Thesis submitted for the degree of Doctor of Philosophy at UCL

Using phylogenetic models to characterise natural selection from molecular data

Asif U. Tamuri

Division of Mathematical Biology
MRC National Institute for Medical Research
London

I, Asif U. Tamuri, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. Specifically:

Chapter 3 This chapter was published in Tamuri *et al.* (2009) and includes contributions by Mario dos Reis (MdR), Alan Hay (AH) and Richard Goldstein (RG). The original idea was developed by Asif Tamuri (AT), MdR and RG. The programs used for the analysis were written by AT. The phylogenetic analysis and characterisation of changes in selective constraints was performed by AT, based on sequence alignments constructed by MdR.

Chapter 4 This chapter was published in dos Reis *et al.* (2011), with contributions by MdR, AH and RG. The original idea was developed by AT, MdR and RG. The programs used for the analysis were written by AT. Analysis was performed by AT, based on sequence alignments constructed by MdR. The curve fit to adaptedness was carried out by RG.

Chapter 5 This chapter was published in Tamuri *et al.* (2012), with contributions by MdR and RG. The original idea was developed by AT, MdR and RG. The mammalian mitochondrial sequence alignment used in section 5.3.2 was provided by MdR. The programs used for analysis were written by AT and analysis was performed by AT.

Abstract

Molecular phylogenetics is the application of mathematical and computational techniques to analyse molecular sequences and make inferences about their evolutionary relationships. There is substantial interest in developing probabilistic models of evolution that effectively detect, locate and characterise different type of selection in genes, driven by the relationship of selection to protein structural constraints and function. In this thesis we propose novel approaches that can be used not only to detect the presence of selection but also to characterise its kind and strength. We first develop a phylogenetic method to identify changes in selective constraints and use it to identify those mutations that allow influenza viruses from avian origin to spread successfully in the human population. The model explicitly takes into account differences in the equilibrium frequencies of amino acids in different hosts and locations. We then use these results to develop a measure of the level of adaptation of any given influenza virus sequence to the selective constraints imposed by avian or human hosts. We show that adaptation to the human host has been gradual when applied to historical data. Our results also indicate that the 1918 influenza virus had undergone a period of pre-adaptation prior to 1918 when compared to the adaptation of other avian influenza viruses. Finally, we develop a codon-based model of mutation-selection to estimate the distribution of selection coefficients and find that we can recover distributions similar to those expected by population genetics theory. We show that the distribution of mammalian mitochondrial proteins is bimodal with the majority of mutations being deleterious. When we apply the model to the PB2 influenza polymerase protein following a host shift from birds to humans, we find a trimodal distribution with a significant proportion of advantageous substitutions.

Acknowledgements

First and foremost, my heartfelt thanks to my supervisor Dr Richard Goldstein. I am indebted to him for his immeasurable support, guidance and encouragement, and for fostering my interest in mathematical biology. He has been a wonderful supervisor, mentor and collaborator.

My sincere gratitude to my colleague and collaborator Dr Mario dos Reis, for guiding me through the ocean of molecular evolution, always being willing to point me in the right direction, and providing ideas and data for analysis.

Thanks to my thesis committee members — Professor Christine Orengo, Dr John McCauley and Dr Alan Hay — for their advice and assistance.

Special thanks to the other lab members of Mathematical Biology at the NIMR for their friendship, making the last three years thought-provoking and enjoyable.

This work was funded by the Wellcome Trust.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Background

1.1.1. Computational molecular evolution

A feature of modern biological research is the continuing collection and analysis of molecular sequences. Databases such as GenBank and EMBL, which hold DNA, RNA and protein sequences from organisms across the tree of life, continue to grow at an exponential rate (Cochrane *et al.*, 2011). They include not only the complete genomes of familiar organisms but also the results of massive worldwide surveillance programs of pathogens such as influenza.

Gene sequences are units of DNA or RNA sequence that are transferred from parent to offspring. Protein-coding genes are those genes which, given the genetic code, are translated from sequences of nucleotides into chains of amino acids which in turn form macromolecular structures called proteins. Proteins are involved in almost all biological processes and serve a wide variety of functions in organisms, including catalysis of metabolic reactions, facilitating immune response, signalling within cells, storage of energy and structural functions giving shape to a cell. However, the replication and inheritance of genes is not perfect and errors are introduced. Changes of single nucleotide bases, as well as insertions, deletions or rearrangement of sections of the parent gene lead to a different, but related, gene for the progeny. These mutations are markers of evolutionary history.

Computational molecular evolution, or molecular phylogenetics, is the application of mathematical, statistical and computational techniques to analyse these differing sequences and make inferences about their evolutionary relationships. For example, one can take sequences from different species and calculate evolutionary distances to assist in the construc-

tion of a phylogenetic tree, providing the genealogy of species and determining which share common ancestors (Yang & Rannala, 2012). Zuckerkandl & Pauling (1965) were the first to use the amino acid sequences from related species to estimate rate of amino acid (and nucleotide) substitution and dates of divergence of pairs of species. Models have become increasingly sophisticated since then and descriptive probabilistic models describing how sequences evolve are now in common use. They operate at the DNA/RNA, codon or protein level and can be general descriptions, specific to a family of proteins (e.g. Adachi & Hasegawa, 1996; Dimmic *et al.*, 2002), specific to structural environments (e.g. Jones *et al.*, 1994; Koshi & Goldstein, 1995; Liò & Goldman, 1998) or allowing every site in every protein to have its own evolutionary model (Halpern & Bruno, 1998). In this way, phylogenetic research seeks to understand the evolutionary processes that are driving changes in gene sequences.

Probabilistic models of evolution are a more explicit description of the evolutionary process than models that use counting methods or sequence similarity measures, and provide a better measure of evolutionary distances, because they allow for the possibility of multiple substitutions at the same site. They are also fundamental to statistical likelihood and Bayesian methods of estimating phylogenetic relationships. These can provide measures of confidence in estimates and measures of significance for patterns present in data that diverge from a given model. Phylogenetic analysis recognises that any analyses of sequences should account for their relatedness and should not be treated as independent observations. Ignoring or reducing the effect of these relationships can lead to an over-significance of similarities (or differences) between sequences. Conservation between a pair of sequences may not necessarily indicate selective pressure as it could simply be due to a short divergence time.

There is substantial interest in methods that effectively detect, locate and characterise different types of selection in genes. This is driven by the relationship of selection to pro-

tein structure and function. Changes in protein sequences can be responsible for changes in protein function and, ultimately, changes in the phenotype of an organism. A mutation can arise in an individual in a population and this will affect the fitness of that individual compared to the rest of the population. As a mutation on a gene can change the protein that is expressed by that gene, the detection and location of selection may give clues to the adaptation of the protein for its function. For example, sites that are conserved in homologous proteins might indicate that those positions are essential for effective functioning of the protein. In contrast, other sites may show repeated amino acid changes, more than would be expected by chance, indicating that those positions might be in an evolutionary arms race with an attacking pathogen (Lam *et al.*, 2010). The underlying framework of most tests of selection is the neutral theory of molecular evolution proposed by Kimura (1983), which states that most of the changes that are fixed at the molecular level are effectively neutral (i.e. having small effect on fitness) and that the fate of those mutations are dominated by the effect of random genetic drift. This leads to chance fixation or loss of that particular mutant. Many tests for selection look for deviation from this state of neutrality (e.g. Akashi, 1995; McDonald & Kreitman, 1991). An example of a popular phylogenetic method for detecting selection is the analysis of the nonsynonymous to synonymous rate ratio (covered in Chapter 2), which is effective in locating those codons in a protein-coding gene that show an increased rate of repeated amino-acid change (diversifying selection) but may not be so useful for detecting other types of positive selection (Yang & dos Reis, 2011; Yang & Nielsen, 2002).

We can split natural selection into two types: a) positive selection that drives fixation of advantageous mutations or b) negative (or purifying) selection that prevents fixation of deleterious mutations and conserves the current sequence. We can further divide positive selection into a) diversifying selection that drives the adoption of frequent amino acid changes (perhaps in an arms race), b) directional selection that drives the rapid fixation of a particu-

lar set of amino-acid changes and c) balancing selection or frequency-dependent selection (for a review, see Anisimova & Liberles, 2012). The proportion and relative strength of different types of selection will be influenced by the particular function of the protein under consideration e.g. sites involved in protein-protein interactions will tend to conserve areas of hydrophobic residues on the surface. Rates of conservation can range from very highly conserved proteins such as histones to rapidly evolving proteins such as fibrinopeptides (Dickerson, 1971). Proteins that bind small molecules may accept the substitution of several similar amino acids, all conferring the same fitness, in the binding region as long as they do not affect the charge or pocket size. We will describe in chapter 2 some of the models that attempt to represent these characteristic properties.

1.1.2. Influenza

As the First World war was coming to an end, the editors of the Journal of the American Medical Association wrote (JAMA, 1918):

> ...1918 has gone: a year momentous as the termination of the most cruel war in the annals of the human race; a year which marked, the end at least for a time, of man's destruction of man; unfortunately a year in which developed a most fatal infectious disease causing the death of hundreds of thousands of human beings. Medical science for four and one-half years devoted itself to putting men on the firing line and keeping them there. Now it must turn with its whole might to combating the greatest enemy of all—infectious disease.

Better understanding of diseases has been one of the primary drivers for better understanding of underlying biological processes. Our knowledge of why certain diseases seem to affect humans much more readily than others, or why a particular strain of a viral patho-

gen spreads more rapidly than another, can lead to more effective strategies of defending against those pathogens.

Influenza is an infectious disease caused by the *Myxovirus influenza* family of viruses, of which there are three distinct genera (A, B and C). Influenza A is the most virulent type and is responsible for both localised outbreaks and pandemics. Although influenza A viruses are most known as causes of significant morbidity and mortality in humans, this genera's viruses are also found in other animals including swine, horses, sea mammals and birds, of which waterfowl are the natural reservoir (Webster *et al.*, 1992). The virus usually attacks cells in the lower respiratory tract (the bronchioles and alveoli) in birds and replicates in the intestinal tract but causes little or no disease (Webster *et al.*, 1992). In humans the virus typically targets the upper respiratory tract (trachea and bronchi) (van Riel *et al.*, 2007).

Genome & replication

The core of the virus holds its genome of around 13,500 bases and is surrounded by the viral envelope. The genome is comprised of 8 segments of negative-sense RNA, which encodes 10 proteins: haemagglutinin (HA) and neuraminidase (NA), two glycoproteins present on the viral envelope, both of which are recognised by antibodies and are also the target of antiviral drugs (Wilson & von Itzstein, 2003); the heterotrimeric polymerase complex (PA, PB1 and PB2) responsible for transcription and replication; the matrix proteins (M1 and M2), nucleocapsid protein (NP) and non-structural proteins (NS1 and NS2) (Baigent & McCauley, 2003) (Figure 1.1, from Taubenberger & Kash (2010)).

Haemagglutinin assists the virus in identifying and binding to the wall of host cells at particular receptor sites allowing the viral genome to enter into the host cell (Kumar & Clark, 1999). The genome and core proteins are then released into the cytoplasm, which then forms a complex for transportation into the cell nucleus. RNA replication, transcription and packaging with the nucleoprotein takes place in the nucleus and is then transported

Figure 1.1.: Cartoon representation of an influenza A virus, from Taubenberger & Kash (2010).

back into the cytoplasm and translated into proteins. The copies of the RNA and other viral proteins are assembled together outside the cell nucleus in a bulge created by haemagglutinin and neuraminidase gathering at the cell membrane. The replicated viruses then leave the infected cell surrounded by haemagglutinin and neuraminidase proteins, and the host cell dies.

The influenza virus, like many RNA viruses, mutates rapidly. Three important evolutionary processes occur with influenza. First, the established 'seasonal' influenza evolves to avoid the immune response. This process of gradual change through RNA point mutations leading to amino acid changes in the two antigens is known as 'antigenic drift'. Second, the segmented nature of the genome allows for mixing of genes when different strains of influenza infect the same host cell, resulting in 'reassortants' that combine genetic segments from separate viral lineages. Having a unique collection of reassorted genes, the new variant may be better in avoiding the host immune response. Third, influenza can undergo a shift of host, possibly through an intermediary species (Figure 1.2). An 'antigenic shift' to humans can result from the transfer of an entire virus strain *in toto*, or it can be combined with re-assortment so that genetic segments from the zoonotic virus combine with other genetic segments already circulating in humans. Antigenic shifts–at least to the human host–are

rare events, suggesting that this process requires the virus to adapt to the new host.

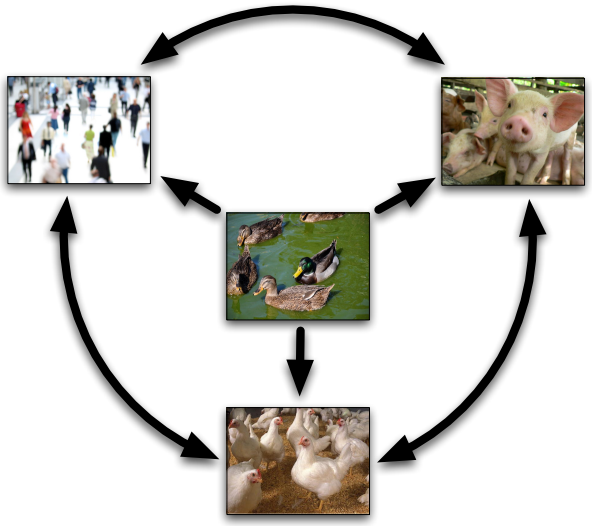


Figure 1.2.: Common host transmissions: influenza can sporadically transmit between hosts, sometimes via an intermediate host.

Human pandemics

Influenza is classified into subtypes based on its two membrane-bound glycoproteins. The types are determined by the response of the immune system to these proteins. There are sixteen known types of haemagglutinin (H1 to H16) and nine of neuraminidase (N1 to N9), all found in waterfowl. Only H1N1, H2N2 and H3N2 are known to have circulated in humans, with H1N1 and H3N2 currently predominant.

When transmission from aquatic birds to other species occurs, the new antigenic strain can lead to large outbreaks and human pandemics. The 'Spanish flu' of 1918 is thought to have infected a third of the world population, causing 50-130 million deaths worldwide (Johnson & Mueller, 2002; Taubenberger & Morens, 2006). It also, unusually, infected young adults (Knobler *et al.*, 2005). It has been suggested that this H1N1 virus was the result of a single host-shift event from birds to humans (Reid *et al.*, 2004; Taubenberger, 2006; Taubenberger *et al.*, 2005) but this remains controversial (Antonovics *et al.*, 2006;

dos Reis *et al.*, 2009; Gibbs & Gibbs, 2006; Smith *et al.*, 2009a). In 1957 three virus segments (HA, NA, and PB1) from an avian-like source were combined with the other five segments already circulating in humans to create the H2N2 'Asian flu' pandemic, while in 1968 two segments (HA and PB1) from an avian-like source were combined with the other six from the already-present human H2N2 virus to form the H3N2 'Hong Kong flu' pandemic (Schäfer *et al.*, 1993). These two pandemics were responsible for around a million and half-a-million deaths each (Salomon & Webster, 2009). The 2009 H1N1 pandemic virus seems to represent a more complex process where genetic segments from human (PB1 gene), avian (PA, PB2), and two different lineages of swine viruses (M1, M2, NA from avian-like European swine; NP, NS, HA from 'classical swine') produced a reassortant in swine, which then presumably underwent a single host shift to humans (Novel Swine-Origin Influenza A H1N1 Virus Investigation Team *et al.*, 2009). It was first identified in April 2009 (Centers for Disease Control and Prevention (CDC), 2009; Smith *et al.*, 2009b) and quickly spread throughout the world, causing the first pandemic of the 21st century (Fraser *et al.*, 2009). In addition to these pandemics, sporadic human infections have been caused by a number of different avian subtypes including H5N1, H7N3, H7N7, and H9N2 (Lin *et al.*, 2000). H5N1 emerged in the 1990s and it was thought that it may lead to a new pandemic. Although the strain is lethal, it does not transmit easily between humans and the genetic changes necessary for widespread transmission between humans have seemingly not occurred. Recent work has demonstrated that as few as five amino acid changes in an avian H5N1 can lead to mammal-to-mammal transmissibility (Herfst *et al.*, 2012; Russell *et al.*, 2012). Molecular phylogenetics is one way to enhance our understanding of what limits the host ranges, inhibiting host shifts, and how these limitations are overcome by the process of molecular evolution.

1.2. Outline

This thesis is concerned with developing models of evolution that can detect and characterise selection. We describe models that account for heterogeneous evolutionary processes across protein sites and across time. Chapter 2 provides some background to phylogenetics theory and methods, focusing on probabilistic models and how these are used within a likelihood framework to estimate evolutionary distances and parameters of interest.

In chapter 3, we develop a phylogenetic method to study influenza host shifts. We describe an approach for identifying which mutations allow viruses from avian origin to spread successfully in the human population. We use a site-wise nonhomogeneous phylogenetic model that explicitly takes into account differences in the equilibrium frequencies of amino acids in different hosts and locations. We identify amino acid sites with varying levels of support for differing selective constraints in human and avian viruses.

Chapter 4 describes how we can use estimates of amino acid equilibrium frequencies from chapter 3 to develop a measure of how well any given virus sequence is adapted to the selective constraints imposed by avian or human hosts. We focus on the 1918 H1N1 pandemic and examine the rate of host adaptation for individual influenza proteins, the degree of human adaptation found in currently circulating strains and how the avian viruses that initiate human pandemics compare with other avian viruses.

In chapter 5, we develop a mechanistic model of codon evolution which we use to estimate the distribution of selection coefficients (or 'fitness effects'), a long-standing issue in molecular evolution. This model is applied to a data set of mammalian mitochondrial genomes and PB2 influenza proteins. We are interested in comparing distributions in systems at equilibrium, such as mammalian mitochondria and influenza proteins evolving in its natural reservoir, with a clear adaptive event: the host shift of influenza proteins from birds to humans. Chapter 6 summarises the work.

2. Theory & methods

Since the advent of large amounts of molecular sequence data there has been a great deal of progress in designing models describing evolutionary processes and estimating evolutionary distances. These range from simple identity measures to sophisticated probabilistic models. In this chapter we give some background to key concepts in probabilistic modelling of molecular sequence evolution and describe how they are used to infer phylogenetic relationships. We show how these models are implemented within the likelihood framework to estimate model parameters, measure how well the model fits the data and how we can use statistical tests to test hypotheses. We provide a brief account of popular tests which look for the effect of natural selection in sequence data and their limitations. Finally, we give some motivations to the methods used in this thesis.

2.1. Probabilistic models of molecular sequence evolution

The gene sequences of extant organisms that we see today are the results of mutations that have accumulated and been selected over time from those organisms' common ancestors. Given the sequences of a particular protein from divergent but related species, we can use various methods to reconstruct phylogenetic relationships and measure the evolutionary distances between the different species. Probabilistic models characterise sequence change by describing the evolutionary process itself rather than using a heuristic, such as the idea that evolution proceeds parsimoniously. They allow repeated substitutions at the same site and the model can be used within a probabilistic and statistical framework (Aris-Brosou & Rodrigue, 2012).

2.1.1. Continuous-time Markov processes

A continuous-time Markov process can be used to model rates of change between nucleotides. The formulation of codon and amino acid models are different but the underlying theory is the same. A Markov process is a model describing how a state probabilistically changes to other states over time. A single position in a DNA sequence can be in one of four possible discrete states (T, C, A or G) and can change to one of the other three states or remain in the same state. We say the Markov process has N character states and for DNA/RNA $N = 4$, for codons $N = 61$ (ignoring stop codons) and for amino acids $N = 20$. The probability that the current character state jumps to a given state only depends on the current state and ignores any ancestral history of states. Markov processes improve upon simple sequence identity because they allow for multiple substitutions at the same site as well as hidden substitutions at conserved sites.

Markov processes are defined by instantaneous rate matrices. For example, we can construct a model that assumes that each state is equally likely to change into any other different state. If we are currently in state 'T', the rates to either state 'C', 'A' or 'G' are equal. This model was first proposed by Jukes & Cantor (1969) and can be stated as a Markov process instantaneous substitution-rate matrix:

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix} \tag{2.1}$$

where λ is the rate of substitution and $q_{ii} = -3\lambda$ is the negative sum of rates which leave state i so each row sums to 0. Given $\mathbf{Q}$, we can calculate the probability that any state i changes

to a state j in time t. This is determined by calculating the matrix exponential:

$$\mathbf{P}(t) = \left\{p_{ij}(t)\right\} = \mathrm{e}^{\mathbf{Q}t} \tag{2.2}$$

This is known as the transition-probability matrix, and $p_{ij}(t)$ is the probability that state i will change to state j in time t. Practically, the transition-probability matrix is calculated via numerical eigenvalue decomposition of the $\mathbf{Q}$ matrix. Because $\mathbf{Q}$ is usually normalised to have an average rate of 1, the units of t will be expected number of substitutions per site.

If the Markov process is given an initial distribution of character states, $\pi_{(0)} = (\pi^T_{(0)}, \pi^C_{(0)}, \pi^A_{(0)}, \pi^G_{(0)})$, then after time t the distribution of characters will be $\pi_{(t)} = \pi_{(0)}\mathbf{P}(t)$. Running the Markov process for an infinite amount of time has the effect that the process completely forgets its initial state (i.e. i could have been any character) and reaches a stationary distribution of states. These are known as the equilibrium frequencies. For the Jukes-Cantor model, the equilibrium frequencies, π_j, are $1/4$ for every character state. This is the probability that the process will be in state j as $t \to \infty$, regardless of the starting state i.

Most evolutionary models are time-reversible. This property of Markov processes is true if $\pi_i q_{ij} = \pi_j q_{ji}$ and means that we can formulate matrix $\mathbf{Q}$ as a symmetric matrix $\mathbf{S}$ times a diagonal matrix Π, $\mathbf{Q} = \Pi\mathbf{S}$, where the symmetric matrix gives the rates of change and the diagonal gives the equilibrium frequencies.

The Jukes-Cantor DNA model is an example of a simple substitution process: each rate of character change in the Markov process is equal. Over time, models of DNA have become increasingly sophisticated to better reflect features of substitution observed in biological systems. For example, Kimura's K2 model (Kimura, 1980), adds two parameters to better express the observation that DNA transitions (purine-to-purine or pyrimidine-to-pyrimidine) occur more frequently than transversions (pyrimidine-to-purine or vice versa) (see Figure 2.1). This model is extended in the Hasegawa-Kishino-Yano model (Hasegawa

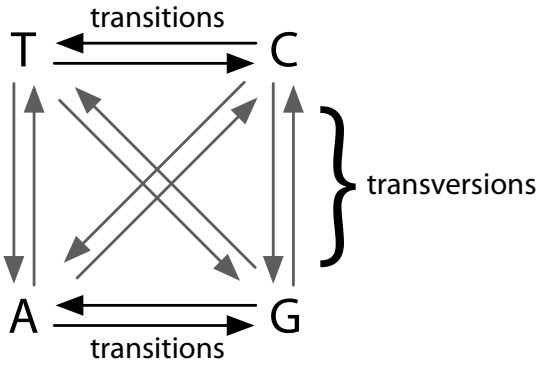


Figure 2.1.: Diagram of possible DNA state changes. Kimura's K2 model allows for different rates depending on whether the change is a transition or transversion.

et al., 1985) (known as HKY85) with the addition of equilibrium frequency parameters (π_i) to indicate that DNA sequences often have biased base composition (e.g. higher GC content). Tavaré (1986) and Yang (1994a) introduced the most general time-reversible form of the model (GTR), which has parameters for the equilibrium frequencies (π) as well a separate rate parameter for every state change. In each of these models, the parameters are used to formulate a new **Q** matrix, which leads to new transition-probabilities for state changes.

2.1.2. Amino acid & codon models

Amino acid and codon substitution models have the same mathematical foundation as DNA/RNA models. A key difference is the number of character states: a 20×20 **Q** matrix for amino acids and 61×61 for codons. Amino acid models are usually empirical, meaning that the rates of substitution between different amino acids are estimated by analysing datasets of protein sequences. The models attempt to reflect the substitutions occurring between these sequences, such as amino acids with similar physiochemical properties having a higher rate of substitution than dissimilar amino acids. The DAYHOFF, JTT and WAG (Dayhoff *et al.*, 1978; Jones *et al.*, 1994; Whelan & Goldman, 2001) models are examples of empirical amino acid models. They specify the symmetric matrix $\mathbf{S} = \{s_{ij}\}$ of amino acid exchangeabilit-

ies and the diagonal matrix $\Pi = \text{diag}\{\pi_1, \pi_2, ..., \pi_{20}\}$ of equilibrium frequencies, so the **Q** matrix can be calculated by $\mathbf{Q} = \Pi\mathbf{S}$. When analysing a given dataset, instead of using the equilibrium frequencies provided by the model, one can estimate the frequencies from the data being analysed. This is usually noted by the suffix '+F' e.g. WAG+F.

Despite the development of empirical codon models (e.g. Kosiol *et al.*, 2007; Schneider & Cannarozzi, 2012), most codon models are descriptive and model the biological process causing substitutions. Codons models can account for the underlying genetic code that governs how DNA/RNA sequences are translated into amino acids. A codon is a triplet of nucleotide bases, each coding for a particular amino acid (or the STOP codon). Each amino acid can be coded by a number of codons, ranging from one (e.g. ATG for methionine in the standard genetic code) to six (TTA/G, CTT/C/A/G for leucine) codons. As codons evolve and substitutions occur, those substitutions may result in a synonymous change (i.e. the codon has changed but amino acid remains the same) or a nonsynonymous change (i.e. the codon has changed and the amino acid it now codes for has also changed). Therefore, nonsynonymous substitutions change the protein sequence and, possibly, protein function and fitness. Observing only synonymous changes at a site, one might infer that the site is under purifying selection and amino acid changing substitutions at this site are deleterious to protein function. On the other hand, if we see many nonsynonymous changes (relative to synonymous changes), we might infer that the site is under diversifying selection. The model proposed by Goldman & Yang (1994) and Muse & Gaut (1994) describes the **Q** matrix

of instantaneous rates as

$$\mathbf{Q} = \{q_{ij}\} = \begin{cases} \pi_j, & \text{for synonymous transversions} \\ \kappa\pi_j, & \text{for synonymous transitions} \\ \omega\pi_j, & \text{for nonsynonymous transversions} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transitions} \end{cases} \tag{2.3}$$

where π_j is the equilibrium frequency of codon j, κ is the transition-transversion bias and ω is the nonsynonymous/synonymous rate ratio.

As shown, each of these models have a number of parameters that need to be estimated. The technique that we use in this thesis is maximum likelihood estimation (MLE).

2.2. Maximum likelihood methods

Maximum likelihood estimation is a statistical technique for estimating parameters in a model. It can be used to estimate evolutionary distances and parameters in phylogenetic models that best explain the data. Maximum likelihood also lays the foundation for statistical tests of models, which we cover in the next section.

Likelihood is a fundamental concept in statistics. The likelihood is the probability of the data given some model with specified values for parameters. This is shown as $L(\theta; D)$, where θ are the parameter values and D is the data. Different parameter values may give different likelihood and it is expected that there are specific values of parameters that best explain informative data. The estimation of the parameters, $\hat{\theta}$, is achieved through maximum likelihood estimation (MLE). Likelihood theory posits that the likelihood curve around the parameter estimate provides confidence in the estimate. In order to use maximum likelihood estimation, we need a function that returns the probability of the data given the para-

meters. In molecular phylogenetics, this function involves calculating the likelihood on a phylogenetic tree.

2.2.1. Likelihood computation on a phylogenetic tree

Imagine we have a molecular sequence alignment and the tree topology (T) describing how these sequences are related. Given a particular model of Markov process, we want to calculate the probability of the data and tree given the model and parameters, $L(\theta; D; T) = P(D|\theta, T)$. When calculating the likelihood of the entire sequence alignment the models treat every site in the alignment as an independent observation. This means the likelihood of every location in the alignment can be calculated separately and then the log-likelihoods can be summed to give the log-likelihood of the entire alignment:

$$\log\{L(\theta; D; T)\} \equiv \ell(\theta; D; T) = \sum_{l} \log(P(D_l|\theta, T)) \tag{2.4}$$

where l specifies the location in the alignment and D_l is the sequence data at that location.

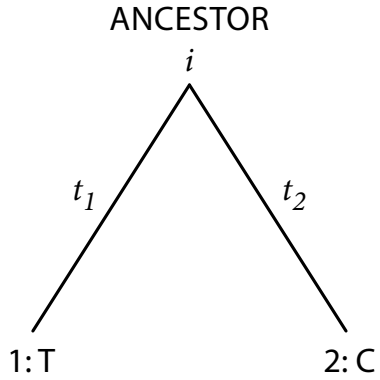


Figure 2.2.: A two species tree to demonstrate likelihood calculation. The nucleotides at the tips diverged time t_x ago from the common ancestor. The nucleotide state at the ancestor is unknown.

Figure 2.2 shows the phylogenetic relationship at a single location of two DNA sequences that diverged from a common ancestor. Taxa 1 diverged from the ancestor time t_1 ago and

is a 'T'. Similarly, taxa 2 diverged from the ancestor time t_2 ago and is a 'C'. Therefore, the likelihood is the probability of going from the ancestral state to 'T' in time t_1 and to 'C' in time t_2. This is calculated by summing over all the possibilities for the ancestral node (i.e. T, C, A or G):

$$L = \sum_{i=(\mathrm{T,\ C,\ A,\ G})} \pi_i P_{i\to T}(t_1) P_{i\to C}(t_2) \tag{2.5}$$

where π_i is the equilibrium frequency of character state i and $P_{i\to j}(t)$ is given by the transition probability matrix of the Markov process, which is calculated by taking the exponential of the **Q** matrix.

Calculating $P(D_l|\theta, T)$ for all but the very simplest cases (say a handful of taxa) involves summing over a huge number of terms. For larger trees, the pruning algorithm can be used to calculate $P(D_l|\theta, T)$ (Felsenstein, 1981). The principle is to start from tips of the tree (i.e. sequences in an alignment) and then, working up the tree, sum over all the possibilities for each of the ancestral nodes from which those sequences diverged. As the models are time-reversible, moving up the tree, from tips to root, is the same as moving down the tree. This is repeated until the entire tree has been traversed, giving the likelihood at that site. We sum the log-likelihoods over all sites to get the total log-likelihood for the data and model.

By maximising the log-likelihood of the data using the pruning algorithm, maximum likelihood estimates parameter values that best fit the data. Using numerical optimisation routines, such as Nelder & Mead (1965) or Newton's method, the likelihood surface can be explored to find parameter estimates. For the trivial example above, this would mean picking different values for t_1 and t_2 repeatedly (the particular order of exploration being dependent on the optimisation method) until no other values can improve the log-likelihood. At this point the optimisation routine has converged and has estimated the parameter values that give the maximum log-likelihood. Further details about various optimisation algorithms can be found in Nocedal & Wright (2006).

2.3. Statistical tests of phylogenetic models

Probabilistic phylogenetic models that operate within the likelihood framework can use various statistical techniques to test hypotheses (Edwards, 1992; Felsenstein, 2003). We described in the previous sections several types of models, each trying to capture some particular aspect of the biological process. Each additionally complex model introduces a number of additional parameters. For example, Kimura's K2 model adds one parameter, representing the ratio of transitions/transversions rates, to the Jukes-Cantor model to account for transition-transversion substitution bias in DNA sequences. The question then is can we justify the addition of this parameter? Does the more complex model provide a significantly better fit than the simpler model? Here we describe two methods to test hypotheses, the likelihood-ratio test and bootstrapping.

2.3.1. Likelihood-ratio test

The likelihood-ratio test (LRT) is used to compare two models, one of which is a special case of the other i.e. they are 'nested'. The Jukes-Cantor model (M_0) is nested in the Kimura K2 model (M_1) when the transition & transversion parameters are equal (i.e. their ratio is 1). Imagine that we want to test whether the more complex model M_1 having likelihood L_1 fits the data better so that we can reject the simpler null model M_0 having likelihood L_0. The likelihood test statistic is twice the difference of the log-likelihoods:

$$2\Delta\ell = 2\log(L_1/L_0) = 2(\ell_1 - \ell_0) \tag{2.6}$$

This test statistic follows the chi-squared distribution with degrees of freedom equal to the increase in number of parameters. Because M_0 has no parameters and M_1 has a single additional parameter, we can check a $\chi^2_{\text{dof}=1}$ distribution to see if the improvement in log-

likelihood when using M_1 is large enough to reject M_0 in favour of M_1 at some level of significance (Felsenstein, 2003; Yang, 2006). For example, if the log-likelihoods of M_0 and M_1 are 10743.03 and 10513.01, respectively, this gives a test statistic score of $2\Delta\ell = 460.04$. Checking the test statistic using the $\chi^2_{\text{dof}=1}$ distribution gives a P value < 0.000001, giving us confidence to reject M_0.

2.3.2. Simulations and bootstrapping

Often it is not possible to use the χ^2 distribution to test the likelihood-ratio test statistic because the two models being tested are not nested or the χ^2 distribution is thought to be too conservative. In these cases, one can use simulations to test the consistency of the model and perform hypothesis testing. Say we have real data D and are comparing two models, M_0 and M_1 having a difference in log-likelihood of $\Delta\ell$ and these models cannot be nested. Parametric bootstrapping can be used to test whether the null hypothesis can be rejected (that M_1 better explains the data than M_0). First, multiple simulated datasets ($D'_1, D'_2, ..., D'_n$) are generated under the M_0 model. We then calculate the maximum likelihood for each of the simulated datasets under the M_0 and M_1 models. This gives a bootstrap distribution of differences of log-likelihoods between the M_1 and M_0 model ($\Delta\ell'_1, \Delta\ell'_2, ..., \Delta\ell'_n$) when M_0 is true (because the synthetic data was generated using M_0). If the difference in log-likelihood from the real data D ($\Delta\ell$) is greater than a significant proportion (say the bottom 95%) of the distribution of log-likelihood differences from the synthetic data sets D', we can reject the hypothesis that the real data was generated by the M_0 model. Conversely, if we find that $\Delta\ell$ is less than the top 5% of bootstrapped log-likelihoods differences, we have no reason to believe that M_1 describes the real data D significantly better than M_0 (Efron & Tibshirani, 1993; Goldman, 1993).

Bootstrapping techniques can also be used to calculate confidence intervals for parameter

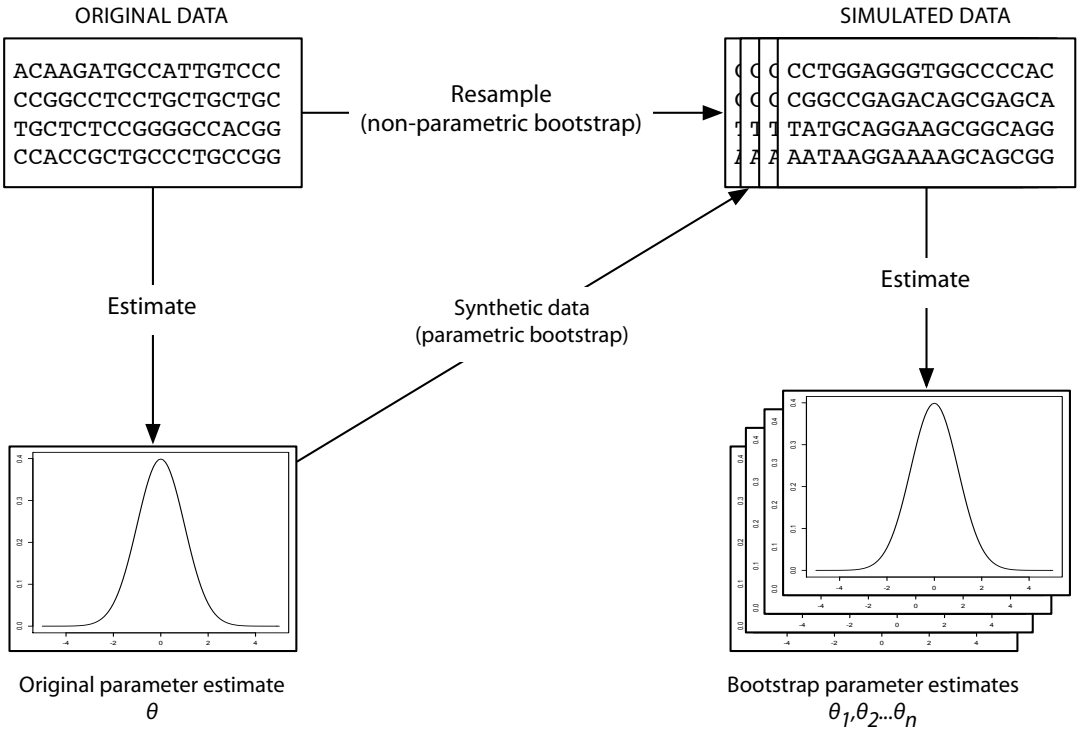


Figure 2.3.: Diagram of parameteric and non-parametric bootstrapping. New data is generated by either resampling the original data (non-parametric) or simulating data based on estimated parameters from the original data (parametric). The parameters are re-estimated from the simulated data to get confidence of the original parameter estimate.

estimates. These can be non-parametric or parametric. In the non-parametric case, say we have an alignment and we have estimated the parameters for a model using maximum likelihood. We sample sites from the alignment (with replacement) to create a new bootstrap alignment of the same size as the original alignment. Model parameters are re-estimated from the bootstrapped data and the procedure is repeated many times. For each parameter, we now have a distribution of parameter estimates that can be used to construct confidence intervals indicating how much trust we have in the parameter estimate from the original alignment. If the data are informative, we expect the confidence intervals to be narrow. If the data are not informative, we expect wide confidence intervals. For parametric bootstrapping, instead of sampling sites from the original alignment, we generate many synthetic

datasets using the maximum likelihood estimates of the parameters. The remainder of the procedure is the same as the non-parametric example. Figure 2.3 shows a diagram of parametric and non-parametric bootstrap procedure to test parameter estimates.

2.4. Phylogenetic methods for detecting selection

The most popular method for detecting adaptation in protein-coding sequences is based on the ω parameter in codon models described in section 2.1.2. which compares nonsynonymous and synonymous substitution rates. It is based on the assumption that although mutations occur at the nucleotide level, selective pressure is applied at the protein level and the corresponding amino acid given by the codon. If a particular site is under strong purifying selection, we would expect to see a higher rate of synonymous substitutions than nonsynonymous substitutions. This is reflected by an ω ($= d_N/d_S$) value less than 1. On the other hand, $\omega > 1$ reflects the increased fixation of nonsynonymous mutations compared to synonymous mutations, reflecting diversifying selection, the repeated fixation of amino acid changes. When $\omega = 1$, this reflects neutrality, where neither nonsynonymous nor synonymous changes dominate. The codon models and statistical tests based on the ω parameter are very well studied and many different approaches are available. For example, one can search for evidence of adaptation on the entire protein sequence, at some particular site, at some particular branch or, combining both, adaptation along a particular branch and site (Yang, 2006; Yang & dos Reis, 2011).

However, there are types of selection that the ω tests are not able to detect as well. If a single, non-repeated, amino acid mutation conferred a significant increase in fitness in a mutant, it may rapidly become fixed in the population. As there are not repeated amino acid changes at that site, the ω test would not recognise it as an adaptive change (Yang, 2007b). It has been suggested that many phenotypic adaptations are of this type (Hughes,

2007). Another type of adaptation that would not be recognised is when the rate of amino acid change remains the same but the types of amino acids being substituted change. This is because the ω tests expect selective pressure to affect every amino acid equally. But it is not unreasonable to believe that in one environment a protein may prefer any hydrophobic residue at a particular location whilst in another it changes this constraint and prefers any hydrophilic residue. The kind of substitution has changed but not necessarily the rate.

The probabilistic models described above uniformly model the evolutionary process at every location in the sequence alignment. They are "site-invariant", meaning the model does not change at different locations. To accommodate dissimilarity of evolutionary change across sites, the models adopt an among-site rate variation using a distribution like the gamma distribution (Yang, 1994b, 1996). This is to recognise that not only do substitution rates vary across sites but also that they vary for different reasons, based on the structure or function of the protein. However, this method only varies the absolute rate of change of amino acids and does not reflect the constraints on the particular amino acids acceptable at a given location. If equilibrium frequencies are allowed as adjustable parameters, the site-invariant models reflect the stationary distribution given all locations in the alignment, averaging over any particular constraints that exist at individual sites. Recognising these limitations, models have been proposed that explicitly account for different amino acid substitutions at different sites. Bruno (1996) and Halpern & Bruno (1998) developed a model that characterised site-wise amino acid frequencies which they argued were, in addition to rate heterogeneity, representative of selection acting at specific locations. Others have proposed mixture models of categories of amino acid substitution processes reflecting physicochemical properties (Koshi & Goldstein, 1997), or patterns in secondary (Goldman *et al.*, 1998; Koshi & Goldstein, 1995) or tertiary (Robinson *et al.*, 2003) structure. There are also methods which use the data to determine the number and kind of amino acid frequencies categories for the mixture (Koshi & Goldstein, 1998; Lartillot & Philippe, 2004).

Previous work has shown that ignoring site-wise amino acid frequencies can lead to serious underestimation of sequence distance, even in those models that allow for variable rate among sites (Halpern & Bruno, 1998). They are also more likely to have an adverse effect on phylogenetic tree estimation such as long-branch attraction, causing highly divergent taxa to tend to group together (Lartillot *et al.*, 2007; Wang *et al.*, 2008).

We are interested in developing site and time heterogeneous models of protein evolution that better reflect the different functional and physicochemical constraints in proteins across sites and time (Halpern & Bruno, 1998; Koshi & Goldstein, 1998, 2001; Lopez *et al.*, 2002). We use these to detect changes in selective constraints and characterise the strength of adaptation in evolutionary processes. In the next chapter we introduce a site-wise non-homogeneous model of substitution that we use to identify changes in selective constraints in influenza viruses that occur during host shifts from avian to human hosts. Instead of looking at change in rate of substitution we focus on the pattern of amino acid change and the propensity of particular amino acids in the different hosts, and use statistical tests to locate positions in proteins where these differ.

3. Identifying changes in selective constraints: host shifts in influenza

3.1. Introduction

The spread of influenza viruses among different hosts is thought to be due to a number of viral factors and all virus proteins may potentially be related to transmissions and greater virulence. Influenza haemagglutinin binds to sialic acid linked to galactose on the surface of the targeted cell; the differing nature of the sialic acid-galactose linkages in birds and humans (α2,3 sialic acid linkages in the bird gut, α2,6 sialic acid linkages of the upper human respiratory tract (Gambaryan *et al.*, 2003)) provides an important barrier to host shift events. A number of amino acid substitutions have occurred in human influenza haemagglutinin (e.g. Q226L and G228S in H2 and H3, E190N/D and G225E/D in H1) to adjust to the different receptors (Connor *et al.*, 1994; Matrosovich *et al.*, 2000; Nobusawa *et al.*, 1991; Rogers *et al.*, 1983; Vines *et al.*, 1998). However, H5N1, which binds α2,3-linked sialic acids, is lethal to humans, and it is unknown whether this strain of virus requires α2,6-linked sialic acid binding to become pandemic (Salomon & Webster, 2009). Recent work demonstrated mutations that allow H5N1 haemagglutinin to bind α2,6-linked sialic acid, leading to mammal-to-mammal transmissibility (Herfst *et al.*, 2012; Russell *et al.*, 2012).

Neuraminidase, the protein responsible for cleaving sialic acid groups from the receptor surface, also seems adapted to the particular sialic acid linkages, as well as for the pH and temperature of the host tissues (Baigent & McCauley, 2003). Proteins in the viral replication complex (PA, PB1, PB2, and NP) have also been implicated in limiting host range by restricting replication and intra-host spread in mammals (for a review, see Naffakh *et al.* (2008).) Of particular note is the PB2 gene, where one specific substitution, E627K, was

identified and characterised experimentally as crucial for replication and intra-host spread in mammals (Hatta *et al.*, 2001; Steel *et al.*, 2009; Subbarao *et al.*, 1993). However, this site remains a glutamate in the 2009 H1N1 pandemic strain. The M1 and M2 matrix proteins also seem to contain amino acid sites that are host-specific and it is thought that compatibility between HA and M2 proteins is required for successful infection (Buckler-White *et al.*, 1986).

As part of the widespread surveillance effort, it is important to understand the process of host shifts, and to identify the important changes that are necessary for the shift to occur or that make the shift more likely. We currently have many examples of both avian and human viruses, so there have been a number of efforts at identifying 'genetic signatures' that characterise the virus as adapted to one or the other host. The most common method is to identify sites where the distribution of amino acids found in the virus in one host are sufficiently different from the distribution of amino acids found in the same site in viruses that affect the other host (Chen *et al.*, 2006; Finkelstein *et al.*, 2007; Miotto *et al.*, 2008). Unfortunately, there are two fundamental problems with this approach.

Firstly, the observed changes could represent the result of neutral drift rather than anything specific to the nature of the different hosts. As the human viruses are more closely related to each other than they are to the avian viruses, it would be expected that there would be characteristic amino acids found in the human lineages that are distinct from those found in the avian lineages because of the 'founder effect' (Mayr, 1942), that is, the maintenance of the idiosyncratic properties of the particular virus that first infected humans. Comparisons of amino acid frequencies in viruses from the two hosts cannot easily distinguish between those that accidentally accompanied the host shift event and those that were actually associated with different selective constraints acting on the viruses in the two hosts.

The second related problem is the use of inappropriate statistical tests to identify when these two distributions are sufficiently different. The statistical tests used generally assume

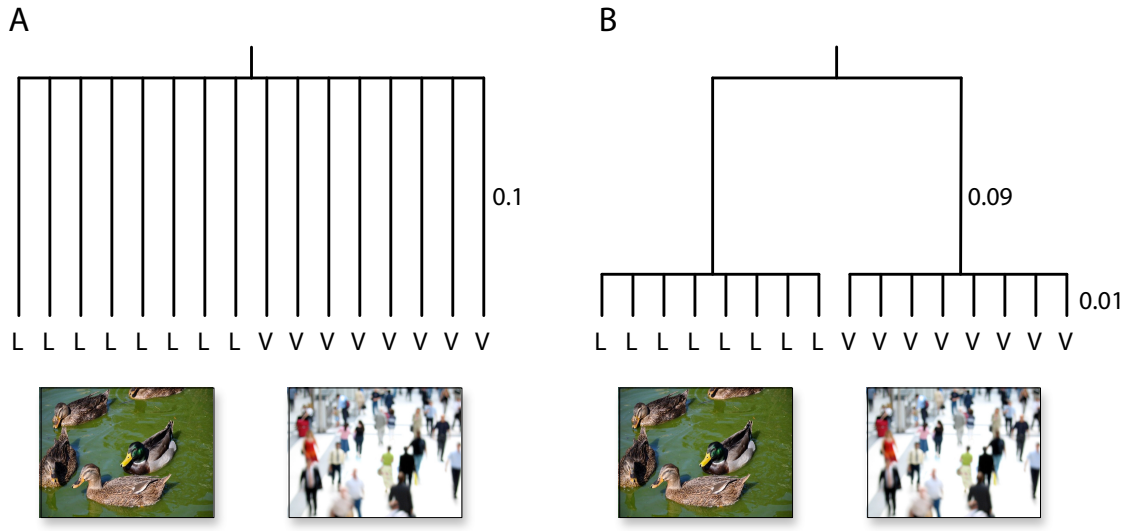


Figure 3.1.: Possible evolutionary scenarios. Two possible phylogenetic trees representing the situation where eight different avian sequences have a L in a given position, while eight different human sequences have a V. (Branch lengths are not to scale.) The situation shown on in 'B' provides much weaker evidence for a shift in selective constraints compared with the situation shown in 'A'.

that each of the observed sequences represent a set of independent measurements. But the underlying phylogenetic relationships will generate correlations in the amino acids at a site, confounding the signal due to the host shift event (Felsenstein, 1985). This can be demonstrated by considering Figure 3.1, which shows two possible situations where the avian viruses all have a leucine in a given position where all of the human viruses have a valine in the same position. In example A the results are statistically significant, in that the positions are independent, and it is unlikely that the simultaneous parallel changes in sequence occurred at random in the human viruses but not in the avian viruses. In example B there is much less statistical signal, as only one change of amino acid on the branch connecting the human and avian viruses is needed to explain the multiple observations. By neglecting the underlying phylogenetic structure, a single change of amino acid can be interpreted as a large number of independent events, grossly exaggerating the statistical significance.

A number of the published approaches to this problem suffer from the above problems.

For example, both Chen *et al.* (2006) and Miotto *et al.* (2008) employed an information-based approach to identify sites where host-specific amino acids can be identified. Their computations of entropy (a measure of sequence diversity) and mutual information (the dependence of the observed residue distribution on host species) are based on considering every observed sequence as an independent data-point, ignoring correlations between the evolutionarily related sequences. Different distributions in the two hosts can be explained due to the founder effect described above, independent of any role these sites have in host adaptation. That is not to say that their results are incorrect, only that these problems make it impossible to determine their statistical significance. Finkelstein *et al.* (2007) looked at sites with a significantly higher degree of conservation in human lineages than avian lineages, and identified 32 markers within the M1, NP, NS, PA, and PB2 genes, 26 of them on the polymerase proteins NP, PA, and PB2. This analysis did not consider the phylogenetic relationships explicitly in their calculation of conservation, choosing instead to base their calculation on the frequency of the different amino acids observed in that site in the different hosts. While they employed strict tests for, for instance, multiple hypothesis testing, it is difficult to determine how much their results were affected by considering only frequencies of amino acids to represent the selective constraints, again ignoring the underlying phylogenetic relationships. It is known, for instance, that such counting methods produce very inaccurate amino acid frequencies compared with phylogenetically-based methods (Bruno, 1996), and can not generally identify the rate of substitutions in the tree, but only the range of acceptable amino acids.

As described above, the differences in the distribution of amino acids at a given site between avian and human viruses might represent neutral drift or, more interestingly, a change in the underlying selective pressure applied to the virus by the host. Rather than characterising only the difference in observed amino acid distributions, we can instead look directly for evidence of changes in the selective constraints by modelling the phylogenetics

explicitly. These selective constraint changes will result in differences in the substitution process, as mutations that arise in one virus or another will have different probabilities of achieving fixation. Thus, changes in selection constraints will manifest themselves as changes in the observed substitution rates. This also allows rigorous statistical methods, such as the likelihood ratio test, to be used to establish statistical significance.

The selective pressure acting on a site can be positive, negative or neutral. Positive selection, also called adaptive (or more misleadingly 'Darwinian' (Freire-Maia, 1979)) refers to the acceptance of advantageous mutations; negative, or purifying selection involves the rejection of deleterious mutations. Neutral selection pressure involves the chance acceptance of mutations that do not have a significant effect on the fitness. The fate of all three types of mutation is also affected by population size (Hartl, 1980). Both positive and negative selection pressure represent strong constraints on the amino acids at a given site; the difference is that during purifying selection the current amino acids generally fulfil these constraints so change is restricted, while during adaptive evolution the current amino acids are not well suited, generally due to changes in the constraints or a selective advantage for diversification, enhancing the rate of evolution until more appropriate residues are found.

