000980, CY001020, CY001028, CY001044, CY001052, CY001087, CY001095, CY001111, CY001135, CY001151, CY001167, CY001175, CY001191, CY001196, CY001204, CY001212, CY001220, CY001228, CY001236, CY001244, CY001260, CY001268, CY001292, CY001300, CY001308, CY001316, CY001324, CY001340, CY001348, CY001380, CY001412, CY001436, CY001444, CY001468, CY001476, CY001519, CY001543, CY001551, CY001631, CY001639, CY001687, CY001727, CY001735, CY001743, CY001951, CY001959, CY002007, CY002015, CY002023, CY002031, CY002039, CY002047, CY002063, CY002071, CY002079, CY002087, CY002111, CY002135, CY002159, CY002183, CY002191, CY002199, CY002207, CY002215, CY002223, CY002231, CY002271, CY002287, CY002359, CY002367, CY002399, CY002415, CY002423, CY002431, CY002447, CY002463, CY002471, CY002479, CY002495, CY002511, CY002527, CY002535, CY002543, CY002575, CY002591, CY002599, CY002607, CY002615, CY002631, CY002639, CY002663, CY002679, CY002687, CY002695, CY002703, CY002719, CY002727, CY002743, CY002783, CY002807, CY002815, CY002823, CY002913, CY002921, CY002929, CY002937, CY002945, CY002953, CY002961, CY002969, CY002976, CY002983, CY002991, CY002999, CY003015, CY003023, CY003047, CY003055, CY003063, CY003079, CY003087, CY003095, CY003103, CY003111, CY003119, CY003132, CY003135, CY003143, CY003175, CY003183, CY003199, CY003207, CY003215, CY003295, CY003303, CY003311, CY003319, CY003335,

CY003343, CY003351, CY003391, CY003399, CY003407, CY003415, CY003431, CY003479, CY003487, CY003647, CY003655, CY003663, CY003671, CY003687, CY003695, CY003703, CY003711, CY003768, CY003776, CY003784, CY006083, CY006091, CY006114, CY006130, CY006138, CY006146, CY006154, CY006162, CY006178, CY006202, CY006298, CY006362, CY006378, CY006394, CY006402, CY006410, CY006426, CY006434, CY006442, CY006682, CY006866, CY006882, CY006930, CY006938, CY006946, CY006978, CY007010, CY007018, CY007050, CY007066, CY007082, CY007090, CY007098, CY007114, CY007122, CY007130, CY007154, CY007178, CY007186, CY007202, CY007210, CY007218, CY007226, CY007250, CY007258, CY007274, CY007298, CY007306, CY007314, CY007322, CY007338, CY007346, CY007354, CY007362, CY007410, CY007434, CY007458, CY007466, CY007474, CY007490, CY007498, CY007506, CY007546, CY007562, CY007570, CY007578, CY007586, CY007594, CY007610, CY007674, CY007682, CY007690, CY007722, CY007730, CY007738, CY007746, CY007754, CY007762, CY007778, CY007802, CY007826, CY007834, CY007866, CY007898, CY007930, CY007938, CY007946, CY007954, CY007962, CY007970, CY008002, CY008034, CY008050, CY008058, CY008066, CY008074, CY008082, CY008090, CY008106, CY008114, CY008155, CY008211, CY008227, CY008235, CY008259, CY008299, CY008307, CY008315, CY008323, CY008331, CY008347, CY008363, CY008371, CY008403, CY008443, CY008451, CY008523, CY008531, CY008547, CY008603, CY008611, CY008619, CY008627, CY008635, CY008643, CY008659, CY008883, CY008907, CY008915, CY008923, CY009251, CY009259, CY009267, CY009403, CY009411, CY009443, CY009867, CY009875, CY009883, CY009891, CY009939, CY009955, CY009987, CY010155, CY010163, CY010171, CY010187, CY010203, CY010219, CY010243, CY010251, CY010259, CY010267, CY010275, CY010283, CY010291, CY010299, CY010307, CY010315, CY010323, CY010331, CY010339, CY010347, CY010403,

CY010419, CY010435, CY010443, CY010459, CY010467, CY010475, CY010483, CY010555, CY010771, CY010779, CY010787, CY010859, CY011087, CY011095, CY011159, CY011199, CY011207, CY011231, CY011239, CY011399, CY011407, CY011415, CY011615, CY011623, CY011631, CY011639, CY011647, CY011655, CY011663, CY011671, CY011679, CY011687, CY011695, CY011703, CY011711, CY011735, CY011743, CY011759, CY011775, CY011975, CY011983, CY011991, CY011999, CY012007, CY012015, CY012023, CY012031, CY012039, CY012055, CY012063, CY012071, CY012079, CY012095, CY012103, CY012111, CY012119, CY012319, CY012327, CY012343, CY012351, CY012375, CY012383, CY012407, CY012431, CY012647, CY012679, CY012695, CY012711, CY012719, CY012727, CY012799, CY013079, CY013087, CY013095, CY013103, CY013111, CY013135, CY013143, CY013151, CY013223, CY013231, CY013239, CY013420, CY013428, CY013444, CY013484, CY013492, CY013500, CY013524, CY013540, CY013564, CY013572, CY013580, CY013588, CY013596, CY013604, CY013812, CY013934, CY013974, CY013990, CY013998, CY014006, CY014014, CY014022, CY014038, CY014046, CY014062, CY014070, CY014086, CY014094, CY014102, CY014118, CY014142, CY014166, CY015555, CY015587, CY015603, CY015611, CY015619, CY015635, CY015683, CY015691, CY015699, CY015707, CY015715, CY015731, CY015739, CY015747, CY015755, CY015763, CY015771, CY015779, CY015787, CY015795, CY015803, CY015811, CY015835, CY015851, CY015859, CY015867, CY015883, CY015891, CY015899, CY015907, CY015915, CY015923, CY015939, CY015963, CY015971, CY015979, CY015995, CY016003, CY016011, CY016019, CY016027, CY016035, CY016043, CY016051, CY016219, CY016442, CY016450, CY016458, CY016466, CY016482, CY016602, CY016610, CY016666, CY016682, CY016690, CY016698, CY016706, CY017002, CY017090, CY017098, CY017106, CY017114, CY017138, CY017378, CY017490, CY017506, CY017546, CY017554, CY017562, CY017570,

CY017586, CY017602, CY017610, CY017618, CY017626, CY017828, CY017900, CY017908, CY017916, CY017924, CY017932, CY017948, CY017956, CY017972, CY017980, CY017988, CY017996, CY018932, CY018948, CY018972, CY018988, CY019004, CY019012, CY019020, CY019028, CY019036, CY019196, CY019332, CY019340, CY019348, CY019754, CY019834, CY019866, CY019882, CY019890, CY019930, CY019938, CY020004, CY020012, CY020036, CY020044, CY020148, CY020156, CY020164, CY020268, CY020276, CY020284, CY020308, CY020348, CY020428, CY020436, CY020900, CY021764, CY021772, CY022188, CY022204, CY022212, CY022532, CY022540, CY022556, CY022564, CY022588, CY022604, CY023065, CY023073, CY023081, CY025057, CY025220, CY025228, CY025236, CY025244, CY025252, CY025268, CY025276, CY025284, CY025292, CY025300, CY025308, CY025316, CY025324, CY025332, CY025340, CY025348, CY025356, CY025364, CY025372, CY025380, CY025388, CY025396, CY025412, CY025420, CY025428, CY025436, CY025444, CY025452, CY025460, CY025476, CY025484, CY025492, CY025500, CY025508, CY025516, CY025524, CY025532, CY025540, CY025562, CY025578, CY025586, CY025594, CY025602, CY025610, CY025618, CY025626, CY025650, CY025666, CY025674, CY025682, CY025698, CY025706, CY025714, CY025722, CY025730, CY025738, CY025746, CY025754, CY025778, CY025786, CY025802, CY025810, CY025818, CY025826, CY025842, CY025850, CY025858, CY025866, CY025890, CY025898, CY025906, CY025914, CY025922, CY025946, CY025954, CY025962, CY025970, CY025978, CY025986, CY025994, CY026002, CY026010, CY026026, CY026154, CY026178, CY026186, CY026194, CY026202, CY026210, CY026226, CY026242, CY026250, CY026258, CY026266, CY026274, CY026282, CY026322, CY026338, CY026362, CY026378, CY026386, CY026394, CY026402, CY026410, CY026506, CY026514, CY026522, CY026530, CY026538, CY026546, CY026554, CY026570, CY026586, CY026594, CY026602, CY026618, CY026634,

CY026650, CY026658, CY026666, CY026674, CY026682, CY026690, CY026698, CY026706, CY026714, CY026722, CY026738, CY026762, CY026770, CY026778, CY026786, CY026794, CY026802, CY026834, CY026842, CY026850, CY026858, CY026866, CY026874, CY026882, CY026890, CY026898, CY026906, CY026914, CY026922, CY026938, CY026946, CY026954, CY026962, CY026970, CY026978, CY026994, CY027002, CY027026, CY027042, CY027050, CY027066, CY027074, CY027098, CY027106, CY027122, CY027138, CY027146, CY027154, CY027170, CY027178, CY027186, CY027194, CY027210, CY027218, CY027226, CY027234, CY027250, CY027266, CY027274, CY027290, CY027322, CY027330, CY027346, CY027354, CY027362, CY027370, CY027394, CY027418, CY027426, CY027434, CY027442, CY027458, CY027474, CY027482, CY027498, CY027546, CY027570, CY027586, CY027602, CY027610, CY027618, CY027634, CY027642, CY027650, CY027658, CY027666, CY027690, CY027698, CY027706, CY027714, CY027722, CY027730, CY027738, CY027746, CY027754, CY027762, CY027770, CY027778, CY027794, CY027810, CY027818, CY027834, CY027842, CY027850, CY027858, CY027874, CY027890, CY027914, CY027922, CY027930, CY027938, CY027946, CY027970, CY027994, CY028018, CY028034, CY028050, CY028066, CY028074, CY028082, CY028098, CY028106, CY028114, CY028122, CY028130, CY028138, CY028146, CY028202, CY028210, CY028218, CY028314, CY028322, CY028338, CY028362, CY028386, CY028410, CY028450, CY028458, CY028466, CY028482, CY028739, CY028747, CY028771, CY028779, CY030060, CY030204, CY030212, CY030734, DQ415286, DQ415287, DQ415288, DQ415289, DQ415290, DQ415291, DQ415292, DQ415293, DQ469955, DQ889682, EF554792, EU399758, AB212050, AB212051, AF036363, AF084261, AF084262, AF084263, AF115290, AF115291, AF258835, AF258836, AF258837, AF258838, AF258839, AF258840, AF258843, AF258844, AF258845, AF258846, AJ404630, AJ404631, AJ404632, AY576380,

AY626149, AY627892, AY627898, AY651718, AY651719, AY651721, AY818126, CY014167, CY014175, CY014269, CY014293, CY014308, CY014320, CY014322, CY014325, CY014329, CY014332, CY014333, CY014337, CY014341, CY014344, CY014348, CY014352, CY014356, CY014360, CY014364, CY014367, CY014391, CY014396, CY014404, CY014420, CY014444, CY014452, CY014476, CY014484, CY014492, CY014500, CY014512, CY014520, CY014526, CY014534, CY015013, CY017635, CY017643, CY017651, CY017659, CY017667, CY017683, CY017685, DQ138181, DQ360837, DQ492894, DQ492896, DQ492900, DQ492902, DQ535731, DQ835310, EF456776, EF467805, EF467806, EF467807, EF587274, EU146647, EU146655, EU146846, EU263988, AAL14087, ABQ45435, ABQ45446, ABQ45542, ABR15840, ABR15873, ABR29624, ABR29604, ABR28635, ABR28679, ABR29594, ABR29574, ABS50130, ABS49953, ABS49964, ABU80209, ABU80264, ABU80297, ABU80429, ABU80419, ABV82605, ABW71530, ABW71491, ABW71502, ABX58678, ABR87897, ABS53350, ABV31977, AAO15319, AAO15321, ABR15829, ABR28547, ABR28558, ABR28646, ABR28690, ABR28701, ABR28767, ABS49931, ABV82594, ABW36365, ABW86584, ABQ41895, ABQ51944, ABS00319, AAF76002, AAG01784, ABR15884, ABR28723, ABU80231, ABW36332, ABW71513, ABW86605, ABW86595, ABX58667, ABS00330, BAG49628, AAO15320, ABQ45468, ABR15851, ABR28613, ABR28624, ABR28668, ABR28745, ABS50120, ABV82583, ABW36354, AAG01748, AAG17437, AAL87930, ABY40436, ACA25365, AAL87934, ABO44045, ABY40446, AAG01757, AAL87932, AAL87936, CAC37000, ABY16782, ABY16783, ABY16784, AAV30835, ABY84693, AAL87929, AAL87935, ABV55858, ABY16781, ACD85164, ABA27429, ABB86920, ABB86940, ABD77118, ACE78138, ACE78139, ABD61561, ABE27174, ABE27163, ABR15862, ABV29600, AAA43650, AAA43652,

AAA43126, AAU25842, AAU25852, AAU05323, AAV30843, ACF22114, ABB86930, ABI54402, ABK00131, ACE78129, ACE78130, ACE78132, ACE78145, ABW38020, ABW36343, AAA43125, ABY51214, ABY51225, AAZ79398, ABA27437, ABB86870, ABB86890, ABB86900, ABB86910, ABD77110, ABF18004, ACE78128, ACE78133, ACE78135, ACE78141, ACE78143, ABD61259, ABD62803, ABY81436, ABF18003, ABF18005, ABI54403, ABJ16472, ACF49407, ACE78126, ACE78134, ACE78140, ACE78142, ACF04401, ABD95721, BAG49739, AAO15318, AAL87931, ABY81853, ACA25355, ACE78127, ACE78131, ABR28712, ABR28734, ABU80242, ABX58656, ABV31978, AAL87933, ACD65210, ACE78137, ABR28657, ABS53360, CAC36999, AAU25862, ABD78114, AAL14086, ABR28580, ABU80220, ABA46960, ACA42425, ABE12644, AAN46834, ACE78136, ACE78144, ABV29534, ABW36387, ABS53370, ABV31957, ABB86880, ACA42435, ABD62842. For the pandemic H1N1 sequences: CY071334, CY071342, CY071350, CY071358, CY071366, CY083013, CY083021, CY083029, CY083037, CY063827, CY063837, CY063845, CY080304, CY080312, CY080328, CY073965, CY073005, CY083532, CY073123, CY062265, CY062273, CY076728, CY080296, CY074979, CY080320, CY080336, CY071142, CY071150, CY071158, CY075066, CY075074, CY071166, CY071174, CY071182, CY071190, CY071270, CY071278, CY071286, CY071294, CY075082, CY071302, CY071310, CY071318, CY071326, CY071382, CY062986, CY066454, CY063186, CY063194, CY063202, CY063210, CY063218, CY066462, CY066470, CY066478, CY066486, CY064994, CY066494, CY066502, CY066510, CY071214, CY071222, GU562465, CY065879, CY071374, CY081058, CY081066, CY081074, CY081082, CY081098, HQ393490, CY063010, CY063018, CY071078, CY072533, CY073740, CY061585, CY061114, CY061593, CY061122, CY061130, CY061138, CY061146, CY061154, CY061162, CY061170, CY061178, CY061186,

CY061194, CY061202, CY061210, CY080571, CY062994, CY071062, CY073780, HQ891286, CY083824, HQ834743, CY062073, CY062081, CY062089, CY062097, CY062105, CY062113, CY062121, CY062129, CY062137, CY062145, CY062153, CY062161, CY062169, CY062177, CY062185, CY062193, CY065002, CY065010, CY065018, CY065026, CY065034, CY065042, CY065050, CY065058, CY065066, CY065074, CY065114, CY072541, HM124380, HM569737, HM189301, CY061802, HM189587, HM189588, HM189589, HM189590, CY081090.

For segment 2: AB166860, AB188814, AB188822, AB256664, AB256672, AB256680, AB256688, AB256696, AB256704, AB256712, AB256720, AB256728, AB256744, AB261851, AF262210, AF457671, AF457673, AF457682, AF457706, AF508627, AF508629, AF508630, AF508632, AF508633, AF508636, AF508639, AJ620348, AM262524, AM262526, AM262527, AM262528, AM262529, AY253751, AY303663, AY590582, AY609310, AY616765, AY646084, AY648293, AY653039, AY653199, AY676027, AY684704, AY724261, AY737287, AY737294, AY818130, AY818131, AY849784, AY849785, AY950272, AY950274, AY950275, AY950276, CY005452, CY005486, CY005505, CY005611, CY005799, CY005837, CY005894, CY005912, CY006041, CY014183, CY014191, CY014655, CY014662, CY014669, CY014677, CY014755, CY014759, CY014770, CY014777, CY014784, CY014792, CY014798, CY014804, CY014827, CY014835, CY014842, CY014886, CY014899, CY014915, CY014997, CY015025, CY015032, CY015052, CY015071, CY015079, CY015095, CY015102, CY015107, CY015113, CY016282, CY016290, CY016298, CY016306, CY016793, CY016817, CY016841, CY016857, CY016873, CY016881, CY016913, CY016921, CY016929, CY016937, CY016945, CY016953, CY017057, CY017065, CY017185, CY017409, CY018955, CY020235, CY020587, CY020595, CY020603, CY020611, CY020619, CY020627, CY020635,

CY020643, CY020651, CY020659, CY020675, CY020683, CY020699, CY021363, CY021371, CY021387, CY021395, CY021491, CY021499, CY021515, CY022083, CY022267, CY022619, CY022771, CY022819, CY022827, CY022835, CY024792, CY024840, CY024848, CY024872, CY025115, CY025139, CY028513, CY028546, CY028578, CY028586, CY028602, CY028634, CY028674, CY028682, CY028690, CY028706, CY028714, CY029096, CY029166, CY029194, CY029208, CY029236, CY029243, CY029257, CY029383, CY029397, CY029432, CY029460, CY029474, CY029481, CY029773, CY029839, CY029911, CY029919, CY029946, CY029962, CY029970, CY029986, CY029994, DQ064517, DQ064518, DQ064519, DQ064520, DQ064521, DQ064522, DQ064523, DQ064527, DQ064529, DQ064530, DQ064531, DQ064532, DQ064533, DQ064534, DQ064537, DQ064538, DQ064540, DQ064542, DQ067443, DQ138145, DQ138151, DQ138152, DQ138153, DQ138154, DQ138155, DQ334758, DQ334766, DQ334774, DQ335777, DQ351873, DQ351874, DQ351875, DQ366328, DQ376835, DQ376836, DQ376837, DQ376838, DQ376839, DQ376842, DQ376844, DQ376846, DQ376847, DQ376848, DQ376849, DQ376850, DQ376851, DQ376852, DQ376853, DQ376854, DQ376855, DQ376856, DQ376857, DQ376858, DQ376860, DQ376861, DQ376863, DQ376864, DQ376865, DQ376868, DQ376869, DQ449638, DQ469996, DQ485206, DQ493343, DQ493348, DQ493351, DQ493352, DQ493365, DQ493367, DQ493368, DQ493369, DQ650669, DQ870886, DQ914811, DQ997129, DQ997161, DQ997180, DQ997192, DQ997269, DQ997297, DQ997313, DQ997316, DQ997322, DQ997543, DQ997545, EF063531, EF063533, EF070739, EF123893, EF123895, EF123898, EF123936, EF123943, EF123946, EF123957, EF123958, EF123959, EF123960, EF123964, EF124007, EF124011, EF124036, EF124038, EF178515, EF205196, EF205197, EF205201, EF205202, EF362424, EF474444, EF523715, EF523718, EF551043,