Changes in the selective constraints can result in changes in the rate of substitutions at that location. If the initial amino acids do not match the current requirements of that site, there may be an adaptive burst of faster substitutions until the constraints are satisfied. Modifications of the stringency of the constraints, causing a given site to be more or less restricted, may cause a longer-term change in the substitution rate without necessarily causing an adaptive burst. Previous phylogenetic methods have generally focused on identifying changes in the substitution rate (Blouin *et al.*, 2003; Dorman, 2007; Gu, 1999, 2001; Gu *et al.*, 1995; Knudsen & Miyamoto, 2001; Kosakovsky Pond *et al.*, 2008; Penn *et al.*, 2008; Pupko & Galtier, 2002) or ratio of nonsynonymous to synonymous changes (Guindon *et al.*, 2004; Yang & Nielsen, 2002; Zhang *et al.*, 2005). The latter method was used, for instance, to identify

twelve sites on the influenza A nucleoprotein that seem to have undergone a change in selective constraints corresponding to the switch from avian to human host (Forsberg & Christiansen, 2003). While these approaches are often useful, transient position specific adaptive bursts are difficult to identify given the short duration of the effect. Sites can also undergo shifts in selective constraints without adaptive bursts or detectable changes in substitution rates, especially if the constraints in the two hosts overlap. Monitoring changes in the nature of the selective constraints has been much less common (Blackburne *et al.*, 2008) and has not been applied to host shift events.

In this chapter we investigate the use of a phylogenetic method to detect changes in selective constraints that considers not only changes in the magnitude of selection constraints, but also changes in its nature, represented as the relative propensity for the different amino acids. We do this by considering two different models for each site, one a homogeneous model where the selective constraints are independent of host, the other a nonhomogeneous model where the selective constraints depend upon the host. The likelihood ratio test can then determine the level of statistical support for rejecting the null hypothesis of no such dependence.

3.2. Methods

3.2.1. Theory

For the following discussion we assume the evolution of a viral protein along a phylogenetic tree with two different host lineages, avian and human, where we consider the root of the tree to exist somewhere in the avian lineage. The evolution of amino acids in a site along a phylogenetic tree can be modelled as a continuous Markov process, described by a 20×20 substitution matrix $\mathbf{Q}$. In order to provide for time reversibility (that is, the expected

number i to j transitions equalling the expected number of transitions from j to i), this is commonly represented as $q_{ij} = v\pi_j S_{ij} \, (i \neq j)$ where **S** is a symmetric matrix representing the exchangeability of amino acids i and j, π_j is the equilibrium frequency of amino acid j ($\sum_i \pi_i = 1$) and v is a scaling parameter that accounts for the overall rate of substitution at the site. **S** encodes the underlying codon structure as well as the relative similarities of the physicochemical properties of the amino acids, while the equilibrium frequencies represent the relative propensities for each of the amino acids at that site. We can calculate the likelihood of the data at this site given the model using Felsenstein's pruning algorithm (Felsenstein, 1973, 1981).

We first consider a standard substitution model where **S** and π are given by the WAG substitution matrix (Whelan & Goldman, 2001), where each site in the set of proteins is characterised by a distinct substitution rate scaling factor v whose value is determined by maximising the log likelihood given the sequence data at that site and the input phylogenetic tree. This we refer to as Model 1. We then considered the appropriateness of modelling each site in the set of proteins with a distinctive set of equilibrium amino acid frequencies (Bruno, 1996), what we refer to as single-site homogeneous Model 2. We adjust the values of π simultaneously with v to maximise the likelihood. To avoid over-parameterisation, we use WAG **S** values for all sites. The tree topology is assumed fixed, and branch lengths are the same for all sites. In order to reduce the number of adjustable parameters, $\pi_i = 0$ for any amino acids not found at that site. We then use parametric bootstrapping to see if site-dependent equilibrium frequencies can be justified with the data.

Now let us imagine that upon inspection of the phylogenetic tree, we notice that amino acid preferences at a particular site seem different in the two host clades. We can incorporate this observation into our model by using two distinct **Q** matrices to describe the evolution of this site in the different hosts, as illustrated in Figure 3.2. For the reservoir avian host we write $q_{ij} = v\pi_j S_{ij}$ and for the new human host $q'_{ij} = v\pi'_j S_{ij}$ where π and π' represent the

equilibrium amino acid frequencies at that site in avian and human viruses, respectively. (In principle we could also have **S** depend upon the host, but this would result in a large increase in the number of adjustable parameters. We will consider host-dependence of ν below.) The host shift event is defined as the midpoint of the branch connecting the common ancestor of the human viruses with its parent node. We can now calculate a new likelihood for this site using the same fixed topology, again adjusting π, π', and ν to maximise the likelihood. We call this the single site nonhomogeneous model, Model 3. The increase in the number of adjustable parameters for Model 3 relative to Model 2 equals the number of amino acid types observed at that site minus one. Because the Model 2 is nested inside Model 3, we can use the likelihood ratio test to test the hypothesis of different selective constraints in different hosts at that site.

For a protein with n variable sites, we could repeat the procedure above for each site in the alignment and perform n likelihood ratio tests. This would generate a list of those sites that show statistically different amino acid compositions, and hence distinctive selective constraints, in the different hosts. Following the calculation of the statistical significance for each site we can then use standard false discovery rate (FDR) methods to account for multiple hypothesis testing (Benjamini & Hochberg, 1995).

Finally, we consider if, in addition to host-dependent equilibrium frequencies, we also have statistical evidence for host-dependent rate scaling factors. We again use $q_{ij} = \nu\pi_j S_{ij}$ for the reservoir avian host but now use $q'_{ij} = \nu'\pi'_j S_{ij}$ for the new human host where ν and ν' represent the rate scaling factors at that site in avian and human viruses, respectively. Again, Model 3 is nested inside Model 4 with an increase of one adjustable parameter, meaning that the statistical support for this extra factor can be evaluated with the likelihood ratio test.

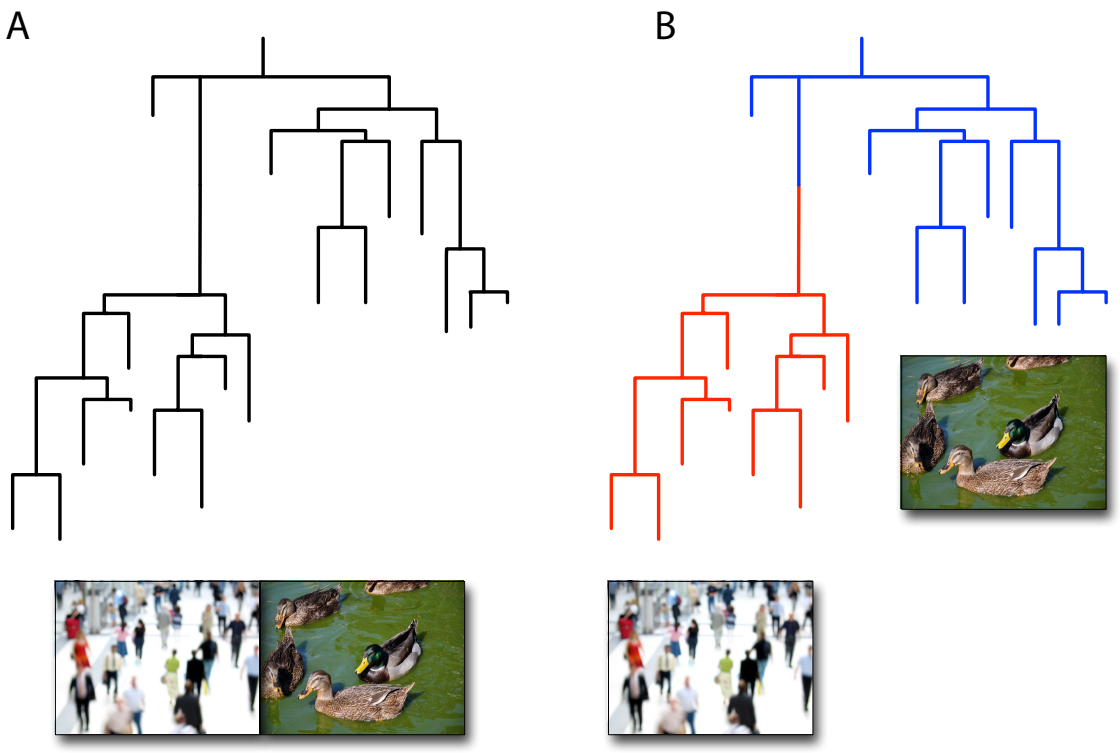


Figure 3.2.: Homogeneous and nonhomogeneous substitution models. Illustrative phylogenetic trees showing set of avian and human influenza sequences. A: In the homogeneous models (Models 1 and 2), the same substitution rates are used throughout the tree. B: In the nonhomogeneous models (Models 3 and 4) different substitution rates are used for the avian (blue) and human (red) lineages. The root of the tree is assumed to be inside the avian lineage. (Because the model is reversible with the avian clade, the exact location of the root within this clade does not affect the calculation.) The host shift event is assumed to occur at the midpoint of the branch connecting the common ancestor of the human strains with its parent.

Protein	Alignment length	Number of human sequences	Number of avian sequences
H1	566	404	30
H2	592	34	37
H3	567	354	56
M1	252	102	291
M2	97	168	273
N1	470	274	232
N2	469	254	166
NP	507	122	308
NS1	305	61	312
NS2	126	101	323
PA	716	60	347
PB1	784	108	284
PB2	759	80	321

Table 3.1.: Protein sequences used in the analysis

3.2.2. Data and data analysis

Human and avian viral sequences were collected from the NCBI Influenza Virus Resource (Bao *et al.*, 2008). Due to the frequency of reassortment, we cannot assume that the phylogenetic relationships for the various genomic segments are similar; they must be treated independently, including creating genetic segment-specific phylogenetic trees. The sequences for the various segments were treated as independent data sets, with separate datasets for the H1, H2, H3, N1, and N2 genes. Clusters of highly similar sequences (approximately >99.5%) were culled as to reduce the overall number of sequences to around 400 per dataset. It is common to find sporadic transmissions between avian, human, and other (e.g. swine) hosts; we eliminated all sequences resulting from such transmissions (e.g. human H5N1 sequences), leaving us with a single connected set of avian sequences and separate monophyletic human clades corresponding to the host shift events of 1918 (H1, N1, internal genes), 1957 (H2, N2, PB1), and 1968 (H3, PB1). The number of sequences used are listed in Table 3.1 and accession numbers in Appendix A.

In order to generate phylogenetic trees, the culled sequences were aligned using MUSCLE (Edgar, 2004) at the amino acid level, with these alignments then used to create nucleotide codon alignments using PAL2NAL (Suyama *et al.*, 2006). The phylogenetic tree topologies were then created for the nucleotide data using PhyML (Guindon & Gascuel, 2003) under the HKY85 model (Hasegawa *et al.*, 1985) and Gamma-distributed rates. Branch lengths representing amino acid evolutionary distances were then optimised for this fixed-tree topology using the corresponding amino acid data using PAML codeml (Yang, 1997, 2007a), the WAG substitution matrix (Whelan & Goldman, 2001) and Gamma-distributed rates. The analysis was then performed with each gene set, based on the phylogenetic tree for the genomic segment in which the gene is located.

The determination of changes in selective constraints at each site is a separate hypothesis to be evaluated, so we must account for multiple hypothesis testing. That is, if we ask a suitably large number of statistical questions we are likely, by chance, to obtain some statistically significant results. We use the false discovery rate method, specifying for each site the false positive rate that would have to be tolerated in order for that result to be statistically significant, following the estimator of Benjamini & Hochberg (1995). This is in contrast to the more conservative Bonferonni correction which tries to eliminate any possibility of false positives, often leading to reduction of true positives (Noble, 2009). We first choose an acceptable false discovery rate δ. If $P(k)$ is the kth smallest P value for a set of n sites, we choose the largest value of k so that $nP(k)/k \leq \delta$. As different genes are evolving in different circumstances, we would not expect the fraction of sites in each gene undergoing changes in selective constraints to be the same. Combining all of the genes together in one dataset would result in an increase in false positives for the genes with fewer changes in selective constraints and an increase in false negatives for the genes with more changes in selective constraints. For this reason we analyse the false discovery rate for each gene individually. Table 3.3 and Appendix C list, for each site, the smallest possible acceptable false discovery

rate that would result in that site being labelled as statistically significant. These should not be interpreted as the probability that that given site is a false positive.

3.2.3. Parametric bootstrapping

To test for statistical error, each site was simulated under the homogeneous (Model 2) and nonhomogeneous (Model 3) models 10 times using the program *Evolver* (Yang, 2007a) using the estimated tree topology and the WAG+F substitution matrix (Whelan & Goldman, 2001). For each site, the tree was scaled according to the site-specific estimated rate-scaling parameter v. Simulation under the nonhomogeneous model was performed in two steps: the avian part of the tree was simulated using a randomly generated root sequence following the avian equilibrium frequencies for that location. The avian subtree contained a host shift tip that served as the root of the human clade. The human subtree was then simulated according the human equilibrium frequencies using the simulated avian sequence at the host shift.

3.2.4. Alternative tree topologies

The PB2 sequence was bootstrapped 10 times and tree topology re-estimated for each boot sample. The homogeneous and nonhomogeneous models were optimised for the observed data at each location, and the LRT was performed again for each one of the 10 new tree topologies in order to assess the effect of tree topology uncertainty on the identification of adaptive sites.

3.2.5. Simple model for relationship between π and v

To better understand the behaviour of the rate scaling parameter, we designed a simple model of substitution. Consider a protein site where two amino acids, A and B, are found.

Let us imagine that A is the more advantageous amino acid, that is, organisms with A at this site have a higher fitness, while organisms with B at this site have relative fitness $1 - s$, $s > 0$. Let us also imagine that the mutation rate from A to B, μ_{AB}, is equal to the reverse mutation rate $\mu_{BA} = \mu$. We imagine a number of different lineages that have diverged, each with effective population size N_e. Assuming that the mutation rate relative to the population is reasonably small, A or B will become fixed in each lineage. For haploid organisms, the probability that A would become fixed in a given lineage is given by Bulmer (1991)

$$P(A) = \pi_A = \frac{e^{2N_e s}}{e^{2N_e s} + 1} \tag{3.1}$$

where we have recognised that this fraction is simply the equilibrium frequency of A in the ensemble of diverged organisms, with $\pi_B = 1 - \pi_A$.

The substitution rate of A to B is the number of mutants in a generation, $N_e\mu$, times the fixation probability, given by Kimura's formula for small s (Crow & Kimura, 1970; Yang & Nielsen, 2008).

$$\begin{aligned} q_{AB} &= N_e\mu \times \frac{-2s}{1 - e^{2N_e s}} \\ q_{BA} &= N_e\mu \times \frac{2s}{1 - e^{-2N_e s}} \end{aligned} \tag{3.2}$$

We can compare these expressions with $q_{ij} = v\pi_j S_{ij}$ as used in phylogenetic analyses. As we are only dealing with two different residues, $S_{AB} = S_{BA}$ is a simple multiplicative constant and can be set equal to one, resulting in $q_{BA} = v\pi_A$. Equating these two expressions for q_{BA} and solving for v yields

$$q_{BA} = \mu\frac{2N_e s}{1 - e^{-2N_e s}} = v\pi_A = v\frac{e^{2N_e s}}{e^{2N_e s} + 1} \tag{3.3}$$

$$\nu = \mu 2N_e s \frac{(e^{2N_e s} + 1)}{(e^{2N_e s} - 1)} \tag{3.4}$$

Similar results are obtained, as would be expected, when we express $q_{AB} = \nu\pi_B$.

We can now consider the cases of neutral, positive and negative selection. Neutral selection is simply the case when $2N_e s$ is small and $\nu \approx \mu \lim_{2N_e s \to 0} 2N_e s(e^{2N_e s}+1)/(e^{2N_e s}-1) = 2\mu$. For negative selection, we can consider the overall rate at which substitutions occur, given by $\Gamma_- = \pi_A q_{AB} + \pi_B q_{BA} = 2\nu\pi_A\pi_B$. Positive selection involves the situation where we are not at equilibrium, but rather, at least in this case, we have the less-fit residue occupying the given position. In this case, assuming again that A is the favoured residue, $\Gamma_+ = q_{BA} = \nu\pi_A$.

3.2.6. Characterising the magnitude of selective constraints

We characterise the selection constraints by how far the equilibrium amino acid frequencies π differ from what would be expected under no selection π^0 through the relative entropy (Kullback & Leibler, 1951), defined as

$$d = \sum_i \pi_i \ln\left(\frac{\pi_i}{\pi_i^0}\right) \tag{3.5}$$

which is zero when π equals π^0. Unfortunately, it is difficult to estimate π^0 as there is little of the virus genome that is not under some degree of selective constraints. We estimate π^0 by averaging the amino acid frequencies over our entire database, with the expectation that specific selection constraints will, at least approximately, average out.

3.3. Results

We start our analysis with a set of human and avian influenza viral sequences and the associated phylogenetic trees for each influenza gene. We consider the different haemagglutinin and neuraminidase serotypes (e.g. H1, H2, H3, N1, N2) separately. For each non-conserved site, we apply increasingly complicated substitution models, using simulations and the likelihood ratio test (LRT) to evaluate the statistical support for each further complication.

The simplest model, Model 1, consists of the WAG exchangeability matrix combined with the associated equilibrium frequencies for the different amino acids (Whelan & Goldman, 2001), with one adjustable parameter per site representing the scaling factor v. We then consider Model 2 where the equilibrium frequencies of the amino acids are optimised individually for each site (Bruno, 1996). Parametric bootstrap simulations showed that the use of site-specific equilibrium frequencies was justified for all sites (see Figure 3.3).

We then created a nonhomogeneous model, Model 3 where virus substitutions are modelled by one set of substitution rates in the avian host, and by a different set of substitution rates in the human host, as illustrated in Figure 3.2. The two different substitution models shared the WAG exchangeability matrix $\mathbf{S}$ and a site-specific rate-scaling factor v, but now the equilibrium amino acid frequencies were both host- and site-specific. We identified sites with statistical support for different substitution rates in the two hosts, using a false discovery rate (FDR) method to account for multiple hypothesis testing (Benjamini & Hochberg, 1995). We identified 172 sites with an FDR<0.05 (i.e. we would expect 5% of these sites to be false positives), and 518 sites with an FDR<0.20. We will refer to the 172 higher-confidence locations as 'A sites' and the remaining 346 lower-confidence locations as 'B sites'. Table 3.2 lists the number of high- and low-confidence sites identified in each protein.

We then considered if modelling differences in the equilibrium amino acid frequencies was adequate, or whether we should include host-dependent rate scaling factors as well. We

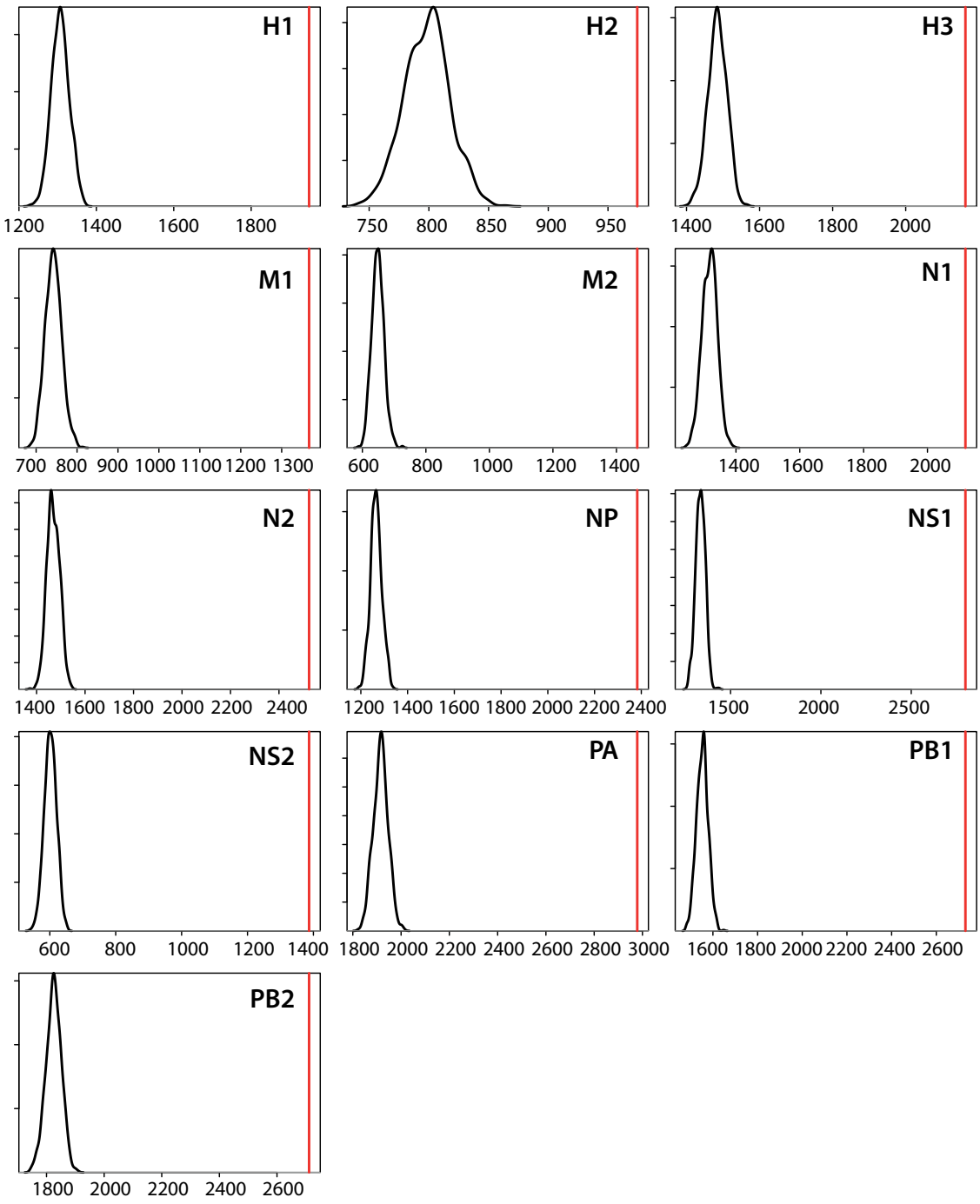


Figure 3.3.: Parametric bootstrap test for model 2 vs. model 1. Density of $\Sigma(\Delta\ell)(=\Sigma_{sites}\ell_{\text{Model2}}-\ell_{\text{Model1}})$ for simulations of influenza proteins. Black curve is distribution of $\Sigma(\Delta\ell)$ from datasets generated using Model 1. Red curve is $\Sigma(\Delta\ell)$ of protein sequence alignment, showing that Model 2 can be justified for all influenza proteins.

Protein	Alignment Length	Conserved positions	Number of sites identified A-sites (FDR < 0.05)	B-sites (FDR < 0.20)
H1	566	370	51	94
H2	592	432	7	32
H3	567	357	11	55
M1	252	143	4	13
M2	97	16	2	13
N1	470	274	33	100
N2	469	266	29	106
NP	507	329	13	47
NS1	305	131	4	21
NS2	126	41	0	0
PA	716	453	2	5
PB1	784	549	3	10
PB2	759	495	13	22
	6210	3856	172	518

Table 3.2.: Number of sites identified having host-specific selective constraints in each protein

implemented a more complicated model (Model 4) where the substitution rates were still defined with the WAG exchangeability matrix, but now both the equilibrium frequencies and the scaling factor v were host- and site-dependent. Of the 2143 sites considered, few (37) had P values less than 0.05; after correcting for multiple hypothesis testing using the false discovery rate method, no site yielded any statistically significant improvement. The results described below will be based on Model 3 above.

The list of 172 'A' sites (FDR<0.05) is shown in Table 3.3. Sites were found on all of the genes considered (haemagglutinin sites are listed using H3 numbering). Appendix C shows the list of the 518 'A' and 'B' sites with FDR<0.20. Sites that have been identified experimentally are detected using this method, notably PB2 627. HA sites H1 190 and 225 and H3 228 are also identified. Sites H2 226 and 228 are significant at the weaker FDR<0.20 level, while H3 226 is not statistically significant.

Table 3.3.: Sites identified as undergoing changes in selective pressure during host shifts from birds to humans. Residues are shown for amino acids with $\pi > 0.5$, $(\pi > 0.1)$ and $((\pi > 0.01))$.

Location	P value	FDR cutoff	Residues		Sel. constraint (d)		Cal09
			Avian	Human	Avian	Human	
H1							
-5	3.05e-04	2.59e-03	E	K	2.74	2.85	K
2	5.09e-05	7.86e-04	F	L	3.30	2.63	L
7	1.58e-03	0.010	V((A))	A	2.59	2.91	T
8	4.47e-03	0.021	L	T	2.63	2.71	A
15	9.36e-03	0.038	V((I))	I	2.68	2.63	I
54	4.90e-05	7.86e-04	N	K	2.79	2.89	R
63	5.80e-03	0.025	K	N	2.89	2.79	K
70	5.19e-03	0.023	L	I((V))	2.63	2.57	I
77	4.58e-03	0.021	D	E(G)	3.07	2.25	E
80	1.86e-03	0.011	T	S((P))	2.73	2.26	T
91	2.57e-03	0.014	S(T)	P	2.06	3.13	P
120	7.69e-04	5.68e-03	K	R	2.86	2.82	R
138	0.011	0.042	A	S((A))	2.91	2.24	A
141	1.55e-04	1.48e-03	Y	H	3.50	4.01	H
154	4.62e-05	7.86e-04	I((L))	L	2.51	2.63	L
155	1.01e-05	7.42e-04	T(I)	T	2.20	2.73	V
159	1.57e-04	1.48e-03	N(T)	G((S))	2.41	2.49	N
160	0.011	0.042	S	L((S))	2.50	2.41	S
163	1.89e-04	1.70e-03	K	N((T,S))	2.89	2.52	K
187	3.47e-03	0.017	T((N))	N((S))	2.61	2.71	T
188	4.16e-05	7.86e-04	T((V,A))	I((S,T,M))	2.23	2.20	S

Location	P value	FDR cutoff	Residues		Sel. constraint (d)		Cal09
			Avian	Human	Avian	Human	
189	3.84e-04	2.96e-03	S(G)((D,N))	G(K,R)((T,E,D))	1.64	1.36	A
190	1.76e-09	2.99e-07	E	D(V)((N))	2.74	2.20	D
192	0.012	0.044	Q	(K,M,R)	3.34	1.98	Q
193	4.10e-03	0.020	N(E)((T,S))	(A,T,N)	1.80	1.84	S
197	5.62e-03	0.025	N	T(K)	2.79	2.15	N
198	3.23e-04	2.62e-03	T((V,A))	E((G,V))	2.28	2.47	A
214	0.011	0.042	T((N))	T	2.49	2.70	K
222	2.46e-03	0.013	K((R))	K	2.62	2.89	K
225	6.46e-05	9.15e-04	G	D((G,N))	2.60	2.83	D
238	1.82e-05	7.42e-04	D	E	3.07	2.70	E
239	4.90e-05	7.86e-04	Q	P	3.34	3.21	P
244	1.50e-03	0.010	T	I((M))	2.73	2.56	T
248	2.08e-03	0.012	T	N((S))	2.73	2.67	T
261	1.21e-03	8.54e-03	N	S((N))	2.79	2.45	E
262	1.46e-04	1.48e-03	K	R	2.89	2.82	R
271A	1.55e-04	1.48e-03	D	N	3.07	2.79	D
272	2.98e-03	0.015	A(T,V)	A	1.97	2.87	T
274	2.18e-05	7.42e-04	V((I))	M	2.73	3.42	V
279	0.011	0.042	T	A((S))	2.73	2.77	T
280	1.90e-03	0.011	R(K)	K	2.18	2.89	T
285	1.48e-05	7.42e-04	H((Y,R))	Q	3.60	3.20	K
288	4.90e-05	7.86e-04	L	I	2.63	2.66	I
300	7.61e-05	9.95e-04	I	V	2.66	2.79	I
309	1.76e-03	0.011	V(I)	V	2.49	2.80	V
310	1.26e-04	1.48e-03	K	R	2.89	2.80	K

Location	P value	FDR cutoff	Residues		Sel. constraint (d)		Cal09
			Avian	Human	Avian	Human	
323	8.95e-03	0.037	V	I	2.83	2.61	I
H1(2)							
72	3.32e-03	0.033	N	K	2.79	2.89	H
77	2.08e-04	4.17e-03	I	M	2.66	3.47	K
116	2.07e-04	4.17e-03	R	K	2.83	2.89	K
127	7.99e-04	0.011	R	K	2.83	2.89	K
H2							
186	5.32e-04	0.018	N	I(N)((K))	2.78	1.95	
197	3.62e-04	0.018	N	E(K,N)	2.78	2.01	
205	1.11e-04	0.011	G	S(V)	2.60	1.96	
H2(2)							
45	5.36e-03	0.042	I(V)	F	2.28	3.30	
130	5.38e-03	0.042	A	V((A))	2.91	2.57	
169	5.47e-03	0.042	N	K	2.79	2.89	
180	2.11e-03	0.042	N((S))	S	2.63	2.50	
H3							
-7	2.50e-03	0.037	C(Y)	Y	2.92	3.43	
0	5.04e-05	5.70e-03	(G,S,C)	A(T)((S))	1.79	2.13	
4	2.10e-03	0.035	S((P))	P(S)	2.30	2.70	
57	1.10e-03	0.028	K((R))	Q(R)	2.60	2.72	
63	7.32e-04	0.024	D	N	3.07	2.79	
67	1.30e-03	0.028	(I,V,M)	I	1.86	2.60	
92	6.95e-05	5.70e-03	N((S))	K((T,N))	2.49	2.48	
145	1.40e-03	0.028	N(R,S)	K(N)((S))	1.98	2.00	
213	1.17e-04	6.50e-03	I	V	2.66	2.83	

Location	P value	FDR	Residues		Sel. constraint (d)		Cal09
		cutoff	Avian	Human	Avian	Human	
228	5.59e-04	0.023	G	S	2.60	2.50	
244	1.80e-03	0.033	V	L((I))	2.83	2.42	
M1							
115	2.05e-04	0.011	V	I((V))	2.79	2.55	V
137	1.30e-04	0.011	T	A((T))	2.71	2.76	T
174	1.06e-03	0.029	R	K(R)	2.80	2.20	R
231	4.64e-04	0.017	D	(N,D,S)	3.02	1.70	D
M2							
10	9.63e-04	0.039	(L,H,P)	P	2.20	3.21	P
93	4.26e-04	0.035	N((Y,I,S))	(S,I,Q)((N))	2.39	1.49	N
N1							
3	5.96e-03	0.040	P	P((T,S))	3.21	3.04	P
29	7.25e-03	0.043	M(I)	I	2.81	2.64	I
34	4.03e-03	0.031	V((G,I,A))	(I,V,A)	2.36	1.71	I
42	3.43e-03	0.028	(G,N)((S,D))	S((N))	1.63	2.35	N
46	1.96e-03	0.017	(A,P,V,T,S)	T	1.35	2.72	I
47	4.34e-03	0.031	E(G)((D))	G	1.92	2.59	E
52	1.42e-03	0.013	S	R((G,N,K))	2.50	2.51	S
59	2.08e-03	0.018	N((K))	S((N,R))	2.65	2.19	N
67	6.62e-03	0.043	(L,I,V)	V	1.63	2.83	V
74	3.23e-04	6.33e-03	(L,F,S,V)((I))	V	1.28	2.81	F
80	7.93e-04	9.32e-03	V((R,A,M))	(I,V,K)((T,S))	2.21	1.32	V
157	2.11e-04	5.47e-03	T	A	2.71	2.87	T
189	3.08e-08	6.04e-06	S((G))	G	2.40	2.60	N
214	4.73e-06	3.09e-04	D	E(G)	3.07	2.15	D

Location	P value	FDR cutoff	Residues		Sel. constraint (d)		Cal09
			Avian	Human	Avian	Human	
220	4.22e-04	6.82e-03	R((G))	K(R)	2.57	2.37	R
221	4.52e-04	6.82e-03	N(G)	K	2.41	2.87	N
264	8.09e-04	9.32e-03	(I,A,V)	T	1.71	2.68	V
274	2.80e-05	9.13e-04	Y	(S,F)((Y))	3.50	2.12	Y
288	5.84e-04	8.18e-03	I(V)	V	2.08	2.83	I
289	8.64e-04	9.41e-03	(I,T,M)	M	1.82	3.47	T
309	6.95e-03	0.043	N(D)	N	2.37	2.79	N
311	2.12e-05	8.30e-04	E((D))	D	2.59	3.07	E
329	7.22e-03	0.043	N	K(E)((R))	2.79	2.43	N
339	1.27e-03	0.012	S((L))	T(Y)((N))	2.40	2.11	S
340	1.18e-05	5.80e-04	(L,S,P)((H))	V((A,H,P))	1.55	2.26	S
341	4.13e-04	6.82e-03	N	D	2.76	3.05	N
351	6.82e-04	8.91e-03	F((Y))	Y	3.01	3.49	F
365	1.20e-03	0.012	T(I,P)	N((S,T))	2.05	2.59	I
382	8.60e-08	8.42e-06	E((G,D))	D(N)	2.35	2.30	G
393	4.35e-03	0.031	I	V(I)	2.63	2.30	I
427	4.44e-03	0.031	I	V(I)	2.66	2.46	I
430	2.23e-04	5.47e-03	R((L))	L((Q,R))	2.59	2.30	R
455	2.60e-04	5.66e-03	G(S,D)	N(D)	1.65	2.47	W
N2							
41	1.58e-03	0.025	E((G))	E	2.43	2.74	
50	4.12e-03	0.039	V(A)((T,I))	V((A))	1.98	2.69	
51	5.36e-04	0.012	V((M,T))	M	2.39	3.43	
60	5.79e-03	0.047	R(K)	R	2.30	2.83	
62	4.28e-03	0.039	(I,T,M)((V))	I((T))	1.59	2.59	

Location	P value	FDR cutoff	Residues		Sel. constraint (d)		Cal09
			Avian	Human	Avian	Human	
70	1.82e-04	6.14e-03	S(H,N)	N	1.96	2.78	
72	2.10e-03	0.027	T(I)((V))	T	2.00	2.70	
77	2.05e-03	0.027	(I,K,T)((V,L))	I((K))	1.36	2.32	
81	2.83e-04	8.22e-03	A((V,M,L,I,T))	(P,L,A,T)	1.92	1.62	
83	7.97e-05	3.24e-03	G(E,D)((K))	E	1.61	2.70	
125	7.62e-04	0.014	(G,S,D)	D	1.63	3.07	
126	3.57e-03	0.038	(L,P,T)((H,S))	P((S))	1.68	3.08	
147	7.60e-04	0.014	G	D((N))	2.60	2.94	
155	4.04e-03	0.039	H	Y(H)	3.96	3.19	
192	3.44e-04	8.74e-03	V(I)	V	2.48	2.83	
216	6.64e-03	0.048	(G,V,A)((S))	V(G,S)	1.45	1.92	
283	8.85e-04	0.015	R(Q)	R	2.36	2.83	
286	2.03e-03	0.027	(I,E,N)((D))	G((D))	1.51	2.45	
315	6.05e-03	0.047	G(S,R)((N))	S(R)	1.53	2.14	
328	4.65e-05	2.36e-03	N	K((R))	2.79	2.70	
331	2.49e-03	0.028	(I,R,G,S)	S(R)((N))	1.38	1.86	
338	4.47e-03	0.039	R(K)	L(Q,W)((K,R))	2.34	1.79	
369	5.46e-03	0.046	D	K(E)	3.07	2.18	
378	2.43e-03	0.028	R(K)	K	2.29	2.89	
381	1.46e-05	9.88e-04	G((D,N))	E(D)	2.36	2.42	
384	4.38e-06	4.45e-04	(A,T,I)((V,N,S))	V(I)	1.18	2.06	
386	2.74e-06	4.45e-04	A((P))	P((S))	2.68	3.09	
396	6.80e-03	0.048	V(I)	V	2.21	2.83	
399	6.69e-03	0.048	D	E	3.06	2.74	
NP							

Location	P value	FDR cutoff	Residues		Sel. constraint (d)		Cal09
			Avian	Human	Avian	Human	
77	1.37e-04	8.30e-03	R(K)	K(R)	2.44	2.52	K
101	1.88e-07	3.45e-05	(E,D,N)	G(N)((D))	1.90	1.96	D
102	3.60e-04	0.013	G	G((R))	2.60	2.44	G
131	2.71e-04	0.012	A	A(R)	2.90	2.51	A
136	6.30e-04	0.019	(L,M,I)	I(M)	1.91	2.32	I
283	3.50e-03	0.049	L	P	2.63	3.21	L
305	1.40e-03	0.032	R(K)	K((R))	2.44	2.80	K
335	9.62e-04	0.025	S	S((F))	2.50	2.38	S
353	3.10e-03	0.047	(V,I,L)((A))	(C,S,F,L)((I,V))	1.45	1.48	I
357	2.00e-03	0.036	Q	K((R))	3.29	2.65	K
375	1.26e-04	8.30e-03	(V,D,E,G)((S,N))	G(V)((E))	1.30	1.97	D
425	1.90e-03	0.036	I	I(V)	2.66	2.11	V
472	2.50e-03	0.042	T	T(A)	2.73	2.31	T
NS1							
81	5.08e-04	0.029	I(T)((V,M))	M((V))	1.93	3.39	I
84	4.84e-04	0.029	(V,M,G,S)((L,A,I,T))	T(A)((V))	1.17	1.97	V
215	8.84e-06	1.55e-03	(P,S,L)((T,A))	T((P))	1.65	2.60	P
227	6.65e-04	0.029	E((G,K))	R(G)((E))	2.43	2.19	-
PA							
356	2.66e-04	0.035	K((R))	R((K))	2.70	2.68	R
552	1.94e-04	0.035	T	S	2.73	2.50	T
PB1							
52	4.10e-04	0.032	K((R))	R(K)	2.74	2.17	K
517	3.63e-04	0.032	I((V))	V(I)	2.58	2.06	V
584	7.67e-07	1.81e-04	R((H))	Q(H)	2.73	2.93	Q

Location	P value	FDR cutoff	Residues		Sel. constraint (d)		Cal09
			Avian	Human	Avian	Human	
PB2							
44	6.18e-04	0.023	A(S)	L(S)	2.53	2.15	A
105	9.68e-05	0.013	T(A)((I,M))	V(M)((I))	2.01	2.42	T
199	2.78e-04	0.023	A	S	2.88	2.50	A
475	5.46e-04	0.023	L((M))	M	2.51	3.50	L
493	1.80e-03	0.039	R((K))	K((R))	2.53	2.70	R
569	2.40e-03	0.049	T((A))	A((S))	2.51	2.69	T
613	1.10e-03	0.035	V(A)((I))	T(I,A)	2.33	1.82	V
627	1.20e-03	0.035	E(K)	K	2.20	2.89	E
661	5.91e-04	0.023	A(T)((V))	T((V))	2.28	2.51	A
682	1.50e-03	0.038	G	S(N)	2.60	1.94	G
684	7.63e-05	0.013	A((T))	S(T)	2.69	1.95	S
702	1.60e-03	0.038	K(R)	R	2.42	2.78	K
740	3.83e-04	0.023	D	D(N)	3.03	2.24	D

To assess the performance of the technique described here, we simulated each one of the 759 sites in the PB2 gene ten times (7,590 simulations in total). All sites were simulated using the same fixed tree topology. The 22 'A' and 'B' sites identified as undergoing selective constraint changes (FDR<0.20) were simulated under the nonhomogeneous model, using the parameters obtained by optimising model 3. Similarly, the 737 locations with no evidence for change in selective constraints were simulated under the homogeneous model (model 2). We then applied the analysis described above to identify locations in the synthetic datasets that had undergone changes in selective pressure. On average, we observed

that 2.1% of the locations identified with FDR<0.05 were false positives (false positive rate of 0.12%); this increased to 12.76% (false positive rate of 1%) for FDR<0.20. This indicates that the FDR values are, at least for PB2, likely to be conservative. Of the 22 locations modelled with changing selective constraints, 13.7 were identified with FDR<0.05 (false negative rate of 37.7%), with 17.1 identified with FDR<0.20 (false negative rate of 22.3%). The 13 'A' sites were identified more consistently, with 7.9 found with FDR<0.05 and 10.2 found with FDR<0.20. This suggests that there remain more locations undergoing changes in selective pressure than are being identified with the procedure described here.

Our approach relies on the prior construction of an appropriate phylogenetic tree. In order to estimate the effect of phylogenetic uncertainty, we repeated the analysis of the PB2 gene segment with ten different phylogenetic trees obtained through nonparametric bootstrapping. The 13 'A' sites were identified on 79% of the bootstrap trees with FDR<0.05 and identified on 90% with FDR<0.20. 85% of the 22 'A' and 'B' sites were similarly identified on the bootstrap trees with FDR<0.20. Conversely, the bootstrap trees identified on average 2% (with FDR<0.05) and 6% (with FDR<0.20) of alternative locations that were not identified on the original tree. These might be false positives for the alternative trees, suggesting a similar amount of false positives on the original tree. Some of these locations, however, may be locations with changes in selective constraints, and thus represent false negatives for the original tree; most of these locations would have been so identified with a higher FDR threshold of 0.50, although these points represent only about 12% of the otherwise unidentified locations.

We constructed a simple model to help explain the lack of statistically significant improvement with adding host-specific scaling factors. This was based on considering a protein site where two amino acids (A and B) are present, where an organism with residue B has a fitness equal to $1 - s$ relative to an organism with residue A. We used Kimura's fixation rate theory (Crow & Kimura, 1970) to calculate the resulting substitution rates between A to B, and for-

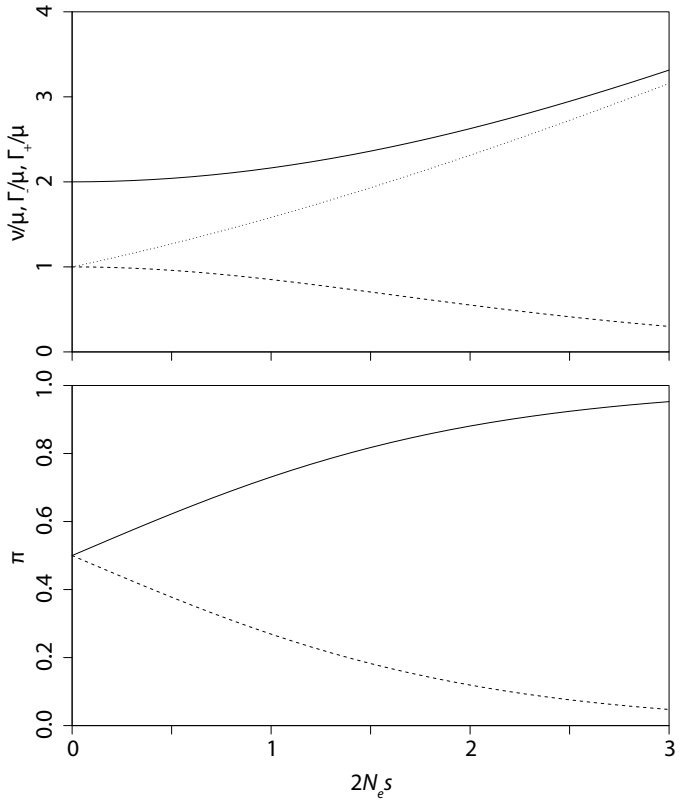


Figure 3.4.: Changing equilibrium frequencies and rates versus selective constraints. Top: Dependence of rate scaling factor ν (solid line) and rate of substitutions for positive selection Γ_+ (dotted line) and negative selection Γ_-, scaled by mutation rate μ, as a function of scaled selective disadvantage of residue B compared with residue A ($2N_e s$). Bottom: Equilibrium frequencies π_A of A (solid line) and π_B of B (dashed line) as a function of scaled selective disadvantage.

mulate these expressions in terms of a rate scaling factor ν and equilibrium frequencies π_A and π_B (= $1 - \pi_A$). We considered how ν, π_A, and π_B change as the relative fitness difference between A and B is altered. We also considered the overall rate at which substitutions occur in both directions, both for negative selection where the residues are at equilibrium (Γ_-) as well as for positive selection (Γ_+) where the location contains the unfavourable residue B. Figure 3.4 shows the dependence of π_A, π_B, ν, Γ_- and Γ_+ (the latter three normalised by the mutation rate μ) on the relative fitness difference s (scaled by the effective population size N_e). As shown, under conditions of negative selection, increasing fitness differences result in a decrease in the overall rate of substitutions, but an increase in the rate-scaling factor. There is a relatively weak dependence of ν on s as long as the latter is not large relative to $1/N_e$. Under conditions of positive selection, both quantities increase with larger fitness differences.

The theoretically predicted weak dependence of ν on selective pressure and the lack of statistical support for host-dependent values of this parameter indicate that ν is not a good measure of the degree of selective constraints. To generate a more appropriate measure, we calculated the relative entropy (d), between the equilibrium frequencies and what would be expected under no selection, π^0, estimating the latter by averaging the amino acid frequencies over our entire database. This measure of selective constraint magnitudes for the various sites in avian and human hosts are presented in Table 3.3, Appendix A, and in Figure 3.5.

3.4. Discussion

As described in the introduction, ignoring the underlying phylogenetic relationship often results in a gross over-estimation of statistical significance, as single evolutionary events are interpreted as a large number of independent measurements. Correspondingly, certain sites that have been identified by other methods that do not model the underlying phylogenetics

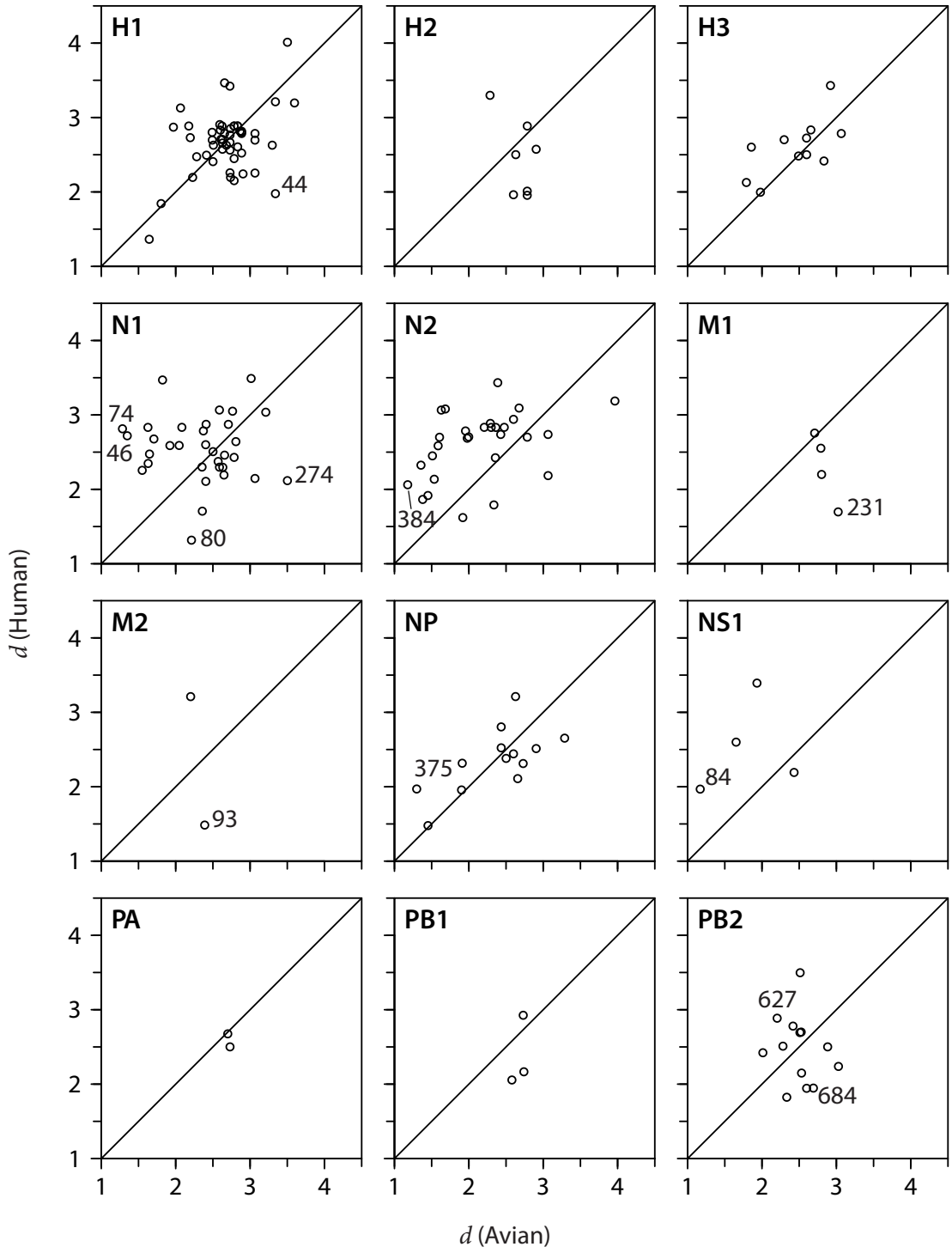


Figure 3.5.: Strength of selective constraints (measured as *d*, described in Eqn. 3.5), for viral sites identified (FDR<0.05) as under different selective constraints in avian and human hosts. Selected sites are labelled.

lose their statistical significance when the phylogenetics is considered. For instance, site 271 in PB2 is identified as a significant site in three previous analyses (Chen *et al.*, 2006; Finkelstein *et al.*, 2007; Miotto *et al.*, 2008); human viral sequences are most commonly alanine at this site, while avian viral sequences are predominantly threonine, although alanine also occurs. When each sequence is interpreted as an independent event, there is strong statistical support for host-specific amino acid distributions at this site. All of the alanines in the human lineage, however, can be explained by a single threonine to alanine substitution. In contrast, in the avian influenza there were at least three independent threonine to alanine substitutions (see Figure 3.6). The single example of the substitution in human influenza is not significant given the relative frequency of this transition in avian influenza. Indeed, the more complex Model 3 incorporating host-dependent substitution rates has a P value of 0.095 compared with Model 2 that assumes no such host dependence, and would need an FDR cutoff of 0.48 to be included in our set of identified sites. More threonine to alanine substitutions in the human lineage, even if that meant more human sequences with a threonine at this site, would have provided more statistical support. The statistical support would also have been larger if the various avian strains with an alanine at this site represented the result of a single substitution.

The sites that are identified are those with a significant statistical signal given the available data; other sites might be undergoing shifts in selective constraints that are not detected for different reasons. As with all appropriate statistical methods applied to this problem, we require adequate evolutionary time and a suitable substitution rate for the substitution patterns to be detectable (Huelsenbeck, 1995; Pollock *et al.*, 2002). In particular, there has to be sufficient evolutionary time in both the avian and human lineages for the parameters in the substitution models to be sufficiently well defined in each so that the differences in selective constraints are detectable. This will require longer evolutionary time when the selective constraint changes are smaller. As shown in the phylogenetic trees (Appendix B), there is

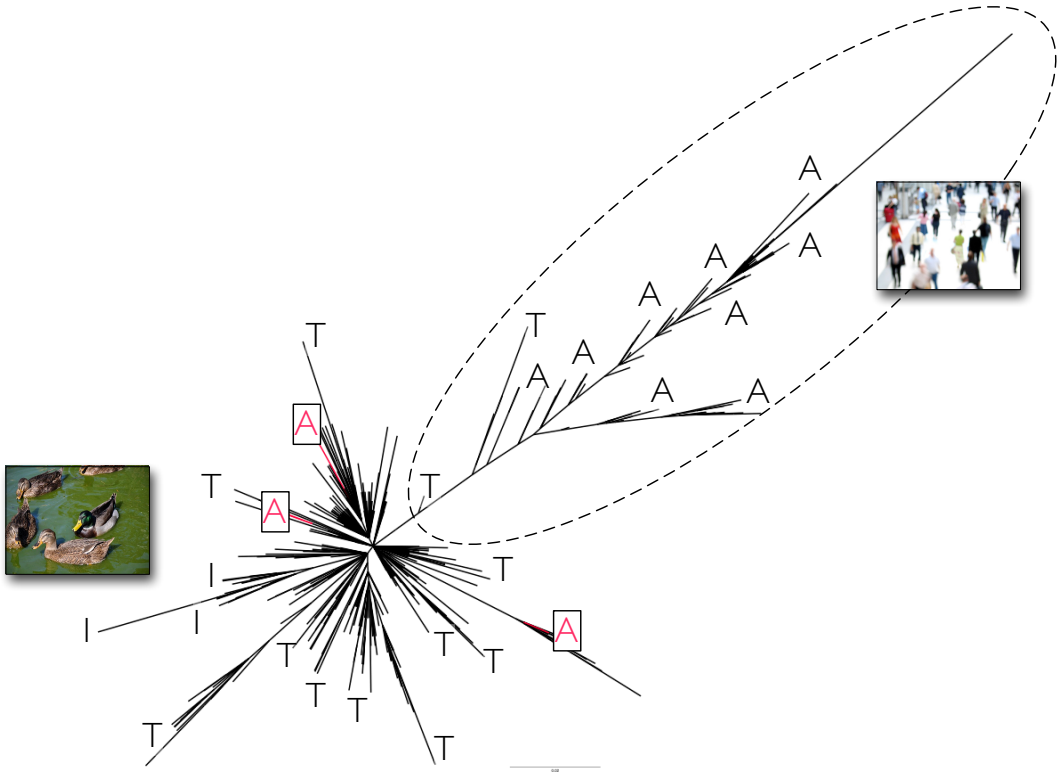


Figure 3.6.: Tree for PB2 protein with residue at site 271. All of the alanines in the human lineage can be explained by a single substitution. In the avian influenza there were at least three independent threonine to alanine substitutions.

relatively little sequence evolution in the human H2 lineage; this is possibly the cause of the relatively few sites identified in this gene subtype. There are more H3 sequences, although most available avian H3 sequences are highly similar, reducing our ability to detect selective pressure changes in this gene subtype. In particular, we do not identify the H3 Q226L mutation whose importance has been determined experimentally, as the strict conservation of glutamine in the avian lineage is not highly informative given the lack of evolutionary divergence among the avian H3 sequences. Finally, the improvement in the log likelihood necessary for a given level of statistical significance is a function in the increase in the number of adjustable parameters between the two models, which is one minus the number of amino acids found in that location. Locations that are highly variable require more adjustable parameters, reducing the power of the likelihood ratio test. In particular, human H3 viruses contain glutamine, leucine, isoleucine, and valine at position 226, making identification of selective constraint changes at this location difficult.

The identified changes in selective constraints may not be the direct result of the host shift event. Selection constraint changes at one site might be a response to substitutions that occur at a different site, even if those changes were themselves the result of neutral drift. We have also assumed that the change in selection constraint occurs simultaneously with the host shift event. In reality this method has limited temporal resolution, and changes in the substitution rate occurring near the host shift event might also be identified.

We do not include 'pre-selection' in the model, that is, that the match between the avian sequence and the selective constraints in the human host does not influence the probability that that particular virus strain will undergo a shift to humans. This could be added to such a phylogenetic-based model by considering the probability that a host shift would occur on a given lineage as a function of the protein sequence. This would greatly increase the complexity of the model, increasing the number of adjustable parameters, reducing the statistical power of the method. This would also increase the number of false negatives, as

these occurrences would look identical to founder effects. It is less likely that this process would produce false positives.

We have included information from the A/California/04/2009 (H1N1) sequences from the 2009 'Swine flu' pandemic in Table 3.3. Considering the locations identified with a false discovery rate of 5%, most segments originally from classical or European swine (NA, M1, NP, NS1) mostly matched the human selective constraints, suggesting a similarity between the constraints in humans and swine. The exception is the HA gene, where many locations seemed to match the avian selective constraints despite its classical-swine origin, possibly reflecting the slow rate of antigenic change of the classical swine haemagglutinin (Sheerar *et al.*, 1989; Vincent *et al.*, 2006). In the segments more recently from the avian lineages (PA, PB2), most locations more closely matched the avian constraints, while PB2 684 and PA 356 more closely matched the human. Interestingly, by comparing with avian sequences, it appears that PB2 A684S and PA K356R substitutions, both involving changes from an avian-like to a human-like amino acid, occurred in the interval between the host shift to swine and the subsequent transfer to humans, suggesting that these changes might be related to the ability of these viruses to infect humans.

3.4.1. Changes in π versus change in ν

Most methods that look for changes in the substitution rates model this as changes in ν, the scaling parameter, or in the related ratio of nonsynonymous to synonymous substitution rates. In our analysis we find that when we allow the equilibrium frequencies π to vary, there is no statistically significant variation in ν. This seems initially counter-intuitive, as there are some sites where there seems to be substantial changes in the degree of conservation; in site 274 in N1, for instance, is almost universally tyrosine in avian viruses, while it varies between tyrosine, serine, and phenylalanine in human viruses. Yet the likelihood ratio test

applied to this site rejects the inclusion of host-dependent scaling factors with a P value of 0.90, suggesting that the relationship between rate scaling factors and site variation are not simply related.

This observation motivated our simple model to try to gain insight into the relationship between equilibrium frequencies and rate scaling factors, by considering a protein site where two different amino acids, A and B, are found. We imagine that organisms possessing residue A at this location have a fitness advantage. Negative purifying selection would occur when the residues at this location are at their equilibrium value, while positive selection would occur when this location was filled by B, such as might occur when the selective pressure on the protein changes. By using Kimura's theory of fixation probability (Crow & Kimura, 1970), we can calculate the values of the rate scaling factor ν, the overall rate of substitutions for purifying (Γ_-) and positive selection (Γ_+), and the equilibrium frequencies of A (π_A) and B (π_B), as a function of the different fitnesses provided to an organism with the two different possible amino acids at that location, as described in the Methods section. Normalised values of ν, Γ_-, and Γ_+ are plotted as a function of $2N_e s$ in Figure 3.4. As shown, ν varies surprisingly little with s as long as s is not much more than $1/N_e$. This explains why including a host-specific ν never yielded statistically significant improvements with our data. When we consider adaptive substitutions, larger values of s correspond to higher selective constraints, larger values of ν, and faster evolution. The situation is quite different with purifying selection. As might be expected, larger values of s (corresponding to larger degree of purifying selection) result in a slower substitution rate, but this actually corresponds to larger values of ν. The reason why most phylogenetic programs use an inverted relationship, where larger values of ν correspond to faster substitution rates, is that they do not consider the value of π appropriate for each site. By assuming that the same values of π apply to all sites, a more extreme distribution of equilibrium frequencies, resulting in a decrease in the number of substitutions, is interpreted as a reduction in ν although this

parameter is, in fact, increasing.

The magnitude of the selective constraints for the various sites in avian and human hosts are presented in Table 3.3, Figure 3.5 and in Appendix C. It is interesting to note the number of positions under changing selection constraints where the magnitudes of the selection constraints are relatively constant. Such sites would be difficult to detect by looking for changes in the substitution rate, especially in cases where the distributions of amino acids found in the two hosts have significant overlap.

The methods described here are applicable for a wide range of problems involving changes in selective constraints. There are two particular factors, however, that make the technique especially well suited for influenza. Firstly, the branch along which the selective pressure changes can be identified a priori. Secondly, it is important to generate appropriate phylogenetic trees for the position under consideration. Generation of such trees can be complicated when there is incongruence between different locations. For influenza, incongruence between the various genomic segments results from the process of reassortment, where chimeric viruses containing genomic segments of different origin result from multiple infections. We are able to address this issue by considering each different genomic segment independently, constructing gene-specific phylogenetic trees. A more difficult problem is intra-gene homologous recombination, where different regions of a single genomic segment have different phylogenies. Such recombination is either extremely rare or non-existent in influenza (as well as other negative-sense RNA viruses), and has never been observed experimentally (Boni *et al.*, 2008; Chare *et al.*, 2003; Krasnitz *et al.*, 2008).

We have assumed that the transitions from avian to human hosts did not go through an intermediate species, such as swine. There is no evidence of involvement of swine in the 1957 Asian flu and 1968 Hong Kong flu host shift events. Based on his analysis of the 1918 Spanish flu sequences and the relative timing of the 1918 influenza outbreaks in swine and humans, Taubenberger concluded that the Spanish flu transferred in toto from birds to humans and

from humans to swine (Reid *et al.*, 2004; Taubenberger, 2006; Taubenberger *et al.*, 2005), although this conclusion has been challenged (Antonovics *et al.*, 2006; dos Reis *et al.*, 2009; Gibbs & Gibbs, 2006; Smith *et al.*, 2009a). If an intermediate host species were involved, it would not be expected to affect the results if the selective constraints at any location in this intermediate host were to resemble either that of avian or human viruses, as this would only change the timing of the shift from one selective constraint to another. If there were an intermediate host and the selective constraints at some locations in this intermediate host were strong and substantially different from either avian or human viruses, the amount of evolutionary time in this intermediate host were sufficiently long, and the evolutionary time in humans sufficiently short so that the new equilibrium is not attained, the results of these calculations could be affected.

There are two other important assumptions made in this work. Firstly, we assume that the selective constraints in human and avian viruses are constant, and that each location can be considered independently. We do not consider, for instance, that there may be different selective constraints in low-pathogenic and high-pathogenic avian viruses, or that compensatory changes can occur elsewhere in the protein or even in other proteins. The observation (both here and experimentally (Connor *et al.*, 1994; Matrosovich *et al.*, 2000; Nobusawa *et al.*, 1991; Rogers *et al.*, 1983; Vines *et al.*, 1998)) that different haemagglutinin subtypes undergo different patterns of change of selective constraints indicates that this assumption is not strictly valid.

4. Charting the host adaptation of influenza viruses

4.1. Introduction

We provided some background to the four influenza pandemics that have struck the human population over the last 100 years in chapter 1. The pandemics were caused by the introduction of a new virus into the human population from an avian or swine host or through the mixing of virus segments from an animal host with a human virus to create a new reassortant subtype virus. These host-shift events can result from the transfer of a complete virus from one host to another or from genetic reassortment, where a chimera is formed by the mixing of genetic segments from a virus of a different host with genetic segments of a virus already circulating in the "new" host.

Around the same time as the 1918 H1N1 pandemic, a panzootic was observed in swine, which is thought to have been the origin of the "classical swine" lineage observed especially in North America. The timing and nature of the host-shift events that caused the near simultaneous human and swine epidemics have been a matter of controversy. Reassortment resulted in two further pandemics in 1957 (H2N2) and 1968 (H3N2) (Kawaoka *et al.*, 1989; Schäfer *et al.*, 1993). After each of these pandemics, the new virus replaced the previously circulating subtype. In 1977, an H1N1 virus reappeared in the human population and co-circulated with H3N2 until 2009. The re-emerging virus closely resembled the H1N1 viruses that had circulated approximately 25 years earlier (dos Reis *et al.*, 2009; Nakajima *et al.*, 1978), suggesting that the virus was a member of the 1957 lineage and had been held in artificial evolutionary stasis during this time (Palese, 2004).

In the late 1970s, an independent "Eurasian swine" H1N1 lineage resulted from a direct

transmission from an avian host to pigs (Pensaert *et al.*, 1981). In the late 1990s, a series of reassortant viruses appeared in pigs in North America that initially combined genetic elements from human H3N2 (PB1, H3, and N2) with classical swine viruses followed by the introduction of genetic elements from avian influenza (PA and PB2) (Zhou *et al.*, 1999). This "triple-reassortant" strain then underwent various reassortments acquiring genetic elements from classical swine (H1) and Eurasian swine (N1 and MP) before undergoing a host shift to humans, resulting in the novel "swine origin" influenza virus (pandemic H1N1 2009). The major events over the last century of relevance to humans are listed in Table 4.1.

In the previous chapter we developed a maximum likelihood phylogenetic method to detect and characterise amino acid locations in influenza virus proteins that evolve under host-specific constraints. In this chapter, we describe how we can use these measures to characterise how well any given virus sequence is adapted to the selective constraints imposed by avian or human hosts. We focus on the host shift that led to the 1918 H1N1 pandemic and the process of adaptation of the viral proteins during the approximately 70 years that the viruses have circulated in the human population.

We show that adaptation to the human host has been gradual with a timescale of decades and that none of the virus proteins have yet achieved full adaptation to the selective constraints. We also find that the 1918 influenza virus is more adapted to human selective constraints compared to the ancestral reconstruction of the avian virus that founded the classical swine and 1918 human influenza lineages. The ancestral virus shows no evidence that it was exceptionally pre-adapted to humans. This indicates that adaptation to humans occurred following the initial host shift from birds to mammals, including a significant amount prior to 1918. It also seems that the 2009 pandemic virus had undergone pre-adaptation to human-like selective constraints during its period of circulation in swine. By analysing the adaptedness of ancestral sequences along the human virus tree, we find that mutations that have increased the adaptation of the virus have occurred preferentially along the trunk of

Year	Event	Segments	Resulting pandemic/panzootic lineage
Pre-1918	Host shift: ? to swine	?	Classical swine (H1N1)
Pre-1918	Host shift: ? to human	?	Spanish flu (H1N1)
1957	Host shift: avian to human	H2, N2, PB1	Asian flu (H2N2)
1968	Host shift: avian to human	H3, PB1	Hong Kong flu (H3N2)
1977	Reintroduction of human H1N1 virus	All	Russian flu (H1N1)
Late 1970s	Avian to swine	All	Eurasian swine (H1N1)
Late 1990s	Host shift: human to swine	H3, N2, PB1 from human	Reassortant swine (H3N2)
Late 1990s	Host shift: avian to swine	PA, PB2	Triple-reassortant swine (H3N2)
Pre-2009	Mixing between swine	H1 from classical swine; N1, M from Eurasian swine; NS, NP, PA, PB1, PB2 from triple-reassortant	H1N1
2009	Swine to human	All	Pandemic H1N1 2009

Table 4.1.: Significant events of relevance to recent human pandemics

the tree.

4.2. Methods

4.2.1. Host adaptation measure

In addition to identifying locations in influenza proteins where there is a change in selective constraints following a host shift from birds to humans, the analysis in the previous chapter also provided us with the expected equilibrium frequency of amino acid A_i at identified location k evolving in host h, $\pi_k^h(A_i)$. We can use these equilibrium frequencies to construct a measure of host adaptation. Consider that we have identified N locations in a given protein where there is a difference in selective constraints in human and avian hosts. If we assume that the selective constraints act at the protein level, we can, following Yang & Nielsen (2008), express the equilibrium frequencies $\pi_k^h(A_i)$ in terms of the "fitness parameters" for those amino acids $F_k^h(A_i)$:

$$\pi_k^h(A_i) \propto \Big(\sum_{I \in A_i} \pi^*_{I_1}\pi^*_{I_2}\pi^*_{I_3} e^{F_k^h(A_i)}\Big) \tag{4.1}$$

where $\pi^*_{I_l}$ represents the background equilibrium frequency for the nucleotide found in position l of codon I, and the sum is over all codons that code for amino acid A_i. With this expression, we can write $F_k^h(A_i) = K(A_i) + \ln(\pi_k^h(A_i))$, where $K(A_i)$ represents the nucleotide biases and the proportionality constant. Assuming that the fitness effects of the different locations are additive, we can create a measure of host adaptation $\theta^h(\{S_k\})$ of a virus

with amino acid sequence $\{S_k\}$, where S_k is the amino acid found at identified location k:

$$\begin{aligned}\theta^h(\{S_k\}) &= \sum_{k=1}^{N} F_k^h(S_k) \\ &= \sum_{k=1}^{N}[\ln(\pi_k^h(S_k)) + K(S_k)] \\ &= \sum_{k=1}^{N}[\ln(\pi_k^h(S_k))] + N\bar{K}\end{aligned} \tag{4.2}$$

where we have replaced the sum of $K(A_i)$ with the average value of $K(A_i)$, N times $\bar{K}$, which is only a function of the background distribution of nucleotides and should not vary significantly from one sequence to another.

Fully adapted proteins that had equilibrated with the selective constraints would have amino acid frequencies at the various sites given by the equilibrium frequencies $\pi_k^h(A_i)$. We can model random proteins as having amino acid frequencies at each location given by $\pi_0(A_i)$, the frequency of amino acid A_i averaged over our influenza sequence database. For convenience, we scale $\theta^h(\{S_k\})$ so that an ensemble of random proteins have an average host adaptedness of 0, whereas an ensemble of fully adapted proteins have an average host adaptedness of 1 by computing

$$H^h = \frac{\theta^h(\{S_k\}) - \langle\theta^h\rangle_{\text{Random}}}{\langle\theta^h\rangle_{\text{Adapted}} - \langle\theta^h\rangle_{\text{Random}}} \tag{4.3}$$

where $\langle\theta^h\rangle_{\text{Random}}$ and $\langle\theta^h\rangle_{\text{Adapted}}$ represent the average value of $\theta^h(\{S_k\})$ for an ensemble of random and adapted sequences, respectively:

$$\begin{aligned}\langle\theta^h\rangle_{\text{Random}} &= \sum_{k=1}^{N}\sum_{i=1}^{20}\pi_0(A_i)\ln(\pi_k^h(A_i)) + N\bar{K} \\ \langle\theta^h\rangle_{\text{Adapted}} &= \sum_{k=1}^{N}\sum_{i=1}^{20}\pi_k^h(A_i)\ln(\pi_k^h(A_i)) + N\bar{K}\end{aligned} \tag{4.4}$$

Note that $N\bar{K}$ drops out of equation 4.3 and does not need to be computed. Our results and conclusions were negligibly affected by our choice of $\pi_0(A_i)$, which was only used to scale the adaptedness values. We call H^h the "human adaptedness" when the host h is human and the "avian adaptedness" when the host is avian.

Individual sequences can have host adaptedness values less than zero or greater than one if the sequences have a greater number of especially unfavourable (low equilibrium frequency $\pi_k^h(A_i)$) residues compared with random sequences or a greater number of favourable (high equilibrium frequency $\pi_k^h(A_i)$ residues compared with fully adapted sequences.

The maximum likelihood estimate $\hat{\pi}_k^h(A_i)$ of $\pi_k^h(A_i)$ is zero for all amino acids not present at identified location k. In order to avoid logarithms of zero in equations 4.2 and 4.4, we incorporated pseudocounts into the calculation of $pi_h^k(A_i)$:

$$\pi_k^h(A_i) = \frac{\hat{\pi}_k^h(A_i) + \delta}{1 + 20\delta} \tag{4.5}$$

where δ was set equal to 10^{-6}. Varying δ did not appreciably change the results.

4.2.2. Example of adaptedness calculation

Consider an aligned set of protein sequence of length 2 where two different residues, A and B, are observed. Imagine our analysis indicates that A is strongly favoured in humans in both sites ($\pi_A^{\text{Human}} = 0.7$ and $\pi_B^{\text{Human}} = 0.3$). Over the entire viral genome, both residue types are found equally often ($\pi_A^0 = \pi_B^0 = 0.5$). Ignoring the effect of $N\bar{K}$ (which drops out at the end of the calculation), we can express the raw fitness of sequences AA, AB, BA and BB as

the sum of logs of the equilibrium frequencies:

$$\theta^{\text{Human}}(AA) = \log(0.7) + \log(0.7) = -0.71$$
$$\theta^{\text{Human}}(AB) = \log(0.7) + \log(0.3) = -1.56$$
$$\theta^{\text{Human}}(BA) = \log(0.3) + \log(0.7) = -1.56$$
$$\theta^{\text{Human}}(BB) = \log(0.3) + \log(0.3) = -2.41$$

An ensemble of random sequences, where each possible sequence is equally likely, would have an average θ^{Human} of $\langle\theta^{\text{Human}}\rangle_{\text{Random}} = 0.25 \times (-0.71) + 0.5 \times (-1.56) + 0.25 \times (-2.41) = -1.56$. In an ensemble of fully adapted sequences, where the proportion of *A*s and *B*s at each location matches the equilibrium frequencies, we would expect to find 49% *AA*, 21% *AB*, 21% *BA* and 9% *BB*. Such an ensemble would have an average θ^{Human} of $\langle\theta^{\text{Human}}\rangle_{\text{Adapted}} = 0.49\times(-0.71)+0.42\times(-1.56)+0.9\times(-2.41) = -1.22$. We scale the human adaptedness values by subtracting the average value of the random ensemble and dividing by the difference between the average of the adapted and random ensembles to yield

$$H^{\text{Human}}(AA) = \frac{\theta^h(AA) - \langle\theta^h\rangle_{\text{Random}}}{\langle\theta^h\rangle_{\text{Adapted}} - \langle\theta^h\rangle_{\text{Random}}} = \frac{-0.71 - (-1.56)}{-1.22 - (-1.56)} = 2.50$$
$$H^{\text{Human}}(AB) = H^{\text{Human}}(BA) = \frac{-1.56 - (-1.56)}{-1.22 - (-1.56)} = 0$$
$$H^{\text{Human}}(BB) = \frac{-2.41 - (-1.56)}{-1.22 - (-1.56)} = -2.50$$

As desired, our random ensemble of sequences (with equal mixtures of *AA*, *AB*, *BA* and *BB*) would have an average human adaptedness value of 0, whereas our adapted ensemble would have an average human adaptedness value of $0.49 \times 2.5 + 0.42 \times 0 + 0.09 \times (-2.5) = 1$.

In this case, *BB* has an adaptedness value less than 0 and *AA* has an adaptedness value greater than 1. This is because *BB* is less adapted than the average of a random ensemble,

Protein	Number of avian sequences	Number of human sequences	Number of swine sequences
H1	30	405	153
N1	232	279	65
NP	308	127	83
NS1	312	66	75
PA	347	65	70
PB2	321	85	71

Table 4.2.: Protein sequences used in the analysis

75% of which have at least one more favoured *A*; conversely, *AA* is better adapted than the average of an ensemble of adapted proteins, 51% of which have at least one less favoured *B*.

4.2.3. Sequence data and analysis

The protein sequence alignment and analysis used to estimate amino acid equilibrium frequencies are described in the previous chapter. To correct for multiple hypothesis testing, we used a false discovery rate (FDR) cutoff of 0.20 on all proteins sites together, rather than each protein separately (Appendix F shows results when FDR is applied per-gene). We identified 294 sites on six different proteins (H1: 84 sites, N1: 68 sites, NS1: 28 sites, NP: 48 sites, PA: 27 sites, and PB2: 39 sites). (M1, M2, and PB1 have relatively few locations undergoing changes in selective constraints and thus do not have sufficiently robust statistics for computing human and avian adaptedness.) We used these 294 sites to calculate host adaptedness for the various human and avian virus sequences as well as for the pandemic H1N1 2009 virus and selected classical swine and Eurasian swine virus sequences (see Table 4.2 and Appendix E) using equations 4.2-4.4 described above. Varying the FDR threshold between 5% and 20% or random resampling of included sites results in different magnitudes of change in adaptedness but has little effect on the qualitative results (see Appendix G for an example).

4.2.4. Reconstructing the host shift sequence

We are also interested in studying the host adaptedness of the ancestor of the 1918 pandemic virus. The host shift was assumed to have occurred at the midpoint of the branch connecting the parent node of the 1918 human H1N1 sequence with its parent (trees are included in Appendix B). (Moving the host shift along this branch did not appreciably affect the results of the calculation.) Using the maximum likelihood of our site-wise nonhomogeneous model, we calculated the posterior probability of every amino acid for every site at the host-shift event (Koshi & Goldstein, 1996). We sampled sequences from the posteriors 1,000 times, calculating the host adaptedness for each reconstruction. The mean and 95% confidence intervals (CIs) of the human adaptedness and avian adaptedness measures were constructed based on this sampling.

4.2.5. Reconstruction the pattern of sequence changes

We performed a reconstruction of the most likely set of substitutions for each protein using the joint reconstruction method of Pupko *et al.* (2000) under the WAG amino acid substitution model (Whelan & Goldman, 2001) with site-optimised rates. We then calculated human adaptedness measures for each node of the phylogenetic tree following the avian-to-human host shift. By traversing the phylogenetic tree for the protein starting at the host-shift node down through the human lineage, we calculated the change in human adaptedness along the trunk of tree, leaf nodes, and the remaining internal branches.

4.2.6. Fits to host adaptedness data

To study the change in host adaptedness with time, we fit the host adaptedness of human virus sequences (ignoring sporadic H5N1 infections) as a function of isolation date to two possible functional forms: (a) an exponential decay to baseline equal to 1.0, where $H(t)$ =

$1 - Ae^{-t/\tau}$ and (b) an exponential decay to an adjustable baseline, where $H(t) = B - Ae^{-t/\tau}$. The adjustable parameters are, as appropriate, the amplitude of change A, the adaptation time τ, and the asymptotic value B. We subtracted 25 years from the isolation date of post-1977 human H1N1 viruses corresponding to the time that these viruses were in artificial evolutionary stasis (dos Reis *et al.*, 2009). We used the likelihood ratio test ($P<0.05$) to test whether model (a) can be rejected in favour of (b). For the chosen model, we calculated CIs for the parameters and the time when the fit matches the human adaptedness at the host-shift sequence through bootstrapping by sampling the residuals.

4.3. Results

Figure 4.1 shows the host adaptedness (human or avian) values computed for the H1, N1, NS1, NP, PA, and PB2 proteins for a variety of avian, human, and swine viruses. Points representing the human pandemic viruses of 1918 and 2009 are indicated. In addition, we represent the position of the reconstructed virus at the host-shift event that gave rise to the 1918 pandemic. This figure highlights that the avian sequences are at equilibrium, clustering around 1.0, whereas mammalian viruses are spread out, suggesting an ongoing adaptation process.

To evaluate whether the virus at the pre-1918 host-shift event was a typical or exceptional avian virus, we calculated the fraction of avian viruses that were less well adapted to avian and human hosts compared with the host-shift virus. As shown in figure 4.2, the avian adaptedness and human adaptedness of the host-shift virus are generally within the distribution of values obtained for other avian viruses, although, interestingly, the polymerase proteins (PA and PB2) have relatively high avian adaptedness. This suggests that the host-shift virus was not exceptionally pre-adapted to humans. Figure 4.2 also shows how the pandemic H1N1 2009 virus proteins compared with the corresponding proteins of the lineage from

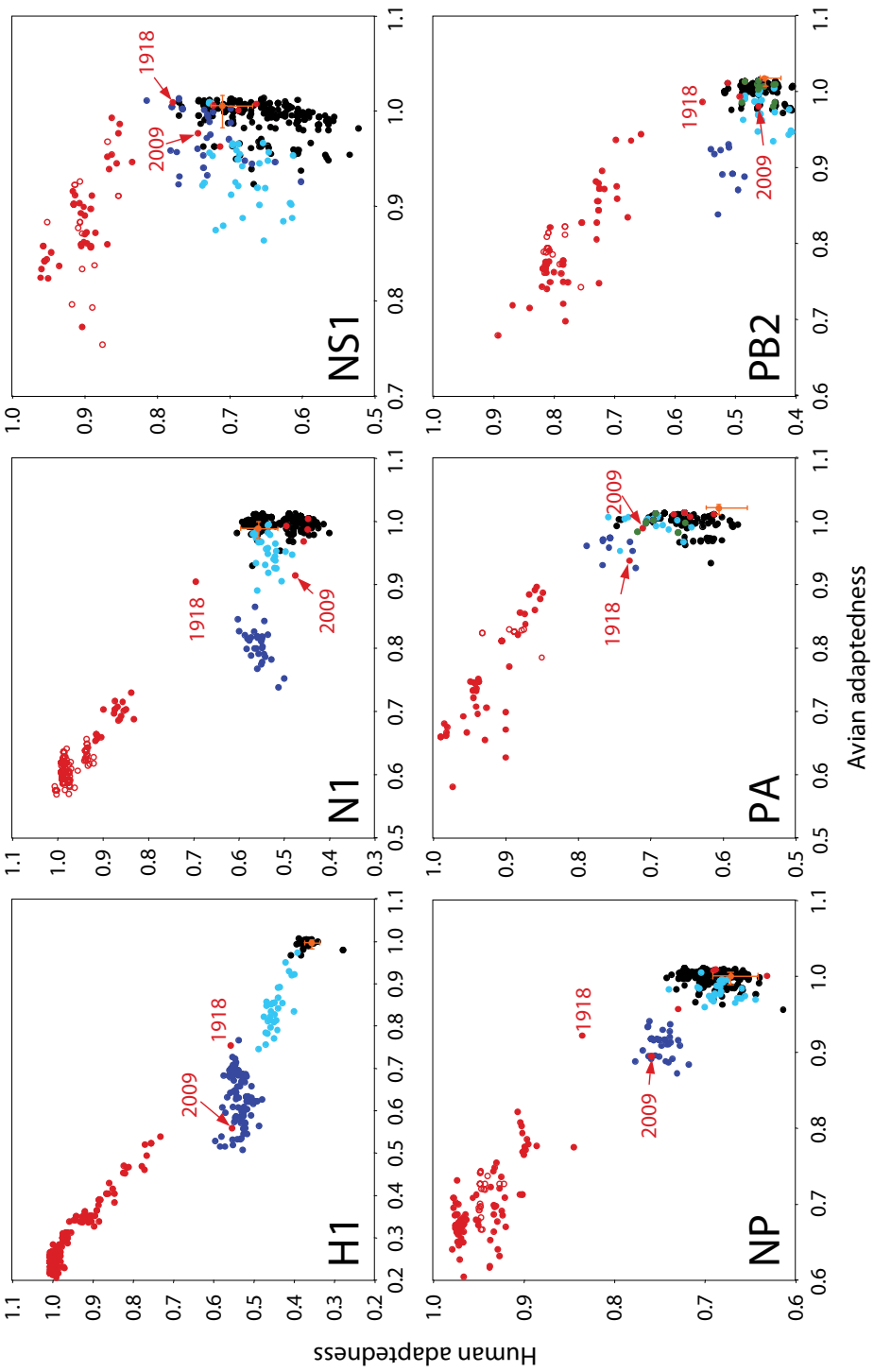


Figure 4.1.: Host adaptedness values for a series of different virus sequences, including avian (black), human (red), classical swine (blue), Eurasian swine (cyan) and the host-shift sequence (orange). Open red circles represent post-1977 human H1N1 viruses whose isolation times were corrected for evolutionary stasis. Error bars for the host-shift sequence represent the 95% CI indicating the uncertainty in the ancestral reconstruction. For PA and PB2, we include triple-reassortant swine sequences (green). 1918 and pandemic H1N1 2009 sequences are labelled. Human sequences inside the distribution of avian sequences represent sporadic H5N1 infections.

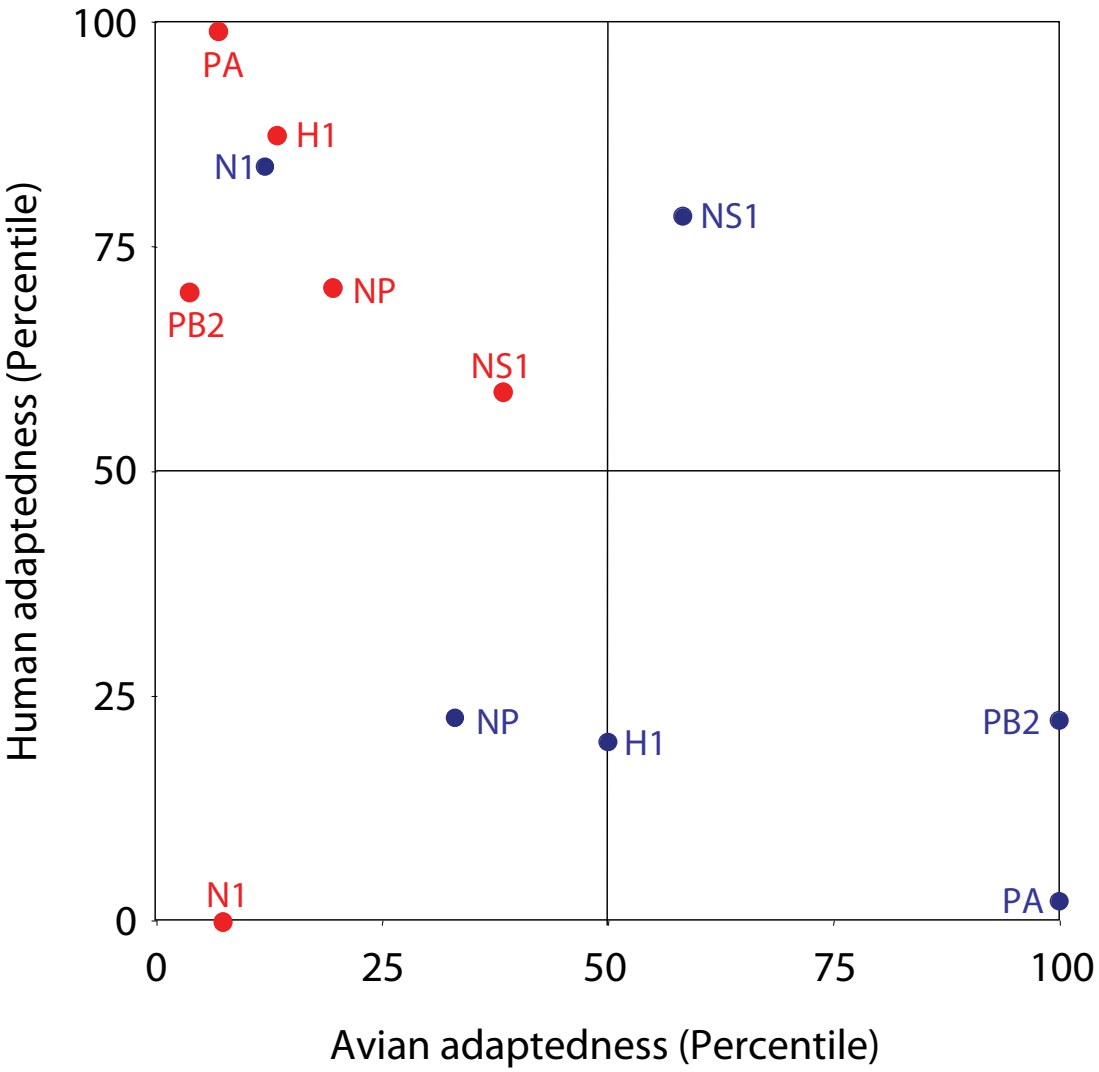


Figure 4.2.: Comparison of various proteins from the pre-1989 host-shift reconstruction and pandemic H1N1 2009 virus with those of the host viruses from which they emerged. Points in blue show the percentage of avian virus protein sequences that have avian and human adaptedness values lower than that of the pre-1918 host-shift reconstruction. Points in red show the percentage of avian (PA and PB2), Eurasian swine (N1) or classical swine (H1, NP and NS1) virus sequences with human or avian adaptedness values lower than the pandemic H1N1 2009 sequences. The human adaptedness values for the pre-1918 host-shift proteins are well within the distribution expected for avian sequences, suggesting that the host-shift virus was not exceptional, whereas the pandemic H1N1 2009 virus proteins, with the exception of N1, have greater than average human adaptedness, indicating pre-adaptation to the new human host.

which the genetic element came (i.e., the human adaptedness and avian adaptedness values for the H1, NS1, and NP proteins are compared with those from classical swine viruses, those for PA and PB2 are compared with avian virus proteins, and N1 is compared with the corresponding protein of Eurasian swine viruses). The pandemic H1N1 2009 virus proteins, with the exception of N1, seem to be more adapted to humans than might be expected. In particular, the human adaptedness of the pandemic H1N1 2009 PA protein is larger than 99% of the corresponding proteins from avian viruses. The N1 protein actually has a lower human adaptedness than the other Eurasian swine N1 proteins, with a human adaptedness value more typical of avian sequences; the latter results from residues V13, A75, and R257, all three of which are rare in human and swine (as well as avian) viruses. The pandemic H1N1 2009 PA and PB2 proteins have high human adaptedness, even relative to the distribution found in the swine triple reassortants. Contributing to this are the PB2 A684S and PA K356R substitutions that have occurred in these two proteins prior to the 2009 pandemic (see page 65).

Figures 4.3 and 4.4 show the changing avian adaptedness and human adaptedness values as a function of isolation year. Waterfowl virus proteins show an average avian adaptedness close to one, agreeing with the notion that waterfowl is the natural reservoir of influenza A. Conversely, human viruses show a trend toward increasing human adaptedness and decreasing avian adaptedness with time of isolation. Interestingly, the 1918 human virus shows intermediate values for both avian adaptedness and human adaptedness, especially for the H1 segment.

Also included in figures 4.3 and 4.4 is a least-squares fit of an exponential to the human adaptedness data for the human virus lineage, performed as described in the Methods section. Fitting parameters are shown in table 4.3. Best fits were obtained with a timescale for adaptation (τ, the time necessary for 63.2% of the adaptation to occur) on the order of 30-70 years, fastest for H1, N1, NP, and PB2 and slowest for NS1. We would expect that

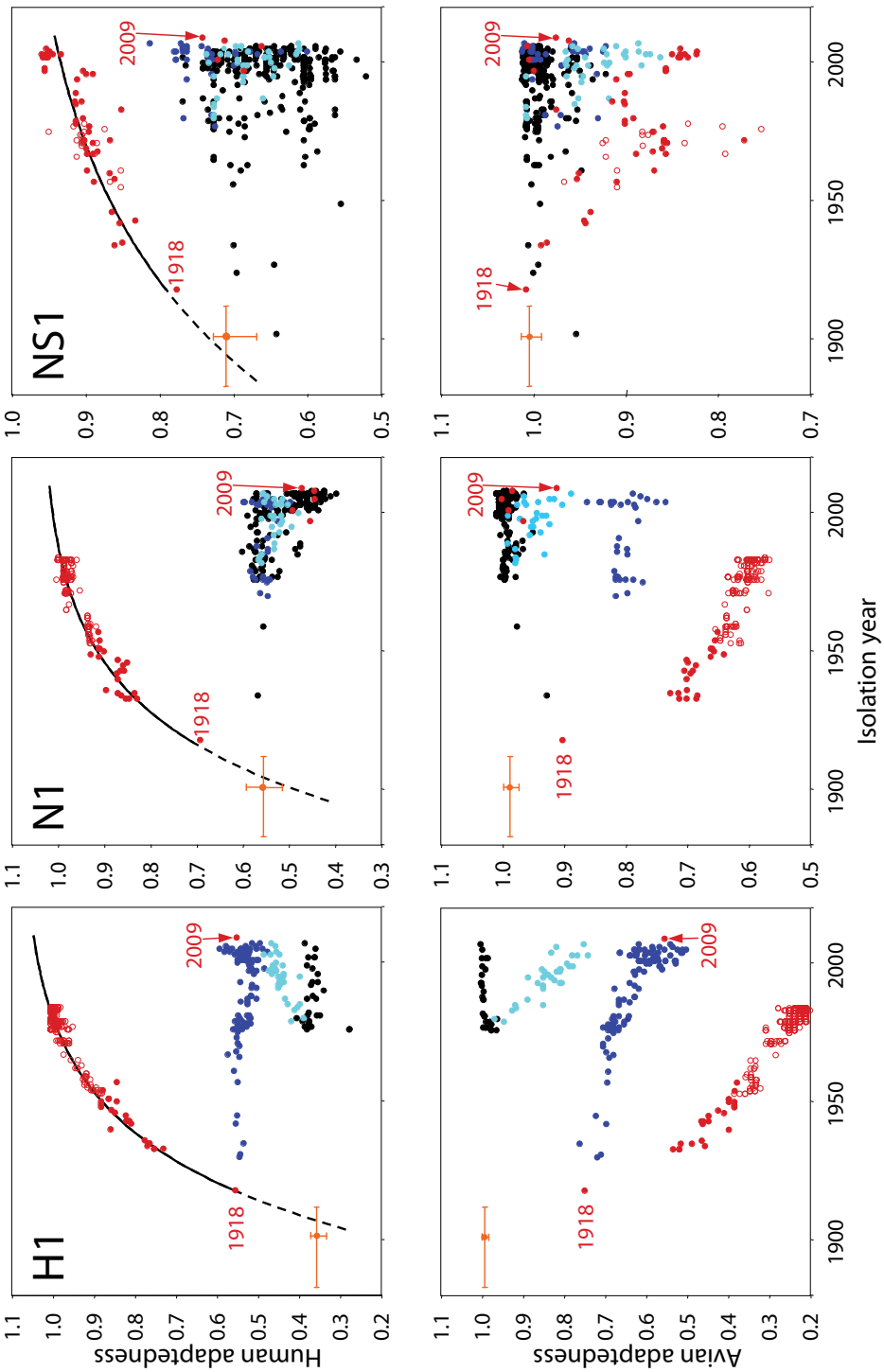


Figure 4.3.: Human and avian adaptedness values for a series of different virus sequences as a function of time. Colour coding is avian (black), human (red), classical swine (blue), Eurasian swine (cyan), triple-reassortant (green) and the host shift sequence (orange). Open red circles represent post-1977 human H1N1 viruses whose isolation times were corrected as described in the text. Human sequences inside the distribution of avian sequences represent sporadic H5N1 infections. Abscissa error bars for the host-shift sequence represent 95% CIs for the timing of this event as determined from an analysis of nucleotide evolution (dos Reis *et al.*, 2009), whereas ordinate error bars indicate the uncertainty of the ancestral reconstruction. Least-squares fits to the human adaptedness of the human virus sequences are included as a solid line, whereas the extrapolation to the host-shift event is shown as a dashed line.

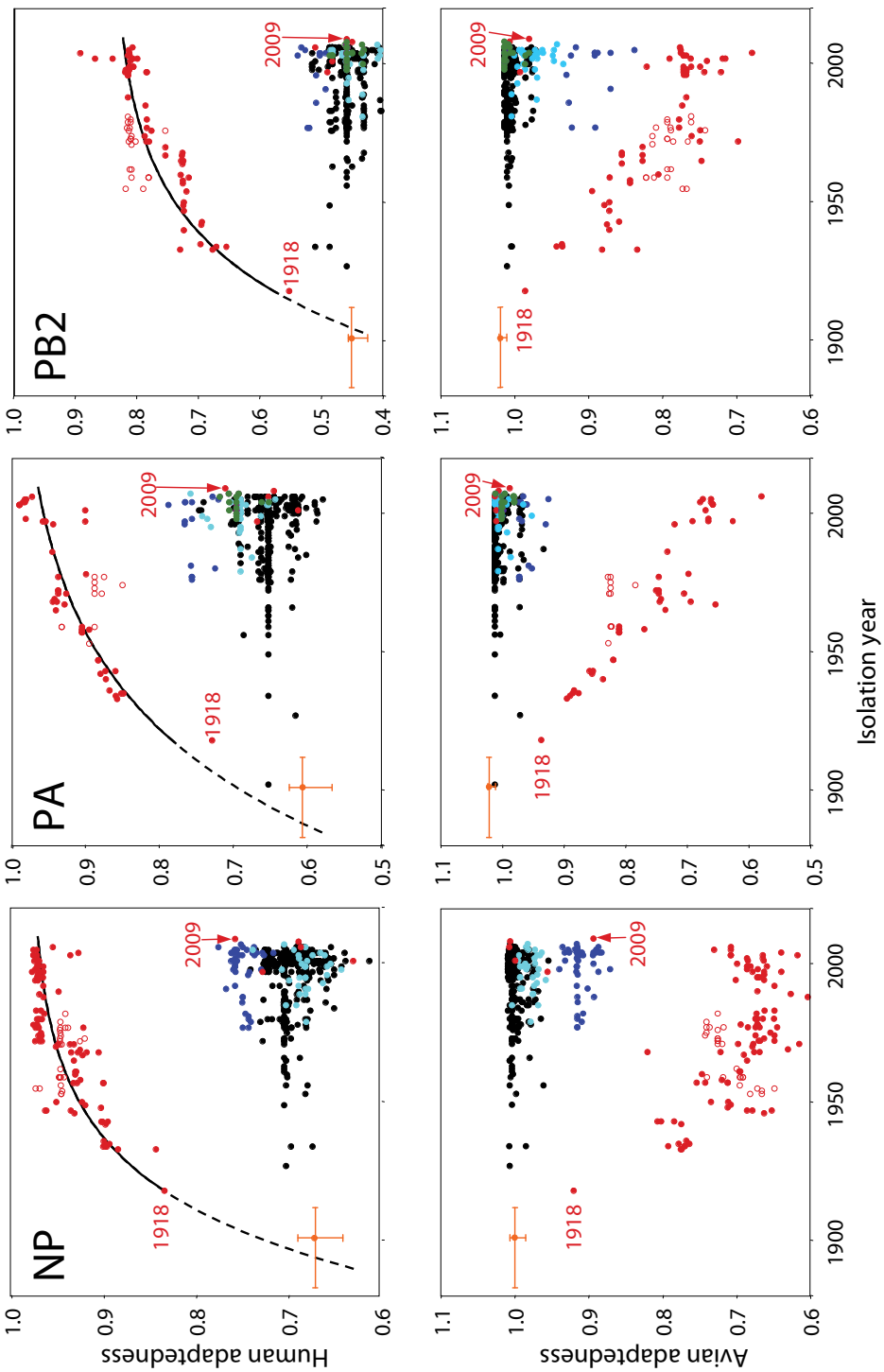


Figure 4.4.: Adaptedness values for proteins as function of isolation time. See Figure 4.3 for details.

Protein	Adaptation Time τ (years)	Equilibrium value (if different from 1.0)	Host-shift year
H1	33.50 (33.21, 35.42)	1.08 (1.08, 1.09)	1907.3 (1906.1, 1907.8)
N1	33.57 (31.10, 35.64)	1.04 (1.03, 1.05)	1905.1 (1903.4, 1906.8)
NS1	71.54 (62.88, 84.03)		1894.6 (1891.5, 1903.5)
NP	31.94 (23.58, 43.29)	0.98 (0.97, 0.99)	1894.9 (1883.0, 1904.9)
PA	50.36 (42.44, 61.76)		1888.2 (1872.5, 1898.0)
PB2	34.15 (24.70, 50.09)	0.84 (0.81, 0.88)	1904.7 (1894.2, 1911.6)

Table 4.3.: Curve-fitting parameters with 95% CI

the asymptotic human adaptedness values for these extrapolations should equal 1.0. In fact, significantly better fits were obtained for four of the proteins when the asymptotic values are larger (H1 and N1) or smaller (NP and PB2) than 1.0. Extrapolation of these fits to the human adaptedness at the host-shift event can provide an estimate of the timing of this host shift. We performed a bootstrap analysis by sampling on the residuals. The estimated host-shift timings are all consistent with previous estimates (1883-1912) based on nucleotide evolution (dos Reis *et al.*, 2009).

In addition to reconstructing the virus at the time of the host shift, we also performed an optimal reconstruction of the various substitutions that occurred in the human lineage following the host-shift event. We separated these into changes that occurred in the "trunk" of the tree connecting the host-shift event directly with recent virus sequences, other interior branches, and exterior branches ending at isolates. As shown in Figure 4.5, we found significant differences in the nature of the sequence changes that occur along these different sets of branches; branches along the trunk of the tree are characterised by a much higher likelihood of an increase in human adaptedness compared with other branches in the tree. This was observed for every gene considered separately.

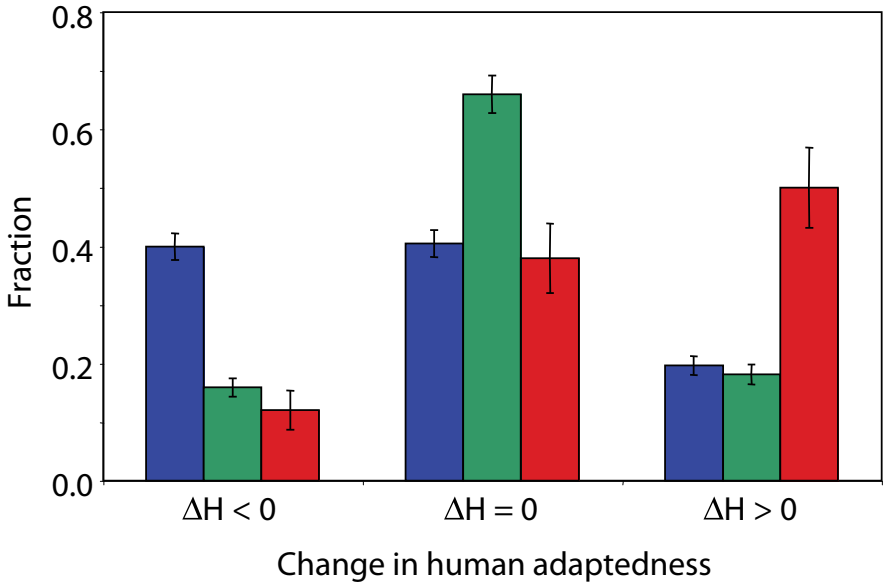


Figure 4.5.: Relative fraction of "trunk" branches (red), other interior branches (green) or exterior branches to isolates (blue) that are characterise by a negative, neutral or positive change in human adaptedness, following the shift from avian-to-human host prior to 1918. Error bars represent standard error based on the number of observations. All genes show a similar distribution.

4.4. Discussion

4.4.1. Properties, limitations and approximations of the model

In the previous chapter, we developed a method for identifying changes in selective constraints acting on influenza virus proteins corresponding to a change in host. We analysed the nature of the substitutions that occur during the evolutionary process and identified when there is statistical support that these substitution patterns are host dependent. In this way, we were able to both identify locations where selective constraints differ and characterise the nature of these differences.

In particular, rather than calculating the observed frequencies of the amino acids found in different positions, our analysis provides the equilibrium amino acid frequencies, given the estimated substitution rates. Observed frequencies are biased by similarities between evolutionarily related viruses and are time dependent as the viruses adapt to the new host following the host-shift event. In contrast, equilibrium frequencies represent the asymptotic value for an ensemble of adapted viruses at equilibrium with the host selective constraints and can be used to describe those constraints. We have used these equilibrium frequencies to develop a measure of how well any virus protein matches the host-specific selective constraints and can compute the corresponding host adaptedness of the viruses to the two hosts. We can then visualise the process of adaptation to the new host following a host shift and provide insight into what might have occurred both prior to and following the host-shift event.

Our evolutionary model assumes that fitness effects at each location are additive and constant within each host only changing at the host shift. Previous work indicates that these assumptions are not strictly valid. Selective constraints can change as the proteins evolve within a host, especially for the HA during changes in antigenic properties (Blackburne

et al., 2008). Adaptation to humans can occur through different sets of substitutions, indicating that the selective constraints at one site are influenced by the amino acids found at other locations. This is clearly seen in HA, where significant differences in structure are reflected in different characteristic substitutions necessary for recognition of receptors on the target human cells (Connor *et al.*, 1994; Matrosovich *et al.*, 2000; Nobusawa *et al.*, 1991; Rogers *et al.*, 1983; Vines *et al.*, 1998). Different substitutions in response to host shifts to human are not confined to these membrane proteins as is clear from considering PB2 627; E627K was experimentally identified as an important substitution necessary for the virus to replicate and spread in mammals (Hatta *et al.*, 2001; Steel *et al.*, 2009; Subbarao *et al.*, 1993; Tarendeau *et al.*, 2008). The pandemic H1N1 2009 virus maintains a glutamic acid at this location, and it appears that a basic amino acid (E) at position 591 compensates for the absence of the basic amino acid at position 627 (Yamada *et al.*, 2010).