EF597448, EF605595, EF605597, EF681868, EF681876, EU084904, EU084910, EU084914, EU084917, EU084921, EU084924, EU084928, EU086232, EU086261, EU086297, EU148362, EU148370, EU148394, EU148402, EU148410, EU148418, EU148426, EU148434, EU148442, EU158162, EU163435, EU182251, EU182263, EU182271, EU182290, EU182294, EU182314, EU182320, EU233681, EU233729, EU277839, EU277847, EU365369, EU402404, EU420032, EU429977, EU429978, EU429979, EU429982, M25925, AB212278, AB239301, AB239308, AB262461, AB264770, AB274964, AB284986, AB285092, AB286117, AB286654, AB284322, AB304145, AF144301, AF468839, AF508638, AY035888, AY233388, AY518366, AY576397, AY585483, AY585484, AY585485, AY585486, AY585487, AY585488, AY585489, AY585490, AY585491, AY585492, AY585493, AY585494, AY585495, AY585496, AY585497, AY585498, AY585499, AY585500, AY585501, AY585502, AY585503, AY676026, AY684880, AY684881, AY724255, AY737302, AY856862, AY950277, AY950278, CY003853, CY003861, CY003869, CY003877, CY003885, CY003893, CY003900, CY003905, CY003920, CY003928, CY003935, CY003942, CY003950, CY003958, CY003966, CY003974, CY003982, CY003990, CY003998, CY004005, CY004010, CY004017, CY004024, CY004033, CY004053, CY004060, CY004070, CY004085, CY004088, CY004093, CY004100, CY004112, CY004120, CY004135, CY004152, CY004168, CY004176, CY004192, CY004200, CY004208, CY004216, CY004224, CY004232, CY004248, CY004256, CY004264, CY004272, CY004280, CY004288, CY004298, CY004314, CY004317, CY004323, CY004328, CY004331, CY004337, CY004344, CY004351, CY004358, CY004365, CY004371, CY004378, CY004385, CY004391, CY004404, CY004418, CY004426, CY004432, CY004449, CY004456, CY004464, CY004472, CY004480, CY004488, CY004496, CY004502, CY004513, CY004521, CY004529, CY004537, CY004544, CY004552, CY004560, CY004566, CY004573, CY004582, CY004586,

CY004590, CY004598, CY004605, CY004612, CY004619, CY004624, CY004629, CY004633, CY004640, CY004648, CY004655, CY004660, CY004668, CY004676, CY004686, CY004690, CY004698, CY004708, CY004715, CY004720, CY004727, CY004748, CY004761, CY004768, CY004775, CY004782, CY004789, CY004796, CY004803, CY004816, CY004823, CY004836, CY004841, CY004852, CY004859, CY004873, CY004879, CY004885, CY004890, CY004895, CY004902, CY004909, CY004917, CY004923, CY004931, CY004938, CY004945, CY004952, CY004959, CY004966, CY004971, CY004975, CY004993, CY005001, CY005007, CY005014, CY005021, CY005037, CY005042, CY005049, CY005056, CY005063, CY005070, CY005077, CY005084, CY005090, CY005104, CY005119, CY005125, CY005132, CY005139, CY005146, CY005152, CY005164, CY005169, CY005172, CY005193, CY005198, CY005205, CY005217, CY005223, CY005230, CY005237, CY005242, CY005248, CY005255, CY005269, CY005279, CY005291, CY005296, CY005303, CY005316, CY005323, CY005329, CY005336, CY005342, CY005349, CY005356, CY005363, CY005370, CY005375, CY005381, CY005386, CY005391, CY005411, CY005419, CY005426, CY005436, CY005467, CY005474, CY005477, CY005499, CY005511, CY005529, CY005558, CY005595, CY005603, CY005617, CY005623, CY005630, CY005637, CY005645, CY005658, CY005664, CY005670, CY005677, CY005689, CY005697, CY005703, CY005708, CY005714, CY005744, CY005763, CY005812, CY005818, CY005821, CY005857, CY005864, CY005872, CY005879, CY005887, CY005901, CY005949, CY011034, CY011046, CY011054, CY011062, CY011118, CY011254, CY012806, CY012814, CY012822, CY012830, CY012838, CY012846, CY013253, CY013261, CY013269, CY014524, CY014554, CY014567, CY014576, CY014626, CY014638, CY014685, CY014692, CY014700, CY014708, CY014715, CY014725, CY014731, CY014737, CY014745, CY014812, CY014819, CY014847, CY014855, CY014863, CY014870, CY014878, CY014894, CY014907, CY014921, CY014927,

CY014935, CY014943, CY014951, CY014966, CY014974, CY014990, CY015063, CY015133, CY015151, CY015154, CY015177, CY015449, CY015457, CY015465, CY015473, CY015482, CY015490, CY015498, CY015506, CY016146, CY016154, CY016162, CY016170, CY016178, CY016186, CY016194, CY016401, CY016417, CY016425, CY016617, CY016625, CY016785, CY016801, CY016809, CY016825, CY016833, CY016889, CY016897, CY016905, CY016961, CY017033, CY017041, CY017049, CY017073, CY017081, CY017281, CY017417, CY017699, CY017707, CY017715, CY017739, CY017747, CY017755, CY017763, CY017771, CY017779, CY017787, CY017795, CY017843, CY017851, CY017859, CY018005, CY018013, CY018021, CY018883, CY018891, CY018899, CY018907, CY018915, CY018923, CY019203, CY020355, CY020723, CY020731, CY020739, CY020747, CY020755, CY020771, CY020787, CY020795, CY020819, CY020827, CY020835, CY020843, CY020851, CY020859, CY020867, CY020875, CY020883, CY020891, CY020907, CY020915, CY020931, CY020939, CY020947, CY020955, CY020963, CY020971, CY020979, CY020987, CY020995, CY021131, CY021139, CY021147, CY021171, CY021179, CY021187, CY021203, CY021211, CY021219, CY021227, CY021243, CY021251, CY021259, CY021267, CY021275, CY021299, CY021307, CY021323, CY021339, CY021347, CY021355, CY021435, CY021451, CY021459, CY021467, CY021475, CY021483, CY021571, CY021587, CY021595, CY021619, CY021627, CY021643, CY021651, CY021659, CY021667, CY021675, CY021683, CY021691, CY021867, CY021875, CY021883, CY021899, CY022611, CY022651, CY022723, CY024800, CY025203, CY028241, CY028249, CY028257, CY028265, CY028273, CY028289, CY028650, CY028698, CY028998, CY029019, CY029040, CY029054, CY029061, CY029075, CY029117, CY029124, CY029131, CY029138, CY029229, CY029278, CY029285, CY029292, CY029299, CY029306, CY029313, CY029320, CY029334, CY029348, CY029355, CY029369, CY029404, CY029418, CY029425, CY029439, CY029453,

CY029488, CY029847, CY029879, CY029887, CY029903, CY029927, CY029935, DQ017492, DQ017500, DQ017507, DQ017511, DQ064535, DQ064536, DQ073403, DQ073404, DQ073405, DQ073406, DQ095740, DQ138161, DQ138162, DQ232605, DQ232606, DQ237955, DQ251450, DQ343505, DQ358749, DQ365000, DQ365008, DQ366304, DQ366312, DQ366336, DQ376834, DQ376841, DQ376843, DQ376845, DQ376859, DQ376862, DQ376866, DQ376867, DQ386304, DQ407244, DQ423612, DQ449646, DQ454102, DQ464361, DQ465398, DQ493359, DQ681202, DQ681214, DQ681219, DQ681224, DQ835792, DQ835793, DQ835794, DQ835795, DQ852606, DQ864506, DQ864714, DQ914810, DQ989959, DQ989967, DQ989975, DQ989985, DQ989993, DQ990001, DQ997168, DQ997177, DQ997281, DQ997403, DQ997415, DQ997518, DQ997527, DQ997529, EF061124, EF112223, EF112224, EF112225, EF112226, EF112227, EF112230, EF112231, EF112236, EF112238, EF112239, EF112240, EF112242, EF112244, EF112245, EF123891, EF123892, EF123897, EF123899, EF123900, EF123901, EF123902, EF123907, EF123917, EF123925, EF123944, EF123945, EF123953, EF123962, EF123965, EF123968, EF123969, EF123995, EF124010, EF124013, EF124015, EF124017, EF124020, EF124021, EF124029, EF178509, EF205198, EF205199, EF491875, EF517399, EF523717, EF523719, EF523720, EF523721, EF523722, EF523723, EF554801, EF592493, EF597440, EF597441, EF597464, EF634330, EF634338, EU026012, EU026020, EU026027, EU026035, EU026043, EU026050, EU026056, EU026064, EU026072, EU026088, EU026096, EU026116, EU030965, EU030973, EU030981, EU084937, EU084941, EU084950, EU094471, EU233713, EU257636, EU263343, EU263351, EU296250, EU329175, EU329188, EU430500, EU430509, EU441923, EU441931, M25926, AF389116, CY009330, CY009450, CY009610, CY010794, CY020451, CY020475, DQ208310, DQ508903, J02151, J02178, AY210008, AY210009,

CY008994, CY009282, CY009338, CY009346, CY009458, CY009602, CY009618, CY013277, CY014982, CY019953, CY019977, CY020291, CY020467, CY021707, CY021715, CY021811, CY021827, CY021907, CY022019, CY022027, DQ508839, EF633441, M25924, M25932, M81574, M81580, M81586, X99037, AF348172, AY210010, AY210011, AY210012, AY210013, AY210014, AY210015, AY210016, AY210017, AY210018, AY210019, AY210020, AY210021, AY210022, AY210023, AY210024, AY210025, AY210271, AY210272, CY006217, CY008162, CY011126, CY015514, CY019897, CY019913, CY020323, CY020379, CY020395, CY020403, CY020419, CY020531, CY020547, CY020555, CY021075, CY021115, CY021795, CY021819, CY021851, CY021939, CY022003, CY022091, DQ508879, J02138, M25935, AX350186, AY210274, AY210275, AY210276, AY210277, AY210278, AY210279, AY210280, AY210281, AY210282, AY210283, AY210284, CY002102, CY002502, CY002750, CY003502, CY003534, CY003558, CY003734, CY006050, CY006225, CY006305, CY006313, CY006689, CY006721, CY006729, CY006817, CY006825, CY006833, CY006889, CY006913, CY007977, CY008466, CY008690, CY008698, CY008706, CY009354, CY009362, CY019905, CY019921, CY021083, CY021091, CY021099, CY021123, CY021603, CY021835, CY021843, CY021947, CY021963, CY022944, CY022952, CY026145, DQ508927, CY002094, CY002758, CY003358, CY003494, CY003510, CY003518, CY003526, CY003542, CY003550, CY003726, CY003742, CY003750, CY006058, CY006105, CY006209, CY006321, CY006329, CY006761, CY006769, CY006857, CY006897, CY007625, CY008178, CY008458, CY008474, CY008674, CY008714, CY008722, CY008730, CY008738, CY008746, CY009058, CY009074, CY009290, CY009298, CY009626, CY010370, CY010762, CY010874, CY010890, CY010898, CY010914, CY010922, CY010962, CY011486, CY012886, CY012894, CY013885, CY015530, CY017209, CY017217, CY017241, CY017249, CY017369, CY017441, CY019051,

CY019059, CY019067, CY019107, CY019227, CY019745, CY019761, CY019777, CY019785, CY019969, CY020171, CY020179, CY020187, CY020195, CY020227, CY020243, CY020299, CY020339, CY020443, CY020459, CY020483, CY020491, CY020571, CY020579, CY021035, CY021043, CY021107, CY021723, CY021731, CY021739, CY021915, CY021979, CY024931, CY026417, CY027537, CY028730, DQ508823, DQ508831, DQ508847, DQ508871, M25934, M25936, M38376, AF037418, AF037419, AF037420, AF037421, AF037422, AF037423, AF258526, AF258527, AF258822, AF258823, AF342823, AF398871, AF483601, CY000456, CY000463, CY000471, CY000599, CY000607, CY000615, CY000623, CY000639, CY000647, CY000655, CY000663, CY000671, CY000679, CY000695, CY000703, CY000711, CY000719, CY000727, CY000735, CY000743, CY000751, CY000807, CY000823, CY000831, CY000995, CY001003, CY001011, CY001142, CY001182, CY001251, CY001283, CY001387, CY001403, CY001419, CY001451, CY001459, CY001483, CY001491, CY001499, CY001510, CY001526, CY001534, CY001574, CY001582, CY001590, CY001598, CY001614, CY001622, CY001662, CY001670, CY001710, CY001750, CY001774, CY001790, CY001814, CY001830, CY001838, CY001846, CY001870, CY001878, CY001894, CY001902, CY001910, CY001918, CY001934, CY001942, CY001974, CY001990, CY002118, CY002142, CY002150, CY002166, CY002174, CY002278, CY002310, CY002342, CY002518, CY002550, CY002566, CY002582, CY002646, CY002654, CY003222, CY003238, CY003246, CY003254, CY003262, CY003270, CY003286, CY003438, CY003446, CY003462, CY003566, CY003574, CY003590, CY003598, CY003606, CY003614, CY003630, CY003638, CY003718, CY003791, CY003807, CY003815, CY003831, CY006066, CY006074, CY006233, CY006241, CY006257, CY006265, CY006273, CY006289, CY006337, CY006345, CY006449, CY006457, CY006473, CY006481, CY006489, CY006497, CY006505, CY006513, CY006521, CY006529, CY006537, CY006545, CY006553, CY006561,

CY006569, CY006577, CY006585, CY006593, CY006617, CY006625, CY006633, CY006641, CY006665, CY006777, CY006793, CY006801, CY006905, CY007985, CY008137, CY008186, CY008506, CY008514, CY008786, CY008794, CY008802, CY008818, CY008826, CY008842, CY008850, CY008858, CY008866, CY008962, CY008970, CY008978, CY008986, CY009082, CY009090, CY009106, CY009114, CY009122, CY009130, CY009138, CY009146, CY009154, CY009170, CY009178, CY009186, CY009194, CY009202, CY009226, CY009322, CY009466, CY009474, CY009482, CY009490, CY009498, CY009506, CY009514, CY009522, CY009530, CY009538, CY009554, CY009650, CY009658, CY009666, CY009682, CY009690, CY009698, CY009706, CY009714, CY009730, CY009738, CY009746, CY009754, CY009778, CY009794, CY009802, CY009818, CY009834, CY009842, CY009850, CY009906, CY009914, CY009946, CY009994, CY010002, CY010010, CY010018, CY010026, CY010034, CY010042, CY010058, CY010074, CY010106, CY010122, CY010130, CY010146, CY010394, CY010490, CY010498, CY010506, CY010514, CY010522, CY010530, CY010538, CY010546, CY010594, CY010602, CY010610, CY010626, CY010634, CY010642, CY010658, CY010690, CY010698, CY010706, CY010722, CY010730, CY010738, CY010746, CY010754, CY010810, CY010818, CY010826, CY010834, CY010842, CY010850, CY010994, CY011142, CY011270, CY011286, CY011326, CY011334, CY011342, CY011350, CY011358, CY011374, CY011382, CY011390, CY011422, CY011430, CY011438, CY011446, CY011462, CY011470, CY011494, CY011502, CY011518, CY011526, CY011534, CY011542, CY011550, CY011558, CY011566, CY011574, CY011582, CY011590, CY011598, CY011782, CY011798, CY011806, CY011814, CY011822, CY011830, CY011838, CY011846, CY011854, CY011862, CY011870, CY011886, CY011894, CY011902, CY011910, CY011926, CY011934, CY011942, CY011958, CY012126, CY012134, CY012142, CY012150, CY012158, CY012166, CY012174, CY012182, CY012190, CY012206, CY012214,

CY012230, CY012238, CY012246, CY012254, CY012262, CY012270, CY012278, CY012294, CY012454, CY012462, CY012470, CY012478, CY012486, CY012494, CY012502, CY012510, CY012526, CY012534, CY012542, CY012550, CY012558, CY012566, CY012574, CY012582, CY012590, CY012598, CY012614, CY012622, CY012630, CY012638, CY012734, CY012742, CY012750, CY012758, CY012766, CY012774, CY012790, CY012862, CY012870, CY012878, CY012902, CY012910, CY012918, CY012942, CY012966, CY012990, CY012998, CY013006, CY013014, CY013030, CY013046, CY013054, CY013062, CY013070, CY013174, CY013182, CY013190, CY013198, CY013206, CY013214, CY013285, CY013293, CY013309, CY013317, CY013333, CY013341, CY013357, CY013365, CY013373, CY013403, CY013611, CY013619, CY013635, CY013651, CY013675, CY013683, CY013691, CY013699, CY013707, CY013723, CY013739, CY013755, CY013763, CY013819, CY013861, CY013877, CY013893, CY013909, CY015522, CY015538, CY015546, CY015658, CY015674, CY016074, CY016082, CY016122, CY016202, CY016210, CY016234, CY016242, CY016266, CY016274, CY016409, CY016433, CY016497, CY016505, CY016521, CY016529, CY016537, CY016545, CY016553, CY016569, CY016585, CY016633, CY016641, CY016649, CY016657, CY016673, CY016713, CY016721, CY016729, CY016745, CY016753, CY016761, CY016969, CY016977, CY017025, CY017129, CY017145, CY017153, CY017273, CY017289, CY017305, CY017329, CY017353, CY017401, CY017449, CY017457, CY017465, CY017473, CY017811, CY017891, CY019115, CY019131, CY019139, CY019793, CY019801, CY019809, CY019873, CY019995, CY020219, CY020259, CY020499, CY020515, CY021011, CY021635, CY021699, CY021747, CY021755, CY021779, CY021787, CY022163, CY022259, CY022507, CY022515, CY022523, CY023032, CY023040, CY023048, CY023056, CY025024, CY025048, DQ415294, DQ415295, DQ415296, DQ487328, DQ487333, DQ508855, DQ508863, U71128, U71129, U71130, U71131,

CY000007, CY000015, CY000023, CY000031, CY000039, CY000047, CY000055, CY000071, CY000079, CY000087, CY000095, CY000103, CY000111, CY000119, CY000127, CY000135, CY000143, CY000167, CY000175, CY000183, CY000191, CY000199, CY000207, CY000215, CY000223, CY000231, CY000239, CY000247, CY000263, CY000271, CY000279, CY000287, CY000295, CY000311, CY000327, CY000335, CY000351, CY000359, CY000375, CY000383, CY000391, CY000399, CY000407, CY000415, CY000423, CY000431, CY000439, CY000447, CY000479, CY000487, CY000511, CY000518, CY000535, CY000543, CY000559, CY000567, CY000575, CY000759, CY000767, CY000783, CY000791, CY000799, CY000871, CY000879, CY000887, CY000895, CY000915, CY000931, CY000955, CY000963, CY000971, CY000979, CY001019, CY001027, CY001043, CY001051, CY001059, CY001070, CY001102, CY001110, CY001134, CY001150, CY001166, CY001174, CY001190, CY001195, CY001203, CY001211, CY001219, CY001227, CY001235, CY001243, CY001259, CY001267, CY001291, CY001307, CY001323, CY001331, CY001339, CY001347, CY001379, CY001411, CY001427, CY001435, CY001475, CY001542, CY001550, CY001630, CY001638, CY001686, CY001726, CY001734, CY001742, CY001950, CY001958, CY002006, CY002014, CY002022, CY002030, CY002038, CY002046, CY002054, CY002062, CY002070, CY002078, CY002086, CY002110, CY002134, CY002158, CY002182, CY002190, CY002198, CY002206, CY002214, CY002222, CY002230, CY002238, CY002246, CY002270, CY002286, CY002358, CY002366, CY002398, CY002406, CY002414, CY002422, CY002430, CY002438, CY002446, CY002454, CY002462, CY002470, CY002478, CY002486, CY002510, CY002526, CY002534, CY002542, CY002574, CY002590, CY002598, CY002606, CY002622, CY002630, CY002638, CY002662, CY002678, CY002686, CY002694, CY002702, CY002710, CY002718, CY002726, CY002734, CY002742, CY002774, CY002806, CY002814, CY002822, CY002912, CY002920, CY002928, CY002952, CY002960,

CY002968, CY002975, CY002982, CY002990, CY003006, CY003014, CY003030, CY003038, CY003046, CY003062, CY003078, CY003086, CY003102, CY003110, CY003118, CY003128, CY003134, CY003142, CY003174, CY003182, CY003190, CY003198, CY003206, CY003214, CY003294, CY003302, CY003310, CY003318, CY003326, CY003334, CY003342, CY003350, CY003374, CY003390, CY003398, CY003406, CY003414, CY003430, CY003478, CY003486, CY003654, CY003662, CY003670, CY003678, CY003686, CY003694, CY003702, CY003710, CY003767, CY003775, CY003783, CY006082, CY006090, CY006098, CY006113, CY006121, CY006129, CY006137, CY006145, CY006153, CY006161, CY006193, CY006201, CY006297, CY006361, CY006377, CY006393, CY006401, CY006409, CY006417, CY006425, CY006433, CY006441, CY006673, CY006681, CY006865, CY006873, CY006921, CY006929, CY006937, CY006945, CY006977, CY007001, CY007009, CY007025, CY007033, CY007065, CY007089, CY007097, CY007105, CY007113, CY007121, CY007129, CY007137, CY007161, CY007169, CY007177, CY007201, CY007209, CY007233, CY007249, CY007257, CY007273, CY007281, CY007297, CY007313, CY007321, CY007337, CY007353, CY007361, CY007409, CY007425, CY007433, CY007441, CY007449, CY007457, CY007473, CY007481, CY007489, CY007497, CY007505, CY007521, CY007545, CY007561, CY007569, CY007585, CY007593, CY007601, CY007609, CY007697, CY007721, CY007737, CY007753, CY007761, CY007769, CY007777, CY007785, CY007801, CY007865, CY007873, CY007929, CY007937, CY007945, CY007953, CY007961, CY007993, CY008017, CY008033, CY008041, CY008049, CY008057, CY008065, CY008073, CY008081, CY008089, CY008105, CY008113, CY008154, CY008170, CY008218, CY008242, CY008258, CY008266, CY008282, CY008306, CY008330, CY008346, CY008362, CY008370, CY008402, CY008410, CY008434, CY008442, CY008450, CY008530, CY008554, CY008578, CY008610, CY008618, CY008626, CY008634, CY008642, CY008658, CY008874,

CY008882, CY008890, CY008906, CY008914, CY009002, CY009242, CY009250, CY009266, CY009402, CY009410, CY009418, CY009442, CY009594, CY009866, CY009874, CY009882, CY009890, CY009954, CY009962, CY009970, CY010154, CY010162, CY010178, CY010202, CY010218, CY010234, CY010258, CY010266, CY010274, CY010282, CY010298, CY010306, CY010314, CY010322, CY010330, CY010338, CY010346, CY010362, CY010402, CY010410, CY010418, CY010426, CY010434, CY010442, CY010482, CY010554, CY010786, CY010858, CY011078, CY011086, CY011094, CY011158, CY011174, CY011182, CY011206, CY011222, CY011406, CY011614, CY011630, CY011646, CY011654, CY011670, CY011686, CY011694, CY011702, CY011734, CY011750, CY011758, CY011766, CY011774, CY011974, CY011982, CY011990, CY012014, CY012022, CY012038, CY012054, CY012062, CY012078, CY012102, CY012110, CY012310, CY012318, CY012350, CY012358, CY012374, CY012398, CY012406, CY012422, CY012430, CY012646, CY012662, CY012678, CY012686, CY012694, CY012710, CY012718, CY012726, CY012798, CY013078, CY013086, CY013094, CY013102, CY013134, CY013222, CY013230, CY013238, CY013427, CY013435, CY013451, CY013459, CY013467, CY013475, CY013483, CY013491, CY013499, CY013507, CY013531, CY013539, CY013547, CY013555, CY013563, CY013571, CY013579, CY013587, CY013603, CY013811, CY013917, CY013949, CY013981, CY013997, CY014013, CY014021, CY014037, CY014045, CY014069, CY014077, CY014085, CY014093, CY014101, CY014109, CY014141, CY014165, CY015554, CY015586, CY015602, CY015610, CY015618, CY015682, CY015690, CY015698, CY015706, CY015714, CY015722, CY015730, CY015738, CY015746, CY015754, CY015762, CY015770, CY015778, CY015786, CY015794, CY015802, CY015810, CY015826, CY015834, CY015842, CY015850, CY015858, CY015866, CY015882, CY015890, CY015898, CY015914, CY015922, CY015930, CY015938, CY015962, CY015970, CY015978, CY015986, CY015994, CY016002,

CY016010, CY016018, CY016026, CY016034, CY016042, CY016050, CY016218, CY016449, CY016457, CY016465, CY016481, CY016601, CY016609, CY016665, CY016681, CY016689, CY016697, CY016705, CY016985, CY017089, CY017097, CY017105, CY017113, CY017121, CY017137, CY017361, CY017377, CY017489, CY017497, CY017505, CY017513, CY017521, CY017529, CY017545, CY017561, CY017569, CY017585, CY017593, CY017601, CY017609, CY017617, CY017625, CY017803, CY017827, CY017899, CY017907, CY017915, CY017923, CY017931, CY017939, CY017947, CY017955, CY017971, CY017987, CY017995, CY018931, CY018947, CY018971, CY018979, CY019019, CY019027, CY019035, CY019195, CY019331, CY019339, CY019347, CY019753, CY019833, CY019865, CY019881, CY019889, CY019929, CY019937, CY020011, CY020035, CY020051, CY020147, CY020155, CY020267, CY020275, CY020283, CY020307, CY020347, CY020427, CY020899, CY021763, CY021771, CY022195, CY022203, CY022211, CY022531, CY022563, CY022587, CY022603, CY023072, CY023080, CY025056, CY025219, CY025227, CY025235, CY025243, CY025251, CY025267, CY025275, CY025283, CY025291, CY025299, CY025307, CY025323, CY025331, CY025339, CY025347, CY025355, CY025371, CY025379, CY025387, CY025395, CY025411, CY025419, CY025427, CY025435, CY025443, CY025451, CY025459, CY025475, CY025483, CY025491, CY025499, CY025523, CY025531, CY025539, CY025553, CY025561, CY025577, CY025593, CY025601, CY025625, CY025641, CY025649, CY025665, CY025673, CY025681, CY025689, CY025697, CY025705, CY025713, CY025729, CY025737, CY025745, CY025753, CY025761, CY025777, CY025793, CY025801, CY025809, CY025817, CY025833, CY025841, CY025849, CY025865, CY025905, CY025913, CY025929, CY025937, CY025945, CY025953, CY025961, CY025969, CY025985, CY025993, CY026001, CY026009, CY026025, CY026153, CY026169, CY026185, CY026193, CY026201, CY026209, CY026225, CY026241, CY026249, CY026257, CY026265,

CY026273, CY026281, CY026321, CY026337, CY026353, CY026361, CY026377, CY026385, CY026401, CY026409, CY026505, CY026521, CY026529, CY026537, CY026545, CY026561, CY026569, CY026585, CY026593, CY026633, CY026641, CY026649, CY026657, CY026673, CY026681, CY026689, CY026697, CY026705, CY026713, CY026721, CY026737, CY026745, CY026769, CY026777, CY026793, CY026801, CY026825, CY026833, CY026841, CY026849, CY026865, CY026873, CY026889, CY026897, CY026905, CY026913, CY026921, CY026929, CY026937, CY026945, CY026953, CY026969, CY026977, CY026993, CY027001, CY027009, CY027017, CY027025, CY027041, CY027049, CY027073, CY027081, CY027089, CY027097, CY027121, CY027137, CY027145, CY027153, CY027177, CY027185, CY027193, CY027209, CY027225, CY027233, CY027265, CY027273, CY027281, CY027289, CY027329, CY027345, CY027353, CY027377, CY027385, CY027393, CY027417, CY027425, CY027433, CY027441, CY027481, CY027497, CY027505, CY027545, CY027569, CY027585, CY027601, CY027609, CY027633, CY027641, CY027649, CY027657, CY027665, CY027673, CY027689, CY027705, CY027713, CY027729, CY027737, CY027753, CY027761, CY027769, CY027777, CY027785, CY027793, CY027801, CY027809, CY027817, CY027825, CY027833, CY027841, CY027849, CY027857, CY027873, CY027889, CY027897, CY027905, CY027913, CY027921, CY027945, CY027953, CY027985, CY027993, CY028009, CY028033, CY028065, CY028081, CY028097, CY028105, CY028113, CY028121, CY028129, CY028137, CY028145, CY028201, CY028209, CY028217, CY028305, CY028321, CY028329, CY028337, CY028345, CY028385, CY028409, CY028457, CY028465, CY028473, CY028481, CY028738, CY028754, CY028762, CY028770, CY028778, CY030067, CY030075, CY030203, CY030211, DQ415297, DQ415298, DQ415299, DQ415300, DQ415301, DQ415302, DQ415303, DQ415304, DQ469956, DQ889683, EF554793, EU399757, AB212052, AY576392, AY626148, AY627891, AY627897,

AY818129, CY014170, CY014176, CY014270, CY014278, CY014286, CY014294, CY014301, CY014309, CY014319, CY014323, CY014324, CY014331, CY014335, CY014336, CY014349, CY014351, CY014355, CY014365, CY014390, CY014395, CY014403, CY014419, CY014427, CY014443, CY014451, CY014459, CY014475, CY014483, CY014491, CY014499, CY014505, CY014527, CY014535, CY015012, CY017636, CY017652, CY017660, CY017668, CY017684, CY017686, CY019414, DQ138159, DQ138160, DQ138163, DQ138164, DQ138165, DQ360838, DQ835311, EF137712, EF467804, EF467808, EF587275, EF619986, EF620010, EU146630, EU146654, EU146662, EU146703, EU146719, EU146772, EU146799, EU146815, EU146831, EU146853, M74899, AAO15323, ABQ45433, ABQ45444, ABQ45541, ABR15827, ABR15871, ABR28556, ABR28567, ABR28633, ABR28655, ABR28666, ABR28710, ABR28732, ABS50129, ABS49962, ABU80240, ABU80295, ABU80428, ABU80418, ABW36341, ABW86582, ABX58654, ABQ41896, ABQ51943, ABV31976, BAG49740, BAG49627, ABQ45466, ABR15838, ABR28589, ABR29623, ABR29603, ABR28611, ABR28622, ABR28677, ABR29583, ABS50119, ABS49940, ABS49951, ABU80262, ABW36385, ABW36396, ABR87896, AAG01751, AAG01787, ABR28578, ABR28699, ABR29573, ABU80229, ABU80284, ABV29532, ABW36330, ABW38018, ABW71511, ABS53361, AAA43640, AAO15324, ABR28545, ABR28644, ABR28688, ABR28721, ABR29593, ABU80251, ABU80218, ABW36363, ABW36374, ABS00328, AAF76001, AAL87927, ACA25356, AAL87921, CAC37001, CAC37003, AAN46833, ABV55857, ABY16780, ACA25366, AAG01744, AAL87923, AAL87924, AAL87925, ABY16777, ACD65218, ACD65219, ABY84692, ABY40445, ABY16778, ABY16779, ABY81661, AAU05322, ACA96538, ABA46964, ABA27430, ABB86881, ABB86921, ABB86941, ABB86954, ABD77119, ABF18009, ABF18010, ACA42436, ACE78091, ACE78112,

ACE78116, ABE27161, ABV82581, ABW86593, AAA43646, AAV30842, ABY51223, ACD85162, ABB86871, ABB86901, ABB86931, ABB86950, ABD77111, ABG67729, ABG67735, ACA42426, ACE78095, ACE78097, ACE78106, ACE78122, ABD79263, ABE12642, ABV82603, AAU25853, ABY81434, ABF18012, ABG67724, ABG67728, ABG67733, ABG67737, ABG67741, ABJ16471, ACE78099, ACE78101, ACE78110, ACE78120, ACE78124, ABD62841, AAV30834, AAV68029, ABY51212, ABA27438, ABB86955, ABB86958, ABG67739, ABI54400, ABJ15718, ACE78102, ACE78114, ABD62802, ABX58676, ABW71529, ABS53351, AAL87926, ACD65217, AAA43636, AAU25843, ABB86891, ABK00142, ABR28765, ABU80207, ABV82592, ABW71489, ABG67731, ACE78108, ABD61559, ABS00317, AAL87922, AAZ79400, ABG67726, AAG17436, AAL87928, ABY40435, ACE78089, ACE78104, AAO15325, ABR28743, ABV31975, ABG67744, ABO44043, ABB86911, ABG67742, ACE78087, AAO15322, ABS49929, ABX58665, CAC37002, AAU25863, ABF18011, ABD78112. For the pandemic H1N1 sequences: CY081091, CY071333, CY071341, CY071349, CY071357, CY071365, CY083014, CY083022, CY083030, CY083038, CY063828, CY063838, CY063846, CY080313, CY080329, CY073966, CY073004, CY083533, CY073124, CY062266, CY076729, CY076737, CY076745, CY080297, CY080321, CY080337, CY071141, CY071149, CY071157, CY075065, CY075073, CY071165, CY071173, CY071181, CY071189, CY071269, CY071277, CY071285, CY071293, CY075081, CY071301, CY071309, CY071317, CY071325, CY071381, CY062985, CY066453, CY063185, CY063193, CY063201, CY063209, CY063217, CY066461, CY066469, CY066477, CY066485, CY064993, CY066493, CY066501, CY066509, CY071213, CY071221, GU562464, CY065878, CY071373, CY081059, CY081067, CY081075, CY081099, HQ393491, CY079542, CY063009, CY063017, CY071077, CY072532, CY073739, CY061584, CY061113,

CY061592, CY061121, CY061129, CY061137, CY061145, CY061153, CY061161, CY061169, CY061177, CY061185, CY061193, CY061201, CY061209, CY080572, CY062993, CY071061, CY073779, HQ891285, CY083823, HQ834744, CY062072, CY062080, CY062088, CY062096, CY062104, CY062112, CY062120, CY062128, CY062136, CY062144, CY062152, CY062160, CY062168, CY062176, CY062184, CY062192, CY065001, CY065009, CY065017, CY065025, CY065033, CY065041, CY065049, CY065057, CY065065, CY065073, CY065113, CY072540, HM124381, HM569738, HM189302, CY061803, CY061811, HM189527, HM189528, HM189529, HM189530.

For segment 3: AB166861, AB188815, AB188823, AB189059, AB256665, AB256673, AB256681, AB256689, AB256697, AB256705, AB256713, AB256721, AB256729, AB256745, AB261852, AF098604, AF098605, AF098606, AF098607, AF098608, AF098612, AF098614, AF098615, AF262211, AF457683, AF457699, AF457716, AF508662, AF508663, AF508665, AF508668, AF508669, AF508670, AF508671, AF508672, AF508673, AF508674, AF508675, AF508677, AF508678, AF508681, AF509197, AJ291397, AJ410511, AJ410512, AJ410513, AJ410514, AJ410515, AJ410516, AJ427307, AJ619677, AM262535, AM262536, AM262537, AM262538, AM262539, AM262540, AM503046, AM503047, AM503048, AM503049, AM503050, AM503051, AM503053, AY253752, AY303660, AY303661, AY551934, AY576410, AY576411, AY576412, AY576415, AY609311, AY616764, AY646083, AY648292, AY651623, AY653198, AY676031, AY684705, AY724260, AY737288, AY737295, AY770082, AY818133, AY818134, AY849786, AY849787, AY862680, AY950265, AY950266, AY950267, CY005451, CY005458, CY005504, CY005610, CY005788, CY005793, CY005830, CY005843, CY005893, CY005911, CY014184, CY014645, CY014654, CY014668, CY014676, CY014769, CY014776, CY014783, CY014791, CY014797,

CY014803, CY014826, CY014834, CY014841, CY014885, CY014914, CY014996, CY015018, CY015024, CY015031, CY015037, CY015051, CY015057, CY015070, CY015078, CY015086, CY015094, CY015101, CY015124, CY016281, CY016289, CY016297, CY016305, CY016792, CY016816, CY016840, CY016856, CY016872, CY016880, CY016912, CY016920, CY016928, CY016936, CY016944, CY016952, CY017056, CY017064, CY017184, CY017408, CY018954, CY020234, CY020586, CY020594, CY020602, CY020610, CY020626, CY020634, CY020642, CY020650, CY020658, CY020674, CY020682, CY020690, CY020698, CY021362, CY021370, CY021378, CY021386, CY021394, CY021410, CY021418, CY021426, CY021490, CY021498, CY021514, CY021530, CY021538, CY021546, CY021554, CY022082, CY022266, CY022618, CY022626, CY022634, CY022658, CY022666, CY022674, CY022682, CY022690, CY022698, CY022706, CY022714, CY022770, CY022818, CY022826, CY022834, CY024743, CY024759, CY024767, CY024775, CY024791, CY024863, CY024871, CY024879, CY024887, CY024895, CY025074, CY025082, CY025114, CY025130, CY025146, CY025162, CY025170, CY025178, CY025186, CY025194, CY028512, CY028545, CY028577, CY028585, CY028601, CY028633, CY028673, CY028681, CY028713, CY029195, CY029237, CY029244, CY029251, CY029258, CY029433, CY029461, CY029468, CY029475, CY029482, CY029910, CY029918, CY029947, CY029963, CY029995, DQ064490, DQ064491, DQ064492, DQ064493, DQ064494, DQ064495, DQ064496, DQ064497, DQ064498, DQ064499, DQ064500, DQ064501, DQ064502, DQ064503, DQ064504, DQ064506, DQ064510, DQ064512, DQ064513, DQ064514, DQ064515, DQ067442, DQ099786, DQ099788, DQ323674, DQ334759, DQ334767, DQ334775, DQ335776, DQ351867, DQ351868, DQ351869, DQ366329, DQ376799, DQ376800, DQ376801, DQ376802, DQ376804, DQ376806, DQ376808, DQ376810, DQ376811, DQ376812, DQ376813, DQ376814, DQ376815, DQ376816,

DQ376817, DQ376818, DQ376819, DQ376820, DQ376821, DQ376822, DQ376824, DQ376825, DQ376827, DQ376828, DQ376829, DQ376832, DQ376833, DQ449637, DQ469997, DQ485207, DQ485215, DQ650668, DQ914813, DQ991303, DQ991311, DQ991327, DQ991335, DQ991342, DQ997106, DQ997107, DQ997118, DQ997128, DQ997137, DQ997160, DQ997186, DQ997191, DQ997274, DQ997303, DQ997321, DQ997328, DQ997551, DQ999885, EF063545, EF063547, EF063548, EF063549, EF063550, EF063551, EF070738, EF124653, EF124658, EF124664, EF124690, EF124691, EF124692, EF124698, EF124701, EF124702, EF124712, EF124714, EF124719, EF124721, EF124761, EF124762, EF124766, EF124780, EF124791, EF124793, EF155272, EF155278, EF155280, EF155283, EF155284, EF155285, EF205189, EF205190, EF205193, EF205194, EF205195, EF362423, EF441266, EF474445, EF523724, EF523727, EF551044, EF597411, EF597413, EF605602, EF681869, EF681877, EU081869, EU084905, EU084911, EU084915, EU084918, EU084922, EU084925, EU084929, EU084947, EU086231, EU086242, EU086258, EU086294, EU146856, EU148361, EU148369, EU148377, EU148385, EU148393, EU148401, EU148409, EU148417, EU148425, EU148433, EU148441, EU148449, EU163434, EU182260, EU182264, EU182272, EU182287, EU182291, EU182295, EU182299, EU182321, EU233415, EU233680, EU233696, EU233728, EU277838, EU277846, EU365370, EU402403, EU408333, EU429971, EU429972, EU429973, EU429975, M21850, M26084, AB049157, AB049158, AB212279, AB239302, AB239309, AB259711, AB262462, AB263751, AB274965, AB284323, AB284987, AB286118, AB286878, AB300228, AB300435, AB300439, AB301914, AB304146, AF098609, AF098610, AF144302, AF250478, AF380163, AF508666, AF508682, AJ243994, AY233389, AY518365, AY585462, AY585463, AY585464, AY585465, AY585466, AY585467, AY585468, AY585469, AY585470, AY585471, AY585472, AY585473, AY585474,

AY585475, AY585476, AY585478, AY585479, AY585480, AY585481, AY585482, AY633121, AY676029, AY676030, AY676032, AY684883, AY703831, AY724254, AY737303, AY856863, AY950270, AY950271, CY003852, CY003860, CY003868, CY003876, CY003884, CY003892, CY003899, CY003904, CY003912, CY003919, CY003927, CY003934, CY003941, CY003949, CY003957, CY003965, CY003973, CY003981, CY003989, CY003997, CY004004, CY004009, CY004016, CY004023, CY004059, CY004099, CY004105, CY004111, CY004119, CY004126, CY004134, CY004140, CY004151, CY004159, CY004167, CY004183, CY004191, CY004207, CY004215, CY004223, CY004231, CY004255, CY004263, CY004271, CY004279, CY004287, CY004294, CY004304, CY004322, CY004336, CY004343, CY004350, CY004357, CY004364, CY004377, CY004384, CY004403, CY004417, CY004425, CY004431, CY004438, CY004443, CY004448, CY004455, CY004463, CY004471, CY004479, CY004487, CY004495, CY004512, CY004520, CY004536, CY004543, CY004551, CY004559, CY004565, CY004572, CY004597, CY004604, CY004611, CY004618, CY004639, CY004647, CY004654, CY004667, CY004675, CY004685, CY004697, CY004707, CY004714, CY004719, CY004726, CY004740, CY004747, CY004754, CY004760, CY004767, CY004774, CY004781, CY004788, CY004795, CY004802, CY004822, CY004828, CY004835, CY004851, CY004858, CY004865, CY004872, CY004884, CY004889, CY004894, CY004901, CY004908, CY004916, CY004930, CY004937, CY004944, CY004951, CY004958, CY004965, CY004970, CY004974, CY004981, CY004986, CY004992, CY004996, CY005000, CY005006, CY005013, CY005020, CY005036, CY005048, CY005055, CY005062, CY005069, CY005076, CY005083, CY005089, CY005118, CY005124, CY005131, CY005145, CY005151, CY005158, CY005168, CY005171, CY005180, CY005185, CY005192, CY005197, CY005204, CY005216, CY005222, CY005229, CY005236, CY005241, CY005247, CY005254, CY005261, CY005268, CY005278, CY005284,