Such violations might explain the asymptotic values for the exponential fits to the human adaptedness with isolation time. According to our model, we would expect this asymptotic value to be 1.0, which is the average adaptedness of viruses at equilibrium with the human selective constraints. For four of the proteins, the asymptotic human adaptedness value was not 1.0, suggesting that the selective constraints on the individual locations might be changing either because of changes in the immunity of the host population or because of interaction between the various locations in the protein. Herd immunity dynamics would tend to increase the asymptotic values over 1.0 as there would be a need for the virus to continue to adapt to the new constraints represented by the adapting host immune response. Correspondingly, H1 and N1, the surface glycoproteins most involved in antigenic recognition, have asymptotic values of 1.08 and 1.04, respectively. Conversely, we might expect that there were a number of different ways that a protein could adapt to its host, and adaptation in some locations might lessen the pressure to adapt in others (as in the example of the complementarity of the basic amino acids at positions 591 and 627 of PB2 as mentioned

above) in violation of our assumption of additivity. In this case, we would expect asymptotic values less than 1.0 as is observed for PB2 (0.84) and NP (0.98).

The magnitudes of the changes in host adaptedness are different for the different proteins, representing the variety of degrees of difference in selective constraints in the two hosts. Locations that undergo a relaxation in selective constraints during the host shift to humans will have a relatively small change in human adaptedness (avian virus sequences are compatible with the human constraints) but a larger change in avian adaptedness (many human viruses will not be compatible with the avian constraints). The opposite relationship would hold for a tightening of selective constraints. The amount of scatter in host adaptedness values for the various proteins mostly reflects the number of significant sites considered, which range from 27 sites in PA to 84 locations in H1.

Our exponential fit to human adaptedness, extrapolated to the host-shift event, is in rough agreement with the estimate of 1883-1912 obtained through the analysis of nucleotide composition changes (dos Reis *et al.*, 2009). These extrapolated values, however, should be treated with caution as they assume that adaptation to the human host occurred in a similar manner prior to and following 1918. If the intermediate host prior to the 1918 pandemic was swine, it is likely that the rate of adaptation was slower before 1918 and the host shift occurred earlier than indicated by the extrapolations. The extrapolation also assumes that the functional form of the adaptation process is correct and that the changing human adaptedness can be represented by an exponential with a single timescale. It might be conjectured that the adaptation was faster immediately following the host shift, suggesting a more recent event. This can be modelled as a mixture of exponentials with different adaptation times; the locations with the shortest adaptation times would equilibrate fastest, leaving locations with longer adaptation times to equilibrate longer after the host-shift event. To test this possibility, the human adaptedness data were fit to an ensemble of exponentials with a Gaussian distribution of adaptation rates. This more complicated model could not be justified by the

data but this does not indicate that some mixture of substitution rates would not give an improved fit.

It is clear that the mathematical model developed here still leaves much unknown about evolution of influenza and host shifts. Our current model should be considered as a basic framework onto which more complete models can be developed. Particularly, modelling variation in selective constraints along time and within hosts could provide a better understanding of the adaptation process. Our assumption of additiveness can also be relaxed, and models that consider interactions among locations could be developed.

4.4.2. How typical was the host shift virus?

It is not clear why a particular virus undergoes a host-shift event. One possibility is that chance mutations result in a "pre-adapted" virus particularly fit for the new host prior to the host transfer event. The other possibility is that the virus is not distinctive, and the host transfer of a particular virus is simply a chance occurrence. The answer to this question has important consequences for our ability to characterise the pandemic potential of zoonotic viruses. To distinguish between these two possibilities, we reconstructed the ancestral sequence of the virus that underwent the shift to humans prior to the 1918 pandemic as well as analysing the 2009 pandemic virus.

We observed that the avian-like pre-1918 host-shift virus, as best shown in Figure 4.2, has human adaptedness values within the distribution of what would be expected for an avian virus, which suggests that the identity of the virus that underwent the host-shift event was a matter of opportunity. In contrast, the pandemic H1N1 2009 virus proteins, with the exception of N1, were more adapted to humans than would be expected, given their origin. The most interesting examples of such pre-adaptation are in PB2 and PA; in both proteins, there was an initial host shift from birds to swine, presumably around 1998, followed by the

host shift to humans in 2009. While circulating in swine, both experienced substitutions identified with increasing human adaptedness (e.g. PB2 A684S and PA K356R) prior to the shift to humans. The resulting increase in human adaptedness for PA is especially large as there are comparatively fewer host-specific locations in this protein compared with PB2. N1 of the 2009 pandemic virus was not as well adapted to humans as N1 from other Eurasian influenza viruses, although it is about as well adapted as a typical avian virus. The relatively lower adaptedness for this particular gene may represent a random fluctuation that is compensated for by the greater adaptedness of the other genes.

4.4.3. Changing adaptedness in the phylogenetic tree

We note that adaptation to the new host has occurred preferentially along the "trunk" of the phylogenetic tree, as noted previously (Bush *et al.*, 1999; Nelson & Holmes, 2007), whereas other branches where the adaptation does not occur as quickly tend to represent evolutionary "dead ends." This would be expected if such sequence changes increase the fitness of these sequences in the new host relative to those viruses experiencing alternative substitutions. This points to the possibility that measures, such as human adaptedness, can be used to provide insight into why certain lineages persisted and others did not.

4.4.4. Ancestral reconstruction methods

Analyses of both the host-shift viruses and the changes along the tree required reconstruction of the evolutionary trajectories. We used marginal reconstruction for the ancestral sequences (Koshi & Goldstein, 1996) and joint reconstruction (Pupko *et al.*, 2000) for the historical changes.

The reconstruction of the ancestral sequence relies on an accurate model of the substitution process, which we observe to depend upon the host, especially for the locations un-

der consideration here. The use of host-specific substitution models is especially important for examining the evidence for pre-adaptation in the host-shift virus as some changes that might reflect the adaptation of the virus to the new host may, with an inappropriate host-independent evolutionary model, appear to be prior to the host shift. We were specifically interested in identifying evidence for pre-adaptation that cannot be explained by such changes in selective constraints, which required the use of host-dependent models and the exclusion of viruses from other than avian and human hosts. Although it is standard, especially for experimental work, to consider the most likely sequence, we generated an ensemble of sequences by sampling from the posterior probabilities of the reconstruction, allowing us to determine unbiased statistical properties of this ensemble (Williams *et al.*, 2006). We recreated an ensemble of sequences representing the virus at the point of host transfer. In this way, we were able to obtain the mean and CIs for the human adaptedness and avian adaptedness at this point.

More accurate ancestral reconstruction could have been achieved by modelling selective constraints in swine. Identification of three sets of selective constraints per location provides computational and statistical challenges. Particularly, with three sets of constraints, alternative models are not nested, and the likelihood ratio test cannot be used. For this reason, in our joint reconstruction, we used a more standard method with substitution models that did not depend on either host or location.

4.4.5. The history of the 1918 pandemic

As is clear in Figures 4.1, 4.3 and 4.4, significant adaptation to human selective constraints had occurred prior to the 1918 pandemic. This is in seeming contrast to the conclusions made by Taubenberger *et al.* (2005), who concluded that the 1918 virus sequences more closely resemble avian than human virus sequences. The difference in conclusions between

earlier work and this work can be explained by a difference in focus; previous work considered all the amino acid changes that had occurred in the virus proteins, whereas our methods allow us to focus on locations involved in host adaptation.

The degree of human adaptation prior to the 1918 pandemic can be explained in three ways: (a) The virus had "pre-adapted" to humans in its avian host, presumably as a result of stochastic fluctuations, perhaps explaining why that particular virus was able to establish itself so readily in humans; (b) the virus had evolved in humans for a period of time prior to 1918; or (c) the virus had evolved in a non-human non-avian host that exerted similar selective pressure on the virus as exerted by a human host. (a) seems unlikely as the human adaptedness values of the 1918 virus are well outside the range of observed avian viruses. In addition, our reconstruction of the sequence of the virus at the host-shift event shows that the host-shift proteins were avian like in their human adaptedness, suggesting that there was little evidence of pre-adaptation. Although we cannot rule out the possibility that the 1918 pandemic virus evolved in humans for a significant period of time prior to the subsequent pandemic, the similarity of avian and porcine cell receptors, the observed successful avian-to-swine host shift in 1979 compared with the lack of precedent for a successful avian-to-human shift, and the difficulty in the virus existing undetected for so long in the human population argue for swine as an intermediate host (dos Reis *et al.*, 2009; Scholtissek, 2008; Smith *et al.*, 2009a).

Adaptation to humans during virus evolution in swine is possible if there are similarities in the selective constraints imposed on viruses in these two species. In fact, human adaptedness values for H1, NP, PA, and PB2 are higher in the classical swine lineage than in avian isolates. The increasing human adaptedness of the Eurasian swine H1 after the initial host shift in 1979 is clear in Figures 4.3 and 4.4. If the evolution of the human virus prior to 1918 occurred mostly in swine, we would expect the human adaptedness values for the 1918 human virus to resemble the human adaptedness values of classical swine. This is true

for most proteins, although the 1918 virus N1 and NP proteins have significantly higher human adaptedness than is observed in later classical swine viruses. Resolution of this issue will require greater availability of early influenza viruses or more sophisticated evolutionary models. We also note that the 2009 virus seems to have pre-adapted to humans during its circulation in swine. This again highlights the ability of swine to preadapt viruses to human hosts, suggesting a potentially similar role for swine in facilitating the 1918 and 2009 human pandemics.

The results described above seem to suggest that, although the virus that underwent the first host-shift event from birds to mammals before the 1918 pandemic seems unexceptional, the virus had substantially adapted to humans prior to the subsequent pandemic. Similarly, we can detect substantial adaptation to humans in five of the virus genes in the triple reassortant prior to the 2009 pandemic. Although the causes of a pandemic are complex, involving a mixture of virus properties, host susceptibilities, and historical contingencies, these results indicate that the degree of human adaptation of the virus plays an important role in host shifts to humans.

5. Estimating the distribution of selection coefficients using mutation-selection models

5.1. Introduction

The last two chapters have demonstrated the use of probabilistic models of sequence change to answer questions regarding the evolution of influenza. In this chapter we introduce a model to estimate the distribution of selection coefficients. The model describes the evolutionary process using a more mechanistic model of evolution, combining the effects of nucleotide change, fitness of amino acids and probability of fixation of a mutant in a population.

When a novel mutation appears in the genome of an organism it may have three different effects on the fitness ($w = 1 + s$) of its carrier: The mutation may be deleterious ($s < 0$), reducing fitness through reduced fertility or survival rate. It may be neutral ($s \approx 0$), that is, having such a small effect on fitness that the fate of the mutant is mostly determined by random drift. Or the mutation may be advantageous ($s > 0$), increasing the fitness of its carrier by increasing its fertility or survival in its environment. The frequency distribution of the different types of mutants and their associated selection coefficients (s, also known as fitness effects) is a key issue in population genetics (Bustamante, 2005; Eyre-Walker & Keightley, 2007). The ultimate fate of a mutation, whether it will become fixed or lost in a population, depends on the strength of selection and on the effect of random drift due to finite population size. In fact, the fitness effect s and the population number N are so closely linked that normally the distribution is expressed in terms of the population scaled

coefficient $S = 2Ns$.

Kimura (1968, 1983), in his neutral theory of molecular evolution, proposed that the dominant fraction (p_-) of all novel mutations would be highly deleterious, with a minority fraction ($p_0 = 1 - p_-$) being neutral. When organisms colonise a new habitat or are subject to environmental change, the opportunity for adaptive evolution would arise, and a fraction ($p_+ = 1 - p_0 - p_-$) of novel mutations would be advantageous. The magnitudes of these fractions for a protein coding gene would depend on the protein in question; functionally important or structurally constrained proteins (such as the histones) would be characterised by a very large fraction of deleterious mutations ($p_- \gg p_0$), while structurally less constrained proteins (such as the fibrinopeptides) would have a larger fraction of neutral mutations ($p_0 > p_-$). Extensions to Kimura's theory have been made, including considering the contribution of nearly neutral mutations to the evolutionary process (Kimura, 1983; Ohta, 1973, 1992). Under this latter extension, there is a spectrum of nearly neutral mutations ranging from slightly deleterious to slightly advantageous, with the neutrality of a given change dependent on the population size; evolutionary trajectories consist of a balance between slightly deleterious and slightly advantageous substitutions. Others have argued that, even under more typical conditions, adaptive substitutions would be frequent, the greater probability of fixation compensating for their relative rarity among mutations (Gillespie, 1994).

Akashi (1999) considered that under a neutral model the distribution of S among novel mutations could be bimodal, with the modes centred around highly deleterious and neutral mutations. During adaptive episodes, the distribution would have three modes, with a small additional mode centred around advantageous mutations. Because deleterious mutations have a vanishingly small probability of becoming fixed in a population, most substitutions (i.e. fixed mutations) would be neutral. In this case, the distribution of S among substitutions would be unimodal and centred around neutral mutations. During an adaptive epis-

ode, natural selection would drive many positively selected mutations quickly to fixation. In this case, the distribution of substitutions would be bimodal, with modes centred around nearly neutral and advantageous substitutions. (See Appendix H for figure.)

While the effect of mutations can be studied experimentally, these studies are difficult to perform on higher organisms and too insensitive to observe any but the largest fitness effects (Eyre-Walker & Keightley, 2007). Due to these limitations, alternative approaches have been developed that estimate the distribution of fitness effects from biological sequence data. Much of the work on estimation of the distribution of S from DNA sequence data has been based at the population level (e.g. Bustamante *et al.*, 2002; Sawyer & Hartl, 1992). These methods usually work with allele data from different individuals within a population, and the level of polymorphism within the population and the number of fixed differences with an outgroup species are used to estimate the distribution. These methods look at the evolutionary process over relatively short periods of time, and thus normally use approximate mutation models such as the infinite alleles model (Kimura, 1969, 1983, p. 43). More recently, phylogenetic methods that look at the evolutionary process over longer periods of time have been used to estimate the distribution of selection coefficients (Nielsen & Yang, 2003; Rodrigue *et al.*, 2010; Yang & Nielsen, 2008). Although these use more realistic mutation models than the population based methods, they ignore polymorphism and assume that all the observed differences among species are fixed. These two approaches sometimes result in different conclusions; population based methods can yield an extremely large fraction of adaptive changes (Fay *et al.*, 2001), especially in Drosophila (Sawyer *et al.*, 2003, 2007), while phylogenetic methods often results in more modest estimates of p_+ (Nielsen & Yang, 2003; Rodrigue *et al.*, 2010). Similarly, population methods find the distribution of slightly deleterious mutations falling off leptokurtically, that is, more rapidly than exponentially (such as in a gamma distribution with $\alpha < 1$) (Eyre-Walker, 2006), while evolutionary models often yield a more rounded distribution ($\alpha > 1$) (Nielsen & Yang, 2003; Rodrigue

et al., 2010). It is not clear if these differences represent the different methodologies and the approximations that they make, or on the details of the particular organisms under study. Worryingly, the evolutionary models fail to yield a substantial amount of lethal mutations (Nielsen & Yang, 2003; Rodrigue *et al.*, 2010) that would be expected based on mutation experiments (Hietpas *et al.*, 2011; Sanjuan *et al.*, 2004; Wloch *et al.*, 2001) and have been obtained by population-based studies (Eyre-Walker, 2006; Piganeau & Eyre-Walker, 2003; Yampolsky *et al.*, 2005).

One of the difficulties in estimating the distribution of selection coefficients is the complex nature of the selective constraints, even within a single protein, representing a range of functional, structural and physiological requirements. Certain locations, such as those involved in protein functionality, may be invariant, while other locations may have a wide latitude in the amino acids compatible with that position. It is not only the magnitude of the selective constraints that vary from one location to another; one position may be constrained to hydrophobic residues, another to residues that can take part in hydrogen bonding interactions, a third requiring a certain degree of flexibility. The types of substitutions that can occur can be substantially different, even among locations that are changing at similar rates. Different approaches have addressed this issue to various degrees. For instance, Nielsen & Yang (2003) considered that the overall rate of substitutions could vary from one location to another, but considered that this rate variation would affect all possible substitutions equally; that is, slowly-varying locations were as unrestricted in the amino acids as rapidly-varying locations. Thorne *et al.* (2007) relaxed the standard assumption of independent sites, considering the selective constraints imposed by the need to maintain a stable well-defined structure; this was estimated using protein structure prediction algorithms, despite their construction being motivated by a quite different problem. Rodrigue *et al.* (2010) adapted a mixture-model approach that group locations under similar selective constraints, and developed more specific models for characterising these different types of locations;

each individual location was then represented by a mixture of these models (Koshi & Goldstein, 1998). The available data determined the number of components in the mixture that could be justified.

The most specific characterisation of the substitution process was developed by Halpern & Bruno (1998), who proposed a site-wise phylogenetic model where evolution at each amino acid residue in a protein is characterised by a location specific set of fitnesses and by the nucleotide level mutation pattern. Although Halpern and Bruno demonstrated its utility for the estimation of evolutionary distances, use of the model has been limited, as the number of adjustable parameters required more data and computational resources than have previously been available. Here we explore the use of this model in the estimation of the distribution of S. We are interested in assessing how the assumption of site specific fitnesses may affect estimates of the shape of the distribution of S among novel mutations and substitutions. We apply a modified version of their model to a data set of 12 mitochondrial proteins in 244 mammalian species. We also apply this model to a data set of a polymerase protein from 401 influenza viruses isolated from avian and human hosts. As the human viruses are the product of a host shift event from an avian host (Taubenberger *et al.*, 2005), this allows us to investigate the distribution of selection coefficients during a well defined adaptive episode.

5.2. Methods

In the following discussion we assume a Wright-Fisher model of random genetic drift (e.g. Wright, 1931). We work with idealised populations where the effective and the real population numbers are the same. Locations in a gene are assumed to evolve independently, and they do not interfere with each other. We assume the selection coefficients (s) involved in the model are small, so that simplifying approximations about relative fixation probabilities can be made. It is also assumed that mutation rates are sufficiently small in relation to

the population size so that polymorphism is negligible and locations remain fixed most of the time (Crow & Kimura, 1970, p. 442–445). The evolutionary process is viewed over long periods so the time from appearance to fixation of a novel mutant is nearly instantaneous. These assumptions are necessary to simplify the mathematical treatment of the model as discussed below.

5.2.1. Basic model

We model the substitution rate of a codon location in a functional protein under the action of selection, mutation and random drift as a time continuous Markov process. We modify the model of Halpern & Bruno (1998), and we use the notation of Yang & Nielsen (2008). Let us write $I = i_1 i_2 i_3$ and $J = j_1 j_2 j_3$ for any two codons ($I \neq J$) where i_k is the nucleotide at the k-th position of I. The Malthusian fitness of codon I at location K of the gene is $f_{I,K}$, so the selection coefficient for a mutant that transforms I into J is $s_{IJ,K} = f_{J,K} - f_{I,K}$. Assuming that the population size remains constant in all lineages, we write $S_{IJ,K} = F_{J,K} - F_{I,K} = 2N(f_{J,K} - f_{I,K})$ for the scaled selection coefficient, where N is the effective chromosomal number and $F_{I,K}$ is the scaled fitness. The substitution rate from I to J ($I \neq J$) at the location is

$$q_{IJ,K} = \begin{cases} \mu_{IJ} \frac{S_{IJ,K}}{1-\mathrm{e}^{-S_{IJ,K}}} & \text{if } S_{IJ,K} \neq 0 \\ \mu_{IJ} & \text{else} \end{cases}, \tag{5.1}$$

where μ_{IJ} is the neutral mutation rate, and $S/(1-\mathrm{e}^{-S})$ is the relative fixation probability of a selected mutation compared with a neutral one (Kimura, 1983, eq. 3.14). If the mutation is advantageous ($S_{IJ} > 0$) then $q_{IJ} > \mu_{IJ}$, and if the mutation is deleterious ($S_{IJ} < 0$) then $q_{IJ} < \mu_{IJ}$. Thus the effect of natural selection is to accelerate or reduce the rate of substitution compared to the neutral mutation rate. The $q_{IJ,K}$ form the off-diagonal elements of a 64×64 rate matrix ($\mathbf{Q}$) whose diagonal elements are $q_{II,K} = -\sum_{J \neq I} q_{IJ,K}$.

The selection coefficients ($S_{IJ,K}$) describe the effect of selection on the amino acid at a given location K, due to the protein's structure and function of the protein. This contrasts with the model of Yang & Nielsen (2008) where the nonsynonymous to synonymous substitution rate ratio, ω, is used to account for the effect of selection at the protein level. When modelling site specific selection, the inclusion of ω is unnecessary.

The location specific fitnesses ($F_{I,K}$) can be modelled at the amino acid or codon levels. We can write $F_{J,K} = F_J^{\text{co}} + F_{J,K}^{\text{aa}}$, where F_J^{co} is the fitness of J due to the effect of selection on codon bias (e.g. Bulmer, 1991) and $F_{J,K}^{\text{aa}}$ for the fitness of the particular amino acid at the location. In this study, we assume that the selective constraints are dominated by selection on the amino acid, and ignore the effect of selection on codon bias. Under this assumption $F_{J,K} = F_{J,K}^{\text{aa}}$.

Mutation at the nucleotide level

Consider a cycle of DNA replication occurring in a tiny time interval τ. The probability of observing a particular nucleotide i mutating into j ($i \neq j$) during interval τ is $p_{ij}(\tau) \simeq g_{ij}\tau$, where $g_{ij}(\geq 0)$ is the rate of change $i \to j$ per time unit. The probability that i will remain unchanged is $p_{ii}(\tau) \simeq 1 + g_{ii}\tau$, where $g_{ii} = -\sum_i g_{ij}$. Note that we are modelling the mutation of DNA *before* natural selection takes place. The probability that a triplet I of nucleotides will change into triplet J ($I \neq J$) is $p_{IJ}(\tau) = \prod_k p_{i_k j_k}(\tau) \simeq \mu_{IJ}\tau$. Because the time interval τ is very small, $p_{ii}(\tau) \approx 1$, so we can ignore these probabilities in the product term and then solve for the mutation rate μ_{IJ} to get

$$\mu_{IJ} \approx \frac{\prod_{k,i_k \neq j_k} p_{i_k j_k}(\tau)}{\tau} = \frac{\prod_{k,i_k \neq j_k} \tau g_{i_k j_k}}{\tau} = \tau^{n-1} \times \prod_{k,i_k \neq j_k} g_{i_k j_k}, \tag{5.2}$$

where n is the number of changing nucleotides.

The rate constants g_{ij} can be defined under any nucleotide substitution model (e.g. Yang,

1994a). Here we use the HKY85 model (Hasegawa *et al.*, 1985), where $g_{ij} = v\kappa\pi^*_j$ for transitions and $g_{ij} = v\pi^*_j$ for transversions, $\pi^*_j (\geq 0,\ \sum_j \pi^*_j = 1)$ is the equilibrium frequency of nucleotide j (achieved under no selection), κ is the transition-transversion rate parameter and v is a scaling constant. The mutation rate $I \rightarrow J$ is thus

$$\mu_{IJ} = \tau^{n-1} v^n \kappa^{n_t} \prod_{k, i_k \neq j_k} \pi^*_{j_k}, \tag{5.3}$$

where n_t is the number of nucleotide transitions necessary to go from I to J. For example, if codons I and J differ by a single transversion, then $\mu_{IJ} = v\pi^*_{j_k}$ while if they differ by two transitions at positions k and l then $\mu_{IJ} = v\tau\kappa^2\pi^*_{j_k}\pi^*_{j_l}$. We can now combine equations (5.3) and (5.1) to get

$$q_{IJ,K} = \left(\tau^{n-1} v^n \kappa^{n_t} \prod_{k, i_k \neq j_k} \pi^*_{j_k} \right) \times \frac{S_{IJ,K}}{1 - \mathrm{e}^{-S_{IJ,K}}}. \tag{5.4}$$

Parameter τ controls the rate at which multiple simultaneous nucleotide substitutions are allowed to occur in $I \rightarrow J$. For example, if $\tau = 10^{-1}$ then triple substitutions occur at a rate in the order of 10^{-2} compared to single substitutions. If $\tau = 0$, simultaneous substitutions are not allowed and equation (5.4) reduces to a site-wise version of equation (2) in Yang & Nielsen (2008). This multiple substitutions model contrasts with that of Halpern & Bruno (1998), which is based on the probability of observing a random mutation in a nucleotide sequence at equilibrium.

We scale the substitution rates based on the expected number of *neutral* mutations per site (Halpern & Bruno, 1998). When there is no selection acting on the sequence, the neutral substitution rate is simply $q^0_{IJ} = \mu_{IJ}$ $(I \neq J)$, and the expected equilibrium frequency of J is $\pi^0_J = \pi^*_{j_1}\pi^*_{j_2}\pi^*_{j_3}$. We thus set $v = 1/\sum_{I=1}^{64}\sum_{J=1}^{64}\pi^0_I\mu_{IJ}$ $(I \neq J)$ so that the expected number of neutral substitutions per codon location is one (i.e. $-\sum_I \pi^0_I q^0_{II} = 1$).

Equation (5.4) describes a reversible process at the codon level. The proof of reversibility

can be obtained by the same argument of Yang & Nielsen (2008) and it will not be shown here. We note that the equilibrium frequency of codon J at location K is

$$\pi_{J,K} = \pi^*_{j_1}\pi^*_{j_2}\pi^*_{j_3}\mathrm{e}^{F_{J,K}}/z, \tag{5.5}$$

where $z = \sum_{J=1}^{64}\pi^*_{j_1}\pi^*_{j_2}\pi^*_{j_3}\mathrm{e}^{F_{J,K}}$.

Maximum likelihood estimation

Equation 5.4 can be used to construct the transition probability matrix $\mathbf{P}_t = p_{IJ}(t) = \mathrm{e}^{t\mathbf{Q}}$ that gives the probability of change $I \to J$ after time t. This matrix can then be used to calculate the likelihood of a sequence alignment under a fixed tree topology using established procedures (Felsenstein, 1981; Yang, 2006). The value of the model parameters that maximise the log-likelihood (ℓ) can then be found by numerical optimisation.

Estimation of branch lengths: Estimation of branch lengths by maximum likelihood is computationally expensive. We estimate individual branch lengths using faster codon based methods (e.g. Yang & Nielsen, 2008), and the estimated tree with fixed topology is then used in the likelihood calculation of the model. During the calculation, the branch lengths are multiplied by a constant c and the value of this constant is chosen as to maximise the likelihood. Therefore, the final tree has branch lengths as expected number of *neutral* substitutions per site, that is, the number of substitutions that would have accumulated if the sequence was a pseudo gene. We can convert the branch lengths to expected number of substitutions per codon in the following manner: for location K, the expected substitution rate at equilibrium is $\lambda_K = -\sum_I \pi_{I,K}q_{II,K}$. The average substitution rate for the whole sequence is $\bar{\lambda} = \sum_K\sum_I \pi_{I,K}q_{II,K}/L_c$. For a pseudo gene $\bar{\lambda} = 1$, while a gene under purifying selection would have $\bar{\lambda} < 1$. For a branch of length b neutral substitutions per site, $\bar{\lambda}b$ represents the usual substitutions per codon.

Adjustable parameters: One of the mutational bias parameters (π_j^*) is redundant as the nucleotide equilibrium frequencies obey the constraint $\sum_{j=1}^{4} \pi_j^* = 1$. Similarly, only relative values of the fitness parameters ($F_{I,K}$) matter, so for each location, one of the fitness parameters can be set equal to zero. For a coding sequence with L_c codon locations, the model has 6 mutation parameters (τ, κ, c and 3 π_j^*) and $(20-1)\,L_c$ values of $F_{I,K}$. Information from all codon locations is used by the likelihood method to estimate the value of the 6 mutation parameters. The variance of these parameters decreases with increased sequence length. The amino acid fitnesses are location specific and can only be reliably estimated for alignments of many sequences under reasonable levels of divergence (see below).

Unobserved amino acids: Only a few amino acid types are usually seen within a given alignment location. For a codon J coding for an unobserved amino acids, the MLE estimate of $F_{J,K}$ tends to $-\infty$ (Exceptions may exist when an unobserved amino acid may help facilitate substitutions between observed amino acids, such as pairs that cannot be connected by a single base change.) Because $S_{IJ,K}/(1-\mathrm{e}^{-S_{IJ,K}}) \to 0$ if $S_{IJ,K} \to -\infty$, then the corresponding columns of the rate matrix are zero. Rather than estimating $F_{J,K}$ for these unobserved amino acids, it is possible to fix these values to $-\infty$ and collapse the rate matrix accordingly. For example, for a location where only two amino acids encoded by two codons each are observed, the corresponding rate matrix would be of size 4×4 and only 2 − 1 amino acid fitness parameters would be found by numerical optimisation (Holder *et al.*, 2008). This approximation greatly reduces computing time.

Distribution of selection coefficients

We calculate the distribution of selection coefficients among novel mutations and among substitutions. At equilibrium, the proportion of expected mutations with a given value of S

among all mutants at all locations is

$$m^0(S) = \frac{\sum_K \sum_{I \neq J} \pi_{I,K}\, \mu_{IJ}\, \delta\left(S - S_{IJ,K}\right)}{\sum_K \sum_{I \neq J} \pi_{I,K}\, \mu_{IJ}}. \tag{5.6}$$

where $\delta\left(S - S_{IJ,K}\right) = 1$ if $S - S_{IJ,K} = 0$ and $= 0$ otherwise. The proportion among substitutions is

$$m(S) = \frac{\sum_K \sum_{I \neq J} \pi_{I,K}\, q_{IJ,K}\, \delta\left(S - S_{IJ,K}\right)}{\sum_K \sum_{I \neq J} \pi_{I,K}\, q_{IJ,K}}. \tag{5.7}$$

Note that the scaled Malthusian fitness, $F = 2Nf$, is related to the Darwinian fitness, w, by $w = \mathrm{e}^{F/2N}$ (Crow & Kimura, 1970, p. 8). For the wild type $w = 1$ and $F = 0$, and for a lethal mutant, $w = 0$ and $F = -\infty$; this means that the distribution of selection coefficients ranges from $-\infty$ to ∞. In experimental studies Darwinian fitnesses are normally used, and the distribution of fitness effects ranges from -1 to ∞. For $S \ll N$, selection coefficients and fitness effects are nearly identical.

Following Li (1978) we define an $I \to J$ mutation (or substitution) as deleterious if $S_{IJ,K} < -2$, as nearly neutral if $-2 < S_{IJ,K} < 2$, and as advantageous if $2 < S_{IJ,K}$. The proportions of the three type of mutations are p_-, p_0 and p_+ respectively. For example, the proportion of advantageous mutations among all substitutions is

$$p_+ = \int_2^\infty m(S)\, \mathrm{d}S. \tag{5.8}$$

The uncertainty in the estimation of the distribution of S can be assessed by classical and parametric bootstrapping. In the classical bootstrap, we sample locations at random (with replacement) from the alignment and then we recalculate the distribution using equations (5.6) and (5.7) in order to generate confidence intervals. In the parametric bootstrap, synthetic data are generated using the ML estimates from the real data set, and then all parameters are re-estimated for the synthetic data using exactly the same procedure as for the

real data (including estimation of the tree topology, branch lengths, global mutation parameters and fitnesses). When the parametric model offers an adequate description of the real data, both the classical and parametric bootstrap lead to similar results (Felsenstein, 2003, ch. 20).

5.2.2. Software implementation

The software implementation of the model is able to utilise multicore and distributed architectures, making the estimation of global and site-specific parameters computationally tractable (see Appendix K for a tutorial). Estimation of the parameters is done in three steps. As our program does not perform branch length estimation, we first optimise branch lengths under one of the codon substitution models available in other software. We use the FMutSel0 model in the program CODEML (PAML package, (Yang, 2007a; Yang & Nielsen, 2008)) using the branch-by-branch optimisation option and empirical codon frequencies (method=1 and estFreq=0 in the CODEML control file). Second, we use our program to estimate the global parameters (π^*, κ, τ and c) using the approximate method described above, where the substitution rate matrix is collapsed by neglecting all unobserved residues. The MLEs of π^* and κ estimated by CODEML in the first step are used as starting values for the site-wise model in the second step. In the third step, the global parameters are fixed, but all fitnesses are re-estimated, this time relaxing the assumption that $F_I = -\infty$ for unobserved amino acids. Likelihood calculation is thus performed using the full 64×64 substitution matrix. The fitness parameter of the most common amino acid at each location is fixed to $F_{I,L} = 0$, while the other fitness parameters are limited to $-20 < F < 20$. The $\mathbf{F}$ are estimated three times for each site, once with all $F = 0$, once with F of unobserved residues set to -20, and once using random starting values (between -3 and 3). We used the MLE with the highest likelihood.

5.3. Results and discussion

5.3.1. Statistical properties of the model

Consistency and normality of fitness estimators

In the general case, the likelihood function L, the probability of observing data $\mathbf{x} = (x_i)$ given parameters $\theta = (\theta_i)$, is given by the joint density $L(\mathbf{x}|\theta) = f(x_1, x_2, \ldots, x_n|\theta)$. When the data (x_i) are independent (and under other regularity conditions), the maximum likelihood estimator of θ, $\hat{\theta}$, is shown to be consistent and asymptotically normally distributed (e.g. Stuart *et al.*, 1999, ch. 18). When the data are not independent, consistency and asymptotic normality may not be guaranteed. Estimation of the fitnesses for a particular location K, when the global parameters (τ, κ, c and π^*) are known, proceeds by maximising a joint likelihood function $L(\mathbf{x}|\theta) = f(x_1, x_2, \ldots, x_n|\theta)$ where the x_i represents the observed codon in species i for the given location. These data are not independent (they are correlated according to the underlying tree structure) and the asymptotic properties of the fitness estimators are unclear.

We can employ Monte Carlo simulations to investigate the asymptotic properties of fitness estimators when the number of species sampled is increased. We follow two simple simulation strategies. For the fixed height tree (FHT) approach, we started with a rooted 64-taxa symmetric tree with branch lengths {7.5, 3.75, 1.875, 0.9375, 0.46875, 0.46875} moving from the root of the tree to the leaves, for a tree height of 15 and total branch length of 105. The next tree in the series with 128 taxa is constructed by inserting a bifurcating node at the midpoint of the terminal branches, resulting in branch lengths of {7.5, 3.75, 1.875, 0.9375, 0.46875, 0.234375, 0.234375}, the tree height unchanged, but the total branch length increased to 120. The 256, 512 and 1024-taxa trees are constructed using the same procedure. For the variable height tree (VHT) approach, we start with a rooted 64-taxa symmetric tree

where all branch lengths are equal to 0.25, for a tree height of 1.5 and a total branch length of 31.5. The 128-taxa tree is constructed by replacing each leaf with a bifurcating node with branch lengths of 0.25 leading to two new leaves. This results in an increase in the tree height to 1.75 and an increase in the total branch length to 63.5. This procedure is repeated to yield 256, 512 and 1024-taxa trees. We consider a location with two possible amino acids with equilibrium frequencies $\{\pi_1, \pi_2\} = \{0.015, 0.985\}$, $\{0.333, 0.667\}$ and $\{0.5, 0.5\}$; the frequencies of all other amino acids are set to zero. The global parameters are set to $\kappa = 2$, $\pi^* = (0.25)$ and $\tau = 0$. For each setup 1,000 sites are simulated, and π_1 is then estimated by ML with global parameters fixed to their true values.

Figure 5.1 shows the results of the simulations for $\pi_1 = 0.333$ for both tree strategies and for 64 to 1024 taxa. In both cases, as the number of taxa is increased, the standard error of $\hat{\pi}_1$ decreases, and the sampling distribution increasingly resembles a normal distribution. The standard error decreases much faster for the VHT than for the FHT strategy. The VHT strategy resembles the case of a biologist who samples additional, more divergent outgroup taxa that roots more deeply in the tree. The FHT strategy resembles the case of a biologist who samples additional, similar species from the same genera, thus adding a modest amount of extra information. Using $\pi_1 = 0.015$ and $\pi_1 = 0.5$ and increasing the number of amino acids observed at the location (4, 7 or 8) yield the same trends (not shown). We note that under the invariance principle of ML, $\hat{\pi} = f(\hat{F})$ (eq. 5.5) therefore estimating $\hat{\pi}$ or $\hat{F}$ leads to the same inference.

Distribution of selection coefficients for simulated data

As seen above, a large number of species of reasonable divergence are necessary to estimate the equilibrium frequencies (and fitnesses) for the codons within each location in a protein. A more important question is whether the distribution of selection coefficients can be estimated adequately for moderate data sets. We tested the robustness of estimates of p_-,

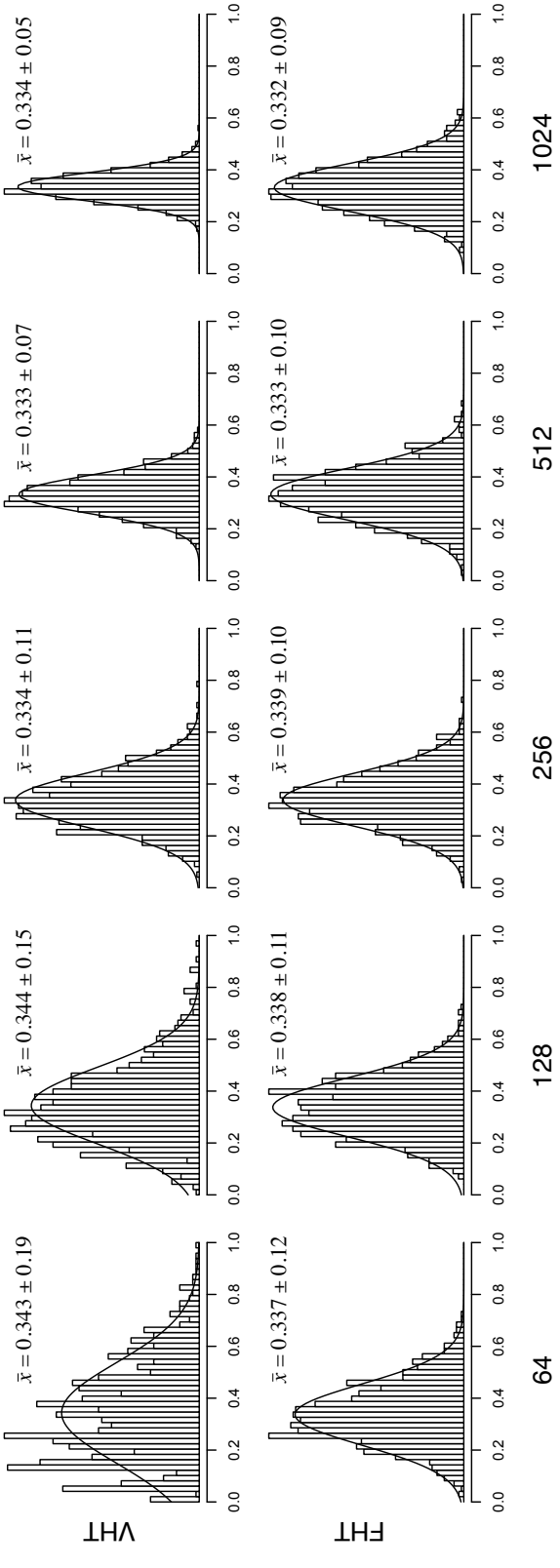


Figure 5.1.: Consistency and normality of fitness estimators. The histogram is the estimated sampling distribution of $\hat{\pi}$ from 1000 simulations. The solid line represents the best-fitting normal distribution to the simulated data.

p_0 and p_+ by generating synthetic data sets that explore the breath of the high-dimensional parameter space of the model. Specifically, we studied how the estimates were affected by different distributions of site-specific fitnesses, varying number of taxa and varying mutation rates.

To ensure that the generated data sets were reasonably realistic, the set of observed residues at each site was determined by randomly choosing a location from a mitochondrial genome alignment (described below). 1000 sites were sampled and, for each site, those residues not observed in the sampled location had their fitnesses fixed to $-\infty$. We then sampled the site-specific fitness for each residue from an underlying distribution. These are the "known" fitnesses. To explore the effect of different distributions of site-specific fitnesses (F), we considered three different distributions: (i) a gamma distribution with $\alpha = 2$ and $\beta = 1$ (ii) a normal distribution with $\mu = 0$ and $\sigma = 2$ and (iii) a normal distribution with $\mu = 0$ and $\sigma = 5$. Each distribution of F leads to a distinct distribution of S. Using these fitness values, we synthesised three data sets on the variable height tree with 256 taxa.

To investigate the effect of varying sample sizes, we created data sets with 64, 128 and 192 taxa by sampling from the 256 sequences generated under the normal distribution ($\sigma = 5$) in the previous step. For each sample, the tree topology and branch lengths were estimated. We also simulated data with the same fitnesses on a 4096 taxa tree to examine the benefit of having many more taxa.

To test the effect of increased or reduced mutation rate, two data sets were synthesised using the original fitnesses drawn from the normal ($\sigma = 5$) distribution. One set was generated with twice the mutation rate of the original 256 taxa tree, while the other had half the mutation rate of the original tree.

The site-specific fitnesses for each of the nine generated sets of sequences were re-estimated by ML using our model, fixing the global parameters to their true values. As

	p_-	p_0	p_+
Gamma distribution ($\alpha = 2,\ \beta = 1$)			
Known	0.864	0.133	0.0034
Estimated	0.882	0.115	0.0029
Normal distribution ($\mu = 0,\ \sigma = 2$)			
Known	0.897	0.099	0.0040
Estimated	0.908	0.089	0.0030
Normal distribution ($\mu = 0,\ \sigma = 5$)			
Known	0.964	0.034	0.0019
Estimated	0.969	0.030	0.0015
64 taxa	0.968	0.032	0.0014
128 taxa	0.967	0.031	0.0013
192 taxa	0.966	0.034	0.0009
4096 taxa	0.966	0.033	0.0019
Half mutation rate	0.968	0.030	0.0014
Doubled mutation rate	0.965	0.033	0.0016

Table 5.1.: Monte Carlo simulation of the distribution of selection coefficients

each synthesised data set was created with known global parameters ($\kappa = 2$, $\pi^* = (0.25)$ and $\tau = 0$) and site-specific fitnesses, the true proportion of deleterious, neutral and advantageous mutations is also known. Table 5.1 shows the proportions p_-, p_0 and p_+ of mutations calculated using the known fitnesses and compares them to the proportions obtained by estimating the fitnesses by ML. We found that in all cases the proportions of different types of mutations can be readily estimated, as well as the general shape of the distribution of S (see Appendix I).

These tests demonstrated the difficulties of estimating the fitnesses for very deleterious mutations. For example, an amino acid with fitness $F = -10$ at a location has an equilibrium frequency of $\pi = 4.5 \times 10^{-5}$ (see eq. 5.5). That is, we would expect to sample sequences from around 22,000 species to see this amino acid once at the location. Therefore, it is not possible to distinguish between $F = -20$ and $F = -10$, and we report the distribution of S from −10 to 10. However, our tests showed that with more taxa and more evolutionary time, we can recover more closely the shape of the curve for very deleterious mutations.

5.3.2. Analysis of real data

We use two real data sets to estimate the distribution of fitness effects. The first data set is an alignment of the 12 protein genes on the heavy strand of the mitochondrial genome of 244 placental mammal species (listed in Appendix J). The alignment is constructed with PRANK (Loytynoja & Goldman, 2008) and edited manually to removed small gappy regions at the end tails of some of the mitochondrial protein genes. The alignment is 3,598 codons long. The tree topology is estimated by ML with RAxML using the GTR+Γ model (Stamatakis *et al.*, 2005; Yang & Kumar, 1996).

The second data set is an alignment of the PB2 gene of 401 influenza viruses isolated from 80 human and 321 avian hosts used in chapter 3. The alignment is 759 codons long. The PB2 gene codes for a subunit of the virus polymerase complex. The polymerase genes seem to be involved in host adaptation, and there is evidence of several amino acid substitutions after the host shift (Taubenberger *et al.*, 2005). We identified 25 locations in PB2 where amino acid equilibrium frequencies are different between the viruses of the two hosts. To accommodate this observation, we first perform estimation of the fitnesses and global parameters for all residues in the protein. In a second step, a nonhomogeneous model that assumes different fitnesses for avian and human viruses is fitted to the 25 adaptive locations. For example, consider a location L that is one of the 25 adaptive locations. The substitution rate between codons I and J along the branches linking the viruses found in the human host (H) is given by

$$q_{IJ,L}^{(H)} = \begin{cases} \mu_{IJ} \frac{S_{IJ,L}^{(H)}}{1-e^{-S_{IJ,L}^{(H)}}} & \text{for } S_{IJ,L}^{(H)} \neq 0 \\ \mu_{IJ} & \text{else} \end{cases}, \tag{5.9}$$

where $S_{IJ,L}^{(H)} = F_{J,L}^{(H)} - F_{I,L}^{(H)}$ are the location and host specific selection coefficients. Similarly, the substitution rate at the adaptive locations and along the branches linking the avian viruses is $q_{IJ,L}^{(A)}$ and the avian specific fitnesses are $F_{J,L}^{(A)}$. Therefore, for each adaptive location,

$2 \times 19 = 38$ fitnesses parameters are estimated, 19 for each host. For a non-adaptive location, $q_{IJ,K}^{(H)} = q_{IJ,K}^{(A)}$ and $F_{IJ,K}^{(H)} = F_{IJ,K}^{(A)}$. The distribution of selection coefficients during evolution in the avian host is calculated using Equations 5.6 and 5.7, with $\pi_{I,K}^{(A)}$, $q_{IJ,K}^{(A)}$ and $S_{IJ,K}^{(A)}$. For the distribution of selection coefficients following the host shift, we consider that, immediately after the host shift, the equilibrium frequencies $(\pi_{I,K}^{(A)})$ will reflect the frequencies characteristic of avian viruses. At this point, however, the substitution rates $(q_{IJ,K}^{(H)})$ and the resulting fitnesses $(F_{IJ,K}^{(H)})$ will reflect the situation in the human host. Therefore, at this host shift instant we have

$$m_{HS}^{0}(S) = \frac{\sum_K \sum_{I \neq J} \pi_{I,K}^{(A)} \mu_{IJ} \delta\left(S - S_{IJ,K}^{(H)}\right)}{\sum_K \sum_{I \neq J} \pi_{I,K}^{A} \mu_{IJ}} \tag{5.10}$$

and

$$m_{HS}(S) = \frac{\sum_K \sum_{I \neq J} \pi_{I,K}^{(A)} q_{IJ,K}^{(H)} \delta\left(S - S_{IJ,K}^{H}\right)}{\sum_K \sum_{I \neq J} \pi_{I,K}^{A} q_{IJ,K}^{H}}. \tag{5.11}$$

Mammalian mitochondrial data

Figure 5.2 shows the distribution of fitness effects for novel mutations and substitutions for the mammalian mitochondria data set. The distribution of S among novel mutations clearly shows a multimodal distribution with one large peak around nearly neutral mutations ($-2 < S < 2$), with another peak corresponding to highly deleterious mutations ($S < -10$). This second peak includes all mutations to amino acids that have not been observed at a given position, and which therefore have the minimum allowed value of $F_{IJ,K} = -20$. Among substitutions, a main peak centred at neutral mutations dominates, and no substantial fraction of highly deleterious or highly advantageous ($10 < S$) substitutions are observed.

We observe that approximately 66% of mutations are deleterious ($S < -2$), similar to the fraction of deleterious mutations estimated in humans (Eyre-Walker *et al.*, 2002; Fay *et al.*, 2001). Approximately 52% of the mutations are strongly deleterious ($S < -10$), comparable

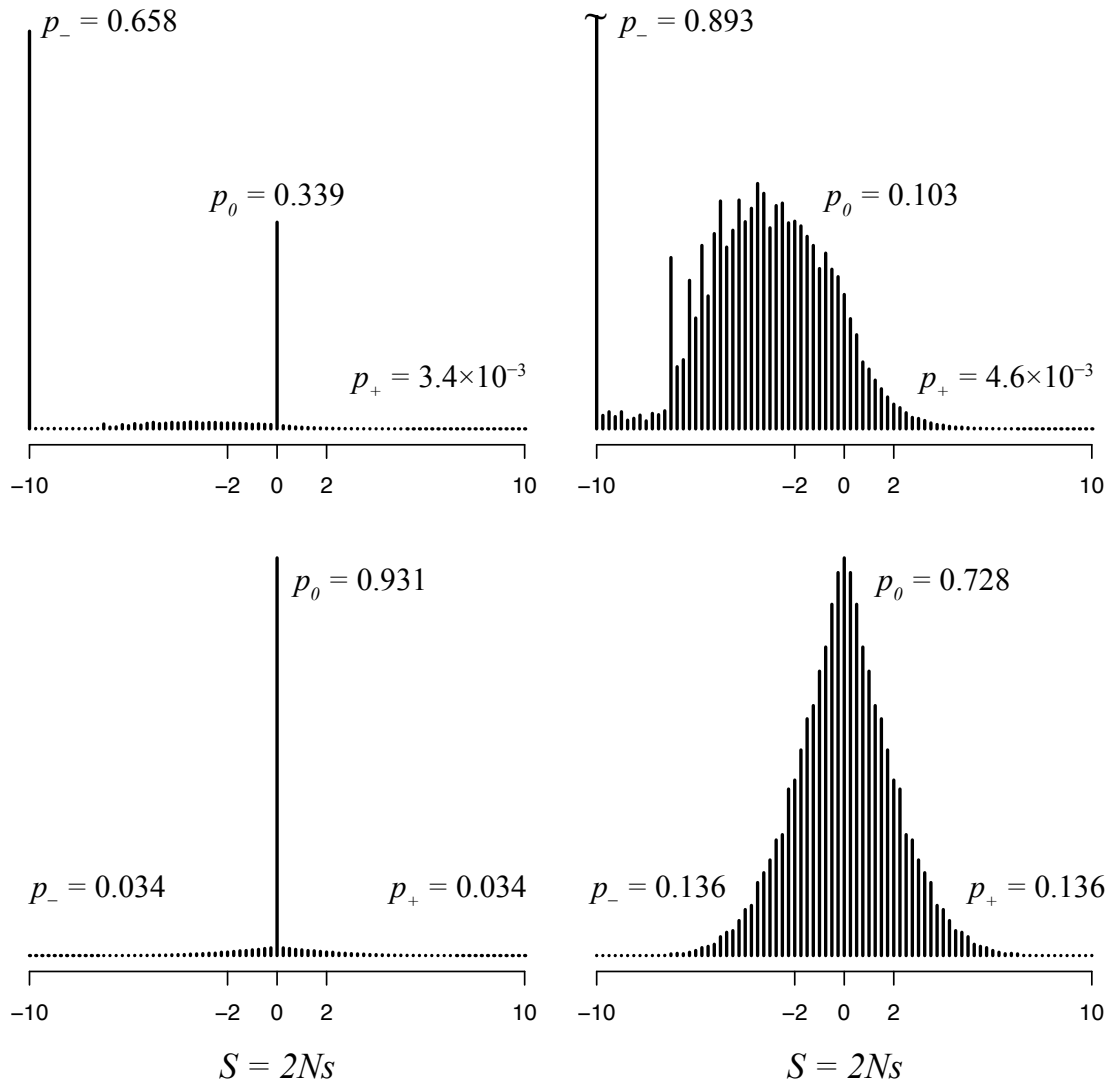


Figure 5.2.: Distribution of selection coefficients in mammalian mitochondrial proteins estimated by ML. The heights of the histogram bars are calculated according to Equations 5.6 and 5.7. Distributions are shown for all mutations (top left), nonsynonymous mutations (top right), all substitutions (bottom left) and nonsynonymous substitutions (bottom right).

with that estimated for humans (Fay *et al.*, 2001) as well as the fraction of mutations observed to be lethal in experimental studies of vesicular stomatitis virus (Sanjuan *et al.*, 2004) and yeast (Wloch *et al.*, 2001). We observe about 34% of mutations to be nearly neutral ($-2 < S < 2$), again similar to the fraction estimated by population-based methods on other data sets (e.g. Eyre-Walker *et al.*, 2002; Subramanian & Kumar, 2006). Our estimates of the number of advantageous changes is modest, representing 0.5% of the nonsynonymous mutations and 14% of the nonsynonymous substitutions. This is in rough agreement with a number of population-based studies of human evolution (e.g. Chimpanzee Sequencing and Analysis Consortium, 2005), although some studies have estimated much larger fractions for humans (Fay *et al.*, 2001) and Drosophila (Sawyer *et al.*, 2003, 2007). In general, our numbers correspond to what would be expected in a nearly-neutral evolutionary model (Akashi, 1999).