CY005290, CY005302, CY005315, CY005322, CY005328, CY005335, CY005341, CY005348, CY005355, CY005362, CY005369, CY005374, CY005380, CY005385, CY005390, CY005397, CY005400, CY005410, CY005418, CY005425, CY005435, CY005442, CY005473, CY005491, CY005498, CY005510, CY005528, CY005535, CY005543, CY005551, CY005557, CY005567, CY005572, CY005580, CY005587, CY005594, CY005602, CY005616, CY005622, CY005629, CY005636, CY005644, CY005650, CY005657, CY005663, CY005669, CY005676, CY005684, CY005688, CY005696, CY005702, CY005707, CY005713, CY005734, CY005739, CY005751, CY005754, CY005758, CY005762, CY005774, CY005781, CY005811, CY005817, CY005820, CY005825, CY005849, CY005856, CY005863, CY005871, CY005878, CY005886, CY005900, CY005906, CY011033, CY011045, CY011053, CY011061, CY011117, CY011253, CY012805, CY012813, CY012821, CY012829, CY012837, CY012845, CY013252, CY013260, CY013268, CY013868, CY014523, CY014553, CY014566, CY014575, CY014589, CY014684, CY014699, CY014707, CY014714, CY014730, CY014736, CY014744, CY014811, CY014818, CY014854, CY014862, CY014877, CY014893, CY014906, CY014920, CY014926, CY014934, CY014942, CY014950, CY014973, CY014989, CY015044, CY015062, CY015132, CY015140, CY015150, CY015153, CY015162, CY015166, CY015172, CY015175, CY015448, CY015456, CY015464, CY015481, CY015489, CY015497, CY015505, CY016129, CY016137, CY016145, CY016161, CY016169, CY016177, CY016185, CY016193, CY016400, CY016416, CY016424, CY016616, CY016784, CY016800, CY016808, CY016824, CY016832, CY016888, CY016896, CY016904, CY016960, CY017032, CY017040, CY017048, CY017072, CY017080, CY017280, CY017416, CY017698, CY017706, CY017714, CY017722, CY017730, CY017738, CY017746, CY017754, CY017762, CY017770, CY017778, CY017786, CY017794, CY017850, CY017858, CY017866, CY018004, CY018012, CY018020, CY018882, CY018890, CY018898, CY018906, CY018914,

CY018922, CY019202, CY020354, CY020722, CY020730, CY020738, CY020746, CY020754, CY020770, CY020786, CY020794, CY020810, CY020818, CY020826, CY020834, CY020842, CY020850, CY020866, CY020874, CY020882, CY020890, CY020906, CY020914, CY020930, CY020938, CY020946, CY020954, CY020962, CY020970, CY020978, CY020994, CY021130, CY021138, CY021146, CY021170, CY021178, CY021186, CY021202, CY021210, CY021218, CY021226, CY021242, CY021250, CY021258, CY021266, CY021274, CY021282, CY021298, CY021306, CY021322, CY021338, CY021346, CY021354, CY021402, CY021434, CY021450, CY021458, CY021466, CY021474, CY021482, CY021562, CY021570, CY021578, CY021586, CY021594, CY021618, CY021626, CY021642, CY021650, CY021658, CY021666, CY021674, CY021682, CY021690, CY021866, CY021874, CY021898, CY022610, CY022642, CY022650, CY022722, CY022746, CY024751, CY024783, CY024799, CY024903, CY025202, CY028240, CY028256, CY028264, CY028649, CY028697, CY029132, CY029230, CY029314, CY029321, CY029405, CY029412, CY029419, CY029440, CY029447, CY029454, CY029489, CY029846, CY029878, CY029894, CY029926, CY029934, DQ017491, DQ017508, DQ017512, DQ064509, DQ073407, DQ073408, DQ073409, DQ073410, DQ095714, DQ095718, DQ095721, DQ138182, DQ138183, DQ234075, DQ234076, DQ251443, DQ251451, DQ343504, DQ358748, DQ363917, DQ363922, DQ366305, DQ366313, DQ366321, DQ366337, DQ376798, DQ376805, DQ376807, DQ376809, DQ376823, DQ376826, DQ376830, DQ376831, DQ394579, DQ399537, DQ407245, DQ434891, DQ449645, DQ454103, DQ464360, DQ465399, DQ525416, DQ681206, DQ681210, DQ681218, DQ681223, DQ822189, DQ822197, DQ835785, DQ835786, DQ835787, DQ835789, DQ835790, DQ835791, DQ852603, DQ864508, DQ864712, DQ978999, DQ989968, DQ989976, DQ989986, DQ990002, DQ997167, DQ997176, DQ997280, DQ997401,

DQ997414, DQ997455, DQ997517, DQ997526, DQ997535, EF061120, EF112247, EF112248, EF112250, EF112251, EF112254, EF112259, EF112262, EF112266, EF112268, EF124645, EF124647, EF124655, EF124656, EF124657, EF124659, EF124660, EF124662, EF124668, EF124670, EF124672, EF124679, EF124680, EF124681, EF124682, EF124683, EF124684, EF124693, EF124694, EF124695, EF124696, EF124699, EF124700, EF124703, EF124705, EF124709, EF124710, EF124711, EF124716, EF124718, EF124720, EF124722, EF124723, EF124742, EF124743, EF124744, EF124745, EF124747, EF124750, EF124752, EF124753, EF124754, EF124757, EF124763, EF124764, EF124765, EF124767, EF124768, EF124771, EF124772, EF124773, EF124774, EF124775, EF124776, EF124778, EF124784, EF124785, EF124786, EF124787, EF124788, EF124789, EF124790, EF124792, EF205191, EF205192, EF210571, EF523725, EF523728, EF523729, EF523730, EF523731, EF523732, EF530048, EF554802, EF592494, EF597404, EF597405, EF597406, EF597407, EF597408, EF597414, EF597415, EF597416, EF597421, EF597429, EF634331, EF634339, EU026019, EU026026, EU026034, EU026042, EU026049, EU026055, EU026063, EU026071, EU026087, EU026095, EU026107, EU026115, EU030967, EU030975, EU030983, EU084938, EU084942, EU084951, EU086259, EU094472, EU158155, EU182323, EU233720, EU263344, EU263352, EU296251, EU329176, EU329187, EU430501, EU430508, EU441924, EU441932, EU486850, M26083, M26085, M26086, M26088, AF389117, CY009329, CY009449, CY009609, CY010793, CY019960, CY020450, CY020474, DQ208311, DQ508904, EF467820, V01106, X17336, AY209990, AY209991, CY008993, CY009281, CY009337, CY009345, CY009457, CY009601, CY009617, CY013276, CY014981, CY019952, CY019976, CY020290, CY020466, CY021706, CY021714, CY021810, CY021826, CY021906, CY022018, CY022026, DQ508840, EF633440, M26078, M81573, M81579, X99039, AF348174, AY209992, AY209993, AY209994,

AY209995, AY209996, AY209997, AY209998, AY209999, AY210000, AY210001, AY210002, AY210003, AY210004, AY210005, AY210006, AY210007, AY210193, AY210194, AY210195, CY006216, CY008161, CY011125, CY015513, CY019896, CY019912, CY020322, CY020378, CY020394, CY020418, CY020522, CY020546, CY021026, CY021074, CY021114, CY021794, CY021818, CY021850, CY021938, CY022090, DQ508880, J02139, M23974, M26079, AX350188, AY210196, AY210197, AY210198, AY210199, AY210200, AY210201, AY210202, AY210203, AY210204, AY210205, AY210206, CY002101, CY002501, CY002749, CY003501, CY003533, CY003557, CY003733, CY006049, CY006224, CY006304, CY006312, CY006688, CY006720, CY006728, CY006824, CY006832, CY006912, CY007976, CY008465, CY008689, CY008697, CY008705, CY009009, CY009353, CY009641, CY019904, CY019920, CY021082, CY021098, CY021122, CY021602, CY021834, CY021842, CY021946, CY022943, CY022951, CY026144, AJ605762, CY002093, CY002757, CY003069, CY003357, CY003493, CY003509, CY003517, CY003525, CY003541, CY003549, CY003725, CY003741, CY003749, CY006057, CY006104, CY006208, CY006320, CY006328, CY006696, CY006704, CY006712, CY006736, CY006760, CY006768, CY006856, CY006896, CY007616, CY007624, CY007632, CY008457, CY008473, CY008673, CY008713, CY008721, CY008729, CY008737, CY008745, CY009057, CY009073, CY009289, CY009297, CY009625, CY010369, CY010377, CY010761, CY010873, CY010881, CY010889, CY010905, CY010913, CY010921, CY010929, CY010937, CY010953, CY010961, CY011301, CY011485, CY012445, CY012885, CY012893, CY015529, CY017208, CY017256, CY017368, CY019058, CY019074, CY019090, CY019098, CY019106, CY019226, CY019744, CY019760, CY019768, CY019776, CY019784, CY019968, CY020170, CY020178, CY020186, CY020194, CY020226, CY020298, CY020330, CY020442, CY020458, CY020482, CY020490, CY020570, CY020578, CY021034, CY021106,

CY021722, CY021730, CY021738, CY021802, CY021914, CY021978, CY024930, CY026416, CY028729, DQ508824, DQ508832, DQ508848, DQ508872, DQ508896, AF037424, AF037425, AF037426, AF037427, AF037428, AF037429, AF257197, AF257198, AF258518, AF258519, AF398864, AF483603, AJ293922, CY000454, CY000462, CY000470, CY000598, CY000606, CY000614, CY000622, CY000638, CY000646, CY000654, CY000662, CY000670, CY000678, CY000686, CY000694, CY000702, CY000718, CY000726, CY000734, CY000742, CY000750, CY000806, CY000814, CY000822, CY000830, CY000838, CY000846, CY000854, CY000862, CY000994, CY001002, CY001250, CY001282, CY001362, CY001370, CY001386, CY001418, CY001450, CY001458, CY001482, CY001490, CY001498, CY001509, CY001525, CY001533, CY001573, CY001581, CY001589, CY001597, CY001613, CY001621, CY001661, CY001669, CY001701, CY001709, CY001749, CY001757, CY001773, CY001781, CY001789, CY001797, CY001805, CY001837, CY001853, CY001861, CY001869, CY001877, CY001885, CY001893, CY001901, CY001909, CY001917, CY001925, CY001933, CY001941, CY001965, CY001973, CY001981, CY001997, CY002117, CY002141, CY002149, CY002165, CY002173, CY002277, CY002301, CY002341, CY002373, CY002389, CY002517, CY002565, CY002581, CY002645, CY002653, CY003229, CY003237, CY003245, CY003261, CY003269, CY003277, CY003285, CY003437, CY003445, CY003461, CY003581, CY003589, CY003597, CY003605, CY003613, CY003621, CY003637, CY003717, CY003790, CY003798, CY003806, CY003814, CY003830, CY006065, CY006073, CY006232, CY006240, CY006248, CY006256, CY006288, CY006344, CY006448, CY006456, CY006480, CY006496, CY006504, CY006512, CY006520, CY006528, CY006536, CY006544, CY006560, CY006568, CY006576, CY006584, CY006592, CY006600, CY006616, CY006624, CY006632, CY006640, CY006664, CY006776, CY006792, CY006800, CY006808, CY006904, CY007984, CY008136, CY008185, CY008481, CY008489, CY008505,

CY008513, CY008753, CY008761, CY008777, CY008793, CY008817, CY008833, CY008841, CY008865, CY008929, CY008945, CY008961, CY008969, CY008985, CY009089, CY009097, CY009105, CY009113, CY009129, CY009137, CY009153, CY009161, CY009169, CY009177, CY009185, CY009193, CY009201, CY009209, CY009217, CY009321, CY009465, CY009473, CY009481, CY009489, CY009505, CY009513, CY009521, CY009529, CY009545, CY009649, CY009657, CY009665, CY009673, CY009681, CY009689, CY009705, CY009713, CY009721, CY009729, CY009737, CY009745, CY009761, CY009817, CY009849, CY009993, CY010001, CY010009, CY010025, CY010033, CY010041, CY010057, CY010065, CY010073, CY010105, CY010129, CY010145, CY010489, CY010497, CY010505, CY010513, CY010521, CY010529, CY010537, CY010545, CY010601, CY010609, CY010617, CY010625, CY010633, CY010641, CY010657, CY010665, CY010673, CY010697, CY010705, CY010721, CY010729, CY010737, CY010745, CY010753, CY010825, CY010833, CY010841, CY010849, CY010993, CY011001, CY011017, CY011325, CY011333, CY011341, CY011349, CY011357, CY011365, CY011373, CY011381, CY011389, CY011421, CY011429, CY011437, CY011461, CY011469, CY011493, CY011501, CY011509, CY011517, CY011525, CY011541, CY011549, CY011557, CY011565, CY011581, CY011589, CY011597, CY011781, CY011789, CY011797, CY011805, CY011821, CY011829, CY011837, CY011853, CY011861, CY011869, CY011877, CY011885, CY011893, CY011901, CY011909, CY011917, CY011925, CY011933, CY011941, CY011949, CY011957, CY012125, CY012133, CY012141, CY012149, CY012173, CY012181, CY012189, CY012205, CY012213, CY012229, CY012237, CY012245, CY012261, CY012269, CY012277, CY012285, CY012437, CY012453, CY012461, CY012477, CY012485, CY012493, CY012533, CY012541, CY012549, CY012557, CY012565, CY012573, CY012581, CY012621, CY012629, CY012637, CY012733, CY012741, CY012765, CY012869, CY012877, CY012901, CY012925, CY012933,

CY012941, CY012957, CY012965, CY012973, CY012989, CY012997, CY013005, CY013013, CY013021, CY013029, CY013069, CY013173, CY013189, CY013197, CY013205, CY013213, CY013284, CY013292, CY013308, CY013316, CY013340, CY013348, CY013364, CY013372, CY013394, CY013610, CY013642, CY013674, CY013714, CY013738, CY013770, CY013794, CY013818, CY013876, CY013900, CY013908, CY014156, CY015521, CY015537, CY015545, CY015657, CY015665, CY015673, CY016057, CY016073, CY016089, CY016097, CY016121, CY016209, CY016233, CY016265, CY016273, CY016408, CY016504, CY016520, CY016528, CY016544, CY016552, CY016568, CY016576, CY016632, CY016640, CY016648, CY016672, CY016712, CY016720, CY016728, CY016744, CY016752, CY016760, CY016968, CY016976, CY017016, CY017024, CY017128, CY017144, CY017152, CY017264, CY017272, CY017288, CY017296, CY017304, CY017328, CY017344, CY017392, CY017448, CY017464, CY017472, CY017834, CY017890, CY018938, CY019114, CY019138, CY019792, CY019800, CY019986, CY019994, CY020218, CY020258, CY020314, CY020498, CY020514, CY021010, CY021698, CY021746, CY021778, CY021922, CY021954, CY022162, CY022170, CY022514, CY022522, CY023015, CY023023, CY023031, CY023039, CY023047, CY023055, CY025023, CY025047, DQ415305, DQ415306, DQ415307, DQ487327, DQ487335, DQ508856, DQ508864, U71136, U71137, U71138, U71139, CY000006, CY000014, CY000022, CY000030, CY000038, CY000046, CY000054, CY000070, CY000078, CY000086, CY000102, CY000110, CY000118, CY000126, CY000134, CY000142, CY000158, CY000166, CY000174, CY000190, CY000198, CY000206, CY000214, CY000222, CY000230, CY000238, CY000246, CY000254, CY000262, CY000286, CY000294, CY000302, CY000310, CY000318, CY000326, CY000334, CY000342, CY000350, CY000374, CY000382, CY000390, CY000398, CY000406, CY000414, CY000422, CY000430, CY000438, CY000446, CY000478, CY000486, CY000494, CY000502, CY000510,

CY000517, CY000534, CY000542, CY000558, CY000566, CY000574, CY000581, CY000590, CY000630, CY000758, CY000766, CY000774, CY000782, CY000790, CY000798, CY000870, CY000878, CY000886, CY000894, CY000914, CY000922, CY000930, CY000938, CY000954, CY000962, CY000970, CY000978, CY001018, CY001026, CY001050, CY001058, CY001069, CY001077, CY001085, CY001133, CY001149, CY001165, CY001173, CY001189, CY001194, CY001202, CY001210, CY001218, CY001226, CY001234, CY001242, CY001258, CY001266, CY001290, CY001306, CY001314, CY001322, CY001346, CY001378, CY001410, CY001434, CY001442, CY001466, CY001474, CY001517, CY001541, CY001549, CY001629, CY001653, CY001685, CY001725, CY001733, CY001741, CY001957, CY002005, CY002013, CY002021, CY002029, CY002037, CY002045, CY002053, CY002061, CY002069, CY002077, CY002085, CY002109, CY002157, CY002181, CY002189, CY002197, CY002205, CY002213, CY002221, CY002229, CY002245, CY002261, CY002269, CY002285, CY002357, CY002413, CY002421, CY002445, CY002453, CY002461, CY002469, CY002477, CY002525, CY002533, CY002541, CY002573, CY002589, CY002605, CY002621, CY002629, CY002637, CY002661, CY002669, CY002677, CY002685, CY002693, CY002701, CY002709, CY002717, CY002725, CY002733, CY002741, CY002805, CY002813, CY002821, CY002911, CY002919, CY002927, CY002935, CY002943, CY002951, CY002959, CY002967, CY002974, CY002981, CY002989, CY003013, CY003021, CY003029, CY003061, CY003093, CY003126, CY003141, CY003157, CY003165, CY003181, CY003189, CY003205, CY003213, CY003293, CY003301, CY003309, CY003317, CY003325, CY003333, CY003341, CY003349, CY003389, CY003397, CY003405, CY003413, CY003421, CY003429, CY003469, CY003477, CY003645, CY003653, CY003661, CY003669, CY003677, CY003685, CY003693, CY003701, CY003709, CY003766, CY003774, CY003782, CY006081, CY006089, CY006097, CY006112, CY006128, CY006160, CY006200, CY006296,

CY006360, CY006368, CY006376, CY006392, CY006400, CY006408, CY006424, CY006432, CY006440, CY006672, CY006680, CY006864, CY006920, CY006928, CY006936, CY006944, CY006952, CY007024, CY007040, CY007056, CY007072, CY007080, CY007104, CY007120, CY007136, CY007144, CY007152, CY007160, CY007168, CY007176, CY007184, CY007200, CY007208, CY007224, CY007232, CY007256, CY007272, CY007280, CY007296, CY007336, CY007344, CY007352, CY007368, CY007376, CY007408, CY007416, CY007424, CY007432, CY007448, CY007456, CY007472, CY007504, CY007520, CY007544, CY007552, CY007560, CY007568, CY007592, CY007664, CY007672, CY007680, CY007704, CY007728, CY007744, CY007752, CY007760, CY007776, CY007800, CY007824, CY007848, CY007864, CY007896, CY007904, CY007912, CY007928, CY007936, CY007944, CY007952, CY007960, CY008008, CY008016, CY008040, CY008048, CY008056, CY008064, CY008072, CY008080, CY008088, CY008104, CY008112, CY008153, CY008201, CY008217, CY008241, CY008249, CY008257, CY008281, CY008297, CY008313, CY008329, CY008345, CY008353, CY008361, CY008393, CY008401, CY008409, CY008425, CY008441, CY008449, CY008529, CY008545, CY008553, CY008577, CY008593, CY008617, CY008625, CY008633, CY008641, CY008657, CY008873, CY008881, CY008889, CY008897, CY008905, CY009001, CY009049, CY009241, CY009249, CY009257, CY009265, CY009273, CY009401, CY009409, CY009441, CY009865, CY009873, CY009881, CY009889, CY009953, CY009961, CY009977, CY009985, CY010201, CY010225, CY010233, CY010273, CY010281, CY010297, CY010321, CY010329, CY010433, CY010457, CY010857, CY011093, CY011157, CY011165, CY011205, CY011221, CY011245, CY011397, CY011413, CY011613, CY011621, CY011637, CY011645, CY011653, CY011661, CY011669, CY011677, CY011685, CY011693, CY011701, CY011717, CY011733, CY011749, CY011765, CY011773, CY011981, CY011989, CY011997, CY012005, CY012021, CY012029, CY012037,