The estimated values for the global mutation parameters for the site-wise mutation selection model (swMutSel0) fit to the mammalian data are listed in Table 5.2. The equilibrium base frequencies (π^*) are similar but not identical to those estimated with the FMutSel0 model by PAML, which neglects changes in base composition resulting from the selective constraints acting at the amino acid level. The value of τ, representing the tendency for simultaneous multiple base substitutions, indicates that the proportions of single, double and triple changes are 99.4%, 0.58% and 0.002% respectively. The optimisation procedure is likely to result in an over-estimation of the frequency of multiple mutations. Mutations between two amino acids that are not convertible by a single base change (e.g. phenylalanine {TTT, TTC} to asparagine {AAT, AAC}) can result either through multiple base changes or through a transient intermediate amino acid (such as Tyrosine {TAT, TAC}). Our procedure, as described above, estimates τ while making the assumption that unobserved amino acids at any location, including possible intermediates, are incompatible with the selection constraints. This increases the requirement for multiple base changes, increasing our estim-

		swMutSel0	FMutSel0
Mammal data	$\hat{\pi^*}$ (T, C, A, G)	0.19, 0.27, 0.48, 0.06	0.17, 0.25, 0.52, 0.06
	$\hat{\kappa}$	7.06	6.97
	ω	—	0.05991
	$\hat{\tau}$	0.06257	—
Influenza data	$\hat{\pi^*}$ (T, C, A, G)	0.24, 0.20, 0.37, 0.20	0.23, 0.19, 0.37, 0.21
	$\hat{\kappa}$	7.86	7.77
	ω	—	0.062
	$\hat{\tau}$	0.09199	—

Table 5.2.: Parameters in swMutSel0 and FMutSel0

ate of τ. Even with this bias, our estimation of the multiple substitution rate is more modest than proportions derived from simpler codon models applied to a more comprehensive protein dataset (Kosiol *et al.*, 2007) and may indicate either differences in the evolutionary process for mitochondrial DNA or biases that result when site-specific selective constraints are inadequately modelled.

Influenza PB2 data

Figure 5.3 shows the distribution of fitness effects for influenza PB2 gene evolving in the avian host, and Figure 5.4 shows the distribution following a well defined adaptive event: the host shift to humans. Like in the mitochondrial case, the distribution of S among mutations at adaptive equilibrium shows a multi-modal distribution, with two main modes centred around nearly neutral ($-2 < S < 2$) and highly deleterious ($S < -10$) mutations. Among substitutions, the distribution is dominated by a main peak centred on neutral mutations. Interestingly, at the host shift event, we find two well defined peaks among substitutions, one peak centred around neutral substitutions and another peak of highly advantageous substitutions ($10 < S$). We estimate that 12% of all substitutions, and 50% of all nonsynonymous substitutions are advantageous at the host shift event. These results are in agreement with an adaptive model as pointed out by Akashi (1999).

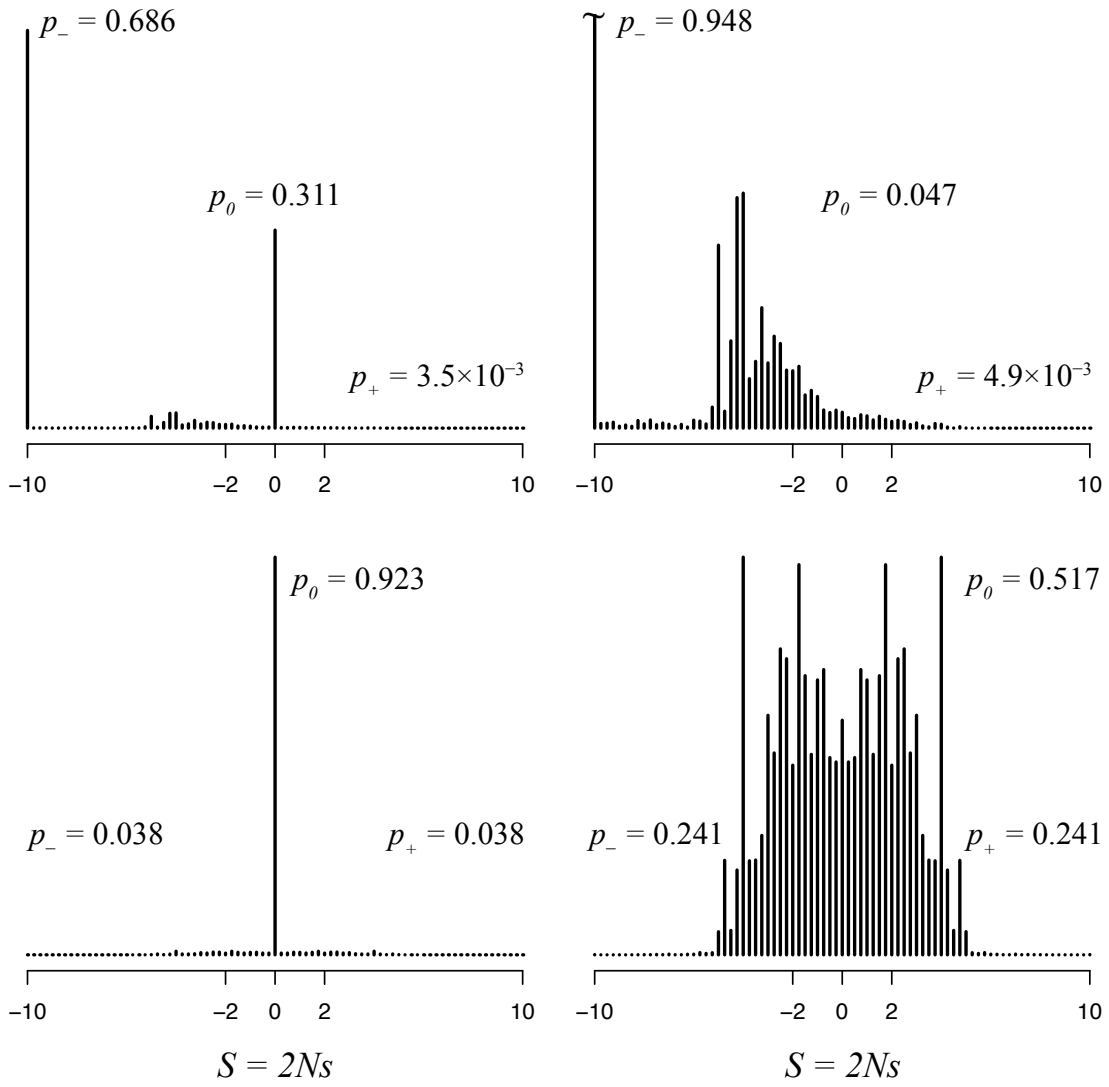


Figure 5.3.: Distribution of selection coefficients in the PB2 gene of influenza for avian viruses at adaptive equilibrium. Distributions are shown for all mutations (top left), nonsynonymous mutations (top right), all substitutions (bottom left) and nonsynonymous substitutions (bottom right).

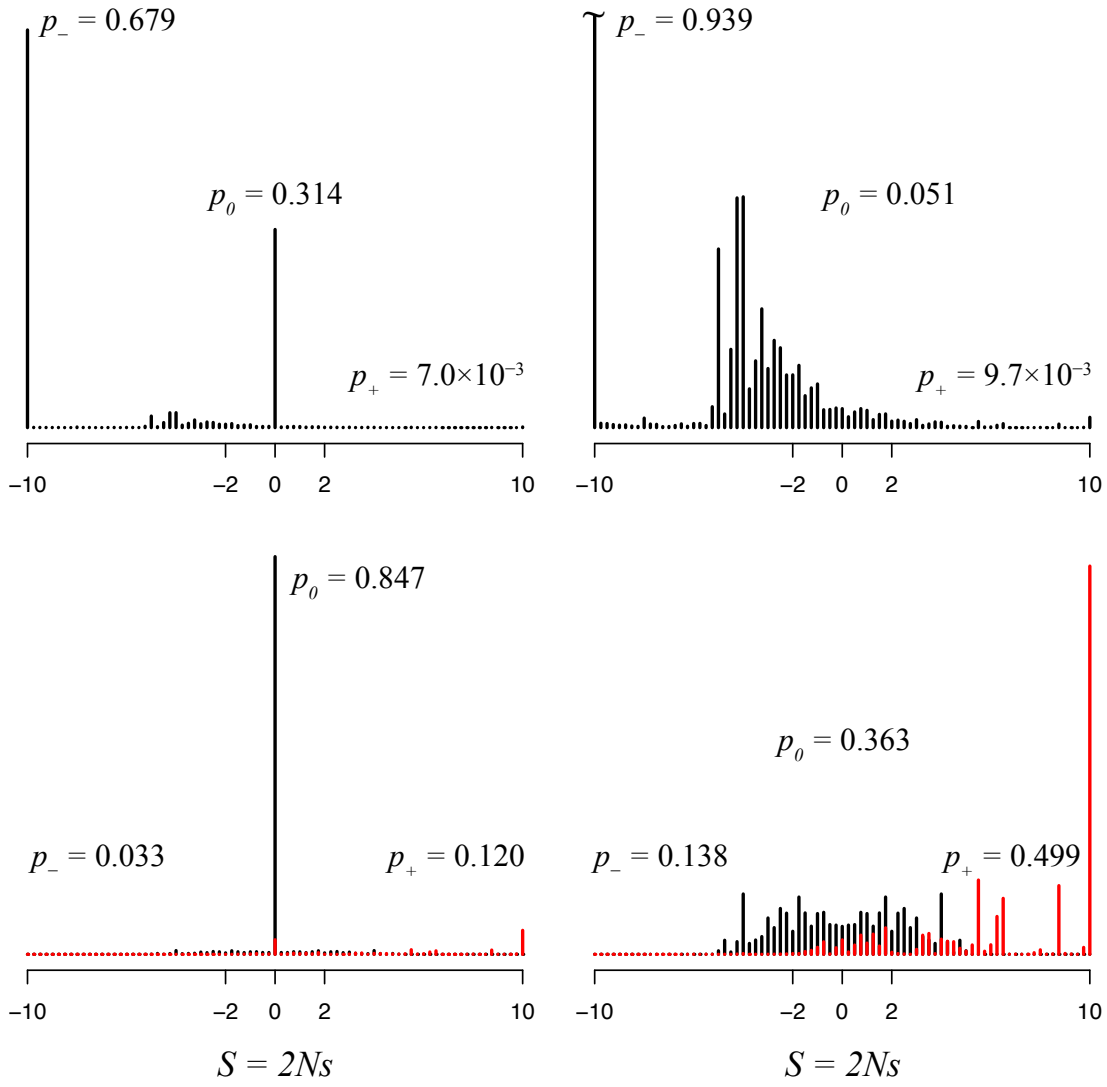


Figure 5.4.: Distribution of selection coefficients in the PB2 gene of influenza for human viruses immediately after host shift from bird. Distributions are shown for all mutations (top left), nonsynonymous mutations (top right), all substitutions (bottom left) and nonsynonymous substitutions (bottom right). The contributions from the 25 sites under different selective constraints in the two hosts are shown in red; the contributions from other sites are shown in black.

There has been much discussion in the literature about the relative contributions of nearly neutral and advantageous substitutions to the evolutionary process (i.e. Kimura (1983) vs. Gillespie (1994)). We suggest that the distribution of S is not constant in time but changes as organisms undergo adaptation through novel environments, with the relative contributions of nearly neutral and advantageous mutations dependent on the particular evolutionary scenario. It seems sensible to think that organisms go through phases of mostly neutral and mostly adaptive episodes.

Estimates for the influenza global mutation parameters are listed in Table 5.2. As for the mitochondrial data, the equilibrium base frequencies (π^*) are similar but not identical to those estimated with the FMulSel0 model. The value of τ is of the same order of magnitude as the mitochondrial case, indicating nearly the same proportions of single, double and triple substitutions.

The parametric form of the distribution of S

Extreme value theory has been used to show that, under a wide range of conditions, the distribution of selection coefficients for advantageous nonsynonymous mutations should be exponential (Gillespie, 1994; Orr, 2003). This prediction has been questioned based on simulations of the evolution of RNA (Cowperthwaite *et al.*, 2005), which yielded a distribution with an overabundance of slightly adaptive mutations. As shown in Figure 5.5, we observe that the distribution of S for advantageous mutations ($S > 0$) matches an exponential distribution for both the mammalian and influenza data; a fit of the data between $0 < S < 5$ to $m_0(S) \sim \exp(-\beta S)$ yields an exponent of $\beta = 0.924$ (95% CI: 0.904 - 0.941) for mammals and $\beta = 0.688$ (95% CI: 0.630 - 0.733) for influenza, both in agreement with the results of extreme value theory.

Previous work analysing intraspecies variation has suggested that the distribution of nonsynonymous deleterious mutations is leptokurtic, that is, having a faster initial fall-off fol-

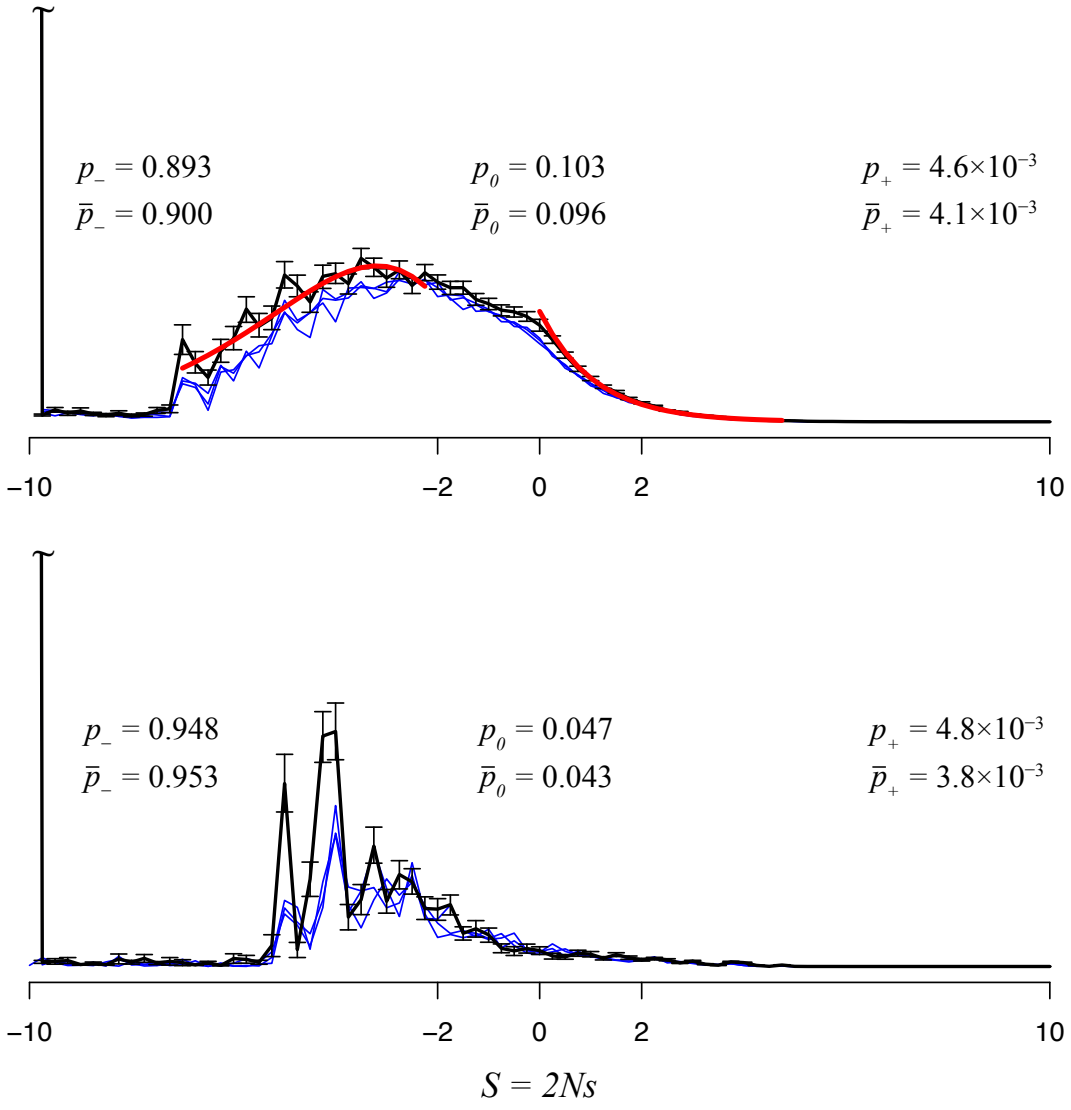


Figure 5.5.: The parametric form of S and bootstrap analysis of the data. The mammalian data are shown at the top and influenza data at the bottom. The black curves display the distribution of selection coefficients for nonsynonymous mutations including error bars obtained by classic bootstrapping. Red curves show best exponential fit for advantageous mutations ($0 < S < 5$) and best gamma distribution fit for moderately deleterious mutations ($-7 < S < -2$). Blue curves represent the distributions obtained from the parametric bootstrap analysis from three synthetic data sets. p and $\bar{p}$ are the proportions for the real data and the average for the parametric bootstraps, respectively.

lowed by a longer tail, such as a gamma distribution with shape parameter $\alpha < 1$. For example, (Eyre-Walker & Keightley, 2007) analysed human SNPs and fit the resulting nonsynonymous deleterious mutations to a gamma distribution with $\alpha = 0.23$. In contrast, Nielsen & Yang (2003) carried out an inter-species study of primate mitochondrial proteins, fitting a reflected gamma to the distribution of S. The reflected gamma distribution around zero is simply $\Gamma_R(S|\alpha,\beta) = \Gamma(-S|\alpha,\beta)$ for $S < 0$. They estimated $\alpha = 3.22$, far from leptokurtic. Their model does not seem biologically realistic, as it suggests that different selective constraints at different locations in the protein act to reduce the overall substitution rate without affecting the resulting equilibrium distribution of amino acids at that location. Our distribution of selective coefficients with $S < 0$ clearly do not fit a reflected gamma distribution. We can, however, fit a reflected gamma distribution to the more limited range of moderately deleterious mutations ($-7 < S < -2$) as shown in Figure **5.5**. Over this range, our results more closely resemble the distribution obtained by Nielsen and Yang, with $\alpha = 3.601$ (95% CI 2.921 - 4.298) and $\beta = 0.817$ (95% CI 0.643 - 0.987). The distribution of S for the influenza data is highly multimodal between $-7 < S < -2$, so we do not attempt to fit these data to a reflected gamma as in the mammalian case.

Although our results on nearly-neutral and advantageous mutations and substitutions roughly correspond to previous results obtained with evolution-based methods, we observe a large fraction of highly deleterious mutations ($S < -10$), better matching the number of experimentally observed lethal mutations. It is not surprising that previous analyses have had trouble estimating these highly deleterious mutations. Nielsen & Yang (2003) explicitly did not allow residues to be less or more favoured at different locations, only allowing changes in the overall substitution rates; all substitutions are allowed at all but perfectly conserved locations. Rodrigue *et al.* (2010) consider models of selection at each location that are mixtures of various components; this averaging effect reduces the ability to identify highly unfavourable amino acids at specific positions.

Uncertainties in the estimation of the distribution of S

We estimate the uncertainties and biases in our approach by using the classical and parametric bootstrap approaches. The classic bootstrap is used to generate error bars for the MLEs obtained from the real data, and the parametric bootstrap is used to generate simulated replicates of the distribution of S. The distribution of selective coefficients for nonsynonymous mutations for three parametric bootstrap datasets are compared with the results for the real data in Figure 5.5. As would be expected, the distributions are extremely similar for the neutral and advantageous substitutions. The general trends for the deleterious mutations are similar, although it appears that the calculations have a tendency to over-estimate the magnitude of S for the deleterious mutations. This is not overly surprising, as this would result if the fitness of the extremely infrequent amino acids were underestimated, and where their omission from the observed data reflects lack of evolutionary time rather than biological impossibility. This discrepancy may also be caused by our optimising the tree branch lengths under the site-invariant FMutSel0 model, rather than our site-wise model (Halpern & Bruno, 1998). These differences have minimal effect on the fraction of mutations and substitutions that are deleterious, neutral and advantageous.

As might be expected given the close correspondence of the distributions for advantageous nonsynonymous mutations, the fit of the distribution of positive ($0 < S < 5$) selective coefficients for the three bootstrap datasets to an exponential yields values ($\bar{\beta} = 0.953$) similar to that obtained with the real mitochondrial data ($\beta = 0.924$). Although there are differences in the reflected gamma distribution fit for deleterious mutations for the mitochondrial data ($\bar{\alpha} = 5.611$ and $\bar{\beta} = 1.551$ in contrast to $\alpha = 3.601$ and $\beta = 0.817$), the results are still far from leptokurtic.

Comparison of the derived distribution of S with the results of parametric bootstrap simulations indicate that our phylogenetic analysis is able to successfully characterise the dis-

tribution of positive and near-neutral changes, although it over-estimates the effect of deleterious mutations. The latter limitation is not unexpected - if an amino acid is rare or not observed at all at a given location, it is difficult to estimate how frequently it would be found at equilibrium. This is not an issue for representing substitutions, as these mutations would be extremely unlikely to occur.

5.3.3. Validations, assumptions and limitations of the model

The model presumes that we know the true alignment for the selected datasets. Both the mammalian mitochondrial genes, which are well conserved, and influenza PB2 datasets produce good quality alignments. We also assume that the true phylogenetic tree is known. It has been shown, however, that small variations in tree topologies have minor impact on the parameters of phylogenetic models (Yang *et al.*, 1994), and we would not expect it to significantly affect our calculations of selection coefficients. Additionally, we tested the effect of tree topology uncertainty during our parametric bootstrap analysis by reestimating the tree topology for each replicate. Although the trees estimated for the bootstrapped datasets were different, but similar, to that of the real datasets, they did not have a major impact on our estimated distribution of S.

The analysis assumes that the various global and location-specific parameters are constants throughout the evolutionary process, with the exception of the host shift event explicitly included in the model for influenza PB2. The assumption that $F_{IJ,K} = 2Nf_{IJ,K}$ is a constant is based on assumptions regarding both the population size N as well as the fitness parameters $f_{IJ,K}$. Our analysis of the mitochondrial dataset and PB2 evolving in an avian host assumes that the amino acid distribution is at equilibrium with respect to fixed selective constraints, resulting in a distribution of selective effects for accepted mutations symmetric around zero. This assumption explicitly eliminates the role of changes in selec-

tion and population size in adaptive evolution. The fitness parameters could change because of a number of effects. Firstly, the structure, function, or physiological context of the protein could change. We have restricted these effects by considering mitochondrial proteins and PB2 from influenza. In neither case is there gene duplication that could lead to neo-functionalisation that might result of changes in function or physiological context. We assume that the gene sequences are related by a single tree and recombination is absent, which is the consensus for both mammalian mitochondria (Lynch, 2007) and influenza genes (Boni *et al.*, 2008). It is well recorded that structural change is extremely slow relative to sequence change (Aronson & Royer Jr, 1994). This does not mean that local structures might not change; these changes are more likely to occur in the exposed loop sections of the proteins, where there is reduced selective constraints. These would, therefore, likely result in small shifts in the neutral and near-neutral parts of the distribution of selection coefficients. Our analysis of the host shift effects on influenza PB2 demonstrates that changes in selective constraints can be explicitly included in the modelling, especially if information about the shift can be obtained independently (as, for instance, when there is a change in the host of a pathogen at a specific branch of the phylogenetic tree). Models of selection that include changes in selective constraints in a more general manner (such as covarion models, see Galtier (2001); Penny *et al.* (2001)) could be used, but would result in even greater computational complexity.

Secondly, the selection constraints at a location might change due to substitutions that occur in other regions of the protein, that is, through the invalidation of our assumption that different locations in the protein evolve independently. There has been significant effort made looking for such correlations between the substitution process at interacting sites (e.g. Bonhoeffer *et al.*, 2004; Lycett *et al.*, 2009). The difficulty of this problem, the rather few examples where such effects have been substantiated, and the overall success of the independent sites assumption compared with models where it is relaxed (Kleinman *et al.*, 2010;

Lakner *et al.*, 2011; Rodrigue *et al.*, 2006), suggest that this effect is not likely to be large. This effect is even less likely to occur in the population-based models, as the timescales relevant to these studies are too short for many substitutions at other sites to occur. These population-based studies, however, are complicated by the interaction between the population dynamics that occur at different locations, such as interference between the fixation of different mutations (Hill & Robertson, 1966; Kirby & Stephan, 1996; Stephan, 1995) and genetic hitch-hiking (Barton, 2000; Maynard-Smith & Haigh, 1974). Simulations of these phenomena in computational models might allow further reconciliation of the results of these types of studies.

It must also be pointed out that codon locations in a protein are tightly linked, and this can have a sizeable effect on the estimation of selection coefficients (Bustamante, 2005). Our condition of independence implicitly assumes free recombination among locations. This is certainly not true either for the influenza or mammal data sets analysed. In particular, selection coefficients involving highly advantageous mutations are expected to be underestimated (Bustamante, 2005). It is not clear at present how phylogenetic models could incorporate the assumption of linkage, and most works that have attempted to estimate the distribution of S from phylogenetic data have worked with the assumption of independence (e.g. Bustamante, 2005; Nielsen & Yang, 2003; Rodrigue *et al.*, 2010; Yang & Nielsen, 2008). Cartwright *et al.* (2011) studied the problem through simulation but using a much simpler substitution model. Their results suggest that accounting for interference between fitness-affecting mutations at linked sites can lead to results that deviate from common assumptions in phylogenetic models (e.g. the Markov assumption).

The assumption that $F_{IJ,K}$ is a constant also assumes that the effective population number has remained constant across lineages in the influenza and mammalian phylogenies. Both humans and Drosophila have undergone recent increases in population and expansion into new evolutionary niches (Glinka *et al.*, 2003; Merriwether *et al.*, 1991), possibly

explaining why some (see Fay *et al.*, 2001; Sawyer *et al.*, 2003, 2007) but not all (Chimpanzee Sequencing and Analysis Consortium, 2005) population-based studies of these groups yield a higher degree of adaptive evolution than observed here. This assumption could be relieved at the expense of additional parameters in the model as suggested by Nielsen & Yang (2003). Influenza viruses evolving in humans present oscillating population numbers, with population bottlenecks of low genetic diversity at the beginning/end of epidemic seasons (Rambaut *et al.*, 2008). However, the estimated distribution of S for human influenza viruses following the host shift event would be affected if the virus population size varies between the avian and human lineages. Because our model currently does not incorporate these variations in effective population number in mammals and influenza, our estimated fitnesses should be interpreted as averages over evolutionary timescales. We are currently exploring ways to incorporate variations in the effective population number in our model, but this is expected to be computationally challenging.

We also assume that the global mutation parameters (τ, κ and the π^*) do not vary across locations and across the tree. This assumption is unlikely to be strictly true; observed base compositions are known to be significantly different in different lineages. Differences in the equilibrium base composition in influenza has been documented by dos Reis *et al.* (2009). Changes in the equilibrium base frequencies of, for instance, 10%, could result in similar changes in the estimate of $q_{IJ,K}$. However, $q_{IJ,K}$ is a steeply varying function of $F_{IJ,K}$, meaning that the changes expected in the latter quantity would be small.

Although a likelihood ratio test for the effect of selection on codon bias is significant in both data sets ($p \ll 0.01$, for details of the test see Yang & Nielsen (2008)), we only estimate fitnesses at the amino acid level and explicitly ignore selection at the synonymous codon level, as estimation of the 60 global codon level fitnesses would be a computationally onerous task. In mammals, selection on codon usage is very weak (dos Reis & Wernisch, 2009; Yang & Nielsen, 2008), similarly, selection on codon bias is negligible in influenza

viruses (Shackelton *et al.*, 2006). For this reason, in these data sets, selection coefficients for codon bias are expected to be small and within the nearly neutral interval, with a negligible effect on the shape of the distribution of selection coefficients among novel mutations and substitutions.

5.4. Conclusion

The dominant method of generating distributions of fitness effects has relied on a combination of intra- and inter-species variation. More recently, these population-based approaches have been joined by phylogenetic analyses that attempt to make a connection between the evolutionary process and population dynamics. These latter analyses offer a few specific advantages. Perhaps the biggest advantage is an ability to look at a different timescale, allowing us to explore the relationship between population variation and evolutionary change. Secondly, the range of different organisms that can be studied is greatly increased, compared with the relatively few species (e.g. humans, Drosophila, yeast) where sufficient data exists to model population variation. Thirdly, although both approaches involve making particular assumptions, the assumptions are different. Comparisons between results obtained with the different methods can provide insight into the nature and validity of these assumptions. Fourthly, the substitution model can be elaborated to include additional effects, such as changes in selective constraints, population size, mutation rates at different points in evolution, or a relaxation of certain assumptions such as independence of sites. These extensions will become increasingly feasible as our sequence data, computational resource and biological understanding continue to increase. Fifthly, it is not necessary to pre-specify a functional form for the distribution of S. This means that it is possible to decompose the evolutionary process and ask specific questions, such as the distribution of fitness effects involving changes to proline in helical regions of the protein.

This approach can be applied to any set of proteins with a sufficiently large and diverse set of homologs. The resulting distribution of fitness effects constitutes a signature of the selective constraints, and could provide interesting perspectives on individual proteins and their physiological context. The relative proportion of deleterious, neutral and advantageous mutations could depend on the protein structure and function, reflecting such distinctions such as whether the protein is globular, membrane, or unstructured; cytosolic or excreted; signalling, enzymatic, or immunological; or solitary or a member of a larger gene family.

Finally, as has been pointed out (Thorne *et al.*, 2007), the connection with population dynamics has the potential to reform our modelling of sequence evolution. Substitution models have predominantly been phenomenological, representing the results of the evolutionary process (an accepted substitution) rather than the mechanics of how those results occurred. The opportunity to provide a firmer basis for these models by connecting it to population processes can result not only in better models, but also ones that can be used to understand biological systems, populations and evolutionary processes. The model presented here bridges the gap between population genetics and substitution models of sequence evolution. Since it was originally introduced by Halpern & Bruno (1998), this site-specific codon-based evolutionary model has seen limited use; the large number of adjustable parameters result in the need for a significant amount of sequence data as well as computational resources. These two limitations are becoming less onerous. With the modifications described here, and with the availability of powerful parallel computing systems, it is now possible to obtain realistic estimates of the distribution of selection coefficients from phylogenetic data.

6. Conclusion

The work presented here is concerned with developing probabilistic models of sequence evolution that can effectively locate and characterise selection acting on proteins.

In chapter 3 we developed a site- and time-heterogeneous model that we used to identify changes in selective constraints in influenza when in avian or human hosts. Given a protein sequence alignment and the corresponding phylogenetic tree, we detected host-specific selective pressure by analysing whether the pattern of amino acid change differed between the two hosts, rather than relying on observed amino acids used by some other non-phylogenetic methods. Previous phylogenetic methods for detecting selection rely on the absolute rate of change or the relative rates of nonsynonymous to synonymous changes. These miss many qualitative changes that might not involve change in rate (for example, a change in the range of acceptable amino acids at a site in a new environment). The site-wise nature of our model allows us to explicitly describe selection pressure across the length of the protein, avoiding the averaging of substitution rates that occurs when using both site-invariant or mixture models. We analysed 13 influenza proteins using sequences spanning avian and human hosts, and found that our nonhomogeneous model of sequence change was able to identify 172 sites with strong statistical support for different substitution patterns in the two hosts. These include sites that have been identified experimentally. To further test the validity of the results, the identified sites can be considered in a structural and functional context. The structures of all the influenza proteins are known, at least partially. We can use this structural understanding to give us insight into the process of sequence change. In particular, we can investigate where locations identified by our method are found on the protein (e.g. see Appendix D). The method can be applied to detecting selective constraint when the timing of the change is known. With respect to influenza, a logical progression would be to expand the model to include the selective constraints in

swine influenza. We could imagine that selective constraints at sites could be the same in all hosts, or differ among birds and mammals, or differ among birds, humans and swine. They could also be different between different swine lineages (classical swine vs. Eurasian swine lineages). The software implementation of our model allows for the construction of such complex nonhomogeneous relationships. These more sophisticated models may be able to test whether the 1918 pandemic was indeed a swine-origin virus. For example, the data may show a significantly better fit to a model that allows for the pre-1918 lineage to evolve in swine before infecting humans.

We then designed a measure for the adaptation of any given influenza protein sequence to the constraint present in the avian or human host. In chapter 4 we used the equilibrium frequencies of amino acids from the identified locations in each host provided by the previous analysis to measure "adaptedness". Analysing an ensemble of avian, human and swine influenza sequences showed gradual adaptation of human strains in the human host. Comparing the 1918 virus with the ancestral reconstruction of the virus that founded the 1918 human and classical swine lineages shows that it had undergone a significant amount of adaptation to the human host prior to 1918. Perhaps that pre-adaptation occurred in swine, similar to the increasing human adaptedness of the 1979 H1 Eurasian swine lineage shown in our results. Reconstructing all the ancestral sequences in the human lineages allowed us to look at the change in human adaptedness along different branches of the tree after the avian-to-human host shift event. We found that that tips have a larger fraction of adaptedness-decreasing changes than other branches and that the trunk connecting ancestral to recent pandemics have a larger fraction of adaptedness-increasing changes. The method should be helpful in assessing the potential of current viruses to found future epidemics or pandemics. This work could be extended to measure adaptedness in swine, and different adaptedness in poultry compared with waterfowl. We could also examine whether the virus founding particular human pandemics are exceptional when compared to their

original host, and compare these to influenza strains that result in sporadic infections (e.g. H5N1).

Finally, we wanted to develop a model that more closely reflected the underlying biological process of sequence change which could be used to measure selection in molecular sequences. In chapter 5 we presented a site-wise mutation-selection model that allowed for mutation at the nucleotide level, fitness effects for each amino acid and fixation of that mutation in population genetic terms. Although these types of mutation-selection models are more complex than traditional models, they can now be used thanks to greater computational resources and increasing amounts of available sequence data. We used this model to estimate the distribution of selection coefficients, a central topic of population genetic research. Our distributions resemble the results from population genetics theory and experimental work, in constrast to previous phylogenetic approaches. We find a bimodal distribution of selection coefficients for mammalian mitochondria and influenza when evolving in the natural avian reservoir. Following a host shift from birds to humans, we find a trimodal distribution with a small proportion of advantageous mutations and significant proportion of advantageous substitutions, again matching theoretical predictions. We have implemented time-heterogeneity in this model, allowing its use in the future to detect changes in selective constraints in codon sequences, as in chapter 3. We hope to continue to rigorously test this model by, for example, measuring the effect of priors on the fitness of residues, giving us some indication of how much data or evolutionary time would be necessary to accurately estimate the distribution of selection coefficients. Because the software can be run on high-performance distributed machines, we can apply our tests and measures of selection to larger genomics datasets now available thanks to high-throughput sequencing.

In summary, we believe that the methods presented here show that site- and time-heterogeneous phylogenetic models, and those that offer a more mechanistic view of molecular evolution, can be used to locate and represent the effects of selection present in

molecular sequences. They highlight the importance of accounting for evolutionary relationships when analysing related sequences and demonstrate that sequences alone can give insight to long standing questions of biological interest.

References

Adachi, J. & Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of Molecular Evolution,* **42** (4), 459–468.

Ahmad, S., Gromiha, M., Fawareh, H., & Sarai, A. (2004). ASAView: database and tool for solvent accessibility representation in proteins. *BMC bioinformatics,* **5**, 51.

Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. *Genetics,* **139** (2), 1067–1076.

Akashi, H. (1999). Within- and between-species DNA sequence variation and the 'footprint' of natural selection. *Gene,* **238** (1), 39–51.

Anisimova, M. & Liberles, D. A. (2012). Detecting and understanding natural selection. In: *Codon Evolution: Mechanisms and Models*, (Cannarozzi, G. M. & Schneider, A., eds) pp. 73–96. Oxford University Press New York, USA.

Antonovics, J., Hood, M. E., & Baker, C. H. (2006). Molecular virology: was the 1918 flu avian in origin? *Nature,* **440** (7088), E9; discussion E9–10.

Aris-Brosou, S. & Rodrigue, N. (2012). The essentials of computational molecular evolution. In: *Evolutionary Genomics*, (Anisimova, M. & Walker, J. M., eds) volume 855 of *Methods in Molecular Biology* pp. 111–152. Humana Press.

Aronson, H. & Royer Jr, W. (1994). Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. *Protein Science,* **3**, 1706–1711.

Baigent, S. & McCauley, J. (2003). Influenza type A in humans, mammals and birds: determinants of virus virulence, host-range and interspecies transmission. *Bioessays,* **25** (7), 657–671.

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., & Lipman, D. (2008). The influenza virus resource at the national center for biotechnology information. *Journal of Virology,* **82** (2), 596–601.

Barton, N. H. (2000). Genetic hitchhiking. *Philos Trans R Soc Lond B Biol Sci,* **355** (1403), 1553–62.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological),* **5** (1), 289–300.

Blackburne, B. P., Hay, A. J., & Goldstein, R. A. (2008). Changing selective pressure during antigenic changes in human influenza H3. *PLoS Pathogens,* **4** (5).

Blouin, C., Boucher, Y., & Roger, A. J. (2003). Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. *Nucleic acids research,* **31** (2), 790–797.

Bonhoeffer, S., Chappey, C., Parkin, N., Whitcomb, J., & Petropoulos, C. (2004). Evidence for Positive Epistasis in HIV-1. *Science (New York, NY),* **306** (5701), 1547–1550.

Boni, M. F., Zhou, Y., Taubenberger, J. K., & Holmes, E. C. (2008). Homologous recombination is very rare or absent in human influenza A virus. *Journal of Virology,* **82** (10), 4807–4811.

Bruno, W. J. (1996). Modeling residue usage in aligned protein sequences via maximum likelihood. *Molecular biology and evolution,* **13** (10), 1368–1374.

Buckler-White, A. J., Naeve, C. W., & Murphy, B. R. (1986). Characterization of a gene coding for M proteins which is involved in host range restriction of an avian influenza A virus in monkeys. *Journal of Virology,* **57** (2), 697–700.

Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics,* **129** (3), 897–907.

Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J., & Fitch, W. M. (1999). Predicting the Evolution of Human Influenza A. *Science (New York, NY),* **286** (5446), 1921–1925.

Bustamante, C. D. (2005). Population genetics of molecular evolution. In: *Statistical methods in molecular evolution*, (Nielsen, R., ed) pp. 63–99. Springer New York, USA.

Bustamante, C. D., Nielsen, R., Sawyer, S. A., Olsen, K. M., Purugganan, M. D., & Hartl, D. L. (2002). The cost of inbreeding in arabidopsis. *Nature,* **416** (6880), 531–4.

Cartwright, R. A., Lartillot, N., & Thorne, J. L. (2011). History Can Matter: Non-Markovian Behavior of Ancestral Lineages. *Systematic Biology,* **60** (3), 276–290.

Centers for Disease Control and Prevention (CDC) (2009). Swine influenza A (H1N1) infection in two children–Southern California, March-April 2009. *MMWR. Morbidity and mortality weekly report,* **58** (15), 400–402.

Chare, E. R., Gould, E. A., & Holmes, E. C. (2003). Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. *The Journal of general virology,* **84** (Pt 10), 2691–2703.

Chen, G., Chang, S., Mok, C., Lo, Y., Kung, Y., Huang, J., Shih, Y., Wang, J., Chiang, C., & Chen, C. (2006). Genomic signatures of human versus avian influenza A viruses. *Emerg Infect Dis,* **12** (9), 1353–1360.

Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature,* **437** (7055), 69–87.

Cochrane, G., Karsch-Mizrachi, I., & Nakamura, Y. (2011). The international nucleotide sequence database collaboration. *Nucleic Acids Research,* **39** (suppl 1), D15–D18.

Connor, R. J., Kawaoka, Y., Webster, R. G., & Paulson, J. C. (1994). Receptor specificity in human, avian, and equine H2 and H3 influenza virus isolates. *Virology,* **205** (1), 17–23.

Cowperthwaite, M. C., Bull, J. J., & Meyers, L. A. (2005). Distributions of beneficial fitness effects in rna. *Genetics,* **170** (4), 1449–57.

Crow, J. F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York: Harper & Row.

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). A Model of Evolutionary Change in Proteins. In: *Atlas of protein sequence and structure* volume 5 pp. 345–352. National Biomedical Research Foundation Washington, D.C.

Dickerson, R. E. (1971). The structure of cytochromec and the rates of molecular evolution. *Journal of Molecular Evolution,* **1**, 26–45.

Dimmic, M. W., Rest, J., Mindell, D. P., & Goldstein, R. A. (2002). rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *Journal of Molecular Evolution,* **55** (1), 65–73.

Dorman, K. S. (2007). Identifying dramatic selection shifts in phylogenetic trees. *BMC evolutionary biology,* **7 Suppl 1**, S10.

dos Reis, M., Hay, A. J., & Goldstein, R. A. (2009). Using non-homogeneous models of nucleotide substitution to identify host shift events: application to the origin of the 1918 'Spanish' influenza pandemic virus. *Journal of Molecular Evolution,* **69** (4), 333–345.

dos Reis, M., Tamuri, A., Hay, A. J., & Goldstein, R. A. (2011). Charting the host adaptation of influenza viruses. *Mol Biol Evol,* **28**, 1755–1767.

dos Reis, M. & Wernisch, L. (2009). Estimating translational selection in eukaryotic genomes. *Mol Biol Evol,* **26** (2), 451–61.

Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research,* **32** (5), 1792–1797.

Edwards, A. W. F. (1992). *Likelihood*. Maryland, USA: John Hopkins University Press.

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. London, UK: Chapman & Hall.

Eyre-Walker, A. (2006). The genomic rate of adaptive evolution. *Trends Ecol Evol,* **21** (10), 569–75.

Eyre-Walker, A. & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nat Rev Genet,* **8** (8), 610–8.

Eyre-Walker, A., Keightley, P. D., Smith, N. G., & Gaffney, D. (2002). Quantifying the slightly deleterious mutation model of molecular evolution. *Mol Biol Evol,* **19** (12), 2142–9.

Fay, J. C., Wyckoff, G. J., & Wu, C. I. (2001). Positive and negative selection on the human genome. *Genetics,* **158** (3), 1227–34.

Felsenstein, J. (1973). Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees From Data on Discrete Characters. *Systematic Zoology,* **22** (3), 240–249.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution,* **17** (6), 368–376.

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist,* , 1–15.

Felsenstein, J. (2003). *Inferring phylogenies*. Sunderland, MA: Sinauer Associates.

Finkelstein, D. B., Mukatira, S., Mehta, P. K., Obenauer, J. C., Su, X., Webster, R. G., & Naeve, C. W. (2007). Persistent host markers in pandemic and H5N1 influenza viruses. *Journal of Virology,* **81** (19), 10292–10299.

Forsberg, R. & Christiansen, F. B. (2003). A codon-based model of host-specific selection in parasites, with an application to the influenza A virus. *Molecular biology and evolution,* **20** (8), 1252.

Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., Griffin, J., Baggaley, R. F., Jenkins, H. E., Lyons, E. J., Jombart, T., Hinsley, W. R., Grassly, N. C., Balloux, F., Ghani, A. C., Ferguson, N. M., Rambaut, A., Pybus, O. G., Lopez-Gatell, H., Alpuche-Aranda, C. M., Chapela, I. B., Zavala, E. P., Guevara, D. M. E., Checchi, F., Garcia, E., Hugonnet, S., Roth, C., & WHO Rapid Pandemic Assessment Collaboration (2009). Pandemic potential of a strain of influenza A (H1N1): early findings. *Science (New York, NY),* **324** (5934), 1557–1561.

Freire-Maia, N. (1979). Note: Neutralist hypothesis is darwinian. *Annals of Human Genetics,* **42** (4), 531–531.

Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol Biol Evol,* **18** (5), 866–73.

Gambaryan, A. S., Tuzikov, A. B., Bovin, N. V., Yamnikova, S. S., Lvov, D. K., Webster, R. G., & Matrosovich, M. N. (2003). Differences between influenza virus receptors on target cells of duck and chicken and receptor specificity of the 1997 H5N1 chicken and human influenza viruses from Hong Kong. *Avian diseases,* **47** (3 Suppl), 1154–1160.

Gibbs, M. J. & Gibbs, A. J. (2006). Molecular virology: was the 1918 pandemic caused by a bird flu? *Nature,* **440** (7088), E8; discussion E9–10.

Gillespie, J. (1994). *The causes of molecular evolution*. USA: Oxford University Press.

Glinka, S., Ometto, L., Mousset, S., Stephan, W., & De Lorenzo, D. (2003). Demography and natural selection have shaped genetic variation in drosophila melanogaster: a multi-locus approach. *Genetics,* **165** (3), 1269–78.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution,* **36** (2), 182–198.

Goldman, N., Thorne, J. L., & Jones, D. T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics,* **149** (1), 445–458.

Goldman, N. & Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution,* **11** (5), 725–736.

Gu, X. (1999). Statistical methods for testing functional divergence after gene duplication. *Molecular biology and evolution,* **16** (12), 1664–1674.

Gu, X. (2001). Maximum-likelihood approach for gene family evolution under functional divergence. *Molecular biology and evolution,* **18** (4), 453–464.

Gu, X., Fu, Y. X., & Li, W. H. (1995). Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Molecular biology and evolution,* **12** (4), 546–557.

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology,* **52** (5), 696–704.

Guindon, S., Rodrigo, A. G., Dyer, K. A., & Huelsenbeck, J. P. (2004). Modeling the site-specific variation of selection patterns along lineages. *Proceedings of the National Academy of Sciences of the United States of America,* **101** (35), 12957–12962.

Halpern, A. L. & Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol,* **15** (7), 910–7.

Hartl, D. L. (1980). *Principles of population genetics*. Sunderland, Mass.: Sinauer Associates.

Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution,* **22** (2), 160–174.

Hatta, M., Gao, P., Halfmann, P., & Kawaoka, Y. (2001). Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. *Science (New York, NY),* **293** (5536), 1840–1842.

Herfst, S., Schrauwen, E. J. A., Linster, M., Chutinimitkul, S., de Wit, E., Munster, V. J., Sorrell, E. M., Bestebroer, T. M., Burke, D. F., Smith, D. J., Rimmelzwaan, G. F., Osterhaus, A. D. M. E., & Fouchier, R. A. M. (2012). Airborne transmission of influenza a/h5n1 virus between ferrets. *Science,* **336** (6088), 1534–1541.

Hietpas, R. T., Jensen, J. D., & Bolon, D. N. (2011). Experimental illumination of a fitness landscape. *Proc Natl Acad Sci U S A,* **108** (19), 7896–901.

Hill, W. G. & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genet. Res.* **8** (269-294).

Holder, M. T., Zwickl, D. J., & Dessimoz, C. (2008). Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. *Philos Trans R Soc Lond B Biol Sci,* **363** (1512), 4013–21.

Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Systematic Biology,* **44** (1), 17–48.

Hughes, A. L. (2007). Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity,* **99** (4), 364–373.

JAMA (1918). Is influenza due to a filtrable virus? *The Journal of the American Medical Association,* **71** (26), 2154–2155.

Johnson, N. P. A. S. & Mueller, J. (2002). Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic. *Bulletin of the history of medicine,* **76** (1), 105–115.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1994). A mutation data matrix for transmembrane proteins. *FEBS letters,* **339** (3), 269–275.

Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian protein metabolism,* **III**, 21–132.

Kawaoka, Y., Krauss, S., & Webster, R. G. (1989). Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. *Journal of Virology,* **63** (11), 4603–4608.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature,* **217** (5129), 624–6.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics,* **61** (4), 893–903.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution,* **16** (2), 111–120.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge, UK: Cambridge University Press.

Kirby, D. A. & Stephan, W. (1996). Multi-locus selection and the structure of variation at the white gene of drosophila melanogaster. *Genetics,* **144** (2), 635–45.

Kleinman, C. L., Rodrigue, N., Lartillot, N., & Philippe, H. (2010). Statistical potentials for improved structurally constrained evolutionary models. *Mol Biol Evol,* **27** (7), 1546–60.

Knobler, S., Mack, A., & Mahmoud, A. (2005). The Story of Influenza. In: *The Threat of Pandemic Influenza: Are We Ready? Workshop Summary*, (Lemon, S., ed) volume 1 , Washington, D.C.: The National Academies Press.

Knudsen, B. & Miyamoto, M. (2001). A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proceedings of the National Academy of Sciences,* **98** (25), 14512.

Kosakovsky Pond, S. L., Poon, A. F. Y., Leigh-Brown, A. J., & Frost, S. D. W. (2008). A maximum likelihood method for detecting directional evolution in protein sequences and its application to Influenza A virus. *Molecular biology and evolution,* **25** (9), 1809.

Koshi, J. M. & Goldstein, R. A. (1995). Context-dependent optimal substitution matrices. *Protein engineering,* **8** (7), 641–645.

Koshi, J. M. & Goldstein, R. A. (1996). Probabilistic reconstruction of ancestral protein sequences. *Journal of Molecular Evolution,* **42** (2), 313–320.

Koshi, J. M. & Goldstein, R. A. (1997). Mutation matrices and physical-chemical properties: correlations and implications. *Proteins,* **27** (3), 336–344.

Koshi, J. M. & Goldstein, R. A. (1998). Models of natural mutations including site heterogeneity. *Proteins,* **32** (3), 289–95.

Koshi, J. M. & Goldstein, R. A. (2001). Analyzing site heterogeneity during protein evolution. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing,* **6**, 191–202.

Kosiol, C., Holmes, I., & Goldman, N. (2007). An empirical codon model for protein sequence evolution. *Mol Biol Evol,* **24** (7), 1464–79.

Krasnitz, M., Levine, A. J., & Rabadan, R. (2008). Anomalies in the influenza virus genome database: new biology or laboratory errors? *Journal of Virology,* **82** (17), 8947–8950.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics,* **22**, 79–86.

Kumar, P. & Clark, M., eds (1999). *Clinical Medicine*. Philadelphia, USA: Saunders Ltd.

Lakner, C., Holder, M. T., Goldman, N., & Naylor, G. J. (2011). What's in a likelihood? simple models of protein evolution and the contribution of structurally viable reconstructions to the likelihood. *Syst Biol,* **60** (2), 161–74.

Lam, T. T.-Y., Hon, C.-C., & Tang, J. W. (2010). Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. *Critical reviews in clinical laboratory sciences,* **47** (1), 5–49.

Lartillot, N., Brinkmann, H., & Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC evolutionary biology,* **7** (Suppl 1), S4.

Lartillot, N. & Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution,* **21** (6), 1095.

Li, W. H. (1978). Maintenance of Genetic Variability under the Joint Effect of Mutation, Selection and Random Drift. *Genetics,* **90** (2), 349–382.

Lin, Y. P., Shaw, M. W., Gregory, V., Cameron, K., Lim, W., Klimov, A., Subbarao, K., Guan, Y., Krauss, S., Shortridge, K., Webster, R. G., Cox, N. J., & Hay, A. J. (2000). Avian-to-human transmission of H9N2 subtype influenza A viruses: relationship between H9N2 and H5N1 human isolates. *Proceedings of the National Academy of Sciences of the United States of America,* **97** (17), 9654–9658.

Liò, P. & Goldman, N. (1998). Models of molecular evolution and phylogeny. *Genome research,* **8** (12), 1233–1244.

Lopez, P., Casane, D., & Philippe, H. (2002). Heterotachy, an important process of protein evolution. *Molecular biology and evolution,* **19** (1), 1–7.

Loytynoja, A. & Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science,* **320** (5883), 1632–5.

Lycett, S. J., Ward, M. J., Lewis, F. I., Poon, A. F. Y., Kosakovsky Pond, S. L., & Leigh-Brown, A. J. (2009). Detection of mammalian virulence determinants in highly pathogenic avian influenza H5N1 viruses: multivariate analysis of published data. *Journal of Virology,* **83** (19), 9901–9910.

Lynch, M. (2007). *The origins of genome architecture*. Sunderland, MA: Sinauer Assoc.

Matrosovich, M., Tuzikov, A., Bovin, N., Gambaryan, A., Klimov, A., Castrucci, M. R., Donatelli, I., & Kawaoka, Y. (2000). Early alterations of the receptor-binding properties of H1, H2, and H3 avian influenza virus hemagglutinins after their introduction into mammals. *Journal of Virology,* **74** (18), 8502–8512.

Maynard-Smith, J. & Haigh, G. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35.

Mayr, E. (1942). *Systematics and the origin of species*. New York: Columbia University Press.

McDonald, J. H. & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature,* **351** (6328), 652–654.

Merriwether, D. A., Clark, A. G., Ballinger, S. W., Schurr, T. G., Soodyall, H., Jenkins, T., Sherry, S. T., & Wallace, D. C. (1991). The structure of human mitochondrial dna variation. *J Mol Evol,* **33** (6), 543–55.

Miotto, O., Heiny, A., Tan, T. W., August, J. T., & Brusic, V. (2008). Identification of human-to-human transmissibility factors in PB2 proteins of influenza A by large-scale mutual information analysis. *BMC bioinformatics,* **9 Suppl 1**, S18.

Muse, S. V. & Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular biology and evolution,* **11** (5), 715–724.

Naffakh, N., Tomoiu, A., Rameix-Welti, M.-A., & Van Der Werf, S. (2008). Host restriction of avian influenza viruses at the level of the ribonucleoproteins. *Annual Review of Microbiology,* **62**, 403–424.

Nakajima, K., Desselberger, U., & Palese, P. (1978). Recent human influenza A (H1N1) viruses are closely related genetically to strains isolated in 1950. *Nature,* **274** (5669), 334–339.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal,* **7** (4), 308–313.

Nelson, M. I. & Holmes, E. C. (2007). The evolution of epidemic influenza. *Nature Reviews Genetics,* **8** (3), 196–205.

Nielsen, R. & Yang, Z. (2003). Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol,* **20** (8), 1231–9.

Noble, W. S. (2009). How does multiple testing correction work? *Nature Publishing Group,* **27** (12), 1135–1137.

Nobusawa, E., Aoyama, T., Kato, H., Suzuki, Y., Tateno, Y., & Nakajima, K. (1991). Comparison of complete amino acid sequences and receptor-binding properties among 13 serotypes of hemagglutinins of influenza A viruses. *Virology,* **182** (2), 475–485.

Nocedal, J. & Wright, S. J. (2006). *Numerical optimization*. New York, USA: Springer Science+Business Media, LLC., second edition.

Novel Swine-Origin Influenza A H1N1 Virus Investigation Team, Dawood, F. S., Jain, S., Finelli, L., Shaw, M. W., Lindstrom, S., Garten, R. J., Gubareva, L. V., Xu, X., Bridges, C. B., & Uyeki, T. M. (2009). Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *The New England journal of medicine,* **360** (25), 2605–2615.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature,* **246** (5428), 96–8.

Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst,* **23** (263-286).

Orr, H. A. (2003). The distribution of fitness effects among beneficial mutations. *Genetics,* **163** (4), 1519–26.

Palese, P. (2004). Influenza: old and new threats. *Nature medicine,* **10** (12 Suppl), S82–7.

Penn, O., Stern, A., Rubinstein, N., Dutheil, J., Bacharach, E., Galtier, N., & Pupko, T. (2008). Evolutionary modeling of rate shifts reveals specificity determinants in HIV-1 subtypes. *PLoS Computational Biology,* **4** (11).

Penny, D., McComish, B. J., Charleston, M. A., & Hendy, M. D. (2001). Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *Journal of Molecular Evolution,* **53** (6), 711–23.

Pensaert, M., Ottis, K., Vandeputte, J., Kaplan, M. M., & Bachmann, P. A. (1981). Evidence for the natural transmission of influenza A virus from wild ducts to swine and its potential importance for man. *Bulletin of the World Health Organization,* **59** (1), 75–78.

Piganeau, G. & Eyre-Walker, A. (2003). Estimating the distribution of fitness effects from dna sequence data: implications for the molecular clock. *Proc Natl Acad Sci U S A,* **100** (18), 10335–40.

Pollock, D. D., Zwickl, D. J., McGuire, J. A., & Hillis, D. M. (2002). Increased Taxon Sampling Is Advantageous for Phylogenetic Inference. *Systematic Biology,* **51** (4), 664–671.

Pupko, T. & Galtier, N. (2002). A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. *Proceedings Biological sciences / The Royal Society,* **269** (1498), 1313–1316.

Pupko, T., Pe'er, I., Shamir, R., & Graur, D. (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular biology and evolution,* **17** (6), 890–896.

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., & Holmes, E. C. (2008). The genomic and epidemiological dynamics of human influenza A virus. *Nature,* **453** (7195), 615–9.

Reid, A. H., Taubenberger, J. K., & Fanning, T. G. (2004). Evidence of an absence: the genetic origins of the 1918 pandemic influenza virus. *Nature Reviews Microbiology,* **2** (11), 909–914.

Robinson, D. M., Jones, D. T., Kishino, H., Goldman, N., & Thorne, J. L. (2003). Protein evolution with dependence among codons due to tertiary structure. *Molecular biology and evolution,* **20** (10), 1692–1704.

Rodrigue, N., Philippe, H., & Lartillot, N. (2006). Assessing site-interdependent phylogenetic models of sequence evolution. *Mol Biol Evol,* **23** (9), 1762–75.

Rodrigue, N., Philippe, H., & Lartillot, N. (2010). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci U S A,* **107** (10), 4629–34.

Rogers, G. N., Paulson, J. C., Daniels, R. S., Skehel, J. J., Wilson, I. A., & Wiley, D. C. (1983). Single amino acid substitutions in influenza haemagglutinin change receptor binding specificity. *Nature,* **304** (5921), 76–78.

Russell, C. A., Fonville, J. M., Brown, A. E. X., Burke, D. F., Smith, D. L., James, S. L., Herfst, S., van Boheemen, S., Linster, M., Schrauwen, E. J., Katzelnick, L., Moster√≠n, A., Kuiken, T., Maher, E., Neumann, G., Osterhaus, A. D. M. E., Kawaoka, Y., Fouchier, R. A. M., & Smith, D. J. (2012). The potential for respiratory droplet,äìtransmissible a/h5n1 influenza virus to evolve in a mammalian host. *Science,* **336** (6088), 1541–1547.

Salomon, R. & Webster, R. G. (2009). The influenza virus enigma. *Cell,* **136** (3), 402–410.

Sanjuan, R., Moya, A., & Elena, S. F. (2004). The distribution of fitness effects caused by single-nucleotide substitutions in an rna virus. *Proc Natl Acad Sci U S A,* **101** (22), 8396–401.

Sawyer, S. A. & Hartl, D. L. (1992). Population genetics of polymorphism and divergence. *Genetics,* **132** (4), 1161–76.

Sawyer, S. A., Kulathinal, R. J., Bustamante, C. D., & Hartl, D. L. (2003). Bayesian analysis suggests that most amino acid replacements in drosophila are driven by positive selection. *J Mol Evol,* **57 Suppl 1**, S154–64.

Sawyer, S. A., Parsch, J., Zhang, Z., & Hartl, D. L. (2007). Prevalence of positive selection among nearly neutral amino acid replacements in drosophila. *Proc Natl Acad Sci U S A,* **104** (16), 6504–10.

Schäfer, J. R., Kawaoka, Y., Bean, W. J., Süss, J., Senne, D., & Webster, R. G. (1993). Origin of the pandemic 1957 H2 influenza A virus and the persistence of its possible progenitors in the avian reservoir. *Virology,* **194** (2), 781–788.

Schneider, A. & Cannarozzi, G. M. (2012). Empirical and semi-empirical models of codon evolution. In: *Codon Evolution: Mechanisms and Models*, (Cannarozzi, G. M. & Schneider, A., eds) pp. 73–96. Oxford University Press New York, USA.

Scholtissek, C. (2008). *Avian influenza* volume 27 chapter History of research on avian influenza, pp. 101–117. Basel (Switzerland): Kager.

Shackelton, L. A., Parrish, C. R., & Holmes, E. C. (2006). Evolutionary basis of codon usage and nucleotide composition bias in vertebrate dna viruses. *J Mol Evol,* , **62** (5), 551–63.

Sheerar, M. G., Easterday, B. C., & Hinshaw, V. S. (1989). Antigenic conservation of H1N1 swine influenza viruses. *The Journal of General Virology,* , **70 (Pt 12)**, 3297–3303.

Smith, G. J. D., Bahl, J., Vijaykrishna, D., Zhang, J., Poon, L. L. M., Chen, H., Webster, R. G., Peiris, J. S. M., & Guan, Y. (2009a). Dating the emergence of pandemic influenza viruses. *Proceedings of the National Academy of Sciences,* , **106** (28), 11709–11712.

Smith, G. J. D., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K., Cheung, C. L., Raghwani, J., Bhatt, S., Peiris, J. S. M., Guan, Y., & Rambaut, A. (2009b). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature,* , **459** (7250), 1122–1125.

Stamatakis, A., Ludwig, T., & Meier, H. (2005). Raxml-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics,* , **21** (4), 456–63.

Steel, J., Lowen, A., Mubareka, S., & Palese, P. (2009). Transmission of influenza virus in a mammalian host is increased by PB2 amino acids 627K or 627E/701N. *PLoS Pathogens,* , **5** (1).

Stephan, W. (1995). Perturbation analysis of a two-locus model with directional selection and recombination. *J Math Biol,* , **34** (1), 95–109.

Stuart, A., Ord, J. K., & Arnold, S. (1999). *Advanced theory of statistics: classical inference and the linear model*, volume 2A. London: Arnold.

Subbarao, E. K., London, W., & Murphy, B. R. (1993). A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. *Journal of Virology,* **67** (4), 1761–1764.

Subramanian, S. & Kumar, S. (2006). Higher intensity of purifying selection on >90revealed by the intrinsic replacement mutation rates. *Mol Biol Evol,* **23** (12), 2283–7.

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research,* **34** (Web Server issue), W609–W612.

Tamuri, A. U., dos Reis, M., & Goldstein, R. A. (2012). Estimating the Distribution of Selection Coefficients from Phylogenetic Data Using Sitewise Mutation-Selection Models. *Genetics,* **190** (3), 1101–1115.

Tamuri, A. U., dos Reis, M., Hay, A. J., & Goldstein, R. A. (2009). Identifying Changes in Selective Constraints: Host Shifts in Influenza. *PLoS Computational Biology,* **5** (11), e1000564.

Tarendeau, F., Crepin, T., Guilligay, D., Ruigrok, R. W. H., Cusack, S., & Hart, D. J. (2008). Host determinant residue lysine 627 lies on the surface of a discrete, folded domain of influenza virus polymerase PB2 subunit. *PLoS Pathogens,* **4** (8).

Taubenberger, J. & Morens, D. (2006). 1918 influenza: the mother of all pandemics. *Emerg Infect Dis,* **12** (1), 15–22.

Taubenberger, J. K. (2006). The origin and virulence of the 1918 "Spanish" influenza virus. *Proceedings of the American Philosophical Society,* **150** (1), 86.

Taubenberger, J. K. & Kash, J. C. (2010). Influenza virus evolution, host adaptation, and pandemic formation. *Cell Host Microbe,* **7** (6), 440–51.

Taubenberger, J. K., Reid, A. H., Lourens, R. M., Wang, R., Jin, G., & Fanning, T. G. (2005). Characterization of the 1918 influenza virus polymerase genes. *Nature,* **437** (7060), 889–893.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *American Mathematical Society: Lectures on Mathematics in the Life Sciences,* **17**, 57–86.

Thorne, J. L., Choi, S. C., Yu, J., Higgs, P. G., & Kishino, H. (2007). Population genetics without intraspecific data. *Mol Biol Evol,* **24** (8), 1667–77.

van Riel, D., Munster, V. J., de Wit, E., Rimmelzwaan, G. F., Fouchier, R. A. M., Osterhaus, A. D. M. E., & Kuiken, T. (2007). Human and avian influenza viruses target different cells in the lower respiratory tract of humans and other mammals. *Am J Pathol,* **171** (4), 1215–23.

Vincent, A. L., Lager, K. M., Ma, W., Lekcharoensuk, P., Gramer, M. R., Loiacono, C., &

Richt, J. A. (2006). Evaluation of hemagglutinin subtype 1 swine influenza viruses from the United States. *Veterinary microbiology,* **118** (3-4), 212–222.

Vines, A., Wells, K., Matrosovich, M. N., Castrucci, M. R., Ito, T., & Kawaoka, Y. (1998). The role of influenza A virus hemagglutinin residues 226 and 228 in receptor specificity and host range restriction. *Journal of Virology,* **72** (9), 7626–7631.

Wang, H.-C., Li, K., Susko, E., & Roger, A. J. (2008). A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC evolutionary biology,* **8** (1), 331.

Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M., & Kawaoka, Y. (1992). Evolution and ecology of influenza A viruses. *Microbiological reviews,* **56** (1), 152–179.

Whelan, S. & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution,* **18** (5), 691.

Williams, P. D., Pollock, D. D., Blackburne, B. P., & Goldstein, R. A. (2006). Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Computational Biology,* **2** (6), e69.

Wilson, J. C. & von Itzstein, M. (2003). Recent strategies in the search for new anti-influenza therapies. *Curr Drug Targets,* **4** (5), 389–408.

Wloch, D. M., Szafraniec, K., Borts, R. H., & Korona, R. (2001). Direct estimate of the mutation rate and the distribution of fitness effects in the yeast saccharomyces cerevisiae. *Genetics,* **159** (2), 441–52.

Wright, S. (1931). Evolution in mendelian populations. *Genetics,* **16**, 97–159.

Yamada, S., Hatta, M., Staker, B. L., Watanabe, S., Imai, M., Shinya, K., Sakai-Tagawa, Y., Ito, M., Ozawa, M., Watanabe, T., Sakabe, S., Li, C., Kim, J. H., Myler, P. J., Phan, I., Raymond, A., Smith, E., Stacy, R., Nidom, C. A., Lank, S. M., Wiseman, R. W., Bimber, B. N., O'Connor, D. H., Neumann, G., Stewart, L. J., & Kawaoka, Y. (2010). Biological and structural characterization of a host-adapting amino acid in influenza virus. *PLoS Pathogens,* **6** (8).

Yampolsky, L. Y., Kondrashov, F. A., & Kondrashov, A. S. (2005). Distribution of the strength of selection against amino acid replacements in human proteins. *Hum Mol Genet,* **14** (21), 3191–201.

Yang, Z. (1994a). Estimating the pattern of nucleotide substitution. *J Mol Evol,* **39** (1), 105–11.

Yang, Z. (1994b). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution,* **39** (3), 306–314.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *TREE,* **11** (9), 367–372.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences,* **13** (5), 555–556.

Yang, Z. (2006). *Computational molecular evolution*. University College London: Oxford University Press.

Yang, Z. (2007a). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution,* **24** (8), 1586–1591.

Yang, Z. (2007b). Adaptive molecular evolution. In: *Handbook of statistical genetics*, (Balding, D., Bishop, M., & Cannings, C., eds). Wiley New York, USA 2 edition.

Yang, Z. & dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. *Molecular biology and evolution,* **28** (3), 1217–1228.

Yang, Z., Goldman, N., & Friday, A. (1994). Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol Biol Evol,* **11** (2), 316–324.

Yang, Z. & Kumar, S. (1996). Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Mol Biol Evol,* **13** (5), 650–9.

Yang, Z. & Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular biology and evolution,* **19** (6), 908–917.

Yang, Z. & Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular biology and evolution,* **25** (3), 568–579.

Yang, Z. & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics,* **13** (5), 303–314.

Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution,* **22** (12), 2472–2479.

Zhou, N. N., Senne, D. A., Landgraf, J. S., Swenson, S. L., Erickson, G., Rossow, K., Liu, L.,

Yoon, K. j., Krauss, S., & Webster, R. G. (1999). Genetic reassortment of avian, swine, and human influenza A viruses in American pigs. *Journal of Virology,* **73** (10), 8851–8856.

Zuckerkandl, E. & Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of theoretical biology,* **8** (2), 357–366.

A. Accession numbers for sequences used in chapter 3

H1

AX56530 AAY78939 AAZ38627 AAZ74374 AAZ79538 AAZ79604 AAZ83253 AAZ83977 ABA08475 ABA06510 ABA08497 ABA06542 ABA08508 ABA08519 ABA12715 ABA42236 ABA43189 ABA42575 ABA87080 ABB02936 ABB03101 ABB03123 ABB03134 ABB04972 ABB19507 ABB19518 ABB19529 ABB19540 ABB19551 ABB19562 ABB19571 ABB19574 ABB19607 ABB19618 ABB19628 ABB19667 ABB20429 ABB21772 ABB53729 ABB53740 ABB79979 ABB79990 ABB83127 ABB83138 ABB82216 ABB96487 ABC40522 ABC42750 ABC86237 ABD15258 ABD60856 ABD60867 ABD62843 ABD60933 ABD60944 ABD60955 ABD62781 ABD61735 ABD60966 ABD79101 ABD79112 ABD77796 ABD77807 ABD77818 ABD77917 ABD77972 ABD78082 ABD94756 ABD94800 ABD94976 ABD95031 ABD95108 ABD95130 ABD95185 ABD95317 ABD95339 ABD95350 ABE11701 ABE11867 ABE11878 ABE11889 ABE11922 ABE11942 ABE12032 ABF47561 ABF47955 ABF47660 ABF47748 ABF47759 ABF47891 ABG88212 ABF82673 ABF82863 ABF82874 ABG26791 ABG26945 ABG37362 ABG47840 ABG80172 ABG80183 ABG88201 ABG88256 ABG88300 ABG88322 ABG88333 ABI20826 ABI20870 ABI21189 ABI21222 ABI21519 ABI21530 ABI21552 ABI84478 ABI84617 ABI84855 ABI84948 ABI85225 ABI85231 ABI92181 ABI92313 ABI95250 ABI96088 ABI96091 ABI96093 ABI96096 ABI96099 ABI96102 ABI96103 ABI96104 ABI96106 ABI96107 ABI96108 ABI96109 ABI96110 ABI96111 ABI96112 ABI96113 ABI96114 ABI96115 ABI96116 ABI96118 ABI96119 ABI96122 ABI96124 ABI96125 ABI96126 ABI96127 ABI96128 ABI96132 ABI96134 ABI96135 ABI96140 ABI96141 ABI96142 ABI96143 ABI96144 ABI96145 ABI96146 ABI96147 ABI96148 ABI96149 ABI96151 ABI96152 ABI96156 ABI96159 ABI96161 ABI96162 ABI96163 ABI96164 ABI96165 ABI96166 ABI96169 ABI96171 ABI96173 ABI96174 ABJ09151 ABJ09184 ABJ16609 ABJ16642 ABJ16653 ABJ16664 ABJ53504 ABK40028 ABK40546 ABK40590 ABK40634 ABL67264 ABM21960 ABM22026 ABM22246 ABM22279 ABN50756 ABN50900 ABN50940 ABN50962 ABN51066 ABN59401 ABN59423 ABN59434 ABO32948 ABO32959 ABO32981 ABO32992 ABO33006 ABO33025 ABO38010 ABO38021 ABO38032 ABO38054 ABO38065 ABO38340 ABO38351 ABO38362 ABO38373 ABO38384 ABO38395 ABO38406 ABO44046 ABO44134 ABO52038 ABO52225 ABO52258 ABO52797 ABP49305 ABP49316 ABP49327 ABP49349 ABP49360 ABP49393 ABP49448 ABP49481 ABQ01322 ABQ44471 ABR15896 ABS49987 ABU80298 ABU80309 ABV29535 ABV29546 ABV29557 ABV29568 ABV29601 ABV29612 ABV29634 ABV29656 ABV29678 ABV29755 ABV29854 ABV29975 ABV30041 ABV30052 ABV30195 ABV30360 ABV30459 ABV30569 ABV30613 ABV45849 ABV45937 ABV82551 ABW36256 ABW36289 ABW36311 ABW39828 ABW39839 ABW39850 ABW39883 ABW39916 ABW39971 ABW40048 ABW40092 ABW40114 ABW40279 ABW40290 ABW40301 ABW40543 ABW40576 ABW71393 ABW86398 ABW86519 ABW86541 ABW91328 ABW91361 ABW91416 ABW91515 ABW91526 ABW91559 ABW91614 ABX58261 ABX58360 ABX58415 ABX58514 ABX58536 ABX58547 ABX58569 ABX58602 ABX58635 ABY51039 ABY51072 ABY51138 ABY81349 ACA03717 ACA03722 ACA03723 ACA03724 ACA03726 ACA03729 ACA03730 ACA03731 ACA03732 ACA03734 ACA03735 ACA03736 ACA03742 ACA03744 ACA03745 ACA03746 ACA03748 ACA03751 ACA03753 ACA03754 ACA03755 ACA03759 ACA03761 ACA03764 ACA03766 ACA03767 ACA04508 ACA28714 ACA28717 ACA28718 ACA28846 ACA35062 ACB05981 ACB05983 ACB05985 ACB05988 ACB05990 ACB05992 ACC61975 ACD13233 ACD13235 ACD13247 ACD37421 ACD37424 ACD37427 ACD37430 ACD37433 ACD37436 ACD37442 ACD37451 ACD37457 ACD37460 ACD37463 ACD37469 ACD37472 ACD37475 ACD37481 ACD37487 ACD37492 ACD37501 ACD37504 ACD45705 ACD45706 ACD45723 ACD45731 ACD45735 ACD45748 ACD45750 ACD45752 ACD45756 ACD45762 ACD45768 ACD45776 ACD45779 ACD45780 ACD45784 ACD45794 ACD45804 ACD45818 ACD45819 ACD45820 ACD45822 ACD45829 ACD45832 ACD45839 ACD56280 ACD85143 ACF41834 ACF41867 ACF54598 ACH69166 ACH69173 ACH69174 ACH69176 ACH69177 ACH69181 ACH69188 ACH69190 ACH69192 ACH69193 ACH69194 ACH69201 ACH69203 ACH69211 ACH69216 ACH69221 ACH69224 ACH69227 ACH69233 ACH69236 ACH69237 ACH69241 ACH69246 ACH69248 ACH69249 ACH69251 ACH69260 ACH88839 ACI26450 ACK99009 ACK99015 ACK99019 ACK99026 ACK99028 ACK99029 ACK99034 ACK99035 ACK99037 ACK99443 ACK99465 ACL12261 ACN32793 ACN33090 AAD17229

H2

ABB17692 ABB17725 ABB18036 ABI84450 ABM21949 ABO38307 ABO44090 ABO52247 ABO52379 ACD56324 ACF54477 ABB17714 ABB18378 ABB17736 ABB17813 ABB18025 ABB18080 ABB19639 ABB20141 ABB20466 ABB20509 ABI84459 ABI84744 ABL67022 ABO38723 ABO44057 ABO52302 ABP49470 ACD85198 ACD85231 ACD85242 ACF41691 ACI26384 ACJ69319 ACJ69324 ABB17670 ABB17756 ABO38296 ABO38734 ABO52236 ABP49459 ABQ01355 ABQ44460 ACD56302 ACD56313 ACD85220 ACF47420 ACF54389 ABB17150 ABB17681 ABB17703 ABB18047 ABB18069 ABB20229 ABB20240 ABI84382 ABI84384 ABI84458 ABI84588 ABI84755 ABI84959 ABI85183 ABO38098 ABO38701 ABQ44438 ACD56291 ACD85187 ACD85209 ACD85253 ACF54488 ACI25724

H3

AAX11455 AAX11475 AAX11485 AAX11495 AAX11515 AAX11575 AAX11635 AAY28571 AAX11625 AAX12771 AAX12781 AAX12791 AAX12801 AAX47525 AAX47515 AAX56440 AAX56460 AAX56510 AAX56540 AAX57644 AAX57904 AAX57944 AAX76623 AAX76653 AAX76703 AAX76733 AAX76743 AAY18126 AAY27863 AAY28325 AAY28335 AAY28405 AAY44775 AAY44755 AAY44661 AAY46371 AAY47023 AAY46416 AAY46426 AAY46436 AAY64192 AAY64212 AAY64252 AAY64372 AAY98117 AAY98127 AAY98137 AAY98329 AAZ38462 AAZ38506 AAZ38528 AAZ38583 AAZ38550 AAZ38572 AAZ43370 AAZ43383 AAZ74397 AAZ74540 AAZ74606 AAZ74584 AAZ79593 AAZ79615 AAZ79985 AAZ80007 AAZ83242 AAZ83266 AAZ83323 AAZ83649 ABA16214 ABA26700 ABA26722 ABA43336 ABA42368 ABA42401 ABB96509 ABB04283 ABB04294 ABB04305 ABB04327 ABB04371 ABB03046 ABB03112 ABB04906 ABB04917 ABB04928 ABB04939 ABB04961 ABB05183 ABB05194 ABB05205 ABB05005 ABB19704 ABB19712 ABB19744 ABB19758 ABB86785 ABB87034 ABB87410 ABB87429 ABB87462 ABB88149 ABB88152 ABB88162 ABB88173 ABB88183 ABB88256 ABB88309 ABB88342 ABB88369 ABB46547 ABB46392 ABB46403 ABB46425 ABB53674 ABB53696 ABB53751 ABB54514 ABB52376 ABB77853 ABB59996 ABB80034 ABB80023 ABB79788 ABB80001 ABB80185 ABB80503 ABB80748 ABB82227 ABB96319 ABB96330 ABB96352 ABB96363 ABB96374 ABB96498 ABB96520 ABB96531 ABC02234 ABC39805 ABC40555 ABC40608 ABC40619 ABC41692 ABC43072 ABC42629 ABC42728 ABC42871 ABC46554 ABC67850 ABC67989 ABC68049 ABC68093 ABC67554 ABC67576 ABC67664 ABC67686 ABC67697 ABC67872 ABC84389 ABC85952 ABC85897 ABC85875 ABC84520 ABC84531 ABD16527 ABD16582 ABD16571 ABD15746 ABD15713 ABD15691 ABD15625 ABD60790 ABD61777 ABD60834 ABE12532 ABE12623 ABD77664 ABD77686 ABD77697 ABD79032 ABD77829 ABD79189 ABD77840 ABD77862 ABD77873 ABD77895 ABD79244 ABD94844 ABD94866 ABE11911 ABE13555 ABE13595 ABE14840 ABF47858 ABI47947 ABI48006 ABF83447 ABG26846 ABG26857 ABG26868 ABG26890 ABG37153 ABG37175 ABG37186 ABG37197 ABG37219 ABG37230 ABG37274 ABG37373 ABG47862 ABG47950 ABG47961 ABG48137 ABG48258 ABG48280 ABG48291 ABG48313 ABG48324 ABG48346 ABG48368 ABG67157 ABG67234 ABG67502 ABG67656 ABG80073 ABG80139 ABG80161 ABG80315 ABG80359 ABG88454 ABG88630 ABG88762 ABG88773 ABI20815 ABI21167 ABI21244 ABI21420 ABI30576 ABI30876 ABI84400 ABI84412 ABI84471 ABI84486 ABI84577 ABI84806 ABI84938 ABI92280 ABI92335 ABI92412 ABI92445 ABI92511 ABI92522 ABI92621 ABI92643 ABI92830 ABI92896 ABI92940 ABI92995 ABI93116 ABI95474 ABJ09118 ABJ09261 ABJ09272 ABJ16587 ABJ16708 ABJ16741 ABJ53460 ABK40623 ABK40667 ABK80014 ABK80179 ABL67110 ABL67132 ABL67165 ABL67220 ABL75563 ABM66853 ABM67029 ABN51121 ABO32803 ABO33069 ABO38230 ABO51862 ABO52071 ABO52126 ABO52313 ABO52335 ABO52357 ABO52566 ABO64343 ABP49184 ABP49404 ABQ01366 ABR37451 ABR68692 ABR68693 ABR68694 ABR68695 ABR68700 ABR68702 ABR68709 ABR87638 ABR87639 ABR87640 ABR87642 ABR87643 ABR87644 ABR87651 ABS00286 ABS00293 ABS00297 ABS00300 ABS11173 ABS11180 ABS11181 ABS11188 ABS11189 ABS11192 ABS50031 ABS50064 ABU80320 ABV29909 ABV30239 ABV30261 ABV72859 ABV72873 ABV72896 ABV72897 ABV72907 ABV72911 ABV72918 ABV72920 ABV72929 ABV72931 ABV72932 ABV72938 ABV72948 ABV72949 ABW40169 ABY51503 ABY51569 ACA24174 ACA24176 ACA24177 ACA24180 ACA24187 ACA24188 ACA28720 ACA28721 ACA65917 ACA65922 ACD13251 ACD13252 ACD13263 ACD62281 ACD62287 ACD62317 ACD62320 ACD62323 ACD62329 ACD62332 ACD62335 ACD62347 ACD62356 ACD62362 ACD62365 ACD69088 ACD69092 ACD69146 ACD69176 ACD69232 ACD69241 ACD69256 ACD69261 ACD69263 ACD69360 ACD69391 ACD85528 ACE76559 ACE76625 ACF22159 ACF41735 ACF41746 ACF41779 ACF41812 ACF41856 ACF41900 ACF48909 ACF48912 ACF48925 ACF48956 ACF48960 ACF54554 ACF54565 ACH56649 ACH56650 ACI26549 ACI26560 ACI26571 ACI26582 ACI89509 ACI89551 ACI89651 ACI90158 ACI90169 ACI90180 ACK99476 ACL12129 ACN32936 ACN32980 ACN43014

M1

BAB39518 BAD02355 BAD89309 BAE48325 BAF46431 BAF46531 BAE94703 BAF37966 BAF37825 BAF56163 AAC63479 AAC63483 AAD25171 AAD25175 AAD25185 AAD25189 AAD25191 AAD25199 AAD25203 AAD25211 AAD25213 AAD25217 AAD25221 AAD51928 AAD49069 AAD49071 AAD49073 AAD49075 AAD49077 AAD49083 AAD49085 AAD49087 AAD49089 AAG01192 AAG01222 AAK14987 AAF87515 AAF87517 AAF87521 AAF87523 AAF87525 AAF87527 AAF87529 AAF87531 AAK18001 AAF99670 AAF99672 AAG09044 AAK51730 AAM09296 AAK26664 AAK70433 AAM75161 AAL60445 AAM69960 AAM49561 AAO33507 AAO33515 AAO33517 AAQ04982 AAQ04985 AAQ04991 AAO52883 AAO52884 AAO52888 AAO52906 AAP49146 AAP49151 AAP49153 CAC19700 CAC09422 CAC84271 CAC84272 CAC95058 CAD30536 CAD30538 CAD30540 CAD30542 CAD30544 CAF33018 CAJ01905 AAL31412 AAL31414 AAL75841 AAL75849 AAO46348 AAO46382 AAO46388 AAO46394 AAO46406 AAO46408 AAO46414 AAO46420 AAO46669 AAO46693 AAO46699 AAO46810 AAO46811 AAO46813 AAO92779 AAO92783 AAO92795 AAO92815 AAO92819 AAO92825 AAO92841 AAO92845 AAP04510 AAP57582 AAP57586 AAP57588 AAP57590 AAR91539 AAR91540 AAR91541 AAR91542 AAR91544 AAR91545 AAT39095 AAT12044 AAT12045 AAT12046 AAT12047 AAT12049 AAT12052 AAT12055 AAT12061 AAT37566 AAT65437 AAT65446 AAT65452 AAT65453 AAT70571 AAT70589 AAT70593 AAT76159 AAU11168 AAU11182 AAU11202 AAT80681 AAU00828 AAU00830 AAV80801 AAV65818 AAX47287 AAW72232 AAW78060 AAX53521 AAX11496 AAY28572 AAX56491 AAX56531 AAX57675 AAX76734 AAY18197 AAY46437 AAY98138 AAZ38639 AAZ43371 AAZ80008 ABB04361 ABB03113 ABB17671 ABB18392 ABB19021 ABB19218 ABB19451 ABB87196 ABB19501 ABB19629 ABB87378 ABB87540 ABB19887 ABB87730 ABB19978 ABB88023 ABB88056 ABB88078 ABB20102 ABB20142 ABB20166 ABB20230 ABB88299 ABB20398 ABB20445 ABB20490 ABB20517 ABB21794 ABB21814 ABB46437 ABB80186 ABB96320 ABC02289 ABC42751 ABD60857 ABD62844 ABD60934 ABD62782 ABD61736 ABD60967 ABD77676 ABD79102 ABD79113 ABD77797 ABD95032 ABE11890 ABF47584 ABF47892 ABI47971 ABG37242 ABG47951 ABG48303 ABG48325 ABG79964 ABG79986 ABG80129 ABG88279 ABI20805 ABI20838 ABI36026 ABI84505 ABI84508 ABI84535 ABI84546 ABI84567 ABI84627 ABI84674 ABI84705 ABI84745 ABI84867 ABI84917 ABI84971 ABI85012 ABI85039 ABI85058 ABI85085 ABI85096 ABI85138 ABI95251 ABI95262 ABI95317 ABI95350 ABJ16566 ABJ16676 ABJ16929 ABK80224 ABL67166 ABL67320 ABM22247 ABN50941 ABO32668 ABO32752 ABO38066 ABO38341 ABO38385 ABO44135 ABO44157 ABO44168 ABO45249 ABO51863 ABO51918 ABO51973 ABO52138 ABO52149 ABO52226 ABO52699 ABO52787 ABO77057 AAY96432 AAY96437 AAY96448 AAY96458 AAY96483 AAY96486 AAY96493 AAY96501 AAY96509 AAY96511 AAY53536 AAY52538 AAY52562 AAY52572 AAY52574 AAZ29590 AAZ29592 AAZ29594 AAZ29596 AAZ72665

AAZ72675 AAZ72693 AAZ16336 AAZ16345 ABB58925 ABB58933 ABB51969 ABB51971 ABB51973 ABB83604 ABB80547 ABC48819 ABC33912 ABC66631 ABC66638 ABC48793 ABC69235 ABC74393 ABC74395 ABC74397 ABC94731 ABD14825 ABD35559 ABD35565 ABD35567 ABD35571 ABD35583 ABD35589 ABD35605 ABD35609 ABD35611 ABD35621 ABD35623 ABD35629 ABD59881 ABF22653 ABF22659 ABF01764 ABF01792 ABF01832 ABF01848 ABF21300 ABF21302 ABG91468 ABH04388 ABK34768 ABJ90229 ABI94739 ABI98901 ABI97329 ABI97333 ABJ52565 ABK00100 ABI98935 ABJ09507 ABK41619 ABI96773 ABJ15714 ABI94774 ABI94773 ABI98944 ABK13777 ABK13776 ABK32097 ABM21862 ABM21866 ABM21868 ABM21874 ABL07963 ABL08039 ABL08183 ABM46019 ABM46026 ABM46031 ABM46035 ABM46042 ABM46044 ABM46045 ABM46051 ABM46064 ABM46066 ABM46079 ABO76647 ABO30350 AAA43092 AAA91323 AAA56804 AAA56806 AAA56808 AAA43313 AAA43251 AAA43252 AAA43256 AAA43311 AAA43258 AAA43307 AAA43294 AAA43347 AAA43302 AAA43286 AAA19193 AAA67337 AAB50990 CAA30886 AAN06597

M2

AAA19192 AAA43091 AAA43249 AAA43255 AAA43257 AAA43273 AAA43274 AAA43279 AAA43281 AAA43285 AAA43293 AAA43295 AAA43301 AAA43306 AAA43577 AAA56807 AAA67336 AAB50989 AAC63486 AAC80156 AAC80162 AAD00134 AAD00138 AAD25172 AAD25174 AAD25190 AAD25192 AAD25200 AAD25210 AAD25212 AAD25216 AAD25218 AAD25222 AAD49068 AAD49074 AAD49078 AAD49084 AAD51929 AAF70407 AAF87510 AAF87514 AAF87516 AAF87518 AAF87522 AAF87524 AAF99671 AAF99673 AAG01193 AAG01203 AAK26663 AAK70438 AAK70446 AAL60446 AAM49562 AAM69961 AAM69992 AAM75162 AAN06598 AAO46345 AAO46361 AAO46365 AAO46369 AAO46395 AAO46409 AAO46674 AAO46682 AAO46702 AAO46704 AAO88262 AAP04511 AAQ77433 AAR99626 AAT37567 AAT70504 AAT70506 AAT70512 AAT70546 AAT70558 AAT70564 AAT70576 AAT70590 AAT70594 AAT70608 AAT76158 AAU00827 AAU00829 AAU11203 AAX11457 AAX11467 AAX11497 AAX12783 AAX35863 AAX47527 AAX53524 AAX56552 AAX76735 AAX76745 AAY28610 AAY46373 AAY52527 AAY52533 AAY52539 AAY52547 AAY52569 AAY52575 AAY64374 AAY98109 AAY98149 AAY98179 AAY98239 AAZ16335 AAZ30555 AAZ38530 AAZ38640 AAZ74619 AAZ79551 AAZ83384 ABA18147 ABA26724 ABB02783 ABB02838 ABB03015 ABB03092 ABB03114 ABB03136 ABB04285 ABB04340 ABB17705 ABB17716 ABB18393 ABB19022 ABB19164 ABB19219 ABB19373 ABB19406 ABB19452 ABB19474 ABB19502 ABB19620 ABB19641 ABB19880 ABB19937 ABB19959 ABB20231 ABB20256 ABB20399 ABB20425 ABB20484 ABB20491 ABB46416 ABB46438 ABB46460 ABB51972 ABB58926 ABB58930 ABB58936 ABB80003 ABB80083 ABB80198 ABB86798 ABB87036 ABB87219 ABB87336 ABB87680 ABB87835 ABB88090 ABB88154 ABB88269 ABB88381 ABB90184 ABB90216 ABB90232 ABB96321 ABC42752 ABC50400 ABC74394 ABC74396 ABC88579 ABD15561 ABD15770 ABD35566 ABD35568 ABD35570 ABD35582 ABD35584 ABD35586 ABD35588 ABD35592 ABD35594 ABD35596 ABD35606 ABD35620 ABD35622 ABD59880 ABD60968 ABD62783 ABD77677 ABD77809 ABD77820 ABD77897 ABD77974 ABD79103 ABD79114 ABD94989 ABE11703 ABE11880 ABE11891 ABE11944 ABE13641 ABE28417 ABF01753 ABF01755 ABF01791 ABF01797 ABF01807 ABF01809 ABF01817 ABF01835 ABF01855 ABF01877 ABF01887 ABF21309 ABF47574 ABF47673 ABF47728 ABF69261 ABG37232 ABG37364 ABG47908 ABG48370 ABG80163 ABG88214 ABG88258 ABG91467 ABI19014 ABI20806 ABI20828 ABI20872 ABI21191 ABI30812 ABI33770 ABI33778 ABI33780 ABI33782 ABI84509 ABI84536 ABI84568 ABI84579 ABI84715 ABI84746 ABI85040 ABI85049 ABI85097 ABI85108 ABI85119 ABI85129 ABI85147 ABI85157 ABI85164 ABI85174 ABI85210 ABI85227 ABI92832 ABI94740 ABI95340 ABI97314 ABI97330 ABJ09120 ABJ09475 ABJ09508 ABJ15715 ABJ16567 ABJ16765 ABJ16820 ABJ16864 ABJ16941 ABK00112 ABK00126 ABL08120 ABL08130 ABL08182 ABL08184 ABL67101 ABL67798 ABL67809 ABM21873 ABN50758 ABO30351 ABO34222 ABO38023 ABO38353 ABO38386 ABO38408 ABO44147 ABO44158 ABO52084 ABO52447 ABO52612 ABO52733 ABO76981 ABO77047 ABO77058 ABO93189 ABP35638 ABP49197 ABP49373 ABP49395 ABQ09750 ABQ12378 ABQ41368 ABQ57382 ABR31776 ABR37321 ABR37519 ABR37607 ABS00912 ABS17547 ABS50732 ABS50750 ABS50756 ABS50778 ABS52597 ABS54064 ABS54066 ABS54104 ABS54144 ABS54192 ABS54206 ABS54232 ABS54246 ABS70307 ABS70440 ABS89312 ABS89367 ABU50607 ABU99114 ABV29559 ABV29647 ABV29702 ABV29911 ABV30142 ABV31844 ABV31854 ABV31871 ABV45840 ABW75846 ABW86444 ABX88823 ABX88834 ABY51505 ABY70957 ABY74985 ABY75003 ABY81638 ABY89676 ABZ04077 ACA33513 ACA33529 ACA47619 ACB47238 ACC60428 ACC60452 ACC60492 ACC60496 ACC60544 ACC60562 ACC60602 ACC60650 ACC60702 ACC60802 ACC60816 ACC61944 ACD12187 ACD12229 ACD35887 ACD37435 ACD37763 ACD37773 ACD47269 ACD47287 ACD47301 ACD47335 ACD47337 ACD47397 ACD47441 ACD56359 ACD62349 ACD62364 ACD88540 ACD88590 ACF08289 ACF25477 ACF25720 ACF25753 ACF25786 ACF33623 ACF33730 ACF37315 ACF47413 ACF47424 ACF47556 ACF54534 ACH43168 ACH68520 ACH68522 ACH88897 ACJ04578 ACJ12601 ACJ14921 ACJ15075 ACJ15108 ACJ15173 ACJ26378 ACJ68620 ACJ68631 ACK43402 ACL12263 ACL12398 ACL79967 ACL79985 ACN33059 ACN37887 ACN39376 ACN41771 ACN42983 ACO36531 ACO36586 ACO36674 BAF37965 BAF46520 BAF57523 BAF57568 BAF63059 BAG32242 BAG80862 BAG84410 CAA24283 CAD30537 CAD30539 CAD30541 CAD30543

N1

AAF77036 AAX56533 AAZ38630 AAZ79607 AAZ85129 AAZ83256 AAZ83980 AAZ83302 ABA08467 ABA08478 ABA08489 ABA08500 ABA08522 ABA12721 ABA42250 ABA42239 ABA43192 ABA42578 ABA87060 ABA87083 ABA87048 ABA87094 ABA87234 ABB02927 ABB02939 ABB03148 ABB18381 ABB17749 ABB19104 ABB19374 ABB19417 ABB87166 ABB19484 ABB19503 ABB19521 ABB19532 ABB19543 ABB19554 ABB19572 ABB19577 ABB19610 ABB19631 ABB19670 ABB87432 ABB19696 ABB19870 ABB87857 ABB87867 ABB87899 ABB88003 ABB20052 ABB20144 ABB20211 ABB20286 ABB20297 ABB20381 ABB20447 ABB20485 ABB21775 ABB53710 ABB53743 ABB79982 ABB80048 ABB80106 ABB82219 ABC41717 ABC42753 ABC86240 ABD15518 ABD15262 ABD60859 ABD60870 ABD60892 ABD61543 ABD62845 ABD60936 ABD60947 ABD60958 ABD62784 ABD61738 ABD60969 ABD77678 ABD79104 ABD77711 ABD79115 ABD77799 ABD77810 ABD77821 ABD77964 ABD77997 ABD78030 ABD78041 ABD94979 ABD95001 ABD95012 ABD95034 ABD95045 ABD95133 ABD95166 ABD95188 ABD95221 ABD95287 ABD95309 ABD95342 ABE11682 ABE11870 ABE11881 ABE11892 ABE11925 ABE11952 ABE26994 ABF47575 ABF47958 ABF47630 ABF47641 ABF47674 ABF47707 ABF47751 ABF47762 ABF47828 ABG88215 ABF82687 ABF82877

ABG37398 ABG47821 ABG47832 ABG80175 ABG88204 ABG88259 ABG88303 ABG88325 ABG88336 ABG88347 ABG88545 ABI20829 ABI20851 ABI20862 ABI20873 ABI21192 ABI21214 ABI21236 ABI22112 ABI21522 ABI30381 ABI30568 ABI84398 ABI84474 ABI84644 ABI84736 ABI85014 ABI85060 ABI85109 ABI85186 ABI92184 ABI92239 ABI92305 ABI92316 ABI95253 ABI95297 ABI95319 ABI95330 ABI95341 ABJ09187 ABJ16612 ABJ16645 ABJ16722 ABJ16821 ABJ16931 ABJ16920 ABJ16832 ABJ16942 ABJ51731 ABJ51698 ABJ51687 ABJ51676 ABJ16953 ABJ16843 ABJ16854 ABJ16876 ABJ16909 ABJ53518 ABJ53540 ABJ53551 ABJ53597 ABJ53562 ABK40009 ABK40549 ABK40593 ABK79962 ABL67025 ABL67124 ABL67157 ABM21952 ABM21996 ABM22007 ABM22249 ABM22260 ABM22293 ABM66889 ABM67054 ABN50759 ABN50903 ABN50943 ABN51069 ABN59404 ABN59437 ABO32951 ABO32962 ABO33009 ABO33028 ABO38013 ABO38024 ABO38035 ABO38057 ABO38266 ABO38321 ABO38343 ABO38354 ABO38365 ABO38387 ABO38398 ABO38409 ABO44049 ABO44126 ABO44137 ABO44203 ABO44225 ABO44236 ABO44280 ABO45251 ABO44291 ABO52228 ABO52261 ABO52459 ABO52756 ABO52778 ABO52789 ABO64346 ABO77059 ABO77070 ABO77081 ABP49308 ABP49330 ABP49341 ABP49352 ABP49385 ABP49484 ABQ44419 ABQ44474 ABR15888 ABR15899 ABR15921 ABR28771 ABR28848 ABR37377 ABR37388 ABR37399 ABS89335 ABS89346 ABS89467 ABS89478 ABS89522 ABU80312 ABV01192 ABV01236 ABV01247 ABV01258 ABV01269 ABV01280 ABV01291 ABV29538 ABV29549 ABV29560 ABV29604 ABV29637 ABV29681 ABV29747 ABV29758 ABV29780 ABV29846 ABV29857 ABV29868 ABV29879 ABV29956 ABV29967 ABV29978 ABV29989 ABV30033 ABV30044 ABV30143 ABV30176 ABV30286 ABV30319 ABV30352 ABV30528 ABV30550 ABV30605 ABV45852 ABV45940 ABV45962 ABW36182 ABW36237 ABW36259 ABW36270 ABW36314 ABW39809 ABW39831 ABW39853 ABW39886 ABW39919 ABW39952 ABW39974 ABW39996 ABW40040 ABW40073 ABW40095 ABW40106 ABW40348 ABW40370 ABW40392 ABW40414 ABW40447 ABW40480 ABW40535 ABW40568 ABW40579 ABW71341 ABW71396 ABW71407 ABW71429 ABW86467 ABW86489 ABW86500 ABW86522 ABW91364 ABW91419 ABW91430 ABW91529 ABW91540 ABX58297 ABX58330 ABX58374 ABX58451 ABX58495 ABX58682 ABY51053 ABY51086 ABY51097 ABY51174 ABY51240 ABY51583 ABY81352 ABY81396 ABY81407 ABY81418 ACA47419 ACA47430 ACA47441 ACA47452 ACA47463 ACA47540 ACA47551 ACA47573 ACA47595 ACA47606 ACA47617 ACA47628 ACA47639 ACA47650 ACA47672 ACA47716 ACA47738 ACA47771 ACA47782 ACA47793 ACA47804 ACA47815 ACA47903 ACA47914 ACA47947 ACA47980 ACA48035 ACA48123 ACB70473 ACB70484 ACB70495 ACB70506 ACB70539 ACB70550 ACB70561 ACB70572 ACB70583 ACB70605 ACB70638 ACB70649 ACB70660 ACB70671 ACB70682 ACB70693 ACB70704 ACB70715 ACB70726 ACB70770 ACB70792 ACB70803 ACB70825 ABY88907 ACA04511 ACA21568 ACA21612 ACA28845 ACA28847 ACJ14864 ACJ14875 ACJ14897 ACJ14908 ACJ14919 ACJ14952 ACJ14963 ACJ14974 ACJ14985 ACJ15007 ACJ15018 ACJ15029 ACJ15040 ACJ15051 ACJ15073 ACJ15106 ACJ15117 ACJ15150 ACJ15171 ACJ15180 ACJ15221 ACJ15242 ACJ15253 ACJ15264 ACB05982 ACB05984 ACB05986 ACB05989 ACB05991 ACB05993 ACB05995 ACH85379 ACH85390 ACH85401 ACH85456 ACH85467 ACH85511 ACH85522 ACC61978 ACC61989 ACD56283 ACD85146 ACE76550 ACE76616 ACE81749 ACF22444 ACF22466 ACF47444 ACF47499 ACF47532 ACF47543 ACF41870 ACF41881 ACF54601 ACF76204 ACF76237 ACH88841 ACH88862 ACH88873 ACH88895 ACH88906 ACH88847 ACH88853 ACI16722 ACI26453 ACI62848 ACJ09832 ACJ26079 ACJ26090 ACJ26101 ACJ26134 ACJ26222 ACJ26244 ACJ26266 ACJ26277 ACJ26288 ACJ26299 ACJ26310 ACJ26332 ACJ26343 ACJ26354 ACJ26376 ACK99270 ACK99292 ACK99303 ACK99446 ACK99468 ACL12264 ACN32504 ACN32515 ACN32840 ACN33126 ACN41780 ACN41791