CY012053, CY012093, CY012109, CY012117, CY012317, CY012325, CY012349, CY012357, CY012373, CY012381, CY012389, CY012405, CY012429, CY012653, CY012661, CY012677, CY012693, CY012701, CY012709, CY012725, CY012797, CY013077, CY013085, CY013101, CY013109, CY013117, CY013125, CY013133, CY013141, CY013221, CY013229, CY013237, CY013418, CY013426, CY013458, CY013466, CY013474, CY013482, CY013514, CY013522, CY013530, CY013546, CY013554, CY013562, CY013570, CY013578, CY013586, CY013594, CY013602, CY013810, CY013916, CY013956, CY013964, CY013980, CY013996, CY014004, CY014020, CY014036, CY014044, CY014060, CY014084, CY014092, CY014100, CY014108, CY014116, CY014164, CY015553, CY015569, CY015577, CY015585, CY015601, CY015609, CY015633, CY015641, CY015681, CY015689, CY015697, CY015713, CY015721, CY015729, CY015745, CY015753, CY015761, CY015769, CY015801, CY015809, CY015817, CY015825, CY015833, CY015849, CY015857, CY015865, CY015881, CY015889, CY015897, CY015905, CY015913, CY015921, CY015929, CY015937, CY015961, CY015969, CY015977, CY015985, CY015993, CY016009, CY016017, CY016025, CY016033, CY016041, CY016217, CY016440, CY016464, CY016480, CY016600, CY016608, CY016664, CY016680, CY016696, CY016704, CY016984, CY017088, CY017096, CY017104, CY017112, CY017120, CY017136, CY017376, CY017488, CY017496, CY017512, CY017520, CY017544, CY017560, CY017568, CY017576, CY017584, CY017592, CY017608, CY017616, CY017624, CY017632, CY017802, CY017898, CY017906, CY017914, CY017922, CY017930, CY017938, CY017946, CY017954, CY017970, CY017986, CY017994, CY018930, CY018970, CY018986, CY019010, CY019018, CY019026, CY019034, CY019194, CY019274, CY019330, CY019338, CY019346, CY019832, CY019864, CY019880, CY019888, CY019936, CY020010, CY020042, CY020138, CY020146, CY020154, CY020162, CY020266, CY020274, CY020306, CY020346, CY020434, CY020898, CY021762,

CY021770, CY022186, CY022194, CY022202, CY022210, CY022530, CY022554, CY022562, CY022570, CY022602, CY023063, CY023079, CY025218, CY025226, CY025234, CY025242, CY025250, CY025266, CY025274, CY025282, CY025290, CY025298, CY025306, CY025322, CY025338, CY025346, CY025354, CY025362, CY025370, CY025378, CY025394, CY025402, CY025410, CY025418, CY025426, CY025434, CY025442, CY025450, CY025466, CY025474, CY025482, CY025490, CY025498, CY025506, CY025514, CY025522, CY025530, CY025538, CY025552, CY025560, CY025568, CY025592, CY025600, CY025608, CY025616, CY025624, CY025640, CY025648, CY025656, CY025664, CY025672, CY025680, CY025688, CY025696, CY025704, CY025712, CY025720, CY025744, CY025752, CY025760, CY025792, CY025808, CY025816, CY025840, CY025848, CY025856, CY025864, CY025888, CY025896, CY025904, CY025912, CY025928, CY025936, CY025952, CY025960, CY025968, CY025976, CY025992, CY026000, CY026008, CY026024, CY026040, CY026152, CY026168, CY026184, CY026200, CY026208, CY026224, CY026240, CY026256, CY026264, CY026280, CY026312, CY026320, CY026336, CY026360, CY026376, CY026392, CY026400, CY026512, CY026520, CY026528, CY026536, CY026544, CY026552, CY026560, CY026568, CY026576, CY026584, CY026592, CY026600, CY026608, CY026616, CY026632, CY026648, CY026656, CY026664, CY026672, CY026680, CY026688, CY026696, CY026704, CY026712, CY026728, CY026736, CY026744, CY026768, CY026784, CY026792, CY026800, CY026832, CY026840, CY026856, CY026864, CY026872, CY026888, CY026896, CY026904, CY026920, CY026928, CY026936, CY026944, CY026952, CY026960, CY026968, CY026976, CY027000, CY027008, CY027024, CY027040, CY027048, CY027072, CY027080, CY027096, CY027112, CY027120, CY027136, CY027144, CY027168, CY027176, CY027192, CY027208, CY027216, CY027232, CY027248, CY027272, CY027320, CY027328, CY027344, CY027352, CY027360, CY027368, CY027384, CY027416,

CY027424, CY027432, CY027456, CY027472, CY027480, CY027488, CY027544, CY027568, CY027584, CY027592, CY027600, CY027608, CY027624, CY027632, CY027640, CY027648, CY027656, CY027664, CY027672, CY027688, CY027712, CY027720, CY027728, CY027736, CY027744, CY027760, CY027768, CY027776, CY027784, CY027792, CY027800, CY027816, CY027832, CY027840, CY027848, CY027856, CY027864, CY027888, CY027896, CY027904, CY027920, CY027936, CY027968, CY027984, CY027992, CY028008, CY028016, CY028024, CY028032, CY028064, CY028080, CY028096, CY028104, CY028112, CY028136, CY028144, CY028160, CY028168, CY028208, CY028304, CY028312, CY028320, CY028328, CY028336, CY028360, CY028368, CY028384, CY028408, CY028424, CY028448, CY028464, CY028737, CY028745, CY028753, CY028761, CY028769, CY028777, CY030058, CY030066, CY030074, CY030202, CY030210, DQ415308, DQ415309, DQ415310, DQ415311, DQ415312, DQ415313, DQ415314, DQ415315, DQ469957, DQ889684, EF554794, EU399756, AB212053, AF084267, AF084268, AF084269, AF084270, AF115294, AF115295, AF257191, AF257192, AF257193, AF257194, AF257195, AF257196, AF257199, AF257200, AF257201, AF257202, AJ289874, AJ291402, AJ404637, AY626147, AY627890, AY627896, AY818132, CY014169, CY014171, CY014271, CY014279, CY014287, CY014310, CY014321, CY014326, CY014328, CY014330, CY014338, CY014340, CY014342, CY014350, CY014353, CY014354, CY014362, CY014366, CY014373, CY014389, CY014394, CY014402, CY014410, CY014418, CY014426, CY014442, CY014482, CY014498, CY014511, CY014519, CY014528, CY014536, CY014542, CY015011, CY017637, CY017653, CY017661, CY017669, CY017677, CY017687, CY019359, CY019383, CY019391, CY019407, DQ099791, DQ099792, DQ138184, DQ138185, DQ138186, DQ138187, DQ138188, DQ360839, DQ372596, DQ835312, EF100817, EF137711, EF587276, EF619995, EU146661, EU146702, EU146710, EU146718, EU146768, EU146857,

EU146858, EU146860, EU146863, EU146864, BAG49626, ABR28555, ABR28566, ABR28621, ABR28632, ABR28654, ABR28709, ABR28742, ABR29582, ABS50118, ABS50128, ABU80239, ABU80427, ABU80217, ABV82591, ABW36340, ABW36373, ABX58653, ABQ51942, ABS53352, ABQ45454, ABQ45540, ABR28643, ABR28687, ABR28720, ABR29592, ABU80250, ABU80294, ABW71488, ABX58675, ABS53371, ABV31973, AAG17435, AAO15326, AAO15328, AAO15329, ABR15848, ABR28544, ABR29612, ABR28610, ABS49928, ABU80195, ABU80272, ABV29597, ABV82580, ABW36329, ABW36351, ABX58664, ABS00316, ABV31954, AAF76000, BAG49741, ABR15826, ABR15837, ABR29622, ABR29602, ABR28731, ABR29572, ABS49950, ABU80417, ABW36384, ABW86592, ABQ41897, ABV31974, AAG01774, ABY40434, AAL87914, AAL87916, CAC85221, CAC37005, CAC84683, ACA25357, AAG01747, CAC84685, ABO44042, ABV55859, ABY40444, ABY16774, ABY16775, ABY16776, ABY51211, AAL87913, AAL87915, AAL87918, CAC85219, CAC85222, CAC37006, ABY16773, ACA25367, AAA43675, AAZ79399, ABA46959, ABB86942, ABF17997, ACE78069, ACE78070, ACE78071, ACE78079, ABE27160, ABQ45465, ABR28577, ABW38017, AAU25844, AAV30841, AAV68019, ABA27439, ABB86892, ABB86951, ACA42427, ACF49406, ACE78078, ACE78082, ACE78084, ACE78085, ABD61558, ABE27171, ABR28665, AAA43681, AAU25864, AAV30833, ABB86882, ABB86912, ABD77112, ABF17999, ABI54398, ACE78067, ACE78072, ACE78081, ACE78086, ABD62840, ABD78111, ABQ45432, ABU80206, ABV82602, ABW71528, ACD85161, ABB86872, ABF18000, ABJ16478, ABJ15725, ABK00129, ACE78068, ACE78080, ABD79262, ABR15870, ABW36362, ABW86581, CAC85217, ABY81693, AAA43617, ABY51222, ACE78077, AAA43669, AAN46832, ABA27431, ABB86922, ACA42437,

ABD95719, ABE12641, AAO15327, ABW71510, AAG01756, ABF17998, ACF04399, AAU05321, ABB86902, ABB86956, ABD61256, ABD62801, ABR28764, ABS49961, ABS00327, ABS53362, ACA25345, CAC37004, AAA43671, ABD77120, ABQ45443, ABR28676, ABR87895, AAL87917, AAL87920, AAU25854, ABY84691, ABY81433. For the pandemic H1N1 sequences: CY073967, CY071140, CY071148, CY071156, CY075064, CY075072, CY071164, CY071172, CY071180, CY071188, CY071268, CY071276, CY071284, CY071292, CY075080, CY071300, CY071308, CY071316, CY071324, CY071332, CY071340, CY071348, CY071356, CY071364, CY071380, CY062984, CY066452, CY063184, CY063192, CY063200, CY063208, CY063216, CY066460, CY066468, CY066476, CY066484, CY064992, CY066492, CY066500, CY066508, CY083015, CY083023, CY083031, CY083039, CY071212, CY071220, GU562463, CY083822, CY073003, CY083534, CY073125, HQ834745, CY065877, CY062071, CY062079, CY062087, CY062095, CY062103, CY062111, CY062119, CY062127, CY062135, CY062143, CY062151, CY062159, CY062167, CY062175, CY062183, CY062191, CY065000, CY065008, CY065016, CY065024, CY065032, CY065040, CY065048, CY065056, CY065064, CY065072, CY065112, CY072539, CY071372, CY081060, CY081068, CY081076, CY081084, CY081092, CY081100, HM569739, HM124382, CY063829, CY063839, HM189303, CY061804, CY061812, HM189625, HM189626, HM189627, HM189628, HQ393492, CY062267, CY062275, CY062315, CY076738, CY076746, CY079543, CY063008, CY063016, CY071076, CY072531, CY073738, CY061583, CY061112, CY061591, CY061120, CY061128, CY061136, CY061144, CY061152, CY061160, CY061168, CY061176, CY061184, CY061192, CY061200, CY061208, CY080306, CY080298, CY074981, CY080322, CY080338, CY080314, CY080330, CY081057, CY062992, CY071060, CY073778, HQ891284.

For segment 5: AB020778, AB166863, AB188817, AB189062, AB256667, AB256675, AB256683, AB256691, AB256699, AB256707, AB256715, AB256723, AB256731, AB256747, AB261854, AF046084, AF098617, AF098618, AF098619, AF098620, AF098624, AF098626, AF098627, AF156403, AF203787, AF261750, AF457676, AF457685, AF457701, AF457709, AF474069, AF509117, AF509118, AF509120, AF509121, AJ291394, AJ410548, AJ410549, AJ410551, AJ410552, AJ427309, AJ620352, AJ627486, AM503029, AM503030, AM503031, AM503032, AM503033, AM503034, AM503035, AY253753, AY268949, AY303658, AY303659, AY496851, AY575911, AY575912, AY590579, AY609313, AY611527, AY650273, AY651511, AY651512, AY653196, AY664717, AY664718, AY664719, AY664722, AY664724, AY664725, AY664727, AY664728, AY664730, AY664731, AY664732, AY676039, AY684707, AY724258, AY737290, AY737297, AY770081, AY818139, AY818140, AY849789, AY849790, AY950251, AY950252, AY950254, AY950255, CY005430, CY005456, CY005481, CY005503, CY005521, CY005608, CY005797, CY005829, CY005841, CY005891, CY005909, CY014186, CY014610, CY014652, CY014666, CY014674, CY014749, CY014753, CY014763, CY014767, CY014774, CY014781, CY014789, CY014795, CY014801, CY014824, CY014832, CY014839, CY014883, CY014898, CY014912, CY014994, CY015022, CY015029, CY015035, CY015049, CY015055, CY015068, CY015076, CY015084, CY015092, CY015099, CY015105, CY015111, CY015118, CY015122, CY016279, CY016287, CY016295, CY016303, CY016790, CY016814, CY016838, CY016846, CY016870, CY016878, CY016910, CY016918, CY016926, CY016934, CY016942, CY016950, CY017054, CY017062, CY017406, CY018952, CY020584, CY020600, CY020608, CY020624, CY020632, CY020640, CY020648, CY020656, CY020672, CY020696, CY021360, CY021368, CY021376, CY021384, CY021488, CY021496, CY021512, CY021528, CY022080, CY022264, CY022616,

CY022624, CY022688, CY022696, CY022704, CY022712, CY022768, CY022816, CY022824, CY022832, CY024789, CY024845, CY025072, CY025080, CY028510, CY028543, CY028575, CY028583, CY028599, CY028631, CY028663, CY028671, CY028679, CY028703, CY028711, CY029196, CY029203, CY029238, CY029245, CY029252, CY029259, CY029434, CY029462, CY029469, CY029476, CY029483, CY029836, CY029949, CY029957, CY029965, DQ023146, DQ064435, DQ064436, DQ064437, DQ064438, DQ064439, DQ064440, DQ064441, DQ064442, DQ064444, DQ064446, DQ064447, DQ064448, DQ064449, DQ064450, DQ064451, DQ064452, DQ064453, DQ064456, DQ064457, DQ064458, DQ064459, DQ064460, DQ067440, DQ076203, DQ099769, DQ099772, DQ099774, DQ208502, DQ211931, DQ227351, DQ321130, DQ321132, DQ334761, DQ334777, DQ335774, DQ351864, DQ351865, DQ351866, DQ376727, DQ376728, DQ376729, DQ376730, DQ376731, DQ376734, DQ376736, DQ376738, DQ376739, DQ376740, DQ376741, DQ376742, DQ376743, DQ376744, DQ376745, DQ376746, DQ376747, DQ376748, DQ376749, DQ376750, DQ376752, DQ376753, DQ376755, DQ376756, DQ376757, DQ376760, DQ376761, DQ449635, DQ469999, DQ470874, DQ485209, DQ485217, DQ493080, DQ493085, DQ493088, DQ493089, DQ493102, DQ493104, DQ493105, DQ650666, DQ792926, DQ870889, DQ870895, DQ914815, DQ997089, DQ997113, DQ997116, DQ997126, DQ997135, DQ997144, DQ997151, DQ997158, DQ997184, DQ997272, DQ997301, DQ997310, DQ997320, DQ997327, DQ997334, DQ997343, DQ997346, DQ997365, DQ997379, DQ997442, DQ997462, DQ997540, DQ997549, EF010524, EF063524, EF063527, EF070735, EF124361, EF124362, EF124390, EF124456, EF124477, EF155133, EF155134, EF155137, EF178518, EF205175, EF205176, EF205180, EF205181, EF362421, EF474447, EF523736, EF593103, EF605593, EF605600, EF681871,

EF681879, EU081867, EU084906, EU084908, EU084912, EU084919, EU084923, EU084926, EU084930, EU084948, EU086229, EU086240, EU086253, EU086289, EU148359, EU148367, EU148391, EU148423, EU148431, EU148439, EU148447, EU158139, EU163432, EU170434, EU170435, EU170436, EU182266, EU182274, EU182292, EU182300, EU233417, EU277836, EU277844, EU365372, EU401754, EU402401, EU414523, EU414524, EU429959, EU429960, EU429961, EU429964, EU443578, M24453, M30768, M30769, M63774, M63778, M76603, M76609, Z26857, AB049161, AB212281, AB239303, AB239310, AB259713, AB262464, AB263753, AB284325, AB284989, AB285095, AB286119, AB286876, AB300229, AB300436, AB300441, AB301915, AB304148, AF079571, AF098621, AF098622, AF098623, AF144303, AF156406, AF250470, AF250471, AF250472, AF250473, AF250474, AF250480, AF370122, AF468842, AF509139, AF523421, AF523423, AJ410555, AJ410556, AJ427298, AJ427301, AY233394, AY518364, AY585420, AY585421, AY585422, AY585423, AY585424, AY585425, AY585426, AY585427, AY585428, AY585429, AY585430, AY585431, AY585432, AY585433, AY585434, AY585435, AY585436, AY585437, AY585438, AY585439, AY585440, AY633119, AY633127, AY633167, AY633215, AY633231, AY633247, AY633279, AY633311, AY633319, AY633343, AY676037, AY676038, AY676040, AY724252, AY737305, AY742258, AY856864, AY862661, AY950256, AY950257, CY003850, CY003858, CY003866, CY003874, CY003882, CY003890, CY003897, CY003917, CY003925, CY003932, CY003939, CY003947, CY003955, CY003963, CY003971, CY003979, CY003987, CY003995, CY004002, CY004008, CY004014, CY004021, CY004051, CY004057, CY004074, CY004083, CY004091, CY004097, CY004103, CY004109, CY004117, CY004124, CY004132, CY004149, CY004157, CY004181, CY004189, CY004205, CY004213, CY004229, CY004253, CY004261, CY004269, CY004277, CY004285, CY004292, CY004296,

CY004302, CY004307, CY004312, CY004320, CY004326, CY004329, CY004334, CY004341, CY004348, CY004355, CY004362, CY004375, CY004382, CY004388, CY004401, CY004415, CY004423, CY004436, CY004446, CY004453, CY004461, CY004469, CY004477, CY004485, CY004493, CY004506, CY004510, CY004518, CY004534, CY004541, CY004549, CY004557, CY004563, CY004570, CY004595, CY004602, CY004609, CY004616, CY004622, CY004637, CY004645, CY004652, CY004665, CY004673, CY004683, CY004695, CY004705, CY004712, CY004718, CY004724, CY004731, CY004738, CY004745, CY004752, CY004758, CY004765, CY004772, CY004786, CY004793, CY004800, CY004814, CY004820, CY004833, CY004845, CY004849, CY004856, CY004870, CY004877, CY004882, CY004888, CY004892, CY004899, CY004906, CY004914, CY004921, CY004928, CY004936, CY004942, CY004949, CY004956, CY004963, CY004973, CY004979, CY004990, CY005004, CY005011, CY005018, CY005030, CY005041, CY005046, CY005053, CY005060, CY005067, CY005074, CY005081, CY005087, CY005092, CY005095, CY005116, CY005122, CY005129, CY005149, CY005156, CY005162, CY005166, CY005174, CY005179, CY005183, CY005190, CY005196, CY005202, CY005221, CY005227, CY005234, CY005239, CY005245, CY005252, CY005259, CY005266, CY005276, CY005282, CY005288, CY005294, CY005300, CY005313, CY005320, CY005326, CY005333, CY005340, CY005353, CY005360, CY005367, CY005372, CY005378, CY005383, CY005395, CY005403, CY005416, CY005423, CY005433, CY005462, CY005465, CY005471, CY005478, CY005496, CY005509, CY005527, CY005533, CY005541, CY005549, CY005555, CY005565, CY005570, CY005578, CY005585, CY005592, CY005600, CY005614, CY005621, CY005627, CY005634, CY005642, CY005648, CY005655, CY005661, CY005667, CY005674, CY005682, CY005687, CY005694, CY005700, CY005705, CY005711, CY005737, CY005742, CY005749, CY005753, CY005760, CY005772, CY005779, CY005806, CY005809, CY005815, CY005847,

CY005854, CY005861, CY005869, CY005876, CY005884, CY005898, CY005904, CY011031, CY011043, CY011051, CY011059, CY011115, CY011251, CY012803, CY012811, CY012819, CY012827, CY012835, CY012843, CY013250, CY013258, CY013266, CY013866, CY014521, CY014551, CY014557, CY014570, CY014578, CY014584, CY014650, CY014658, CY014690, CY014697, CY014705, CY014712, CY014728, CY014742, CY014809, CY014816, CY014845, CY014852, CY014860, CY014868, CY014875, CY014891, CY014904, CY014918, CY014924, CY014932, CY014940, CY014948, CY014964, CY014971, CY014987, CY015042, CY015130, CY015138, CY015144, CY015170, CY015446, CY015454, CY015462, CY015479, CY015487, CY015495, CY015503, CY016135, CY016143, CY016151, CY016159, CY016167, CY016175, CY016183, CY016191, CY016398, CY016414, CY016422, CY016614, CY016622, CY016782, CY016798, CY016806, CY016822, CY016830, CY016886, CY016894, CY016902, CY016958, CY017030, CY017038, CY017046, CY017070, CY017078, CY017414, CY017696, CY017704, CY017712, CY017736, CY017744, CY017752, CY017760, CY017768, CY017776, CY017784, CY017792, CY017840, CY017848, CY017856, CY018002, CY018010, CY018018, CY018880, CY018888, CY018896, CY018904, CY018912, CY018920, CY019200, CY020352, CY020720, CY020728, CY020736, CY020744, CY020752, CY020760, CY020768, CY020784, CY020792, CY020816, CY020824, CY020832, CY020840, CY020848, CY020864, CY020872, CY020880, CY020888, CY020904, CY020928, CY020936, CY020944, CY020952, CY020960, CY020968, CY020976, CY020984, CY020992, CY021128, CY021136, CY021144, CY021168, CY021176, CY021184, CY021200, CY021208, CY021216, CY021224, CY021240, CY021248, CY021256, CY021264, CY021272, CY021296, CY021304, CY021320, CY021336, CY021344, CY021352, CY021400, CY021432, CY021448, CY021464, CY021472, CY021480, CY021568, CY021584, CY021592, CY021616, CY021624, CY021640, CY021648, CY021656, CY021664, CY021672,

CY021680, CY021688, CY021864, CY021872, CY021880, CY021896, CY022608, CY022640, CY022648, CY022720, CY024749, CY024797, CY025200, CY028238, CY028254, CY028262, CY028286, CY028647, CY028695, CY029084, CY029231, CY029315, CY029322, CY029329, CY029406, CY029448, CY029455, CY029844, CY029884, CY029924, D00050, DQ017489, DQ017497, DQ017505, DQ021832, DQ064454, DQ064455, DQ073411, DQ073412, DQ073413, DQ073414, DQ095675, DQ099777, DQ099778, DQ232609, DQ232610, DQ251444, DQ251452, DQ321074, DQ321076, DQ321077, DQ321078, DQ321081, DQ321082, DQ321084, DQ321086, DQ321087, DQ321088, DQ321109, DQ321111, DQ321122, DQ358751, DQ364998, DQ365006, DQ366307, DQ366315, DQ366323, DQ366339, DQ376726, DQ376733, DQ376735, DQ376737, DQ376754, DQ376758, DQ376759, DQ407246, DQ449643, DQ454104, DQ464359, DQ465401, DQ486129, DQ493096, DQ525414, DQ676841, DQ681204, DQ681208, DQ681222, DQ792924, DQ792925, DQ792927, DQ835778, DQ835779, DQ835780, DQ835782, DQ835783, DQ835784, DQ852602, DQ861294, DQ864509, DQ916293, DQ989962, DQ989980, DQ989988, DQ989996, DQ997096, DQ997165, DQ997174, DQ997278, DQ997407, DQ997412, DQ997515, DQ997524, DQ997533, EF061123, EF112293, EF112294, EF112298, EF112300, EF112301, EF112312, EF124343, EF124344, EF124376, EF124397, EF124406, EF124409, EF124445, EF124447, EF124448, EF124466, EF124472, EF124474, EF124476, EF178511, EF205177, EF205178, EF210573, EF418187, EF523738, EF523739, EF592496, EF597328, EF597330, EF597331, EF597334, EF597335, EF597352, EF634333, EF634341, EU026010, EU026017, EU026032, EU026040, EU026047, EU026053, EU026061, EU026077, EU026085, EU026093, EU026105, EU026113, EU030969, EU030977, EU030985, EU084939, EU084943, EU084952, EU086251, EU094474, EU158133, EU158134, EU158140, EU263346,

EU263354, EU329178, EU329185, EU430498, EU430504, EU443573, EU443574, EU443576, EU443577, EU443580, M22344, M22573, M22574, M27298, M27519, M27521, M30752, M30753, M30755, M30756, M30757, M30760, M30761, M30762, M30763, M30764, M30765, M30766, M30767, M36812, M63773, M63775, M63776, M63777, M63780, M63781, M63782, M63783, M63784, M63785, Z26855, AF389119, AY744935, CY009327, CY009447, CY009607, CY010791, CY020448, CY020472, J02147, M38279, V01084, AY210066, AY210067, AY210068, AY210069, CY008991, CY009279, CY009335, CY009343, CY009455, CY009599, CY009615, CY013274, CY014979, CY019950, CY019974, CY020288, CY020384, CY020464, CY021704, CY021712, CY021824, CY021904, CY022016, CY022024, CY022096, D00601, DQ508842, EF633439, M63750, M63751, M63752, M76604, U02086, AF348180, AY210070, AY210071, AY210072, AY210073, AY210074, AY210075, AY210076, AY210077, AY210079, AY210080, AY210081, AY210082, AY210083, AY210084, AY210085, AY210086, AY210087, AY210088, AY210089, AY210090, AY210091, AY210092, AY210093, AY210094, AY210095, AY210096, AY210097, AY210098, AY210099, AY210100, AY210101, AY210102, AY210103, AY210104, AY210207, AY210208, AY210209, AY210210, AY210211, AY210212, AY210217, AY210218, AY210219, AY210220, CY006214, CY008159, CY011123, CY015511, CY019894, CY019910, CY020320, CY020392, CY020416, CY020520, CY020528, CY020544, CY021024, CY021072, CY021112, CY021816, CY021936, DQ508882, J02137, M23976, M63753, X15890, AF072545, AY210221, AY210223, AY210224, AY210225, AY210226, AY210227, AY210228, AY210229, AY210230, AY210231, AY210232, AY210233, AY210234, AY210235, AY210236, AY210237, AY210238, CY002099, CY002499, CY002747, CY003499, CY003531, CY003555, CY003731, CY006047, CY006222, CY006302,

CY006310, CY006686, CY006718, CY006726, CY006830, CY006886, CY006910, CY007974, CY008463, CY008679, CY008687, CY009007, CY009351, CY009639, CY019918, CY021080, CY021088, CY021096, CY021120, CY021600, CY021832, CY021944, CY022941, CY022949, CY026142, L07345, L07346, L07347, L07349, L07358, M63754, M76605, M76606, AJ628066, CY002091, CY002755, CY003067, CY003355, CY003491, CY003507, CY003515, CY003523, CY003539, CY003547, CY003723, CY003739, CY003747, CY006055, CY006206, CY006318, CY006326, CY006694, CY006702, CY006710, CY006734, CY006758, CY006766, CY006846, CY006854, CY006894, CY007614, CY007622, CY007630, CY008175, CY008455, CY008671, CY008711, CY008727, CY008743, CY009055, CY009071, CY009287, CY009295, CY009623, CY010367, CY010871, CY010879, CY010887, CY010903, CY010911, CY010919, CY011483, CY012883, CY012891, CY015527, CY017206, CY017230, CY017438, CY019056, CY019064, CY019096, CY019104, CY019240, CY019742, CY019758, CY019766, CY019774, CY019782, CY019966, CY020168, CY020176, CY020184, CY020192, CY020224, CY020296, CY020440, CY020456, CY020480, CY020488, CY020568, CY020576, CY021032, CY021040, CY021104, CY021720, CY021728, CY021736, CY021800, CY021912, CY021976, CY024928, CY028727, D00599, D00600, D00602, DQ508826, DQ508834, DQ508850, DQ508874, L07350, L07351, L07352, L07353, L07354, L07357, L07359, L07360, L07361, L07364, L07365, L07366, L07367, L07369, L07372, L07373, L07374, M22577, M59329, M59330, M59331, M59332, M59334, M63755, M76602, M76610, X51972, Z54290, Z54291, AB019358, AB019359, AB019360, AB019361, AF038254, AF038255, AF038256, AF038257, AF038258, AF038259, AF255748, AF255749, AF258516, AF258517, AF342819, AF483604, AJ293924, AJ458276, AJ458277, CY000452, CY000460, CY000468, CY000596, CY000620, CY000644, CY000668, CY000676, CY000692, CY000700, CY000708, CY000716, CY000724, CY000740, CY000748,

CY000804, CY000984, CY000992, CY001000, CY001008, CY001179, CY001248, CY001272, CY001416, CY001456, CY001480, CY001488, CY001496, CY001507, CY001571, CY001579, CY001603, CY001619, CY001659, CY001699, CY001707, CY001747, CY001787, CY001875, CY001883, CY001899, CY001907, CY001915, CY001939, CY001971, CY001979, CY001995, CY002115, CY002123, CY002139, CY002163, CY002171, CY002275, CY002371, CY002515, CY002555, CY002563, CY002579, CY002643, CY002651, CY003219, CY003243, CY003251, CY003259, CY003283, CY003443, CY003619, CY003715, CY003755, CY003804, CY003812, CY006063, CY006071, CY006166, CY006230, CY006238, CY006262, CY006286, CY006350, CY006446, CY006462, CY006478, CY006494, CY006502, CY006510, CY006534, CY006542, CY006550, CY006558, CY006566, CY006590, CY006598, CY006614, CY006662, CY006774, CY006790, CY006806, CY006902, CY007982, CY008134, CY008183, CY008479, CY008503, CY008511, CY008767, CY008791, CY008815, CY008831, CY008839, CY008951, CY008959, CY008967, CY008983, CY009079, CY009103, CY009111, CY009151, CY009167, CY009175, CY009183, CY009191, CY009199, CY009207, CY009319, CY009463, CY009471, CY009479, CY009487, CY009519, CY009527, CY009647, CY009655, CY009663, CY009671, CY009679, CY009687, CY009711, CY009727, CY009735, CY009743, CY009807, CY009815, CY009847, CY009943, CY009991, CY009999, CY010007, CY010015, CY010023, CY010031, CY010039, CY010055, CY010071, CY010103, CY010111, CY010143, CY010391, CY010487, CY010495, CY010503, CY010511, CY010519, CY010527, CY010535, CY010543, CY010591, CY010599, CY010623, CY010631, CY010639, CY010647, CY010655, CY010663, CY010671, CY010687, CY010703, CY010719, CY010727, CY010735, CY010743, CY010831, CY010839, CY010847, CY010991, CY010999, CY011015, CY011139, CY011323, CY011331, CY011339, CY011371, CY011379, CY011435, CY011451, CY011459, CY011491, CY011499, CY011507, CY011515,

CY011523, CY011531, CY011539, CY011555, CY011563, CY011571, CY011587, CY011595, CY011795, CY011803, CY011811, CY011819, CY011851, CY011859, CY011867, CY011883, CY011891, CY011899, CY011915, CY011923, CY011931, CY011939, CY011947, CY012123, CY012131, CY012139, CY012147, CY012155, CY012171, CY012187, CY012203, CY012211, CY012227, CY012235, CY012243, CY012251, CY012259, CY012275, CY012283, CY012451, CY012459, CY012475, CY012483, CY012515, CY012523, CY012531, CY012539, CY012547, CY012555, CY012571, CY012587, CY012595, CY012603, CY012627, CY012635, CY012731, CY012747, CY012755, CY012763, CY012771, CY012859, CY012867, CY012875, CY012915, CY012939, CY012947, CY012979, CY012987, CY012995, CY013011, CY013027, CY013171, CY013195, CY013203, CY013282, CY013306, CY013314, CY013322, CY013338, CY013370, CY013400, CY013624, CY013632, CY013664, CY013672, CY013680, CY013688, CY013720, CY013784, CY013816, CY013874, CY013898, CY014154, CY015535, CY015647, CY015663, CY016071, CY016079, CY016087, CY016119, CY016199, CY016207, CY016231, CY016263, CY016406, CY016430, CY016502, CY016518, CY016526, CY016534, CY016542, CY016550, CY016566, CY016630, CY016638, CY016646, CY016670, CY016710, CY016718, CY016726, CY016742, CY016750, CY016966, CY017022, CY017126, CY017142, CY017150, CY017158, CY017166, CY017270, CY017294, CY017302, CY017342, CY017382, CY017390, CY017446, CY017454, CY017462, CY017470, CY017478, CY017832, CY019136, CY019790, CY019806, CY019870, CY019984, CY019992, CY020256, CY020312, CY021008, CY021696, CY021744, CY021776, CY021784, CY021920, CY022152, CY022160, CY022176, CY023013, CY025021, CY025037, CY025045, DQ249264, DQ415327, DQ415328, DQ415329, DQ487330, DQ508890, EF633615, L24394, U71144, U71145, U71146, U71147, Z54292, CY000004, CY000012, CY000020, CY000028, CY000036, CY000044, CY000052, CY000068, CY000076,

CY000084, CY000092, CY000116, CY000132, CY000140, CY000156, CY000164, CY000172, CY000188, CY000196, CY000204, CY000212, CY000220, CY000228, CY000236, CY000244, CY000252, CY000260, CY000268, CY000276, CY000284, CY000292, CY000300, CY000308, CY000316, CY000332, CY000340, CY000348, CY000364, CY000388, CY000396, CY000404, CY000428, CY000436, CY000444, CY000484, CY000492, CY000508, CY000520, CY000540, CY000548, CY000556, CY000564, CY000572, CY000579, CY000588, CY000756, CY000764, CY000788, CY000868, CY000928, CY000936, CY000952, CY000960, CY000968, CY001016, CY001040, CY001067, CY001075, CY001091, CY001099, CY001131, CY001163, CY001171, CY001187, CY001192, CY001200, CY001208, CY001224, CY001240, CY001264, CY001288, CY001312, CY001376, CY001408, CY001424, CY001432, CY001563, CY001635, CY001683, CY001723, CY001955, CY002003, CY002011, CY002019, CY002043, CY002051, CY002059, CY002067, CY002075, CY002155, CY002179, CY002187, CY002195, CY002211, CY002219, CY002227, CY002259, CY002267, CY002355, CY002363, CY002395, CY002403, CY002411, CY002419, CY002427, CY002443, CY002459, CY002467, CY002491, CY002523, CY002531, CY002539, CY002571, CY002595, CY002603, CY002627, CY002635, CY002659, CY002667, CY002675, CY002683, CY002691, CY002699, CY002715, CY002739, CY002763, CY002779, CY002803, CY002811, CY002819, CY002909, CY002917, CY002925, CY002933, CY002949, CY002957, CY002965, CY002972, CY003011, CY003019, CY003043, CY003051, CY003059, CY003083, CY003099, CY003115, CY003124, CY003147, CY003187, CY003195, CY003203, CY003211, CY003291, CY003299, CY003307, CY003315, CY003323, CY003331, CY003371, CY003395, CY003411, CY003475, CY003643, CY003651, CY003659, CY003667, CY003675, CY003683, CY003691, CY003699, CY003707, CY003764, CY003780, CY003836, CY006079, CY006095, CY006110, CY006118, CY006134, CY006150, CY006158, CY006174, CY006198,

CY006294, CY006366, CY006374, CY006382, CY006390, CY006406, CY006430, CY006438, CY006670, CY006678, CY006862, CY006918, CY006926, CY006934, CY006974, CY006998, CY007006, CY007022, CY007118, CY007126, CY007150, CY007158, CY007174, CY007182, CY007198, CY007222, CY007230, CY007238, CY007262, CY007302, CY007350, CY007398, CY007406, CY007422, CY007430, CY007454, CY007462, CY007470, CY007478, CY007486, CY007494, CY007534, CY007550, CY007558, CY007566, CY007582, CY007678, CY007726, CY007750, CY007774, CY007782, CY007790, CY007798, CY007806, CY007822, CY007862, CY007870, CY007902, CY007910, CY007926, CY007942, CY007950, CY007958, CY008022, CY008046, CY008054, CY008062, CY008070, CY008078, CY008086, CY008110, CY008151, CY008199, CY008215, CY008255, CY008279, CY008295, CY008303, CY008311, CY008327, CY008335, CY008343, CY008359, CY008367, CY008407, CY008439, CY008447, CY008607, CY008623, CY008631, CY008639, CY008655, CY008879, CY008887, CY008895, CY008903, CY008919, CY009023, CY009031, CY009247, CY009255, CY009263, CY009271, CY009399, CY009407, CY009431, CY009439, CY009863, CY009871, CY009887, CY009935, CY009951, CY009959, CY010079, CY010087, CY010151, CY010215, CY010223, CY010239, CY010255, CY010271, CY010287, CY010295, CY010303, CY010311, CY010319, CY010335, CY010343, CY010359, CY010399, CY010407, CY010415, CY010447, CY010455, CY010775, CY011075, CY011091, CY011163, CY011403, CY011411, CY011611, CY011619, CY011627, CY011643, CY011651, CY011659, CY011683, CY011691, CY011731, CY011747, CY011755, CY011763, CY011771, CY011995, CY012003, CY012027, CY012035, CY012051, CY012059, CY012067, CY012075, CY012099, CY012107, CY012315, CY012323, CY012371, CY012379, CY012387, CY012643, CY012675, CY012683, CY012699, CY012707, CY012795, CY013075, CY013083, CY013091, CY013099, CY013107, CY013115, CY013139, CY013219, CY013227, CY013235,

CY013243, CY013424, CY013432, CY013504, CY013528, CY013536, CY013544, CY013560, CY013568, CY013576, CY013584, CY013600, CY013808, CY013962, CY013970, CY013994, CY014002, CY014018, CY014034, CY014042, CY014058, CY014066, CY014082, CY014090, CY014114, CY014162, CY015559, CY015583, CY015599, CY015607, CY015631, CY015679, CY015687, CY015695, CY015711, CY015727, CY015735, CY015743, CY015751, CY015759, CY015767, CY015775, CY015799, CY015807, CY015823, CY015839, CY015855, CY015863, CY015887, CY015911, CY015927, CY015935, CY015943, CY015959, CY015967, CY015975, CY015983, CY015991, CY015999, CY016007, CY016015, CY016023, CY016031, CY016039, CY016438, CY016446, CY016454, CY016462, CY016478, CY016598, CY016678, CY016694, CY016702, CY016982, CY017086, CY017094, CY017102, CY017110, CY017134, CY017374, CY017486, CY017494, CY017502, CY017510, CY017534, CY017550, CY017558, CY017566, CY017606, CY017614, CY017622, CY017630, CY017912, CY017920, CY017928, CY017944, CY017960, CY017984, CY017992, CY018928, CY018944, CY018968, CY018976, CY019008, CY019016, CY019024, CY019032, CY019192, CY019328, CY019344, CY019750, CY019830, CY019838, CY019846, CY019862, CY019878, CY019886, CY019926, CY019934, CY020008, CY020040, CY020144, CY020152, CY020264, CY020272, CY020280, CY020344, CY020424, CY020432, CY020896, CY021760, CY022184, CY022192, CY022536, CY022552, CY022560, CY022600, CY023061, CY023069, CY023077, CY025053, CY025216, CY025224, CY025232, CY025240, CY025264, CY025272, CY025280, CY025288, CY025304, CY025336, CY025344, CY025352, CY025360, CY025376, CY025408, CY025416, CY025424, CY025432, CY025440, CY025448, CY025480, CY025488, CY025496, CY025504, CY025520, CY025528, CY025536, CY025558, CY025582, CY025590, CY025598, CY025606, CY025646, CY025662, CY025670, CY025694, CY025702, CY025710, CY025734, CY025742, CY025750, CY025758, CY025774,