N2

AAX11458 AAX11478 AAX11498 AAX11518 AAX11588 AAX11638 AAX12734 AAX11468 AAX11628 AAX12764 AAX12804 AAX35824 AAX35844 AAX47528 AAX35874 AAX56523 AAX56543 AAX56553 AAX56563 AAX56603 AAX57777 AAX57817 AAX57827 AAX57847 AAX57937 AAX57947 AAX76736 AAY18109 AAY18129 AAY18169 AAY18199 AAY28318 AAY28621 AAY44799 AAY44664 AAY46374 AAY47016 AAY46394 AAY46419 AAY64195 AAY78942 AAZ38509 AAZ38520 AAZ38608 AAZ38553 AAZ43373 AAZ43386 AAZ74355 AAZ74377 AAZ74532 AAZ74609 AAZ79519 AAZ79552 AAZ80010 AAZ83245 AAZ83315 ABA16395 ABA26725 ABA43339 ABA42457 ABA42490 ABB96512 ABB02850 ABB04286 ABB04297 ABB04308 ABB04319 ABB04330 ABB04341 ABB04374 ABB03071 ABB04909 ABB04931 ABB04997 ABB86514 ABB18394 ABB18988 ABB19066 ABB19165 ABB87045 ABB87348 ABB19588 ABB87422 ABB87564 ABB87618 ABB19899 ABB87921 ABB87942 ABB20034 ABB20188 ABB90185 ABB90195 ABB88250 ABB20221 ABB20232 ABB88259 ABB20265 ABB88292 ABB88301 ABB20475 ABB88382 ABB90233 ABB21806 ABB21822 ABB46428 ABB46439 ABB53688 ABB53754 ABB54517 ABB77856 ABB59999 ABB60010 ABB80037 ABB80151 ABB80188 ABB80495 ABB80243 ABB96322 ABB96344 ABB96355 ABB96366 ABB96377 ABB96523 ABC39808 ABC40558 ABC41695 ABC42150 ABC43064 ABC42742 ABC50214 ABC50236 ABC67992 ABC67667 ABD38137 ABC85922 ABC85878 ABC84534 ABD16574 ABD15738 ABD15584 ABD60793 ABE12626 ABD77612 ABD77667 ABD77689 ABE12650 ABD79170 ABD77832 ABD79203 ABD77854 ABD77898 ABD79247 ABD94847 ABD94891 ABE13598 ABE13609 ABE13642 ABE13668 ABE14055 ABF47861 ABI48009 ABI48020 ABO52008 ABF83450 ABG88248 ABG26849 ABG26860 ABG26871 ABG37189 ABG37200 ABG37233 ABG37332 ABG37376 ABG37420 ABG37442 ABG37475 ABG47865 ABG47964 ABG47975 ABG48019 ABG48294 ABG48316 ABG48360 ABG67138 ABG67160 ABG67171 ABG67538 ABG80087 ABG80230 ABG80395 ABG80439 ABG88237 ABG88270 ABG88457 ABG88798 ABG88820 ABI20807 ABI21027 ABI21071 ABI21104 ABI21247 ABI21357 ABI30359 ABI30546 ABI30612 ABI30879 ABI84467 ABI84468 ABI84496 ABI84526 ABI84652 ABI84747 ABI84758 ABI84768 ABI84830 ABI84841 ABI84850 ABI84858 ABI84886 ABI84962 ABI85098 ABI85130 ABI85158 ABI92283 ABI92349 ABI92514 ABI92580 ABI92591 ABI92602 ABI92635 ABI92712 ABI92723 ABI92833 ABI92866 ABI93119 ABI95176 ABI95187 ABI95242 ABJ09220 ABJ09264 ABJ09374 ABJ16568 ABJ16711 ABK39976 ABK40670 ABK80116 ABK80138 ABL67234 ABL75566 ABM22095 ABM22150 ABO32793 ABO33072 ABO38299 ABO38310 ABO38704 ABO38737 ABO44060 ABO44071 ABO44093 ABO51920 ABO52019 ABO52074 ABO52338 ABO52349 ABO52371 ABO52437 ABO52569 ABO77026 ABO77048 ABP49187 ABP49440 ABQ01259 ABQ01358 ABQ01369 ABR28529 ABR37344 ABR37520 ABR37586 ABR37608 ABS50023 ABV46329 ABV46340 ABV46351 ABV46384 ABV46395 ABV46406 ABV46417 ABV46428 ABV46450 ABV46461 ABV46472 ABV46483 ABV46498 ABV46509 ABV46642 ABV46664 ABV46679 ABV46781 ABV46801 ABV46812 ABV46823 ABV46867 ABV46878 ABV46941 ABV46962 ABV46973 ABV46984 ABV46995 ABV47006 ABV47017 ABV47027 ABV47038 ABV47060 ABV47115 ABV47269 ABV47291 ABV47346 ABV47368 ABV47390 ABV47434 ABV47445 ABV47456 ABV47500 ABV47511 ABV47544 ABV47566 ABV47588 ABV47643 ABV47665 ABV47676 ABV47731 ABV47753 ABV47786 ABV47797 ABV47819 ABV47841 ABV47874

ABV47885 ABV47907 ABV48006 ABV48039 ABV48083 ABV48094 ABV48127 ABV48160 ABV48303 ABV48389 ABV48400 ABV48466 ABV48488 ABV48510 ABV48554 ABV48565 ABV48587 ABV48609 ABV48631 ABV48642 ABV48653 ABV48664 ABV46318 ABU40955 ABU40956 ABU40959 ABV29725 ABV29912 ABV30121 ABV30253 ABV30264 ABV30396 ABV30418 ABV30484 ABW39941 ABX88802 ABX88813 ABY51407 ACA24692 ACA24703 ACC61879 ACC61890 ACD12174 ACD56294 ACD56305 ACD56316 ACD56327 ACD56393 ACD85190 ACD85201 ACD85234 ACD85245 ACD85256 ACD93582 ACD85443 ACD85487 ACD85509 ACD85542 ACD85564 ACE76594 ACF22162 ACF22433 ACF47565 ACF41859 ACF54392 ACF76435 ACI16700 ACI25727 ACI26365 ACI26519 ACI26552 ACI26563 ACI26585 ACJ10019 ACK99358 ACL12231 ACL12253 ACN32286 ACN32477 ACN32537 ACN32917 ACN86439

NP

BAA86066 BAA86067 BAA86068 BAA35109 BAB39513 BAB39514 BAD02348 BAD89306 BAE07156 BAF46428 BAF46438 BAF46458 BAF46468 BAF46478 BAF46508 BAF37963 BAE96962 BAF37822 BAF56168 BAF56434 BAF56439 AAC32084 AAD12236 AAF02400 AAF02404 AAD51925 AAD49012 AAD49013 AAD49016 AAD49018 AAD49019 AAD49020 AAD49021 AAD49022 AAD49023 AAF70405 AAG01194 AAG01204 AAF87498 AAF87500 AAF87501 AAF87502 AAK18005 AAK18006 AAF99666 AAF99667 AAG09040 AAK51722 AAK60145 AAM75159 AAL59144 AAL60436 AAM69958 AAM69969 AAM49560 AAQ04894 AAQ04896 AAQ04898 AAQ04900 AAQ04901 AAQ04903 AAO52960 AAO52961 AAO52964 AAP49078 CAC19696 CAC84249 CAD20330 CAD30200 CAD30201 CAF33011 CAF33013 CAF31359 CAI29280 CAD22812 AAL31398 AAL31400 AAO46422 AAO46424 AAO46427 AAO46431 AAO46432 AAO46443 AAO46457 AAO46459 AAO46539 AAO46540 AAO46544 AAO46824 AAO46830 AAO46832 AAP29980 AAQ77444 AAQ77445 AAS18236 AAT39107 AAT12086 AAT12087 AAT12088 AAT12090 AAT12094 AAT12095 AAT12096 AAT12097 AAT12099 AAT12100 AAT12101 AAT12105 AAT65358 AAT70618 AAT70643 AAU11205 AAU11211 AAU11214 AAU11218 AAV91222 AAV91224 AAX07774 AAU00815 AAW59391 AAW59409 AAU93405 AAV48837 AAV48549 AAX47284 AAX47285 AAW72231 AAW78284 AAW78285 AAW78291 AAW78295 AAX11459 AAX11499 AAX38241 AAX56534 AAX76747 AAY18200 AAZ38631 AAZ74434 AAZ80011 AAZ83303 ABA42295 ABB04287 ABB04298 ABB04910 ABB17718 ABB86515 ABB18062 ABB18989 ABB19067 ABB19458 ABB19485 ABB87338 ABB19621 ABB87359 ABB87554 ABB87565 ABB87608 ABB87744 ABB19939 ABB19980 ABB88092 ABB20089 ABB88134 ABB88142 ABB20145 ABB20155 ABB90186 ABB90196 ABB88260 ABB20276 ABB20298 ABB20317 ABB20391 ABB20400 ABB20448 ABB20476 ABB20493 ABB90234 ABB21765 ABB46407 ABB53744 ABB53755 ABB96323 ABB96524 ABC68004 ABC67569 ABD15263 ABD61533 ABD62846 ABD60937 ABD60948 ABD62785 ABD61739 ABD77668 ABD79105 ABD79116 ABD77800 ABD77811 ABE14056 ABF47959 ABF47862 ABG37157 ABG37223 ABG37234 ABG37300 ABI20808 ABI20874 ABI20885 ABI21171 ABI84510 ABI84518 ABI84538 ABI84549 ABI84570 ABI84591 ABI84645 ABI84677 ABI84687 ABI84716 ABI84727 ABI84748 ABI84759 ABI84769 ABI84974 ABI84983 ABI85015 ABI85032 ABI85041 ABI85061 ABI85088 ABI85099 ABI85110 ABI85121 ABI85140 ABI85187 ABI92196 ABI92251 ABI92284 ABI95210 ABJ09243 ABJ09276 ABJ16822 ABJ16844 ABJ16866 ABJ53552 ABJ53587 ABK79963 ABL67092 ABL67114 ABM22052 ABO32820 ABM22250 ABM66857 ABN50760 ABN50921 ABN51014 ABN51147 ABN59405 ABN59438 ABO33073 ABO38058 ABO38267 ABO38344 ABO38355 ABO38366 ABO38377 ABO38388 ABO51965 ABO51976 ABO52339 ABO52350 ABO52394 ABO52449 ABO52460 ABO52504 ABO52647 ABO52724 ABO52735 ABO52790 ABO76983 ABO77027 ABO77049 ABO77082 BAA00475 BAA00477 BAA00478 AAY98855 AAY98865 AAY98870 AAY98871 AAY98877 AAY98878 AAY98894 AAY98898 AAY52604 AAY52605 AAY52607 AAY52615 AAY52618 AAY52619 AAY52625 AAY52631 AAZ29586 AAZ29587 AAZ29588 AAZ72756 ABA62302 ABB51966 ABB83593 ABC66718 ABC66719 ABC66729 ABC66736 ABC74404 ABC94735 ABD14811 ABD14812 ABD14813 ABD14815 ABD35666 ABD35667 ABD35668 ABD35669 ABD35670 ABD35676 ABD35677 ABD35679 ABD35680 ABD35681 ABD35682 ABD35685 ABD35686 ABD35688 ABD35689 ABD35690 ABD35692 ABD35693 ABD35694 ABD35695 ABD35696 ABD35697 ABD35698 ABD35699 ABD35701 ABD59858 ABD59859 ABD59860 ABD92956 ABD91842 ABE27339 ABF56633 ABF69260 ABG27057 ABG36720 ABG36722 ABH03490 ABH03498 ABG88890 ABH04380 ABK34765 ABI94743 ABI94749 ABI94756 ABI96743 ABI96752 ABI96758 ABI97325 ABI97337 ABJ52569 ABJ52576 ABJ80587 ABJ09513 ABJ09500 ABI96779 ABK13772 ABK32095 ABM21889 ABM21892 ABM21895 ABL08491 ABL08507 ABL08516 ABM46161 ABM46165 ABM46167 ABO31430 AAA43467 AAA51496 AAA51499 AAA51501 AAA51503 AAA51504 AAA51505 AAA51483 AAA51488 AAA51508 AAA51514 AAA51493 AAA43116 AAA43663 AAA43097 AAA43241 AAA43451 AAA43666 AAA43657 AAA43460 AAA43461 AAA43463 AAA43472 AAA43483 AAA43484 AAA43486 AAA43129 AAA43459 AAA43474 AAA43475 AAA43477 AAA52249 AAA52252 AAA52234 AAA52239 AAA52242 AAA52245 AAA73105 AAA73106 AAA67339 CAA36234 CAA81460 CAA81462

NS1

BAB39522 BAD02353 BAD89311 BAE07161 BAE48327 BAF46433 BAF46443 BAF46453 BAF46463 BAF46473 BAF46483 BAF46513 BAE94705 BAF37968 BAF02318 BAF33065 BAF38378 BAF41921 BAF56064 BAF56173 BAF56178 AAB93958 AAB93960 AAB93962 AAC24236 AAC17974 AAC17976 AAC32082 AAC14267 AAC14269 AAC14273 AAD23278 AAD23290 AAD23292 AAD23306 AAD51930 AAD52943 AAD52945 AAD52949 AAD52953 AAD52957 AAD52959 AAD52961 AAF89557 AAF89575 AAG01190 AAG48235 AAG48236 AAF87537 AAF87545 AAF87549 AAF87551 AAK18009 AAG09043 AAK14368 AAK51750 AAM75163 AAM69956 AAM69967 AAM69997 AAM49563 AAQ04995 AAQ04996 AAQ04997 AAQ04999 AAQ05000 AAQ05001 AAQ05002 AAQ05003 AAQ05008 AAQ05010 AAO52909 AAO52932 AAP49160 AAP49161 AAP49163 AAP49169 AAP49170 AAP49175 CAC09428 CAC85091 CAC85094 CAC85097 CAC85098 CAC85099 CAC85107 CAC85109 CAD20334 CAD58607 CAD58614 CAJ01906 CAD22816 AAK38762 AAL31410 AAO46567 AAO46577 AAO46585 AAO46589 AAO46633 AAO46635 AAO46759 AAO46763 AAO46769 AAO46771 AAO92853 AAO92856 AAO92860 AAO92870 AAO92878 AAO92882 AAO92908 AAO92914 AAP20761 AAP20762 AAP20771 AAP20772 AAP57591 AAP57596 AAP57612 AAQ77416 AAS57538 AAR88850 AAR88856 AAR88860 AAR88862 AAR88864 AAR88870 AAR88876 AAR88883 AAR88885 AAR88887 AAR88889 AAR88893 AAR88895 AAT39023 AAT12106

AAT12112 AAT12114 AAT12115 AAT12116 AAT12120 AAT65460 AAT65464 AAT65472 AAT65481 AAT65486 AAT73368 AAT73372 AAT73401 AAT73413 AAT73423 AAT73431 AAT73435 AAT73437 AAT73441 AAT73453 AAT73457 AAT73461 AAT76163 AAU11253 AAU11255 AAU11259 AAV91230 AAV91232 AAX07776 AAU00824 AAW59385 AAW59404 AAV48547 AAV63987 AAX47283 AAY16309 AAY16312 AAY16313 AAX53538 AAX53549 AAX76911 AAX78819 AAX56535 AAX76738 AAZ43375 AAZ83247 ABA16445 ABB04288 ABB03139 ABB04933 ABB04966 ABB17719 ABB86516 ABB19024 ABB19105 ABB87297 ABB19486 ABB87339 ABB19590 ABB87424 ABB19961 ABB20006 ABB20015 ABB88093 ABB88115 ABB88126 ABB88135 ABB20146 ABB88188 ABB20201 ABB20223 ABB20234 ABB88261 ABB20401 ABB20432 ABB88345 ABB20477 ABB20494 ABB88373 ABB20519 ABB21766 ABB90219 ABB53745 ABB79736 ABB80039 ABB96324 ABC50315 ABC68316 ABC84547 ABD16734 ABD62847 ABD60960 ABD79106 ABD79128 ABD79117 ABD77823 ABE14057 ABG67140 ABI20875 ABI36028 ABI84511 ABI84550 ABI84653 ABI84668 ABI84749 ABI84760 ABI84832 ABI84843 ABI84910 ABI84975 ABI84984 ABI85042 ABI85052 ABI85062 ABI85122 ABI85141 ABI85177 ABI85218 ABJ09123 ABJ51744 ABJ53564 ABL67247 ABM21976 ABM22251 ABO33011 ABO38026 ABO38389 ABO44172 ABO51922 ABO52351 ABO52791 BAA06344 AAY96570 AAY96595 AAY96625 AAY52638 AAY52646 AAY52658 AAY52666 AAY52668 AAY52680 AAZ29599 AAZ29601 AAZ14195 AAZ14205 ABB69706 ABB71851 ABB51975 ABB80550 ABC68546 ABC68577 ABC68588 ABC69220 ABD23025 ABD36820 ABD36827 ABD36829 ABD36847 ABD36849 ABD36851 ABD36853 ABD36859 ABD36863 ABD36867 ABD36869 ABD36875 ABD36879 ABD36881 ABD36883 ABD36885 ABD36891 ABD61023 ABD65977 ABD65984 ABD65992 ABD59905 ABD59907 ABD59917 ABD59919 ABF01940 ABF01948 ABF01968 ABF01970 ABF01972 ABF02060 ABF02076 ABF21198 ABF69259 ABG76021 ABG76059 ABG76063 ABK34770 ABI94757 ABI96753 ABI97317 ABI97326 ABK00084 ABI98932 ABJ09514 ABJ09469 ABJ09491 ABI98941 ABM21911 ABM21917 ABM21923 ABL08545 ABL08577 ABL08583 ABL08599 ABL08703 ABL08711 ABL08713 ABL08769 ABL08773 ABL75552 ABM46335 ABM46349 ABM46359 ABM46361 ABM46371 ABM46407 ABM46435 AAA56810 AAA56814 AAA43504 AAA43512 AAA43548 AAA43557 AAA43545 AAC35564 AAC35576 AAB51015 AAB51017 AAC40657 AAC40663 AAC40667 AAC40669

NS2/NEF

AAA19198 AAA21581 AAA43085 AAA43490 AAA43508 AAA43521 AAA43524 AAA43535 AAA43539 AAA43544 AAA43550 AAA43556 AAA43560 AAB93934 AAB93936 AAB93937 AAB93942 AAB93950 AAB93961 AAC14268 AAC14270 AAC24237 AAC32083 AAC40504 AAC40652 AAC40654 AAC40656 AAC40666 AAC40668 AAC40670 AAC40672 AAC40678 AAC40682 AAC63488 AAD23281 AAD23295 AAD23303 AAD23309 AAF02344 AAF87550 AAK71695 AAL75852 AAM69957 AAM69998 AAO46568 AAO46570 AAO46574 AAO46576 AAO46580 AAO46582 AAO46590 AAO46596 AAO46630 AAO46764 AAO46768 AAO46772 AAO46776 AAO65609 AAO92869 AAO92871 AAO92907 AAO92913 AAO92927 AAP04513 AAR88851 AAR88863 AAR88877 AAR88878 AAR88880 AAR88886 AAR88900 AAR88922 AAR99624 AAT37569 AAT38825 AAT73381 AAT73389 AAT73391 AAT73430 AAT73434 AAT73436 AAT73442 AAT73446 AAT73452 AAT73456 AAT73460 AAT73472 AAT76162 AAT90837 AAU00816 AAU00823 AAU11234 AAU11238 AAU11244 AAV41219 AAV80806 AAV97626 AAV97628 AAW59403 AAX07777 AAX11461 AAX47531 AAX53540 AAX53548 AAX56546 AAX57850 AAY52633 AAY52635 AAY52639 AAY52655 AAY52657 AAY52659 AAY52661 AAY52681 AAY64308 AAY87436 AAZ14204 AAZ30536 AAZ38633 AAZ43376 AAZ74458 AAZ74513 AAZ80036 ABA12790 ABA26739 ABA42984 ABB04967 ABB17687 ABB17698 ABB19000 ABB19420 ABB19513 ABB19653 ABB19807 ABB19817 ABB19962 ABB20147 ABB20197 ABB20202 ABB20224 ABB20309 ABB20402 ABB20433 ABB20450 ABB20461 ABB20488 ABB21747 ABB53658 ABB80191 ABB87298 ABB87673 ABB87984 ABB88094 ABB88189 ABB88346 ABB90188 ABB90220 ABB90226 ABC40637 ABC48795 ABC48804 ABC48834 ABC50195 ABC67670 ABC67878 ABC74398 ABC85881 ABD15477 ABD15620 ABD16478 ABD16588 ABD16735 ABD36821 ABD36822 ABD36828 ABD36846 ABD36848 ABD36858 ABD36874 ABD36876 ABD36880 ABD36882 ABD36884 ABD36890 ABD59916 ABD59918 ABD59922 ABD60961 ABD62787 ABD62848 ABD79250 ABD91843 ABD95158 ABD95246 ABE11966 ABE13561 ABE14872 ABE96874 ABF21211 ABF47556 ABF56637 ABF82836 ABF82858 ABG26951 ABG26962 ABG37368 ABG37511 ABG47824 ABG48055 ABG48143 ABG72674 ABG76020 ABG76022 ABG76026 ABG88240 ABG88328 ABI20876 ABI21195 ABI21250 ABI30527 ABI84512 ABI84654 ABI84718 ABI84750 ABI84844 ABI84872 ABI84976 ABI85034 ABI85043 ABI85053 ABI85219 ABI85229 ABI85242 ABI92605 ABI92759 ABI95201 ABI95256 ABI95300 ABI97327 ABI98906 ABI98942 ABJ09470 ABJ09492 ABJ09515 ABJ15712 ABJ51679 ABJ53565 ABK00092 ABK32091 ABK41617 ABK80009 ABK80053 ABK80130 ABK80207 ABL08546 ABL08584 ABL08620 ABL08704 ABL08772 ABL08788 ABL67094 ABL67776 ABL67800 ABM21910 ABM21916 ABM21922 ABM46312 ABM46354 ABM46382 ABM46434 ABM46442 ABN51149 ABO38357 ABO38707 ABO44228 ABO51967 ABO52396 ABO76683 ABO76685 ABP64755 ABQ41370 ABQ57384 ABQ84572 ABQ84612 ABQ84628 ABR37424 ABR37523 ABR37611 ABR53856 ABS50828 ABS52599 ABS70353 ABU63962 ABU97272 ABU97320 ABU97330 ABU97360 ABU97444 ABV29541 ABV29552 ABV29739 ABV29849 ABV29860 ABV30036 ABV31848 ABV46966 ABV47284 ABV47372 ABV47570 ABV48492 ABV48536 ABV54011 ABV54027 ABV54063 ABV54093 ABV54115 ABW71333 ABW73769 ABW91104 ABW91279 ABW97466 ABW97477 ABX24581 ABX80149 ABX80165 ABX80175 ABX88816 ABX88827 ABX88838 ABY50579 ABY51553 ABY81636 ABZ91683 ACA03208 ACA25335 ACA47412 ACA47797 ACA47907 ACA47940 ACA61631 ACB70477 ACB70488 ACB70521 ACB70642 ACB72447 ACC61893 ACD35889 ACD40206 ACD40220 ACD40226 ACD40230 ACD40234 ACD40238 ACD40240 ACD40250 ACD40258 ACD40262 ACD40266 ACD40268 ACD40270 ACD40274 ACD40276 ACD40282 ACD40284 ACD40292 ACD56341 ACD56374 ACD88532 ACD88558 ACD88598 ACE76631 ACE78955 ACE78986 ACF08039 ACF25463 ACF25485 ACF25634 ACF25675 ACF36831 ACF40973 ACF41774 ACF76218 ACG58423 ACI01657 ACI01659 ACI01661 ACI01663 ACI01667 ACI01669 ACI22229 ACI25718 ACI90142 ACJ03961 ACJ14857 ACJ15055 ACJ15121 ACJ15143 ACJ25120 ACJ26083 ACJ26226 ACJ26380 ACJ68618 ACJ68771 ACL11947 ACL12400 ACN29488 ACN37885 ACN37891 ACN39390 ACN39394 BAD02352 BAD89310 BAE07160 BAE48344 BAF38377 BAF46462 BAG80904 BAG80914 BAG80916 BAG84422 CAA24287 CAA81474 CAC09427 CAQ58511

PA

AAA19207 AAA43098 AAC63417 AAC63419 AAC63456 AAC63459 AAL31438 AAO65607 AAT38883 AAW78266 AAW80715 AAX07772 AAX11542 AAY52692 AAY52702 AAY52703 AAY52712 AAY52713 AAZ38524 AAZ38634 ABB19954 ABB19969 ABB19992 ABB20092 ABB20134 ABB20138 ABB20182 ABB20331 ABB20349 ABB20379 ABB20394 ABB20428 ABB20472 ABB87384 ABB87417 ABB87469 ABB87622 ABB87736 ABB87807 ABB87914 ABB87935 ABC48789 ABC59714 ABC85838 ABD14801 ABD77847 AAA43246 AAT39059 AAY52689 AAY52690 AAY52696 AAY52699 AAY52700 AAY52701 AAY52705 AAY52709 AAY52710 AAY52715 AAZ43377 AAZ43390 ABB19981 ABB19985 ABB20116 ABB20279 ABB20340 ABB20403 ABB20462 ABB87395 ABB87436 ABB87644 ABB87925 ABC46561 ABC87317 ABD14804 ABD15266 ABD77803 ABD79108 ABG47847 AAF02384 AAF02391 AAF02393 AAF02394 AAK55426 AAO46314 AAO46507 AAO46516 AAT12129 AAT12130 AAT12132 AAT12139 AAT12140 AAT12141 AAT12145 AAT12147 AAQ04958 AAQ04959 AAQ77448 AAT74476 AAT76164 AAU00825 ABI20811 AAF99668 AAF99669 AAG01207 AAO46849 AAO46850 AAT39061 AAT39064 AAW59389 AAW59397 ABB21768 ABB21779 ABB21790 ABB21810 ABB20521 ABB21800 ABB88263 ABB90221 AAW78273 AAY47082 AAY52691 AAY52704 AAY64199 ABB88128 ABB88285 ABB88375 ABB90237 ABI20866 ABI84710 ABI84751 ABI84762 ABI84834 ABI85206 ABI85220 AAK18011 AAK18012 AAO46309 AAO46320 AAR99631 AAT12127 AAT12134 AAT12135 AAT12137 AAT12144 AAV65822 AAV91208 AAV97632 ABH03488 ABI84719 ABI85179 ABI85239 ABB20108 ABB20148 ABB20269 ABB20290 ABB87674 ABB87758 ABC86244 ABD14802 ABD14803 AAZ29582 AAZ29583 ABB19070 ABB19114 ABB19224 ABB19790 ABJ09542 AAO46307 AAO46308 AAV48550 ABF56622 ABB19202 ABB19378 ABB19765 ABG88340 AAG09042 ABB19883 ABB83592 ABB87256 ABD60951 ABI84541 ABI84573 ABI85113 ABI85124 ABI85168 ABI94760 ABI98918 AAQ04952 AAQ04957 AAQ04966 AAQ04967 ABK79944 ABL08854 ABL08860 ABL08863 ABL08864 ABL08870 ABL08874 AAA43619 AAD49057 AAD49058 AAD49067 AAM69966 AAM78512 AAX47281 ABD62849 ABF69263 ABG27053 ABG88263 AAY18574 ABB04378 ABI84418 ABI84670 ABI85134 ABI97328 ABI97346 ABK40641 ABL08857 ABL08866 AAD49055 AAD49064 AAD51924 ABB18406 ABB90189 ABB90227 ABC68106 ABC68115 ABC68142 ABD35706 ABD35713 ABD35714 ABD35716 ABD35723 ABD35724 ABD35733 ABD35734 ABD35736 ABI20877 ABI84772 ABI85243 ABD59838 ABD59846 AAO46518 AAT12128 AAT12136 AAT12142 AAM69996 AAM75157 AAX47282 ABB18136 ABI85044 ABI85054 ABI95323 ABJ16825 ABJ51680 ABJ90247 ABL08969 ABL08979 ABL08985 ABL08991 ABC68129 ABC68145 ABC68152 ABC68155 ABD35737 ABI85073 ABL08995 ABJ09481 ABJ09503 ABJ09523 ABJ09533 ABK80010 ABJ53566 ABL08972 ABO38105 ABO44196 BAF46466 ABL67161 ABM21930 ABM21967 ABM46457 ABN50907 ABN51050 ABO77074 ABO77085 BAF40415 BAF46476 ABL08988 ABL08989 ABM21956 ABM46454 ABM46458 ABM46459 ABB19098 ABB19367 ABO52056 ABO52067 ABO52177 ABO52397 BAF56437 ABO51979 ABO52353 BAE93478 CAA67500 CAB56290 ABL08902 ABL08912 BAB39510 ABI84986 ABI96772 ABM21924 ABB87341 ABC69223 ABO52485 ABO52738 ABO38358 CAF02293 ABG88889 ABI20833 ABI84530 ABI84552 ABK13787 ABK32092 AAD49060 AAD49066 ABB80041 BAF37961 ABC68132 ABD35705 ABD35708 ABD35715 ABD35727 ABD35732 ABI84845 ABI84873 ABI85091 ABJ53555 ABJ98939 ABL08967 ABB19400 ABB19614 ABK34763 ABO77030 BAF46456 ABM21928 ABM21978 ABB86792 ABC42647 ABC74408 ABD62788 ABO52309 BAF56167 ABI84977 ABI92199 ABI92254 ABJ09103 ABK00124 ABI30363 ABL67848 ABO52782 CAC84869 ABI84500 ABI84513 ABI84680 ABI85102 AAA43665 AAD49061 AAD49063 AAM49559 ABB17157 ABB19547 ABD35702 ABD35703 ABD35711 ABD35720 ABD35729 ABD35730 ABI96746 AAO46843 AAO46846 AAO46848 AAY52697 ABA55040 ABB19925 ABB20225 ABB20434 ABD77814 ABI92518 ABI98943 AAQ04954 AAQ04955 AAQ77446 AAT74501 AAU00818 ABB20479 ABB88201 ABH03496 ABI36022 ABI84912 ABO38061 ABO76975 ABO77052 ABJ51735 ABO51858 ABO52144 ABO38391 CAC84865

PB1

AAC63412 AAK18014 AAK51714 AAO46324 AAO46325 AAO46334 AAS89191 AAT12148 AAT12155 AAT12156 AAT12157 AAV91204 AAY28559 AAQ04916 AAQ81637 AAR05983 AAA19211 AAC63449 AAC63452 AAO46330 AAO46339 AAO46563 AAO46565 AAO46566 AAT12150 AAT12152 AAV91206 AAY28343 AAT76165 AAT78590 AAU00826 AAY18204 AAZ14154 AAZ74471 AAZ79612 ABA12704 ABB18970 ABB19126 ABB19170 ABB19427 ABB19449 ABB80031 ABB80500 ABC39813 ABB19045 ABB19236 ABB19368 ABB19379 ABB19568 ABB19675 ABB19720 AAC63410 AAK18013 AAO46328 AAA43641 AAM69995 AAO46558 AAR99632 AAT12151 AAY28413 AAX53561 AAY64300 AAZ29581 ABB87289 ABB87342 ABB87353 AAA43644 AAL60434 ABB04357 ABB19063 ABB19625 ABB19658 ABB19664 ABB19828 ABB20039 AAM49557 AAM75156 AAX47279 ABB88019 ABB88107 ABD35752 ABD35753 ABD35762 ABD35763 ABD35772 ABD35773 ABD60941 AAO46326 AAO46340 ABB86519 ABB87214 ABB87246 ABB87374 ABC42758 ABB20496 ABB21769 ABB46411 AAT12153 AAT12154 AAY28626 ABB88096 ABB88202 ABB88376 ABB88387 ABB90190 ABB90228 ABB90238 ABC66802 ABC66819 ABC66820 AAA43581 AAA43582 AAD51923 AAM69975 ABD35743 ABD35749 ABD35751 ABD35759 ABD35766 ABD60974 AAG01208 AAG09041 AAX47280 AAX56539 AAZ38602 AAQ04925 ABA16472 ABB19040 AAZ83320 ABC40552 ABB19489 ABB19741 ABB20009 ABC74415 ABC84430 ABD59824 AAO46857 AAO46859 AAW72229 ABB04302 AAY52721 AAY52723 AAY52731 AAY52733 ABB20173 ABB21780 ABB46422 ABB88041 ABB20421 ABB87547 ABD59826 ABD62850 AAT12163 AAW80712 AAX11513 AAY52720 AAY52727 AAY52730 AAY52734 AAY52735 AAY52737 AAY52743 ABC86004 ABD14797 ABD77947 ABD79120 ABG67176 ABB90200 ABG79960 ABI21230 AAT12159 AAT38884 ABC69222 ABC74413 ABC74414 ABD15743 ABD59818 ABB20216 ABB20280 ABB20350 ABB20369 ABB20404 ABB20463 ABB87396 ABE11908 ABB87664 ABB87841 ABB87926 ABG27037 ABG88253 AAY52722 AAY64220 ABD16579 ABI84711 ABI84773 ABI84987 ABA87088 ABB19918 ABB20093 ABB20395 ABB87470 ABJ09482 ABJ16958 ABK80022 ABM22254 ABI85081 ABI85180 ABI92266 ABJ16815 ABJ16848 ABJ53534 ABL07796 ABL07828 ABL07842 ABL07861 ABD77848 ABD77815 ABL07909 ABL07940 ABO44175 ABO44186 BAF37818 BAF46455 ABH04383 ABI20878 ABI84408 ABI84531 ABI84563 ABI95192 ABI97342 ABJ09126 ABL67195 ABO52684 ABD59820 ABD59834 BAE07153 ABG27049 ABA42332 ABB17689 ABB18055 ABB18386 ABI84752 ABI84978 ABC02340 ABD62789 ABG37227 ABD95347 ABE11919 ABE11930 ABE73104 ABG88825 ABI85103 ABI85125 ABI92255 ABI92838 ABJ09472 ABJ09504 ABG88275 ABI84852 ABI94772 ABJ98940 ABK80011 ABC97382 ABJ52559 ABE97466 ABO51903 ABO52013 ABL67052 ABO52123 ABG67143 ABO52387 ABI84935 ABI92179 ABI84542 ABL07802 ABI95214 ABI97322 ABL67162 ABM22012 ABO44142 ABB19053 ABB19080 ABB19636 ABB19647 BAF41912 BAF46465 BAF46475 CAF04464 ABD35741 ABD35742 ABD35748 ABD35750

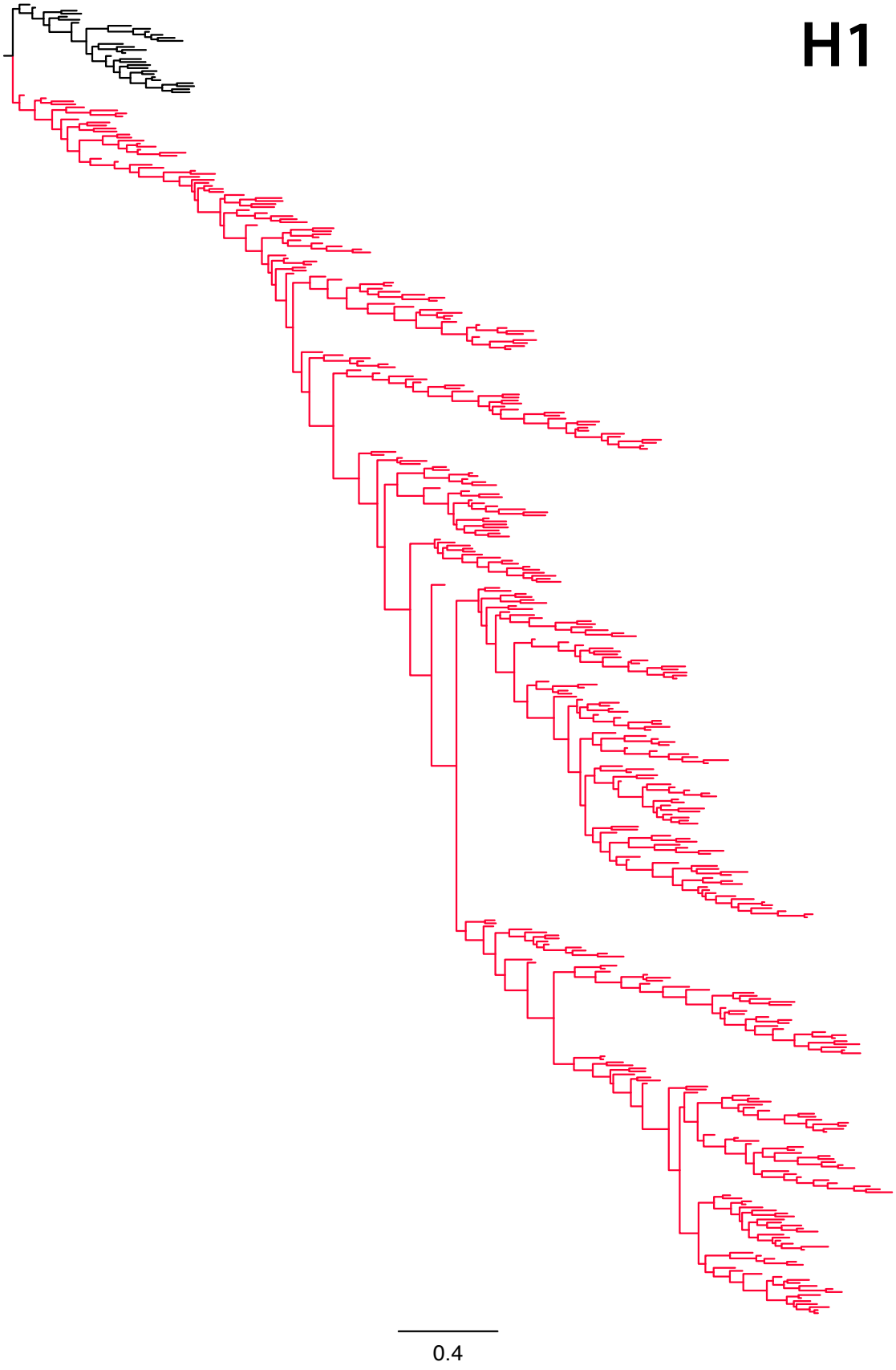
ABD35758 ABD35760 ABD35767 ABD61537 ABI36020 ABI97312 ABI84396 ABI84655 ABI84691 ABK13768 ABO31426 ABO52354 ABO52486 ABO52618 BAE48320 ABO33017 ABG67165 ABB19986 ABG88887 ABI84835 ABI85065 ABK32096 ABL07809 ABL07912 ABO38359 BAF37960 BAF46425 BAF46445 CAA67498 ABO51969 ABO52057 ABO52090 ABO52178 ABO52475 ABO52519 ABO52596 AAF99676 AAF99677 AAT90830 AAU00819 ABB05213 ABB17722 AAA43579 AAA43631 AAY87428 ABB21801 ABB88191 ABB88295 ABD35738 ABD35739 ABD35740 ABD35745 ABD35746 ABD35747 ABD35754 ABD35756 ABD35761 ABD35764 ABD35770 ABD35771 ABG37238 AAT12162 AAT12166 AAT12167 AAW59396 AAY52729 ABA55039 ABB20047 ABB20083 ABB20149 ABB20384 ABB87819 ABD77804 ABD79109 ABC43179 ABD59828 ABD59831 ABE13614 ABI84720 ABI84792 ABI85212 ABL07846 ABL07899 ABL67118 ABM22056 ABO38392 BAF40413 ABI48014 ABI84492 ABI84553 ABI84595 ABI84702 ABI98912 ABK34761 ABO52706 ABO51859 ABG38195 ABK13767 ABM21896 BAD02349

PB2

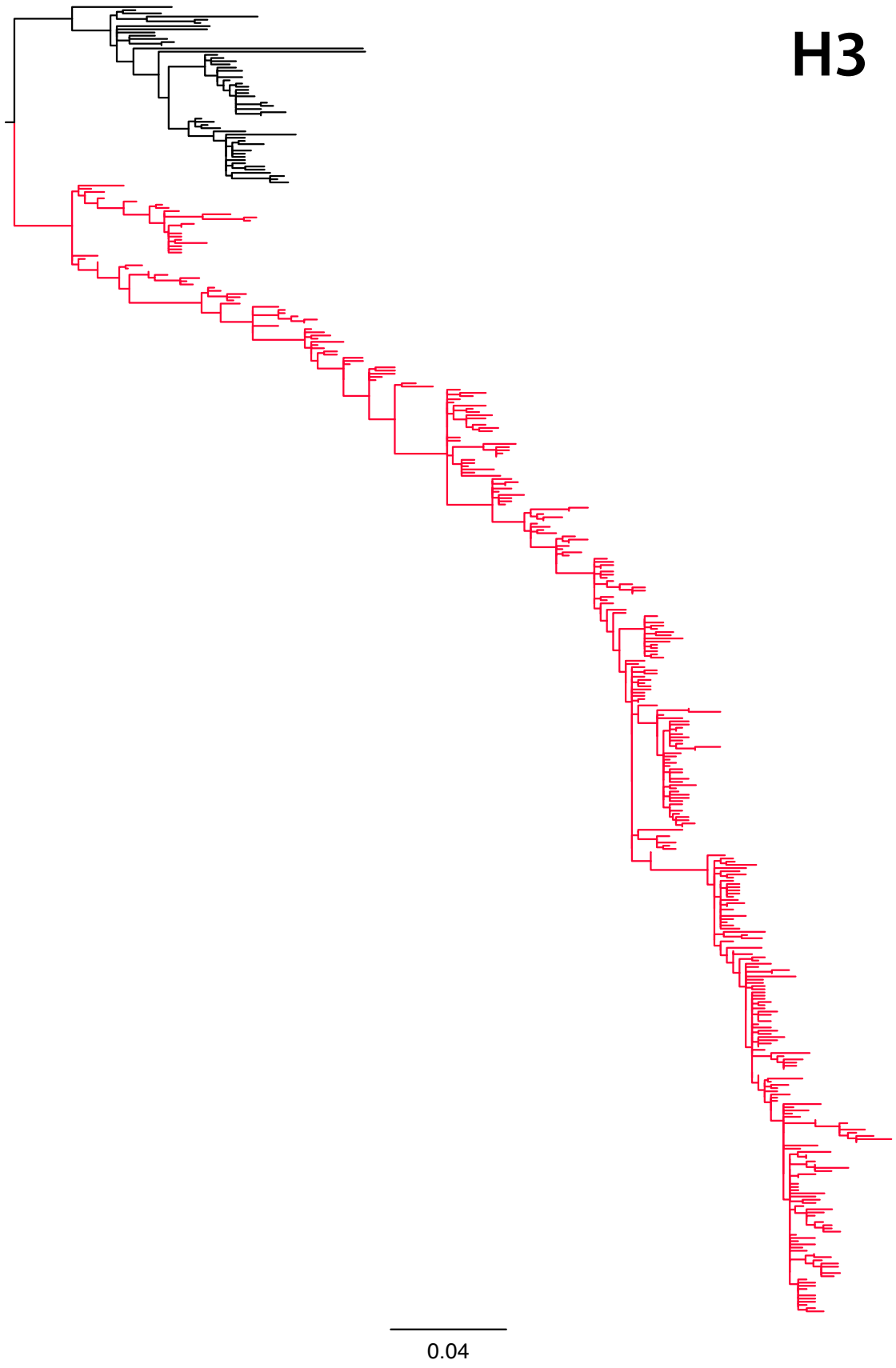
AAA43594 AAA43648 AAA43651 AAA43654 AAL60435 AAM69974 AAM75155 AAM78511 AAT69356 AAW72228 AAX11484 AAX11494 AAY52746 AAY52749 AAY52756 AAY52759 AAY52766 AAY52768 AAY59044 AAY87429 ABB02813 ABB02957 ABB04293 ABB17713 ABB18035 ABB18139 AAO46257 AAO46259 AAQ04931 AAQ04938 ABB20272 ABB20304 ABB20482 ABB54524 ABB87739 ABB87966 ABB88098 ABB90224 ABB90230 ABC66451 ABC66458 ABC66491 ABC66511 ABC67586 AAR99633 AAT73536 AAT73538 AAT73579 AAZ14139 ABD32121 ABD63073 ABD77806 ABD78059 ABD79166 ABD95063 ABF01677 ABG37163 ABB19392 ABB19414 ABB19429 ABB19472 ABB19595 ABB19830 ABB19866 ABB83600 ABB86521 ABB87344 ABC40554 ABC67871 ABD59809 ABF69265 ABG27055 ABI30388 AAA43595 AAA43611 AAA43653 AAD49042 AAD51922 ABG88354 AAO46265 AAM69964 AAW59387 AAX53569 AAY52747 AAY52751 AAY52755 AAY52757 AAY52765 AAY64201 AAY87440 AAF99675 AAG09038 AAO46501 AAO46506 AAT12007 AAT12008 AAT12009 AAT12014 AAT12017 AAT76157 AAU00820 AAZ29576 ABB19082 ABB19117 ABB19194 AAO46266 AAQ04930 AAV97607 AAY98236 ABC86247 ABA55038 ABB19205 ABB19491 ABB19528 ABB19774 ABB19846 ABB86806 ABB88297 ABB19370 ABB19381 ABB19539 ABB19703 ABB19793 ABB19877 ABB83590 ABB87259 ABD79111 ABD94986 ABF21232 ABH03486 ABI20836 ABI84837 ABI93038 ABI94746 ABI94753 ABI94763 AAA43122 AAA43123 AAA43131 AAA43134 AAA43137 AAL31426 ABD14793 ABD59810 ABB88277 ABD59815 ABD61745 ABI84565 ABI84625 AAF99674 ABG88277 AAW59395 AAZ43415 ABB20059 ABB20443 ABB20456 ABB88193 AAO46252 AAT12001 AAT12003 AAT12010 AAT12011 AAT12021 AAU00814 AAY28414 AAW59405 AAY52752 AAY52753 AAY52761 ABB03100 ABB21771 ABB21793 ABB21803 ABB88378 ABB17746 ABB20020 ABB20151 ABB20175 ABB20282 ABB20293 ABB20498 ABB53750 ABB90213 ABC42760 ABD35774 ABD35779 ABD35781 ABD35786 ABD35787 ABD35796 ABD35797 ABD35806 ABD35807 ABD77839 ABC66508 ABD62852 ABD92952 ABF50828 ABF50829 ABF56629 ABI84657 ABI84704 ABI84926 ABI85214 ABI97497 ABI98909 ABC74412 ABJ16949 ABJ52573 AAQ04929 ABG38196 ABG88343 ABG88888 AAK18015 AAK18016 AAT12005 AAT12006 AAT90829 AAX76752 AAZ83322 ABB04315 ABB17274 ABB20386 ABB21813 ABB87450 ABB87636 ABB87666 ABB87832 ABB87843 ABB88054 ABB88341 ABC48790 ABC74411 ABI84498 ABI85057 ABI92279 ABB19754 ABB19896 ABB86795 ABD14794 ABG27050 ABH85397 ABI84485 ABI84544 ABI85028 ABI85047 ABI85094 ABI92268 ABI95161 AAL31435 ABI84854 ABI96728 ABJ51705 AAD49040 AAP04505 AAY52745 AAY52769 AAX47277 ABB18090 ABB21761 ABB88065 ABB90192 ABC66472 ABC66473 ABC66474 ABC66490 ABB03045 ABB19965 ABJ96580 ABI85116 ABI85145 ABI98937 ABJ16575 ABJ51683 ABJ96552 ABJ96553 ABD35799 ABD35808 ABD35809 ABF56620 ABD91845 ABF01673 ABF01683 ABF22672 ABI84865 ABI85137 ABI85162 ABI96708 ABI84754 ABJ90244 ABK34762 ABJ96507 ABJ96513 ABJ96547 ABJ96574 ABM21933 ABM21934 ABK59030 ABN59444 BAF46494 BAF46444 BAF46454 BAF46464 CAA23855 ABL67164 CAD20321 ABJ96522 BAB39505 AAO46866 BAF46424 ABJ96611 ABJ96617 ABL67153 ABL67142 CAF04463 CAF33010 ABL67241 ABN50766 AAO65605 BAF38379 BAF41911 CAC85079 ABB20095 ABB88137 ABD35777 ABD35778 ABD35790 ABD35793 ABD35794 ABD35800 ABD35803 ABD59806 ABD59808 ABD60976 ABN51020 BAE48319 BAE48328 CAC85077 ABI30366 ABI97343 ABJ16696 ABI84722 ABI84775 ABK13788 ABK13789 ABB04304 ABB20313 ABB20529 ABB90239 CAA67496 CAC85074 ABC66460 ABC66486 AAO46251 AAO46253 AAZ29574 ABB20361 ABB20371 ABB51965 ABB87560 ABB87625 ABD62791 ABH03494 ABI84533 ABI84826 ABI85067 ABI92312 ABI92994 AAK72397 AAL31431 AAO46864 AAT37560 AAX47278 ABB20239 ABD35795 ABD35801 ABD35802 ABD35805 ABE28410 ABI85182 ABI85193 ABJ09510 BAE07152 ABI36019 ABI84848 ABI85105 ABI85127 ABI85171 BAF46474 ABJ96615 ABJ96623 ABM22256 ABI84464 ABI84765 ABI84980 ABI84989 ABI92235 ABI92257 ABI95205 ABK32093 BAF37959

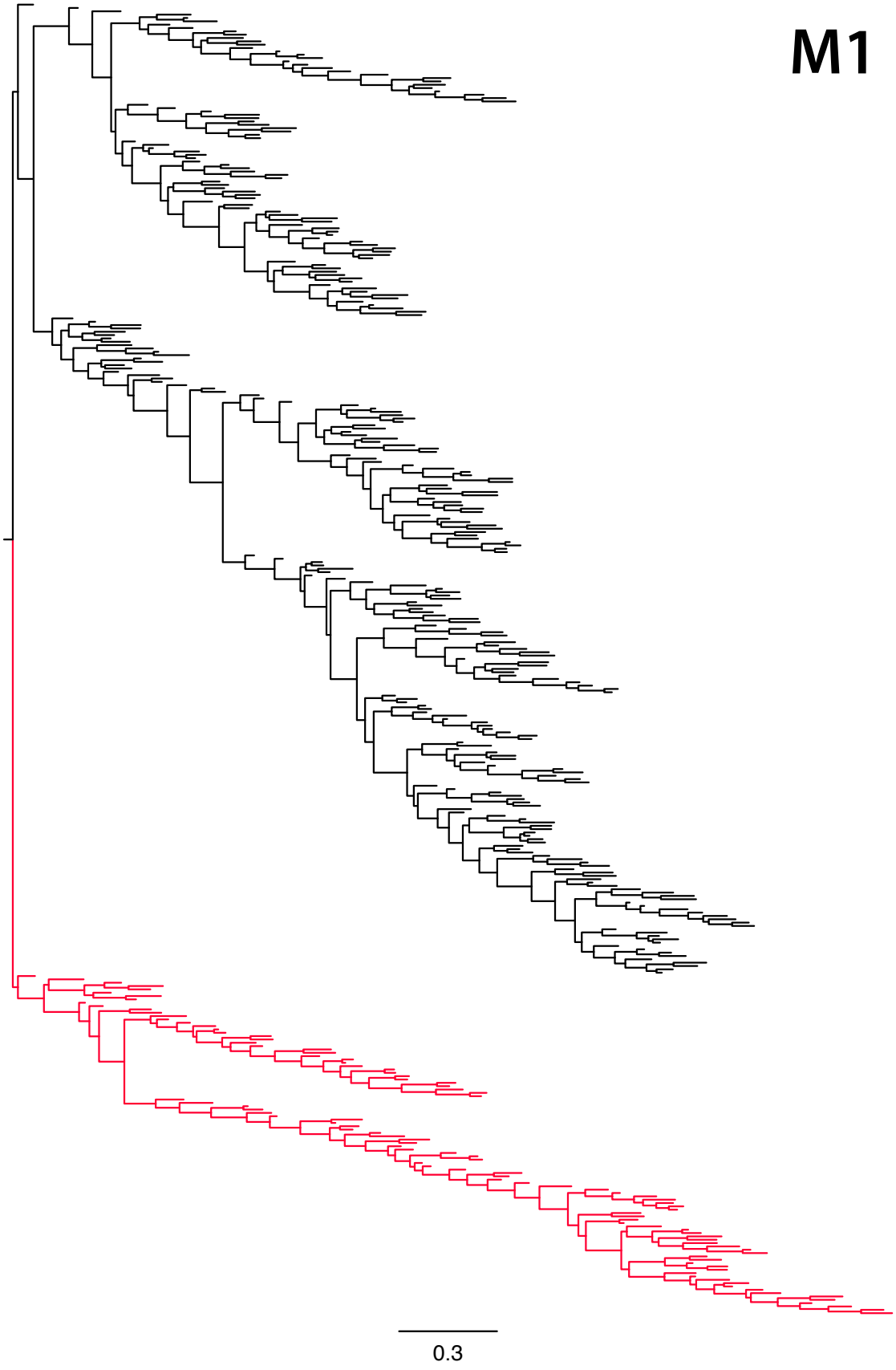
B. Trees used for analyses in chapter 3

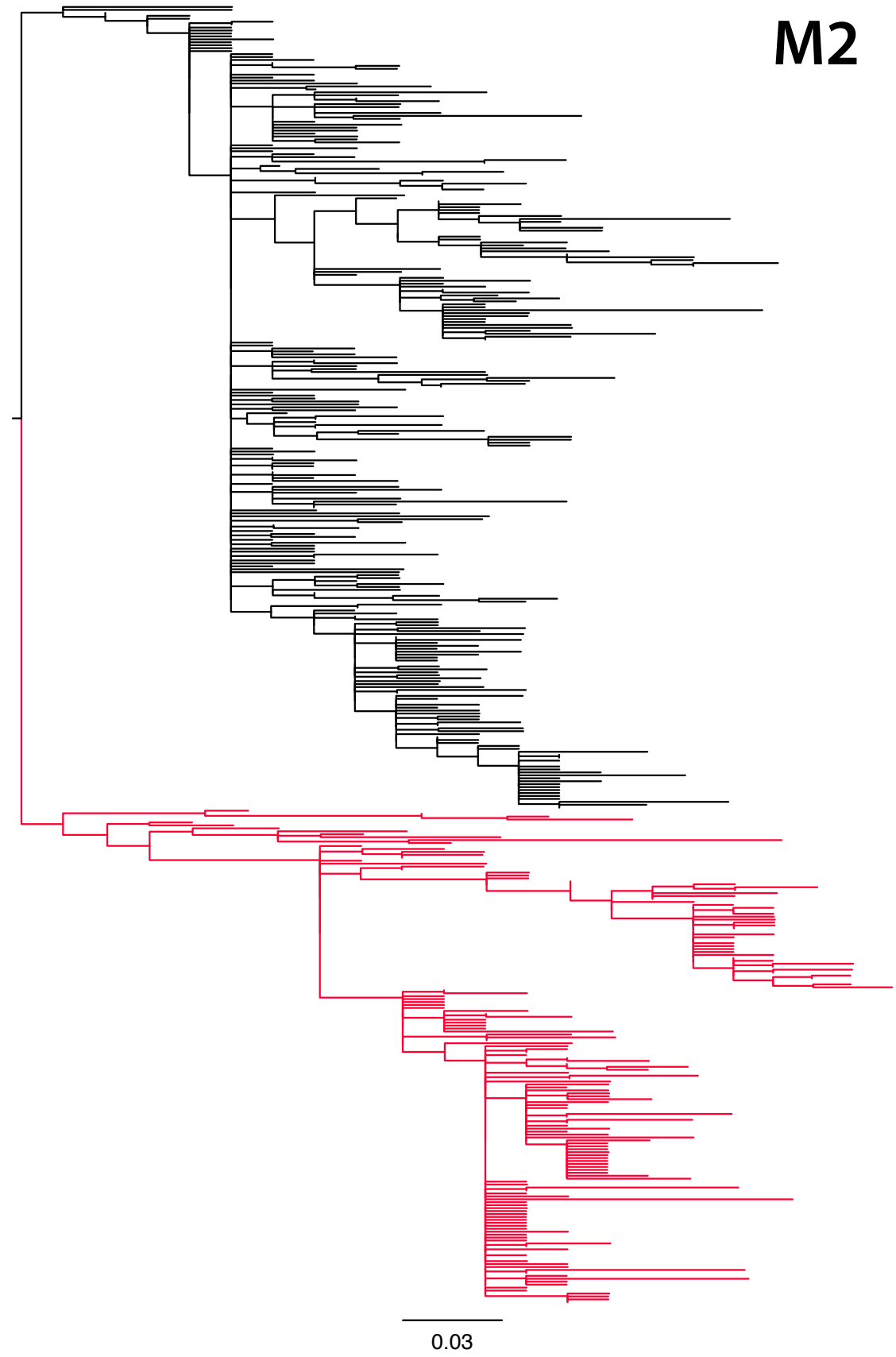
In the nonhomogeneous model, black branches are evolving under avian-specific constraints and red branches are evolving under human-specific constraints. The host-shift is assumed to have occured at the midpoint of the connecting avian-to-human branch.