CY025790, CY025806, CY025814, CY025838, CY025846, CY025854, CY025862, CY025870, CY025894, CY025910, CY025942, CY025990, CY026006, CY026038, CY026150, CY026182, CY026190, CY026198, CY026206, CY026222, CY026238, CY026246, CY026254, CY026262, CY026270, CY026278, CY026318, CY026334, CY026350, CY026366, CY026374, CY026398, CY026406, CY026510, CY026518, CY026526, CY026534, CY026542, CY026566, CY026582, CY026590, CY026606, CY026614, CY026630, CY026646, CY026654, CY026662, CY026686, CY026694, CY026710, CY026718, CY026726, CY026734, CY026750, CY026766, CY026774, CY026790, CY026798, CY026838, CY026846, CY026854, CY026862, CY026870, CY026886, CY026894, CY026910, CY026918, CY026926, CY026934, CY026942, CY026966, CY026974, CY027006, CY027014, CY027022, CY027046, CY027062, CY027070, CY027094, CY027126, CY027142, CY027150, CY027230, CY027238, CY027246, CY027318, CY027326, CY027334, CY027342, CY027350, CY027358, CY027382, CY027398, CY027406, CY027414, CY027446, CY027454, CY027470, CY027494, CY027558, CY027582, CY027590, CY027598, CY027606, CY027654, CY027662, CY027702, CY027710, CY027726, CY027742, CY027774, CY027782, CY027790, CY027806, CY027814, CY027830, CY027846, CY027854, CY027862, CY027870, CY027878, CY027902, CY027910, CY027918, CY027934, CY027942, CY027982, CY027990, CY028030, CY028046, CY028062, CY028078, CY028110, CY028206, CY028302, CY028310, CY028318, CY028326, CY028342, CY028350, CY028358, CY028406, CY028446, CY028454, CY028462, CY028735, CY028751, CY028775, CY030200, CY030208, CY030730, DQ415330, DQ415331, DQ415332, DQ415333, DQ415334, DQ415336, DQ415337, DQ889686, AB212055, AF028710, AF036359, AF046092, AF084276, AF084277, AF084278, AF115284, AF255742, AF255743, AF255744, AF255745, AF255746, AF255747, AF255750, AF255751, AF255752, AF255753, AJ289871, AJ289872, AJ289873, AJ291400, AY575905, AY626145,

AY627889, AY627895, AY818138, CY014172, CY014178, CY014241, CY014242, CY014243, CY014244, CY014245, CY014246, CY014247, CY014249, CY014251, CY014252, CY014273, CY014297, CY014304, CY014312, CY014387, CY014399, CY014407, CY014423, CY014431, CY014447, CY014471, CY014487, CY014495, CY014503, CY014530, CY014538, CY015009, CY017639, CY017655, CY017663, CY017671, CY017689, CY019353, CY019361, CY019385, CY019409, DQ009920, DQ099775, DQ099776, DQ099779, DQ099780, DQ099781, DQ226139, DQ360840, DQ835314, EF467815, EF587278, EU146691, EU146716, EU263985, AAO15346, AAO15349, ABR15856, ABR28585, ABR28618, ABR28662, ABR29579, ABS50115, ABU80280, ABV29528, ABW36370, ABW36381, ABW36403, ABQ41899, ABR87894, ABS53364, AAF73881, AAF73887, ABR15867, ABR15878, ABR29619, ABR28717, ABR29589, ABR28739, ABU80203, ABU80236, ABV82599, ABW38014, ABW71525, ABW86578, ABX58650, ABX58661, ABS00325, ABS53354, AAG01789, BAG49743, AAF73878, AAF73882, AAF73883, AAF73888, ABQ45462, ABR28574, ABR28695, ABS49925, ABS49947, ABU80192, ABW36359, ABW36392, ABW86589, ABQ51939, ABS00314, ABS53373, ABV31967, AAA51481, AAA43668, AAF73879, AAF73885, AAF73886, ABR29599, ABR28651, ABR28673, ABR28728, ABR29569, ABS50125, ABU80414, ABV29594, ABV82577, ABW71507, AAG01746, AAG01753, AAG01762, CAC85241, ABO44039, ABY40429, ABY16766, ABY16767, ACA25354, ACA25364, AAL87896, ABY40441, ABY16765, AAK69308, AAL87893, AAL87894, CAN89845, ABV55861, ACA25341, AAA52256, AAA52259, AAA52260, CAA81461, AAU25860, AAG17432, CAC85238, ABY16768, AAA52255, AAA52263, AAA52269, AAU25867, ABY81430, ABI54394, ABJ15722, ACE78009, ACE78014, ACF17145, ABR28541, AAA62802, AAA43667, AAA43453, AAA43455, AAA52268,

AAA73110, AAC57416, ABY84688, ABB86875, ABB86885, ABB86905, ABB86935, ACA42424, ACE78019, ACE78022, ABD61253, AAO15347, AAO15348, ABR28552, AAA52258, AAA52262, AAA52265, AAA52270, ABB86925, ACA42434, ACE78008, ACE78024, ACE78026, ABE12638, AAF73884, ABQ45451, ABW36326, ABX58672, ACD85158, AAZ79397, ABB86945, ABJ16475, ABK00128, ABJ52556, ABI96717, ACE78010, ACE78012, ACE78025, ABD61555, ABD79259, ABD78108, ABE27168, ABR15823, ABR28640, ABR28706, AAA43676, AAA73109, AAU25847, AAT72507, ABY51219, ABR28684, ABR28761, ABV82588, AAG01780, CAC85236, AAA43456, AAA52266, ACF49404, AAA74749, AAL26994, ABB86915, ACF17150, ABD62837, ABD62798, ABV31948, ABV31968, AAA43670, AAA43679, AAF75997, AAL87889, ABY51208, AAA52264, AAB50974, ACE78007, ACE78018, ABQ45537, CAC85229, ACE78020, ABE27157, AAL26993, AAA74750, AAA52271, ABR28607, AAG01771. For the pandemic H1N1 sequences: CY081094, CY071330, CY071338, CY071346, CY071354, CY071362, CY083017, CY083025, CY083033, CY083041, CY063841, CY063849, CY080308, CY080316, CY080332, CY073968, CY073001, CY083536, CY073127, CY062293, CY080300, CY074983, CY080324, CY080340, CY071138, CY071146, CY071154, CY075062, CY075070, CY071162, CY071170, CY071178, CY071186, CY071266, CY071274, CY071282, CY071290, CY075078, CY071298, CY071306, CY071314, CY071322, CY071378, CY062982, CY066450, CY063182, CY063190, CY063198, CY063206, CY063214, CY066458, CY066466, CY066474, CY066482, CY064990, CY066490, CY066498, CY066506, CY071210, CY071218, GU562461, CY065875, CY071370, CY081062, CY081070, CY081078, CY081102, HQ393494, CY063006, CY063014, CY071074, CY072529, CY073736, CY061581, CY061110, CY061589, CY061118, CY061126, CY061134, CY061142, CY061150, CY061158, CY061166, CY061174, CY061182,

CY061190, CY061198, CY061206, CY080573, CY062990, CY071058, CY073776, HQ891283, CY083820, HQ834747, CY062069, CY062077, CY062085, CY062093, CY062101, CY062109, CY062117, CY062125, CY062133, CY062141, CY062149, CY062157, CY062165, CY062173, CY062181, CY062189, CY064998, CY065006, CY065014, CY065022, CY065030, CY065038, CY065046, CY065054, CY065062, CY065070, CY065110, CY072537, HM124384, HM569741, HM189305, CY061806, CY061814, HM189644, HM189645, HM189646, HM189647, CY086913.

BIBLIOGRAPHY

1. Barciszewski, J. and B.F.C. Clark, eds. *RNA Biochemistry and Biotechnology*. 1998, Kluwer Academic Publishers: Dordrecht.
2. Sigler, P.B., *An analysis of the structure of tRNA.* Annu Rev Biophys Bioeng, 1975. **4**(00): p. 477-527.
3. Lilley, D.M.J. and F. Eckstein, eds. *Ribozymes and RNA catalysis*. 2008, RSC Publishing: Cambridge, UK.
4. Matzke, M.A. and J.A. Birchler, *RNAi-mediated pathways in the nucleus.* Nat Rev Genet, 2005. **6**(1): p. 24-35.
5. Mohajeri, A. and F.F. Nobandegani, *Detection and evaluation of hydrogen bond strength in nucleic acid base pairs.* J Phys Chem A, 2008. **112**(2): p. 281-95.
6. Jaeger, J.A., D.H. Turner, and M. Zuker, *Improved predictions of secondary structures for RNA.* Proc Natl Acad Sci U S A, 1989. **86**(20): p. 7706-10.
7. Adams, D., *The Hitchhiker's Guide to the Galaxy*. 1979, London: Pan Books.
8. Bompfunewerer, A.F., et al., *Variations on RNA folding and alignment: lessons from Benasque.* J Math Biol, 2008. **56**(1-2): p. 129-44.
9. Hofacker, I., et al., *Fast Folding and Comparsion of RNA secondary structures.* Monatshefte f. Chemie, 1994. **125**(2): p. 167-188.
10. Hofacker, I.L. and P.F. Stadler, *Memory efficient folding algorithms for circular RNA secondary structures.* Bioinformatics, 2006. **22**(10): p. 1172-6.
11. Mathews, D.H., et al., *Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.* J Mol Biol, 1999. **288**(5): p. 911-40.
12. McCaskill, J.S., *The equilibrium partition function and base pair binding probabilities for RNA secondary structure.* Biopolymers, 1990. **29**(6-7): p. 1105-19.
13. Walter, A.E., et al., *Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding.* Proc Natl Acad Sci U S A, 1994. **91**(20): p. 9218-22.
14. Zuker, M. and P. Stiegler, *Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.* Nucleic Acids Res, 1981. **9**(1): p. 133-48.
15. Ding, Y., C.Y. Chan, and C.E. Lawrence, *RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble.* RNA, 2005. **11**(8): p. 1157-66.
16. Ding, Y. and C.E. Lawrence, *A statistical sampling algorithm for RNA secondary structure prediction.* Nucleic Acids Res, 2003. **31**(24): p. 7280-301.
17. Bollenbach, T.J. and D.B. Stern, *Secondary structures common to chloroplast mRNA 3'-untranslated regions direct cleavage by CSP41, an endoribonuclease belonging to the short chain dehydrogenase/reductase superfamily.* J Biol Chem, 2003. **278**(28): p. 25832-8.

18. Simons, R.W. and M. Grunberg-Manago, eds. *RNA Structure and Function*. 1998, Cold Spring Harbor Press.
19. Vogel, C., et al., *Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line.* Mol Syst Biol, 2010. **6**: p. 400.
20. Kudla, G., et al., *Coding-sequence determinants of gene expression in Escherichia coli.* Science, 2009. **324**(5924): p. 255-8.
21. Voges, D., et al., *Analyzing and enhancing mRNA translational efficiency in an Escherichia coli in vitro expression system.* Biochem Biophys Res Commun, 2004. **318**(2): p. 601-14.
22. Tuller, T., et al., *Translation efficiency is determined by both codon bias and folding energy*. Proc Natl Acad Sci U S A, 2010. **107**(8): p. 3645-50.
23. Babitzke, P., C.S. Baker, and T. Romeo, *Regulation of translation initiation by RNA binding proteins.* Annu Rev Microbiol, 2009. **63**: p. 27-44.
24. Morris, K.V., ed. *RNA and the Regulation of Gene Expression: A hidden layer of complexity*. Vol. Norfold, UK. 2008, caister Academic Press.
25. Belasco, J.G. and G. Brawerman, eds. *Control of Messenger RNA Stability*. 1993, Academic Press, Inc. : San Diego.
26. Bernstein, J.A., et al., *Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays.* Proc Natl Acad Sci U S A, 2002. **99**(15): p. 9697-702.
27. Plotkin, J.B. and G. Kudla, *Synonymous but not the same: the causes and consequences of codon bias.* Nat Rev Genet, 2011. **12**(1): p. 32-42.
28. Curtis, T.P. and W.T. Sloan, *Prokaryotic diversity and its limits: microbial community structure in nature and implications for microbial ecology.* Curr Opin Microbiol, 2004. **7**(3): p. 221-6.
29. Curtis, T.P., W.T. Sloan, and J.W. Scannell, *Estimating prokaryotic diversity and its limits.* Proc Natl Acad Sci U S A, 2002. **99**(16): p. 10494-9.
30. Knipe, D.M., et al., eds. *Feilds' Virology*. Vol. 1. c2007, Wolters Kluwer Health/Lippincott Williams & Wilkins: Philadelphia.
31. Ward, B.B., *How many species of prokaryotes are there?* Proc Natl Acad Sci U S A, 2002. **99**(16): p. 10234-6.
32. Doty, P., et al., *Secondary Structure in Ribonucleic Acids.* Proc Natl Acad Sci U S A, 1959. **45**(4): p. 482-99.
33. Mathews, D.H., *Revolutions in RNA secondary structure prediction.* J Mol Biol, 2006. **359**(3): p. 526-32.
34. Fox, G.E. and C.R. Woese, *The architecture of 5S rRNA and its relation to function.* J Mol Evol, 1975. **6**(1): p. 61-76.
35. Fox, G.W. and C.R. Woese, *5S RNA secondary structure.* Nature, 1975. **256**(5517): p. 505-7.
36. Moore, P.B., *The RNA Folding Problem,* in *The RNA World*. 1999, Cold Spring Harbor Laboratory Press.
37. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Res, 2003. **31**(13): p. 3406-15.
38. Hofacker, I.L., *Vienna RNA secondary structure server.* Nucleic Acids Res, 2003. **31**(13): p. 3429-31.

39. Ding, Y., *RNA Secondary Structure Prediction and Gene Regulation by Small RNAs*, in *Frontiers in Computational and Systems Biology*, J. Feng, Editor. 2010, Springer-Verlag: London.
40. Jacobson, A.B., et al., *Some simple computational methods to improve the folding of large RNAs.* Nucleic Acids Res, 1984. **12**(1 Pt 1): p. 45-52.
41. SantaLucia, J., Jr. and D.H. Turner, *Measuring the thermodynamics of RNA secondary structure formation.* Biopolymers, 1997. **44**(3): p. 309-19.
42. Ding, Y. and C.E. Lawrence, *A bayesian statistical algorithm for RNA secondary structure prediction.* Comput Chem, 1999. **23**(3-4): p. 387-400.
43. Pipas, J.M. and J.E. McMahon, *Method for predicting RNA secondary structure.* Proc Natl Acad Sci U S A, 1975. **72**(6): p. 2017-21.
44. Bonhoeffer, S., et al., *RNA multi-structure landscapes. A study based on temperature dependent partition functions.* Eur Biophys J, 1993. **22**(1): p. 13-24.
45. Wuchty, S., et al., *Complete suboptimal folding of RNA and the stability of secondary structures.* Biopolymers, 1999. **49**(2): p. 145-65.
46. Ding, Y., C.Y. Chan, and C.E. Lawrence, *Sfold web server for statistical folding and rational design of nucleic acids.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W135-41.
47. Ding, Y., C.Y. Chan, and C.E. Lawrence, *Clustering of RNA secondary structures with application to messenger RNAs.* J Mol Biol, 2006. **359**(3): p. 554-71.
48. Coleman, J.R., *The PB1-F2 protein of Influenza A virus: increasing pathogenicity by disrupting alveolar macrophages.* Virol J, 2007. **4**: p. 9.
49. Ghedin, E., et al., *Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution.* Nature, 2005. **437**(7062): p. 1162-6.
50. Joshi-Tope, G., et al., *Reactome: a knowledgebase of biological pathways.* Nucleic Acids Res, 2005. **33**(Database issue): p. D428-32.
51. Knipe, D.M., et al., eds. *Feilds' Virology*. Vol. 2. c2007, Wolters Kluwer Health/Lippincott Williams & Wilkins: Philadelphia.
52. Parrish, C.R. and Y. Kawaoka, *The origins of new pandemic viruses: the acquisition of new host ranges by canine parvovirus and influenza A viruses.* Annu Rev Microbiol, 2005. **59**: p. 553-86.
53. Fong, I. and K. Alibek, eds. *New and Evolving Infections of the 21st Centuary*. 2007, Springer: New York, NY.
54. Boni, M.F., et al., *Homologous recombination is very rare or absent in human influenza A virus.* J Virol, 2008. **82**(10): p. 4807-11.
55. Webster, R., et al., *Evolution of Influenza A Viruses in Wild Birds.* Journal of Wildlife Diseases, 2007. **43**(3_Supplement).
56. Crawford, P.C., et al., *Transmission of equine influenza virus to dogs.* Science, 2005. **310**(5747): p. 482-5.
57. Gibbs, M.J. and A.J. Gibbs, *Molecular virology: was the 1918 pandemic caused by a bird flu?* Nature, 2006. **440**(7088): p. E8; discussion E9-10.
58. Taubenberger, J.K., et al., *Characterization of the 1918 influenza virus polymerase genes.* Nature, 2005. **437**(7060): p. 889-93.
59. Brown, I.H., et al., *Multiple genetic reassortment of avian and human influenza A viruses in European pigs, resulting in the emergence of an H1N2 virus of novel genotype.* J Gen Virol, 1998. **79 (Pt 12)**: p. 2947-55.

60. Castrucci, M.R., et al., *Genetic Reassortment between Avian and Human Influenza-a Viruses in Italian Pigs.* Virology, 1993. **193**(1): p. 503-506.
61. Horimoto, T. and Y. Kawaoka, *Pandemic threat posed by avian influenza A viruses.* Clin Microbiol Rev, 2001. **14**(1): p. 129-49.
62. Scholtissek, C., *Molecular evolution of influenza viruses.* Virus Genes, 1995. **11**(2-3): p. 209-215.
63. Zhou, N.N., et al., *Genetic reassortment of avian, swine, and human influenza A viruses in American pigs.* J Virol, 1999. **73**(10): p. 8851-6.
64. Kuiken, T., et al., *Host species barriers to influenza virus infections.* Science, 2006. **312**(5772): p. 394-7.
65. Tarendeau, F., et al., *Host determinant residue lysine 627 lies on the surface of a discrete, folded domain of influenza virus polymerase PB2 subunit.* PLoS Pathog, 2008. **4**(8): p. e1000136.
66. *Pandemic (H1N1) 2009- upate 112.* Global Alert and Response, 6 August 2010.
67. Smith, G.J., et al., *Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic.* Nature, 2009. **459**(7250): p. 1122-5.
68. Garten, R.J., et al., *Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans.* Science, 2009. **325**(5937): p. 197-201.
69. CDC. *http://www.cdc.gov/h1n1flu/*. 2009 H1N1 Flu 2010 [cited 2011 November 22].
70. WHO, *H5N1 avian influenza: Timeline of major events*. 2008.
71. WHO, *Cumulative Number of Confirmed Human Cases of Avian Influenza A/(H5N1).* . 2008.
72. Holmes, E., *Novel and Re-Emerging Respiratory Viral Diseases*. 2008, Chichester, UK: John Wiley & Sons, Ltd.
73. Smith, G.J., et al., *Evolution and adaptation of H5N1 influenza virus in avian and human hosts in Indonesia and Vietnam.* Virology, 2006. **350**(2): p. 258-68.
74. Chen, G.W. and S.R. Shih, *Genomic signatures of influenza A pandemic (H1N1) 2009 virus.* Emerg Infect Dis, 2009. **15**(12): p. 1897-1903.
75. Zimmer, S.M. and D.S. Burke, *Historical perspective--Emergence of influenza A (H1N1) viruses.* N Engl J Med, 2009. **361**(3): p. 279-85.
76. Squires, B., et al., *BioHealthBase: informatics support in the elucidation of influenza virus host pathogen interactions and virulence.* Nucleic Acids Res, 2008. **36**(Database issue): p. D497-503.
77. Boone, D. and R. Castenholz, eds. *The Archaea and the Deeply Branching and Phototrophic Bacteria*. 2 ed. Bergey's Manual of Systematic Bacteriology. Vol. 1. 2000, Springer: New York.
78. Snyder, L. and W. Champness, *Molecular Genetics of Bacteria*. 1997, Washington, DC: ASM Press.
79. Galtier, N. and J.R. Lobry, *Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes.* J Mol Evol, 1997. **44**(6): p. 632-6.
80. Wang, H.C., E. Susko, and A.J. Roger, *On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors.* Biochem Biophys Res Commun, 2006. **342**(3): p. 681-4.

81. Hurst, L.D. and A.R. Merchant, *High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes.* Proc Biol Sci, 2001. **268**(1466): p. 493-7.
82. Chen, S.L., et al., *Codon usage between genomes is constrained by genome-wide mutational processes.* Proc Natl Acad Sci U S A, 2004. **101**(10): p. 3480-5.
83. Hershberg, R. and D.A. Petrov, *Selection on codon bias.* Annu Rev Genet, 2008. **42**: p. 287-99.
84. Hildebrand, F., A. Meyer, and A. Eyre-Walker, *Evidence of selection upon genomic GC-content in bacteria.* PLoS Genet, 2010. **6**(9).
85. Kimura, M., *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.* J Mol Evol, 1980. **16**(2): p. 111-20.
86. Sharp, P.M., et al., *Variation in the strength of selected codon usage bias among bacteria.* Nucleic Acids Res, 2005. **33**(4): p. 1141-53.
87. Sharp, P.M. and W.H. Li, *The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias.* Mol Biol Evol, 1987. **4**(3): p. 222-30.
88. Sharp, P.M. and W.H. Li, *The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications.* Nucleic Acids Res, 1987. **15**(3): p. 1281-95.
89. Dobzhansky, T., *Nothing in Biology Makes Sense Except in the Light of Evolution.* The American Biology Teacher, 1973. **35**: p. 125-.
90. Huang, T.S., P. Palese, and M. Krystal, *Determination of influenza virus proteins required for genome replication.* J Virol, 1990. **64**(11): p. 5669-73.
91. Kimura, N., et al., *An in vivo study of the replication origin in the influenza virus complementary RNA.* J Biochem, 1993. **113**(1): p. 88-92.
92. Gabriel, G., et al., *The viral polymerase mediates adaptation of an avian influenza virus to a mammalian host.* Proc Natl Acad Sci U S A, 2005. **102**(51): p. 18590-5.
93. Wasilenko, J.L., et al., *NP, PB1, and PB2 viral genes contribute to altered replication of H5N1 avian influenza viruses in chickens.* J Virol, 2008. **82**(9): p. 4544-53.
94. Hatta, M., et al., *Growth of H5N1 influenza A viruses in the upper respiratory tracts of mice.* PLoS Pathog, 2007. **3**(10): p. 1374-9.
95. Naffakh, N., et al., *Genetic analysis of the compatibility between polymerase proteins from human and avian strains of influenza A viruses.* J Gen Virol, 2000. **81**(Pt 5): p. 1283-91.
96. Subbarao, E.K., W. London, and B.R. Murphy, *A single amino acid in the PB2 gene of influenza A virus is a determinant of host range.* J Virol, 1993. **67**(4): p. 1761-4.
97. Kawaoka, Y., S. Krauss, and R.G. Webster, *Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics.* J Virol, 1989. **63**(11): p. 4603-8.
98. Taubenberger, J.K. and D.M. Morens, *1918 Influenza: the mother of all pandemics.* Emerg Infect Dis, 2006. **12**(1): p. 15-22.
99. Gensheimer, K.F., et al., *Influenza pandemic preparedness.* Emerg Infect Dis, 2003. **9**(12): p. 1645-8.
100. Peiris, J.S., et al., *Re-emergence of fatal human influenza A subtype H5N1 disease.* Lancet, 2004. **363**(9409): p. 617-9.
101. Webster, R.G. and E.A. Govorkova, *H5N1 influenza--continuing evolution and spread.* N Engl J Med, 2006. **355**(21): p. 2174-7.

102. Paddison, P.J., et al., *Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells.* Genes Dev, 2002. **16**(8): p. 948-58.
103. Johnson, N.P. and J. Mueller, *Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic.* Bull Hist Med, 2002. **76**(1): p. 105-15.
104. Greenbaum, B.D., et al., *Patterns of evolution and host gene mimicry in influenza and other RNA viruses.* PLoS Pathog, 2008. **4**(6): p. e1000079.
105. Rabadan, R., A.J. Levine, and H. Robins, *Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes.* J Virol, 2006. **80**(23): p. 11887-91.
106. Coleman, J.R., et al., *Virus attenuation by genome-scale changes in codon pair bias.* Science, 2008. **320**(5884): p. 1784-7.
107. Nakamura, Y., T. Gojobori, and T. Ikemura, *Codon usage tabulated from international DNA sequence databases: status for the year 2000.* Nucleic Acids Res, 2000. **28**(1): p. 292.
108. Sharp, P.M., et al., *DNA sequence evolution: the sounds of silence.* Philos Trans R Soc Lond B Biol Sci, 1995. **349**(1329): p. 241-7.
109. Brown, I.H., et al., *Recent epidemiology and ecology of influenza A viruses in avian species in Europe and the Middle East.* Dev Biol (Basel), 2006. **124**: p. 45-50.
110. Davison, S., R.J. Eckroade, and A.F. Ziegler, *A review of the 1996-98 nonpathogenic H7N2 avian influenza outbreak in Pennsylvania.* Avian Dis, 2003. **47**(3 Suppl): p. 823-7.
111. Longini, I.M., Jr., et al., *Containing pandemic influenza at the source.* Science, 2005. **309**(5737): p. 1083-7.
112. Webster, R.G., *Virology. A molecular whodunit.* Science, 2001. **293**(5536): p. 1773-5.
113. Marsh, G.A., et al., *Highly conserved regions of influenza a virus polymerase gene segments are critical for efficient viral RNA packaging.* J Virol, 2008. **82**(5): p. 2295-304.
114. Tumpey, T.M., et al., *A two-amino acid change in the hemagglutinin of the 1918 influenza virus abolishes transmission.* Science, 2007. **315**(5812): p. 655-9.
115. Webby, R.J. and R.G. Webster, *Are we ready for pandemic influenza?* Science, 2003. **302**(5650): p. 1519-22.
116. Hagan, J.J., et al., *Stimulation of 5-HT1B receptors causes hypothermia in the guinea pig.* Eur J Pharmacol, 1997. **331**(2-3): p. 169-74.
117. Prozesky, O., *Body temperature of birds in relation to nesting habits.* Nature, 1963. **197**: p. 401-402.
118. Kiley, J.P., W.D. Kuhlmann, and M.R. Fedde, *Respiratory and cardiovascular responses to exercise in the duck.* J Appl Physiol, 1979. **47**(4): p. 827-33.
119. Bao, Y., et al., *The influenza virus resource at the National Center for Biotechnology Information.* J Virol, 2008. **82**(2): p. 596-601.
120. Kawaoka, Y., et al., *Molecular basis for the generation in pigs of influenza A viruses with pandemic potential.* Journal of Virology, 1998. **72**(9): p. 7367-7373.
121. Taubenberger, J.K., *The origin and virulence of the 1918 "Spanish" influenza virus.* Proc Am Philos Soc, 2006. **150**(1): p. 86-112.
122. Fukuyama, S. and Y. Kawaoka, *The pathogenesis of influenza virus infections: the contributions of virus and host factors.* Curr Opin Immunol, 2011.
123. Boivin, G., et al., *Predicting influenza infections during epidemics with use of a clinical case definition.* Clinical Infectious Diseases, 2000. **31**(5): p. 1166-1169.

124. Carrat, F., et al., *A 'small-world-like' model for comparing interventions aimed at preventing and controlling influenza pandemics.* Bmc Medicine, 2006. **4**.
125. degli Atti, M.L.C., et al., *Mitigation Measures for Pandemic Influenza in Italy: An Individual Based Model Considering Different Scenarios.* PLoS One, 2008. **3**(3).
126. Ferguson, N.M., et al., *Strategies for containing an emerging influenza pandemic in Southeast Asia.* Nature, 2005. **437**(7056): p. 209-214.
127. Ferguson, N.M., et al., *Strategies for mitigating an influenza pandemic.* Nature, 2006. **442**(7101): p. 448-452.
128. Flahault, A., et al., *Strategies for containing a global influenza pandemic.* Vaccine, 2006. **24**(44-46): p. 6751-5.
129. Fraser, C., et al., *Pandemic potential of a strain of influenza A (H1N1): early findings.* Science, 2009. **324**(5934): p. 1557-61.
130. Germann, T.C., et al., *Mitigation strategies for pandemic influenza in the United States.* Proc Natl Acad Sci U S A, 2006. **103**(15): p. 5935-40.
131. Ginsberg, J., et al., *Detecting influenza epidemics using search engine query data.* Nature, 2009. **457**(7232): p. 1012-4.
132. Halloran, M.E., et al., *Modeling targeted layered containment of an influenza pandemic in the United States.* Proc Natl Acad Sci U S A, 2008. **105**(12): p. 4639-44.
133. Longini, I.M., *A Mathematical-Model for Predicting the Geographic Spread of New Infectious Agents.* Mathematical Biosciences, 1988. **90**(1-2): p. 367-383.
134. Rvachev, L.A. and I.M. Longini, *A Mathematical-Model for the Global Spread of Influenza.* Mathematical Biosciences, 1985. **75**(1): p. 1-1.
135. Schmidt, R., T. Waligora, and L. Gierl, *Predicting influenza waves with health insurance data.* Computational Intelligence, 2006. **22**(3-4): p. 224-237.
136. Thursky, K., et al., *Working towards a simple case definition for influenza surveillance.* J Clin Virol, 2003. **27**(2): p. 170-9.
137. Chen, L.M., et al., *Genetic compatibility and virulence of reassortants derived from contemporary avian H5N1 and human H3N2 influenza A viruses.* PLoS Pathog, 2008. **4**(5): p. e1000072.
138. Kimble, J.B., et al., *Compatibility of H9N2 avian influenza surface genes and 2009 pandemic H1N1 internal genes for transmission in the ferret model.* Proc Natl Acad Sci U S A, 2011. **108**(29): p. 12084-8.
139. Sun, Y., et al., *High genetic compatibility and increased pathogenicity of reassortants derived from avian H9N2 and pandemic H1N1/2009 influenza viruses.* Proc Natl Acad Sci U S A, 2011. **108**(10): p. 4164-9.
140. Mitchell, T., *Machine Learning*. 2010: McGraw Hill
141. Brower-Sinning, R., et al., *The role of RNA folding free energy in the evolution of the polymerase genes of the influenza A virus.* Genome Biol, 2009. **10**(2): p. R18.
142. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software, 2010. **33**(1): p. 1-22.
143. Simon, N., et al., *Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent.* Journal of Statistical Software, 2011. **39**(5): p. 1-13.
144. Webster, R.G., et al., *Evolution and ecology of influenza A viruses.* Microbiol Rev, 1992. **56**(1): p. 152-79.
145. Kinney, R.M., et al., *Avian virulence and thermostable replication of the North American strain of West Nile virus.* J Gen Virol, 2006. **87**(Pt 12): p. 3611-22.

146. LaDeau, S.L., A.M. Kilpatrick, and P.P. Marra, *West Nile virus emergence and large-scale declines of North American bird populations.* Nature, 2007. **447**(7145): p. 710-3.
147. Rappole, J.H., S.R. Derrickson, and Z. Hubalek, *Migratory birds and spread of West Nile virus in the Western Hemisphere.* Emerg Infect Dis, 2000. **6**(4): p. 319-28.
148. Reed, L.M., et al., *Declining mortality in American crow (Corvus brachyrhynchos) following natural West Nile virus infection.* Avian Dis, 2009. **53**(3): p. 458-61.
149. Lloyd-Smith, J.O., et al., *Epidemic dynamics at the human-animal interface.* Science, 2009. **326**(5958): p. 1362-7.
150. Fooks, A.R., et al., *European bat lyssaviruses: an emerging zoonosis.* Epidemiol Infect, 2003. **131**(3): p. 1029-39.
151. Padula, P.J., et al., *Hantavirus pulmonary syndrome outbreak in Argentina: molecular evidence for person-to-person transmission of Andes virus.* Virology, 1998. **241**(2): p. 323-30.
152. Carbone, K.M., *Borna disease virus and human disease.* Clin Microbiol Rev, 2001. **14**(3): p. 513-27.
153. Lundstrom, J.O., *Mosquito-borne viruses in western Europe: a review.* J Vector Ecol, 1999. **24**(1): p. 1-39.
154. Hoogstraal, H., *The epidemiology of tick-borne Crimean-Congo hemorrhagic fever in Asia, Europe, and Africa.* J Med Entomol, 1979. **15**(4): p. 307-417.
155. Charrel, R.N., et al., *Arenaviruses and hantaviruses: from epidemiology and genomics to antivirals.* Antiviral Res, 2011. **90**(2): p. 102-14.
156. Delgado, S., et al., *Chapare virus, a newly discovered arenavirus isolated from a fatal hemorrhagic fever case in Bolivia.* PLoS Pathog, 2008. **4**(4): p. e1000047.
157. Schlegel, M., et al., *Dobrava-belgrade virus spillover infections, Germany.* Emerg Infect Dis, 2009. **15**(12): p. 2017-20.
158. Schmaljohn, C. and B. Hjelle, *Hantaviruses: a global disease problem.* Emerg Infect Dis, 1997. **3**(2): p. 95-104.
159. Banyard, A.C., M. Hartley, and A.R. Fooks, *Reassessing the risk from rabies: a continuing threat to the UK?* Virus Res, 2010. **152**(1-2): p. 79-84.
160. Charrel, R.N. and X. de Lamballerie, *Arenaviruses other than Lassa virus.* Antiviral Res, 2003. **57**(1-2): p. 89-100.
161. Jay, M.T., C. Glaser, and C.F. Fulhorst, *The arenaviruses.* J Am Vet Med Assoc, 2005. **227**(6): p. 904-15.
162. Mackenzie, J.S., et al., *Emerging viral diseases of Southeast Asia and the Western Pacific.* Emerg Infect Dis, 2001. **7**(3 Suppl): p. 497-504.
163. Valassina, M., M.G. Cusi, and P.E. Valensin, *A Mediterranean arbovirus: the Toscana virus.* J Neurovirol, 2003. **9**(6): p. 577-83.
164. Chatziandreou, N., et al., *Relationships and host range of human, canine, simian and porcine isolates of simian virus 5 (parainfluenza virus 5).* J Gen Virol, 2004. **85**(Pt 10): p. 3007-16.
165. Wang, F., et al., *Environmental adaptation: genomic analysis of the piezotolerant and psychrotolerant deep-sea iron reducing bacterium Shewanella piezotolerans WP3.* PLoS One, 2008. **3**(4): p. e1937.
166. Angilletta, M., *Thermal Adaption: A Theroretical and Empirical Synthesis*. 2009, Oxford: Oxford University Press. 289.

167. Wada, A. and A. Suyama, *Local stability of DNA and RNA secondary structure and its relation to biological functions.* Prog Biophys Mol Biol, 1986. **47**(2): p. 113-57.
168. Drummond, D.A. and C.O. Wilke, *Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution.* Cell, 2008. **134**(2): p. 341-52.
169. Eyre-Walker, A., *Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy?* Mol Biol Evol, 1996. **13**(6): p. 864-72.
170. Ikemura, T., *Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes.* J Mol Biol, 1981. **146**(1): p. 1-21.
171. Stoletzki, N. and A. Eyre-Walker, *Synonymous codon usage in Escherichia coli: selection for translational accuracy.* Mol Biol Evol, 2007. **24**(2): p. 374-81.
172. Zhou, T., M. Weems, and C.O. Wilke, *Translationally optimal codons associate with structurally sensitive sites in proteins.* Mol Biol Evol, 2009. **26**(7): p. 1571-80.
173. Brenner, D., N. Krieg, and J. Staley, eds. *The Proteobacteria: Part B The Gammaproteobacteria*. 2 ed. Bergey's Manual of Systematic Bacteriology. Vol. 2. 2000, Springrt: USA.
174. Williams, D.L., et al., *Molecular basis of the defective heat stress response in Mycobacterium leprae.* J Bacteriol, 2007. **189**(24): p. 8818-27.
175. Slayers, A. and D. Whitt, *Bacterial Pathogenesis: A Molecular Approach*. 2 ed. 2002, Washington, DC: ASM Press.
176. Franzblau, S.G. and E.B. Harris, *Biophysical optima for metabolism of Mycobacterium leprae.* J Clin Microbiol, 1988. **26**(6): p. 1124-9.
177. Geissmann, T., S. Marzi, and P. Romby, *The role of mRNA structure in translational control in bacteria.* RNA Biol, 2009. **6**(2): p. 153-60.
178. Henkin, T., in *Riboswitches*. p. 207-214.
179. Bulmer, M., *Coevolution of codon usage and transfer RNA abundance.* Nature, 1987. **325**(6106): p. 728-30.
180. Gouy, M. and C. Gautier, *Codon usage in bacteria: correlation with gene expressivity.* Nucleic Acids Res, 1982. **10**(22): p. 7055-74.
181. Varenne, S., et al., *Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains.* J Mol Biol, 1984. **180**(3): p. 549-76.
182. Kawamoto, T., S. Ochiai, and H. Kagi, *Changes in the structure of water deduced from the pressure dependence of the Raman OH frequency.* J Chem Phys, 2004. **120**(13): p. 5867-70.
183. Hurst, C., ed. *Viral Ecology*. 2000, Academic Press: San Diego.
184. Grantham, R., et al., *Codon catalog usage and the genome hypothesis.* Nucleic Acids Res, 1980. **8**(1): p. r49-r62.
185. Ikemura, T., *Codon usage and tRNA content in unicellular and multicellular organisms.* Mol Biol Evol, 1985. **2**(1): p. 13-34.
186. Kurland, C.G., *Codon bias and gene expression.* FEBS Lett, 1991. **285**(2): p. 165-9.
187. Robinson, M., et al., *Codon usage can affect efficiency of translation of genes in Escherichia coli.* Nucleic Acids Res, 1984. **12**(17): p. 6663-71.
188. Van Dover, C., *The Ecology of deep-sea hydrothermal vents*. 2000, Princeton, NJ: Princeton University Press. 424.

189. Takai, K., et al., *Deferribacter desulfuricans sp. nov., a novel sulfur-, nitrate- and arsenate-reducing thermophile isolated from a deep-sea hydrothermal vent.* Int J Syst Evol Microbiol, 2003. **53**(Pt 3): p. 839-46.
190. Takami, H., et al., *Thermoadaptation trait revealed by the genome sequence of thermophilic Geobacillus kaustophilus.* Nucleic Acids Res, 2004. **32**(21): p. 6292-303.
191. De Vos, P., et al., eds. *The Firmicutes*. Bergey's Manual of Systematic Bacteriology. Vol. 3. 2009, Springer: Dordrechtm Heidelberg, London, New York.
192. Brenner, D., N. Krieg, and J. Staley, eds. *The Proteobacteria: Part C The Alpha-, Beta-, Delta- and Epsilonproteobacteria*. Bergey's Manual of Systematic Bacteriology. Vol. 2. 2005, Springer: USA.
193. Hou, S., et al., *Genome sequence of the deep-sea gamma-proteobacterium Idiomarina loihiensis reveals amino acid fermentation as a source of carbon and energy.* Proc Natl Acad Sci U S A, 2004. **101**(52): p. 18036-41.
194. Mehta, M.P. and J.A. Baross, *Nitrogen fixation at 92 degrees C by a hydrothermal vent archaeon.* Science, 2006. **314**(5806): p. 1783-6.
195. Nakagawa, S., et al., *Deep-sea vent epsilon-proteobacterial genomes provide insights into emergence of pathogens.* Proc Natl Acad Sci U S A, 2007. **104**(29): p. 12146-50.
196. Lu, J., Y. Nogi, and H. Takami, *Oceanobacillus iheyensis gen. nov., sp. nov., a deep-sea extremely halotolerant and alkaliphilic species isolated from a depth of 1050 m on the Iheya Ridge.* FEMS Microbiol Lett, 2001. **205**(2): p. 291-7.
197. SS Bae, Y.K., SH Yang, JK Lim, JH Jeon, HS Lee, SG Kang, SJ Kim, JH Lee, *Thermococcus onnurineus sp. nov., a hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent area at the PACMANUS field.* Journal of microbiology and biotechnology **16**(11): p. 1826-1831.
198. Jannasch, H., et al., *Thiomicrospira crunogena sp. nov., a Colorless, Sulfur-Oxidizing Bacterium from a Deep-Sea Hydrothermal Vent.* Int J Syst Bacteriol, 1985. **35**: p. 422-424.
199. Qin, Q.L., et al., *The complete genome of Zunongwangia profunda SM-A87 reveals its adaptation to the deep-sea environment and ecological role in sedimentary organic nitrogen degradation.* BMC Genomics, 2010. **11**: p. 247.
200. Holden, M.T., et al., *Complete genomes of two clinical Staphylococcus aureus strains: evidence for the rapid evolution of virulence and drug resistance.* Proc Natl Acad Sci U S A, 2004. **101**(26): p. 9786-91.
201. Cole, S.T., et al., *Massive gene decay in the leprosy bacillus.* Nature, 2001. **409**(6823): p. 1007-11.
202. Bork, P., et al., *Predicting function: from genes to genomes and back.* J Mol Biol, 1998. **283**(4): p. 707-25.
203. Moreno-Hagelsieb, G. and K. Latimer, *Choosing BLAST options for better detection of orthologs as reciprocal best hits.* Bioinformatics, 2008. **24**(3): p. 319-24.
204. Tatusov, R.L., E.V. Koonin, and D.J. Lipman, *A genomic perspective on protein families.* Science, 1997. **278**(5338): p. 631-7.