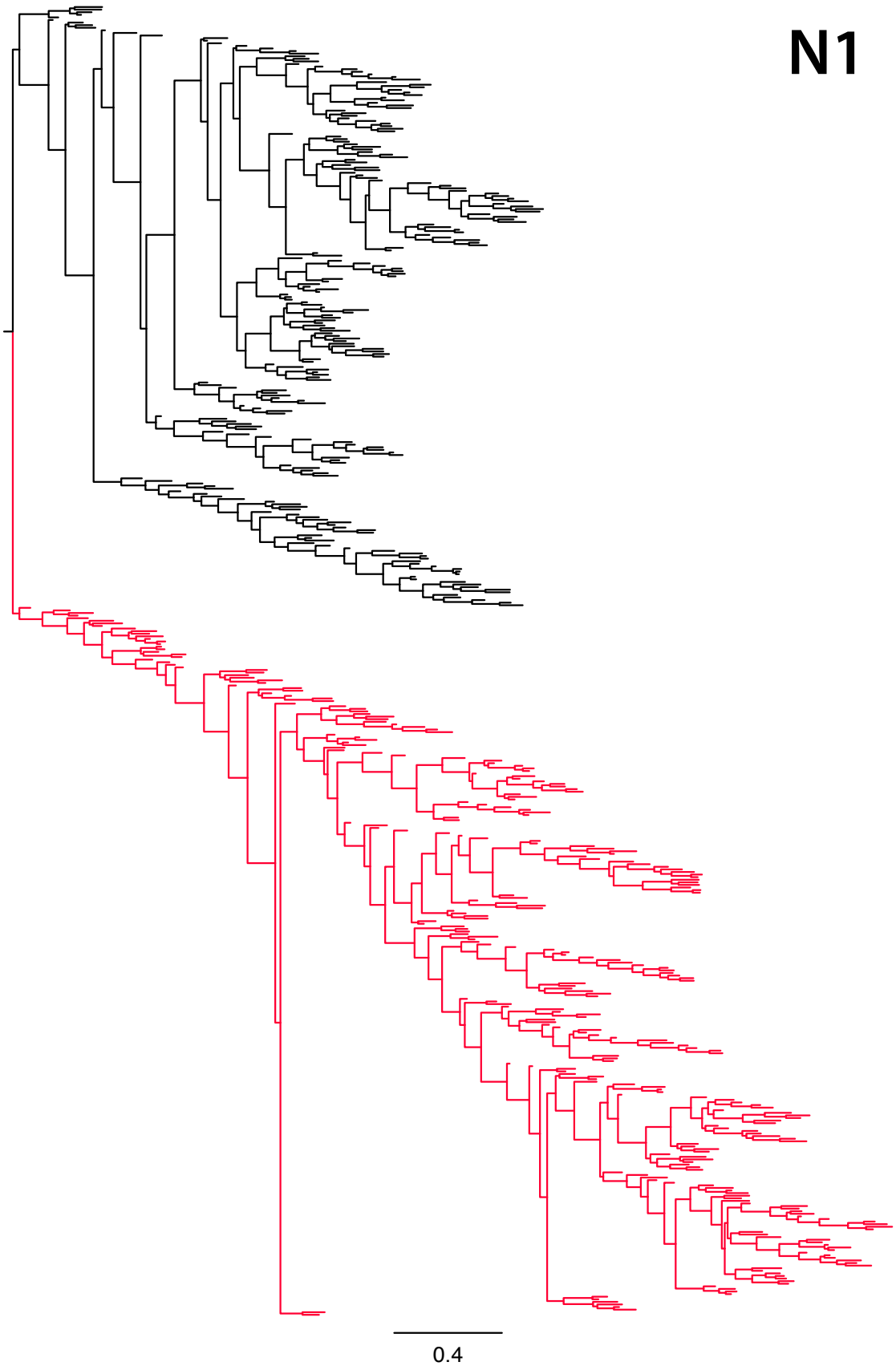

H1
0.4

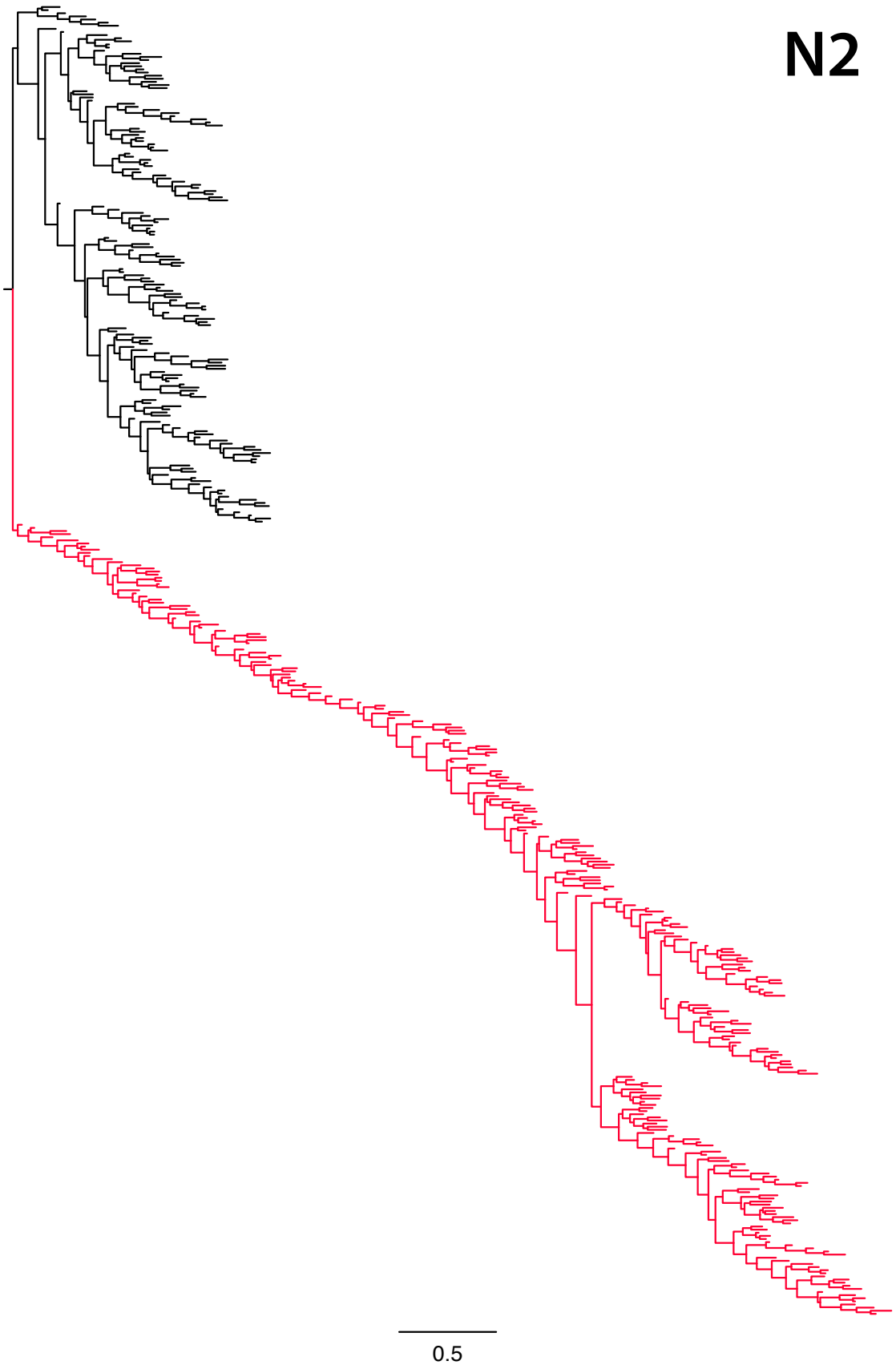

H2
0.02

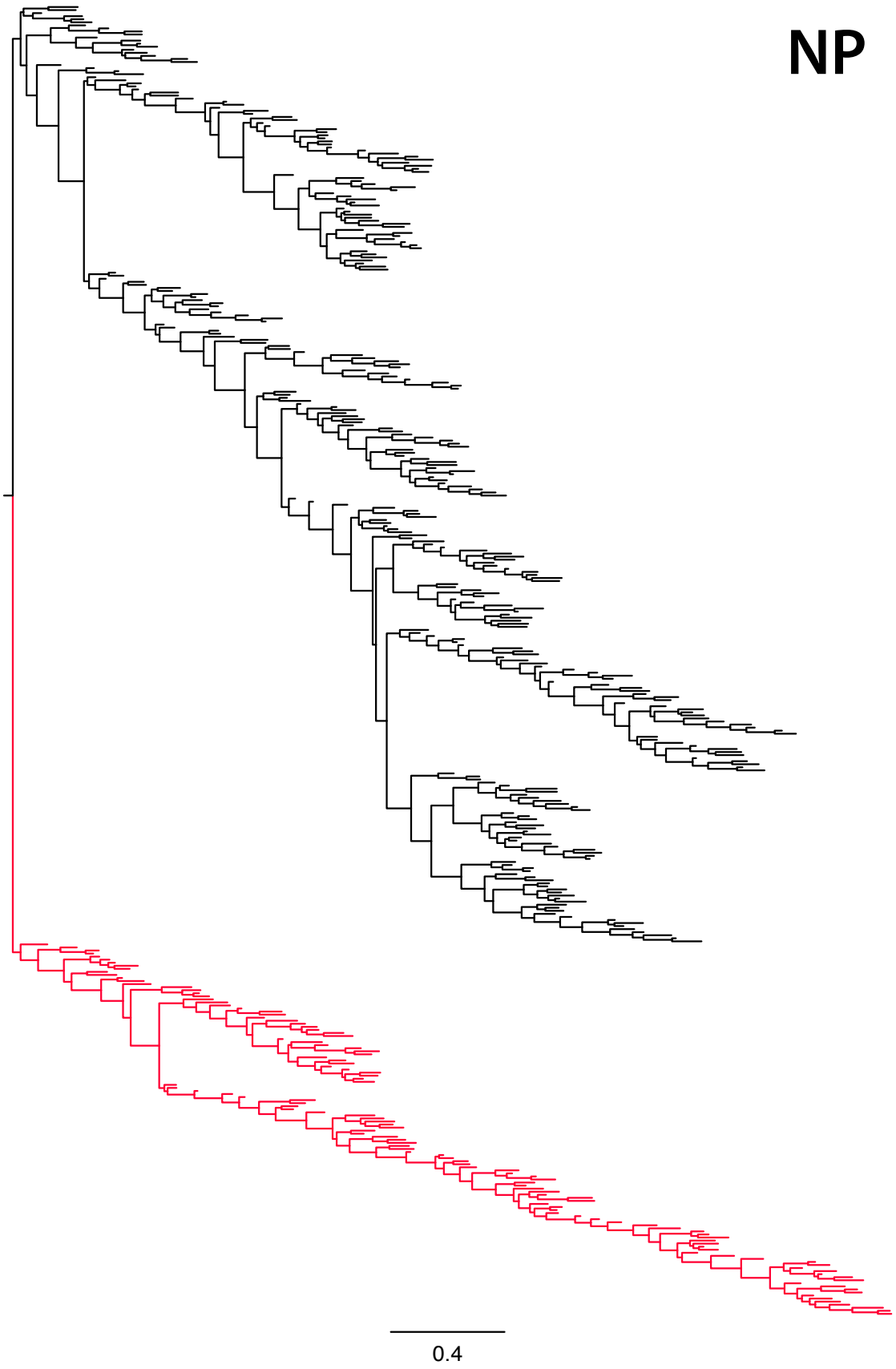

H3
0.04

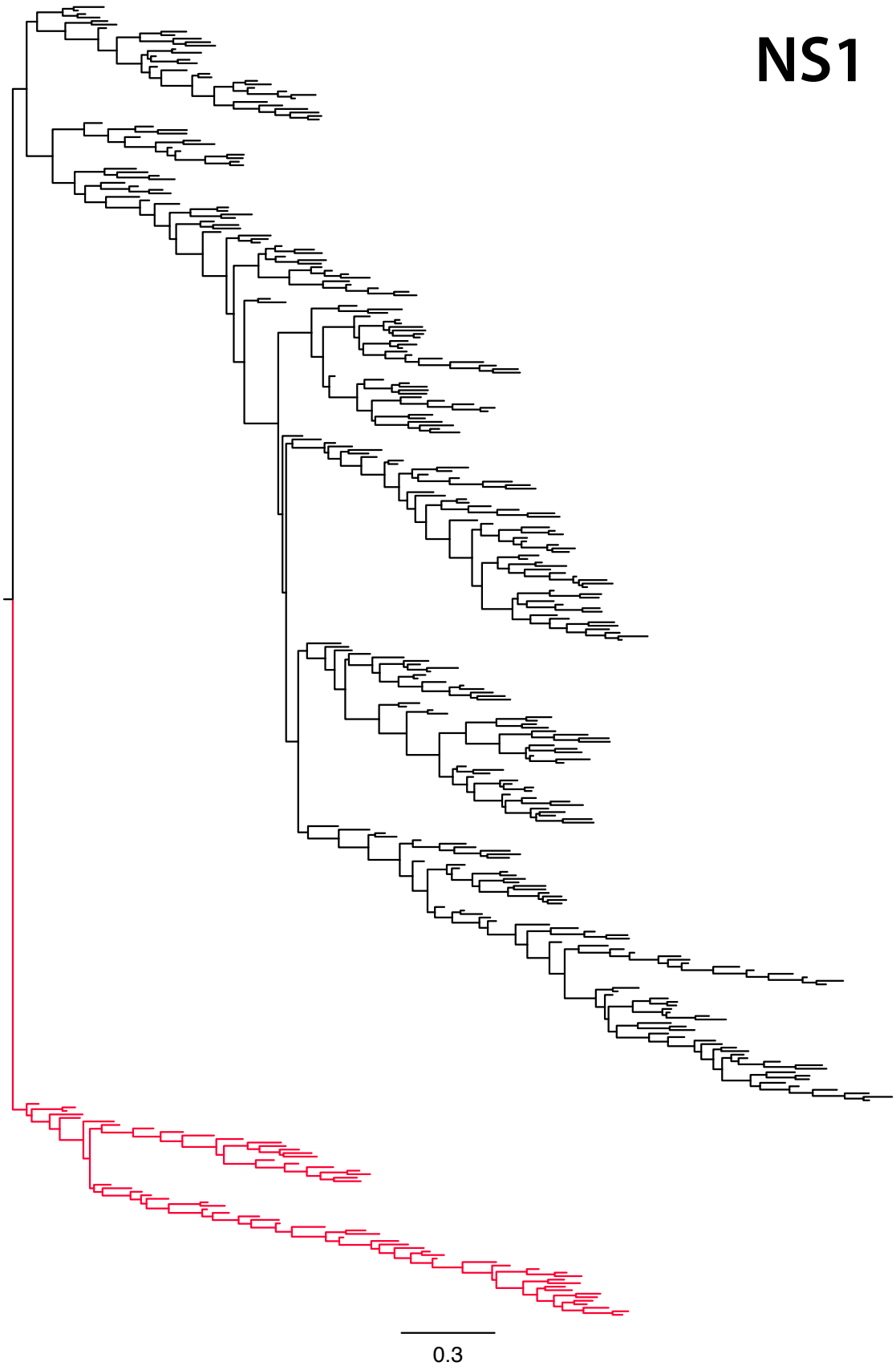

M1
0.3

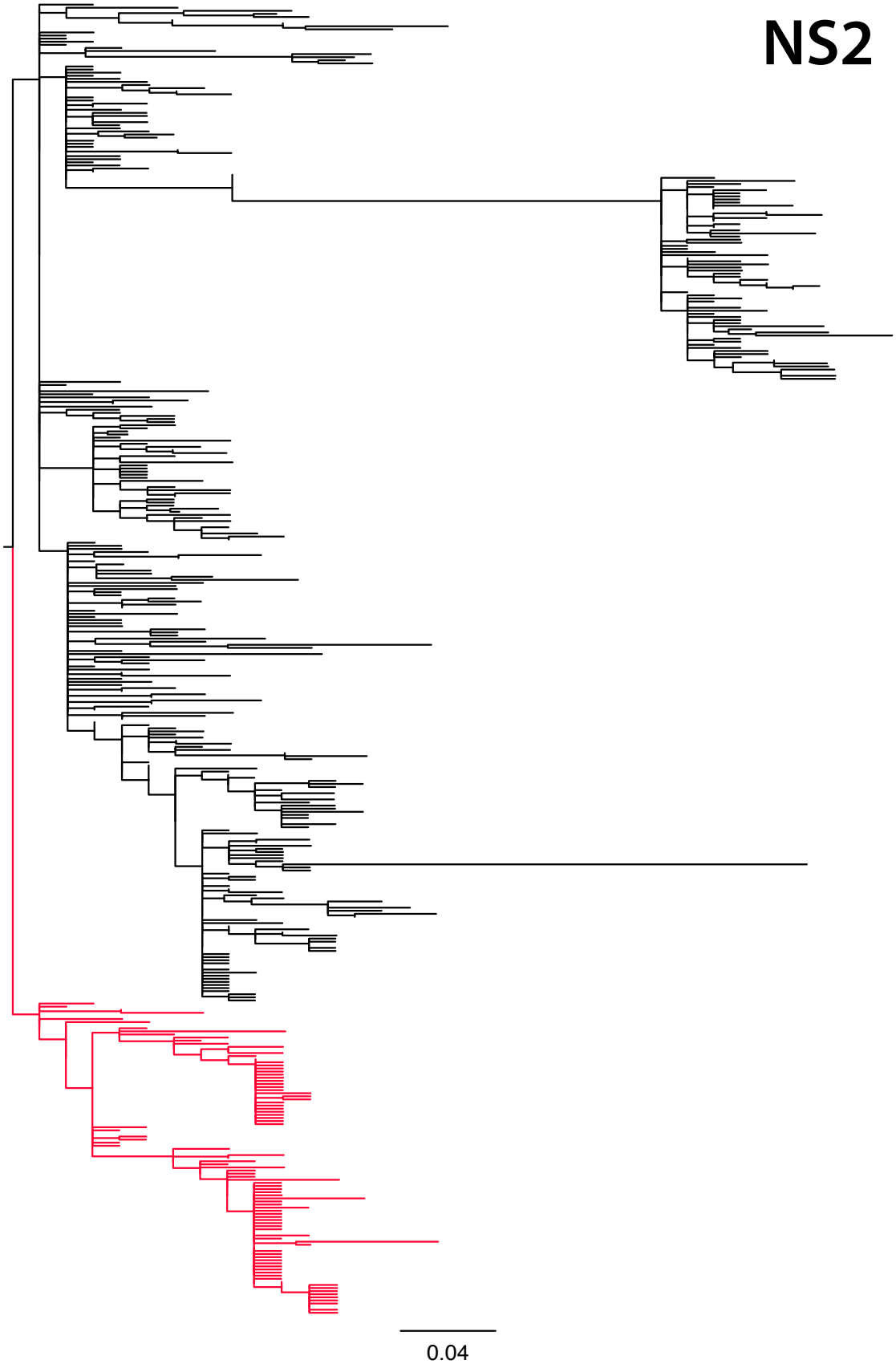

M2
0.03

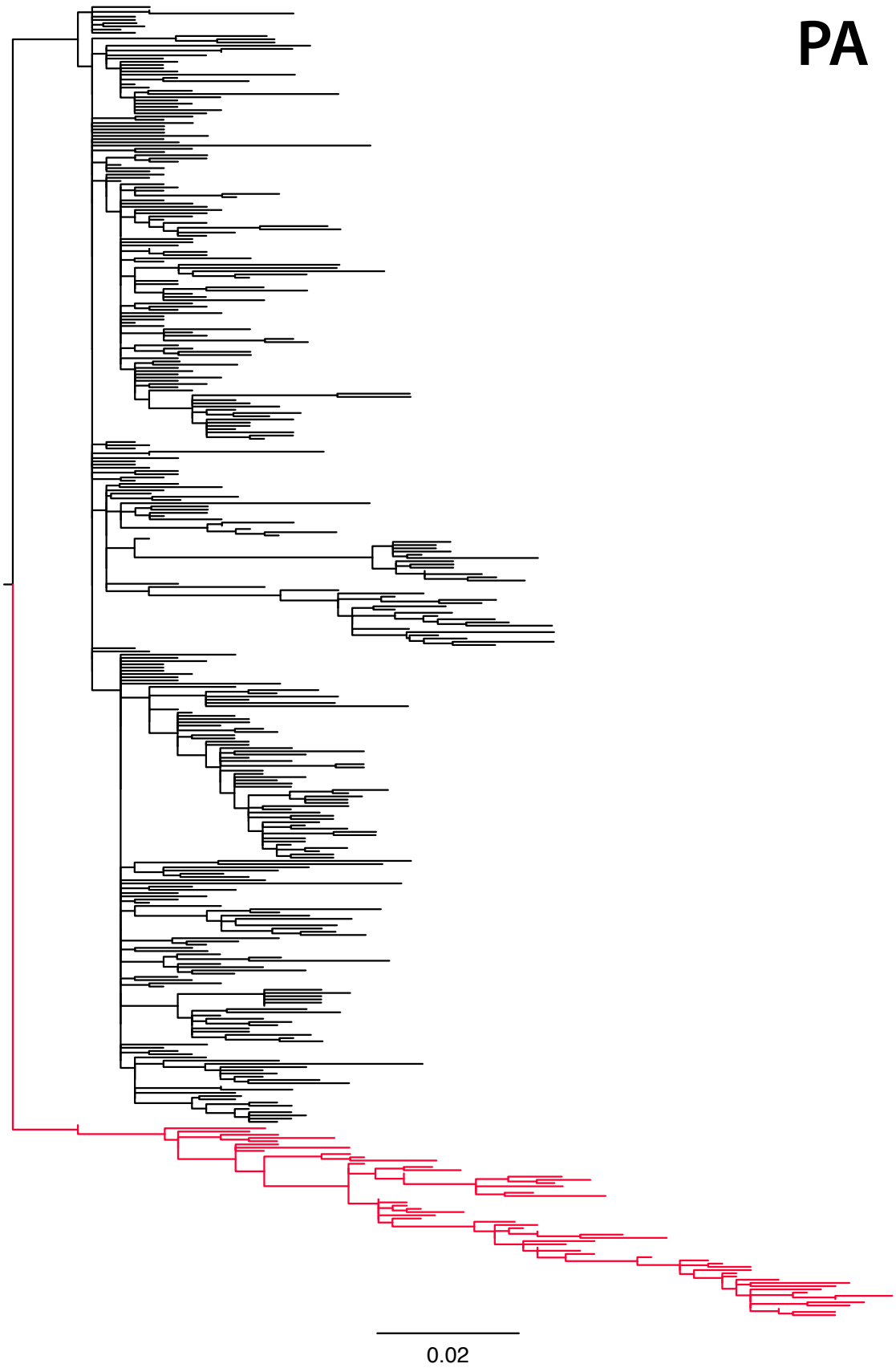

N1
0.4

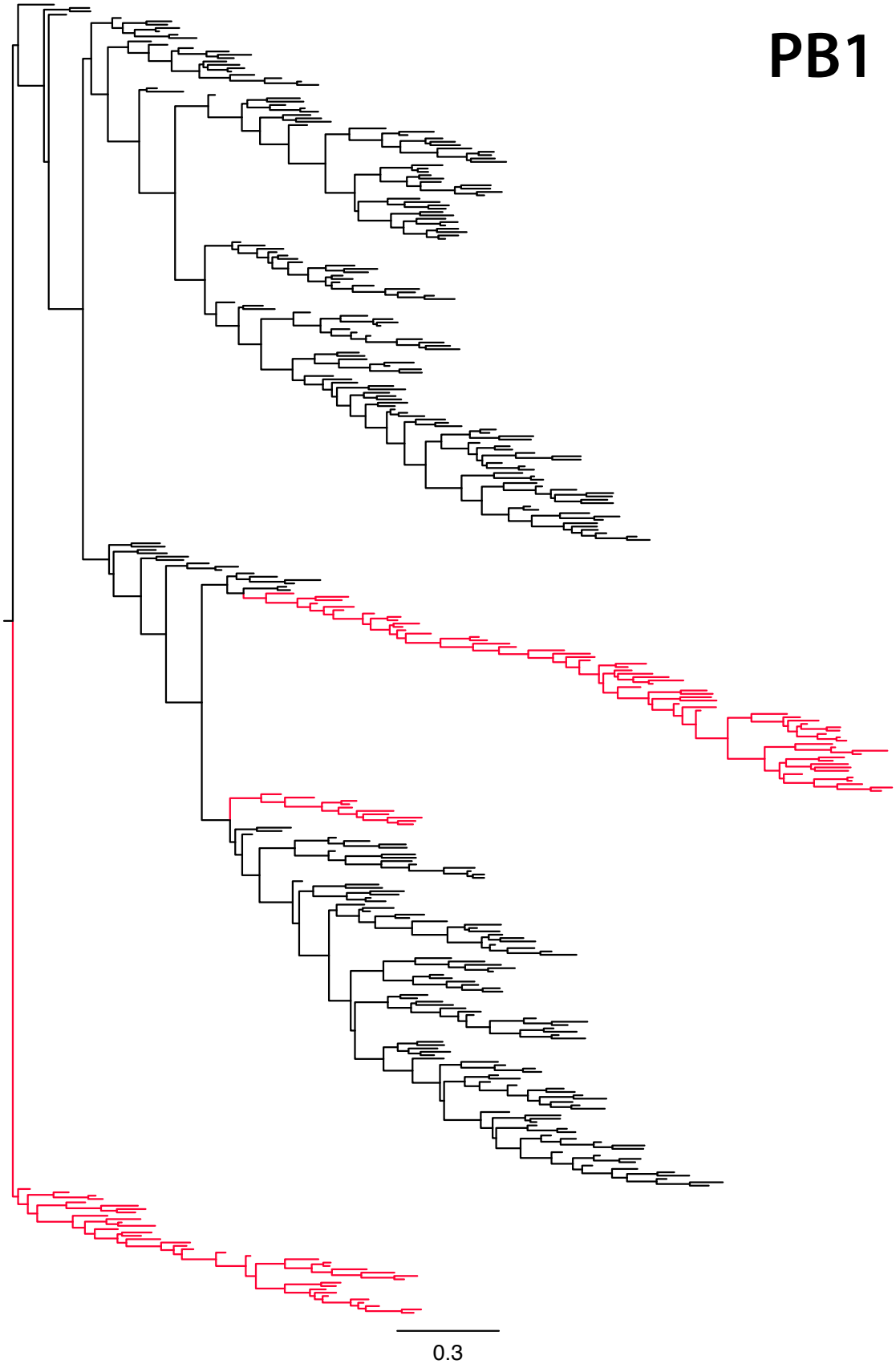

N2
0.5

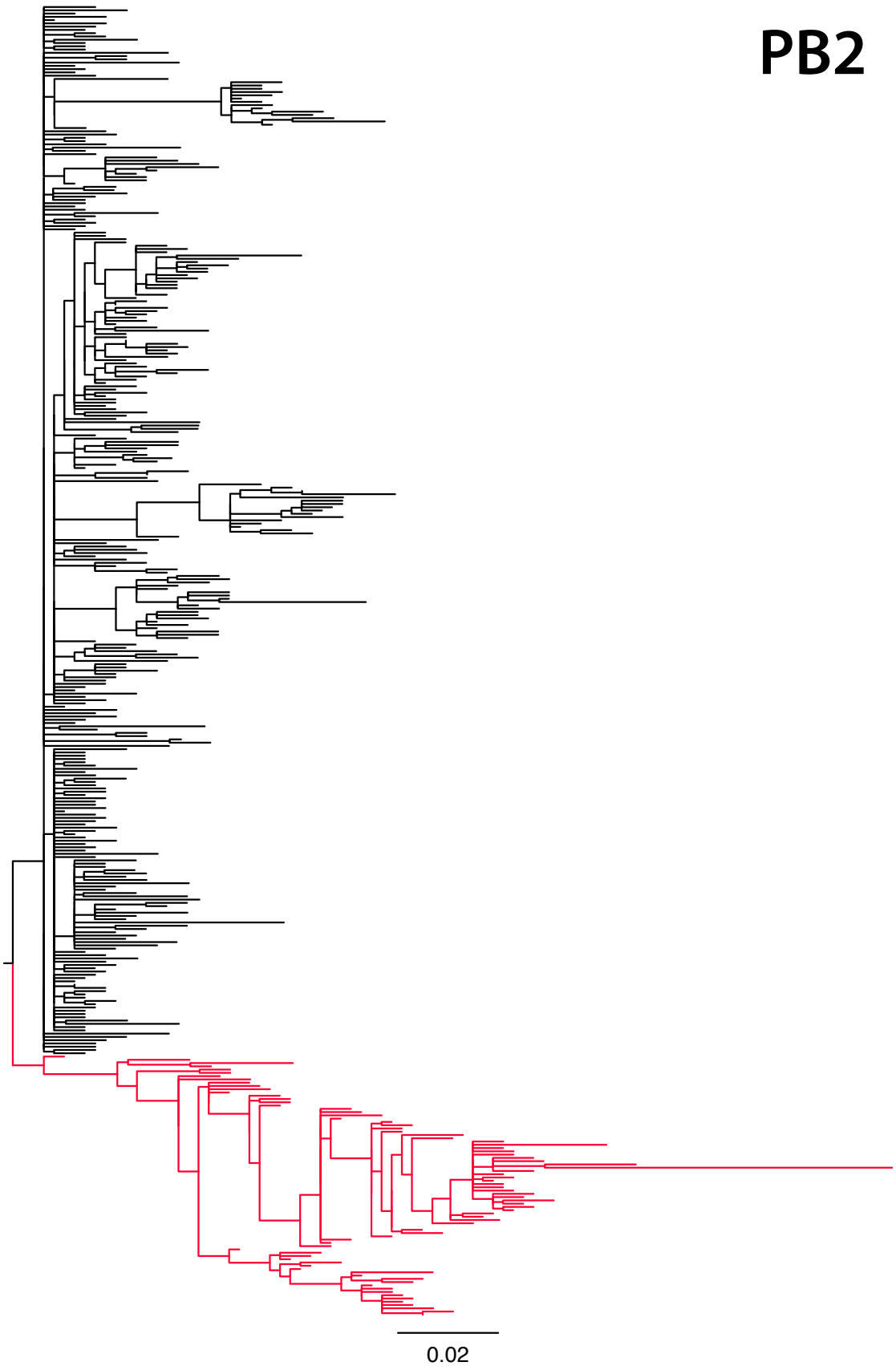

NP
0.4


NS1
0.3


NS2
0.04


PA
0.02


PB1
0.3


PB2
0.02

C. Influenza class 'B' sites (FDR < 0.02)

Table C.1.: Sites identified as undergoing changes in selective pressure during host shifts from birds to humans (FDR < 0.20). Residues are shown for amino acids with $\pi > 0.5$, $(\pi > 0.1)$ and $((\pi > 0.01))$.

Location	P value	FDR cutoff	Residues		Sel. constraint (d)	
			Avian	Human	Avian	Human
H1						
-5	3.05e-04	2.59e-03	E	K	2.74	2.85
2	5.09e-05	7.86e-04	F	L	3.30	2.63
7	1.58e-03	0.010	V((A))	A	2.59	2.91
8	4.47e-03	0.021	L	T	2.63	2.71
15	9.36e-03	0.038	V((I))	I	2.68	2.63
40	0.081	0.168	V(I)	V	2.34	2.82
53	0.070	0.148	S	L(R,K)	2.48	1.86
54	4.90e-05	7.86e-04	N	K	2.79	2.89
63	5.80e-03	0.025	K	N	2.89	2.79
65	0.023	0.065	N	S((N))	2.79	2.40
70	5.19e-03	0.023	L	I((V))	2.63	2.57
77	4.58e-03	0.021	D	E(G)	3.07	2.25
79A	0.052	0.121	L	I(F,V)	2.63	1.99
80	1.86e-03	0.011	T	S((P))	2.73	2.26
88	0.025	0.068	I	V((A,I))	2.65	2.42
91	2.57e-03	0.014	S(T)	P	2.06	3.13
93	0.023	0.065	S	P	2.50	3.21
103	0.024	0.066	I	A((T))	2.66	2.82
120	7.69e-04	5.68e-03	K	R	2.86	2.82
122	0.023	0.065	E(V)	E	2.29	2.74
125C	0.042	0.100	N(S)	S((R,N))	2.32	2.30
131	0.022	0.065	E	T((I,N))	2.74	2.36
133A	0.034	0.084	K	I((R,K))	2.89	2.19
136	0.034	0.084	T	S	2.73	2.50
138	0.011	0.042	A	S((A))	2.91	2.24
140	0.018	0.064	S((P))	S	2.30	2.47
141	1.55e-04	1.48e-03	Y	H	3.50	4.01
142	0.022	0.065	L(S,H)((A))	N((R,E,S,K))	1.76	1.78
144	0.019	0.065	A(G)	(E,K)((R))	2.40	1.92
154	4.62e-05	7.86e-04	I((L))	L	2.51	2.63
155	1.01e-05	7.42e-04	T(I)	T	2.20	2.73
156	0.025	0.066	K((E))	(E,R,G,K)	2.74	1.40
158	0.066	0.146	G	N((E,G,D))	2.60	2.39

Location	P value	FDR cutoff	Residues		Sel. constraint (d)	
			Avian	Human	Avian	Human
159	1.57e-04	1.48e-03	N(T)	G((S))	2.41	2.49
160	0.011	0.042	S	L((S))	2.50	2.41
163	1.89e-04	1.70e-03	K	N((T,S))	2.89	2.52
164	0.023	0.065	L(I)	L	2.17	2.63
167	0.023	0.065	S(T)	S	2.08	2.50
173	0.064	0.143	G((E))	E((G,K))	2.53	2.31
187	3.47e-03	0.017	T((N))	N((S))	2.61	2.71
188	4.16e-05	7.86e-04	T((V,A))	I((S,T,M))	2.23	2.20
189	3.84e-04	2.96e-03	S(G)((D,N))	G(K,R)((T,E,D))	1.64	1.36
190	1.76e-09	2.99e-07	E	D(V)((N))	2.74	2.20
192	0.012	0.044	Q	(K,M,R)	3.34	1.98
193	4.10e-03	0.020	N(E)((T,S))	(A,T,N)	1.80	1.84
196	0.070	0.148	Q	H(S,R)((Y,Q))	3.34	2.58
197	5.62e-03	0.025	N	T(K)	2.79	2.15
198	3.23e-04	2.62e-03	T((V,A))	E((G,V))	2.28	2.47
199	0.054	0.123	N((D))	N	2.52	2.73
205	0.081	0.168	G	V((L))	2.60	2.68
210	0.023	0.065	N	S	2.78	2.50
214	0.011	0.042	T((N))	T	2.49	2.70
217	0.023	0.065	I(L)	I	2.19	2.66
219	0.039	0.094	A	K((E,R))	2.91	2.62
222	2.46e-03	0.013	K((R))	K	2.62	2.89
225	6.46e-05	9.15e-04	G	D((G,N))	2.60	2.83
227	0.036	0.088	A	E(P,H)	2.90	2.11
230	0.030	0.078	M(I)	I((M))	2.85	2.59
238	1.82e-05	7.42e-04	D	E	3.07	2.70
239	4.90e-05	7.86e-04	Q	P	3.34	3.21
244	1.50e-03	0.010	T	I((M))	2.73	2.56
248	2.08e-03	0.012	T	N((S))	2.73	2.67
255	0.064	0.143	W	R(M)	3.95	2.47
261	1.21e-03	8.54e-03	N	S((N))	2.79	2.45
262	1.46e-04	1.48e-03	K	R	2.89	2.82
264	0.018	0.064	S	F(P)((L))	2.50	2.75
265	0.067	0.146	G(D)((E))	G((E))	1.92	2.50
271A	1.55e-04	1.48e-03	D	N	3.07	2.79
272	2.98e-03	0.015	A(T,V)	A	1.97	2.87
274	2.18e-05	7.42e-04	V((I))	M	2.73	3.42
275	0.023	0.065	H	D(G)((N))	4.01	2.26
279	0.011	0.042	T	A((S))	2.73	2.77
280	1.90e-03	0.011	R(K)	K	2.18	2.89
285	1.48e-05	7.42e-04	H((Y,R))	Q	3.60	3.20
288	4.90e-05	7.86e-04	L	I	2.63	2.66
300	7.61e-05	9.95e-04	I	V	2.66	2.79

Location	P value	FDR cutoff	Residues		Sel. constraint (d)	
			Avian	Human	Avian	Human
309	1.76e-03	0.011	V(I)	V	2.49	2.80
310	1.26e-04	1.48e-03	K	R	2.89	2.80
311	0.023	0.065	S(T)	S	2.08	2.50
312	0.032	0.082	T(V)	A(T)	2.28	2.13
317	0.029	0.077	A	V((A))	2.91	2.54
323	8.95e-03	0.037	V	I	2.83	2.61
H1(2)						
46	0.048	0.161	D	N(D)	3.07	2.43
72	3.32e-03	0.033	N	K	2.79	2.89
77	2.08e-04	4.17e-03	I	M	2.66	3.47
91	0.023	0.117	V	I((V))	2.83	2.52
110	6.38e-03	0.051	F(L)	F	2.71	3.27
116	2.07e-04	4.17e-03	R	K	2.83	2.89
123	8.64e-03	0.058	R(K)	K	2.19	2.86
127	7.99e-04	0.011	R	K	2.83	2.89
172	0.011	0.064	E	K((R))	2.74	2.75
198	0.033	0.120	V(I)	V	2.34	2.83
199	0.033	0.120	L(W)	L	2.40	2.63
205	0.028	0.120	I((L))	I	2.40	2.66
H2						
9	0.035	0.147	(E,K,R)	R	1.81	2.83
80	3.48e-03	0.070	(S,K,N,T)	R((K))	1.38	2.72
94	9.21e-03	0.084	N(H)	Y(D)((N))	2.54	2.80
95	0.023	0.116	G	S	2.60	2.50
116	0.014	0.102	T((R,K))	K	2.26	2.79
121	0.026	0.118	I(V)	V	2.05	2.83
122	3.33e-03	0.070	K((R))	R(K)	2.77	2.33
125A	6.74e-03	0.083	R((K))	K((R))	2.57	2.74
126	0.043	0.157	Q(R,E)	G(R)	2.07	2.15
131	0.023	0.116	T	E(K)((T))	2.73	2.07
137	0.010	0.085	Q(R)	M(K,R)	2.63	2.37
144	0.040	0.154	N(G)((K))	E(K)	1.86	2.12
152	0.034	0.147	V	I	2.83	2.66
159	0.023	0.116	S	P	2.50	3.21
182	6.43e-03	0.083	I(V)	V	2.11	2.83
186	5.32e-04	0.018	N	I(N)((K))	2.78	1.95
197	3.62e-04	0.018	N	E(K,N)	2.78	2.01
204	0.024	0.116	V	A	2.83	2.90
205	1.11e-04	0.011	G	S(V)	2.60	1.96
216	0.024	0.116	E	K(D)	2.74	2.31
219	0.023	0.116	T	A	2.73	2.91
226	7.02e-03	0.083	Q	L((Q))	3.34	2.45
228	0.014	0.102	G	S(G)	2.60	2.13

Location	P value	FDR cutoff	Residues		Sel. constraint (d)	
			Avian	Human	Avian	Human
252	0.023	0.116	I	V	2.66	2.83
275	8.27e-03	0.083	E(D)	G(E)	2.43	2.06
297	0.040	0.154	I(V)	V	2.26	2.83
313	8.17e-03	0.083	R(K)	K	2.35	2.89
H2(2)						
45	5.36e-03	0.042	I(V)	F	2.28	3.30
130	5.38e-03	0.042	A	V((A))	2.91	2.57
169	5.47e-03	0.042	N	K	2.79	2.89
180	2.11e-03	0.042	N((S))	S	2.63	2.50
202	0.018	0.111	I((M))	M(V,I)	2.45	2.13
H3						
-14	0.023	0.121	K((R))	K	2.57	2.86
-13	0.040	0.166	T(A)	T	2.29	2.71
-12	0.030	0.148	I(V)	I	2.16	2.66
-7	2.50e-03	0.037	C(Y)	Y	2.92	3.43
-6	0.045	0.166	(I,F,L)	I	1.78	2.58
-5	5.20e-03	0.066	L(F)	S(L)	2.25	2.02
-2	0.020	0.112	A(V,T)	V(G)((A,T))	2.12	1.92
0	5.04e-05	5.70e-03	(G,S,C)	A(T)((S))	1.79	2.13
2	0.059	0.199	N(D)	K(E,N)	2.23	2.19
3	0.023	0.121	(P,Y,L)((H))	L((T,I,P,F))	1.92	1.67
4	2.10e-03	0.035	S((P))	P(S)	2.30	2.70
6	0.016	0.102	N(S)	N((I))	2.27	2.67
7	0.013	0.094	N(G,D)	D	1.82	3.04
9	0.039	0.166	(N,S,D)	S	1.70	2.47
31	0.017	0.102	D((N))	N(S)((D))	2.93	2.20
50	3.70e-03	0.050	K((R))	G(E,R)((K))	2.56	1.66
53	0.016	0.102	N	(D,G,N)	2.79	1.78
57	1.10e-03	0.028	K((R))	Q(R)	2.60	2.72
62	0.029	0.148	R((I))	(G,E,M)((I,K))	2.67	1.43
63	7.32e-04	0.024	D	N	3.07	2.79
67	1.30e-03	0.028	(I,V,M)	I	1.86	2.60
81	0.011	0.089	D((N))	N((D))	2.75	2.73
82	0.043	0.166	A(E)	K(E)	2.27	2.42
83	0.043	0.166	T	K(E,N)	2.73	1.88
92	6.95e-05	5.70e-03	N((S))	K((T,N))	2.49	2.48
94	0.011	0.089	F	H(Y)	3.30	3.21
121	0.039	0.166	I	(T,N,F)((K))	2.66	1.60
126	8.40e-03	0.087	T	N((D))	2.73	2.56
137	0.011	0.089	G(S)((N))	Y(S,F)	1.84	2.15
143	0.016	0.102	P	S	3.21	2.50
144	0.045	0.166	A(T)((V))	(D,N,G)((V,A,T,I))	2.17	1.10
145	1.40e-03	0.028	N(R,S)	K(N)((S))	1.98	2.00

Location	P value	FDR cutoff	Residues		Sel. constraint (d)	
			Avian	Human	Avian	Human
146	0.040	0.166	G	S(G)	2.60	2.15
160	6.30e-03	0.069	(A,T,R)	K(A)((R))	1.73	2.26
163	0.012	0.092	L(V)	A((V))	2.31	2.62
190	5.60e-03	0.066	E	D	2.74	3.06
192	0.031	0.151	T	I(T)	2.73	2.04
196	0.044	0.166	V((I))	A(V)((I))	2.68	2.28
213	1.17e-04	6.50e-03	I	V	2.66	2.83
214	0.055	0.192	I	T((I))	2.66	2.67
222	0.035	0.166	W	R	3.95	2.83
228	5.59e-04	0.023	G	S	2.60	2.50
244	1.80e-03	0.033	V	L((I))	2.83	2.42
248	0.043	0.166	N	T((S))	2.79	2.43
269	0.014	0.101	R(K)	R((K))	2.17	2.76
275	9.50e-03	0.089	D	G(D)((S))	3.07	2.02
307	0.059	0.199	K	R(K)	2.89	2.38
312	0.048	0.171	(G,N,S)((D))	N((K,T))	1.46	2.42
313	0.019	0.112	T(S)	T	2.18	2.73
H3(2)						
55	8.30e-03	0.188	V	L(I)	2.83	2.09
57	0.024	0.188	E	G	2.74	2.60
117	0.018	0.188	K((R))	K	2.77	2.89
147	4.30e-03	0.188	A(S)	A((T))	2.31	2.76
161	0.013	0.188	I(V)	V	2.05	2.83
196	0.025	0.188	F(L)	F	2.83	3.27
M1						
77	0.018	0.190	R((K))	R((I,T))	2.77	2.67
101	0.011	0.190	K(R)	R	2.45	2.79
115	2.05e-04	0.011	V	I((V))	2.79	2.55
121	0.011	0.190	T(A)	A	2.27	2.87
137	1.30e-04	0.011	T	A((T))	2.71	2.76
144	0.022	0.190	F(L)	F((L))	2.32	3.21
147	0.013	0.190	V	V(I)	2.83	2.13
167	0.021	0.190	T((A,I))	(I,T,A)	2.45	1.70
174	1.06e-03	0.029	R	K(R)	2.80	2.20
208	0.019	0.190	Q	Q((K))	3.34	3.23
222	0.021	0.190	H(Q,R)	P(R,H)	3.07	2.37
224	0.022	0.190	S(N)	S	1.94	2.50
231	4.64e-04	0.017	D	(N,D,S)	3.02	1.70
M2						
10	9.63e-04	0.039	(L,H,P)	P	2.20	3.21
20	0.030	0.197	S(N)	N((S))	1.96	2.62
36	0.017	0.184	L	V(M)((L))	2.63	2.43
44	0.013	0.184	D((N))	D	2.88	3.07

Location	P value	FDR cutoff	Residues		Sel. constraint (d)	
			Avian	Human	Avian	Human
48	0.019	0.184	F	S	3.30	2.50
54	0.024	0.184	R((C))	(L,H,T,F)((I,C))	2.75	1.52
56	3.00e-03	0.082	K((R))	(E,K,N,R)	2.67	1.53
65	0.018	0.184	T((M,K))	T	2.55	2.73
78	8.30e-03	0.168	Q((R,H))	(E,N,K,R)((Q))	3.07	1.32
86	0.025	0.184	V((A,I))	A(T)((V))	2.64	2.34
92	0.032	0.197	V	V(C)	2.79	2.51
93	4.26e-04	0.035	N((Y,I,S))	(S,I,Q)((N))	2.39	1.49
96	0.023	0.184	L(K)((M))	A(Q)	2.04	2.41
N1						
3	5.96e-03	0.040	P	P((T,S))	3.21	3.04
5	0.077	0.171	Q((R))	Q	3.27	3.34
8	0.015	0.075	I(T)((V))	I	2.24	2.63
12	0.098	0.195	S(Y)	S	2.25	2.50
13	9.91e-03	0.055	I(V,T)	I	1.91	2.63
14	0.016	0.079	C	S	3.69	2.50
16	0.023	0.093	V(A)((T,I))	A(T)((V))	1.99	2.15
20	0.045	0.128	V(I)	I((L))	2.18	2.50
26	0.075	0.171	(T,I,L)((V))	I	1.51	2.65
29	7.25e-03	0.043	M(I)	I	2.81	2.64
34	4.03e-03	0.031	V((G,I,A))	(I,V,A)	2.36	1.71
40	0.047	0.128	T	T	2.68	2.64
41	0.046	0.128	G((E,R))	G	2.44	2.56
42	3.43e-03	0.028	(G,N)((S,D))	S((N))	1.63	2.35
43	0.091	0.187	Q((L,R))	Q	3.01	3.34
46	1.96e-03	0.017	(A,P,V,T,S)	T	1.35	2.72
47	4.34e-03	0.031	E(G)((D))	G	1.92	2.59
51	0.039	0.118	Q((P))	Q	3.24	3.29
52	1.42e-03	0.013	S	R((G,N,K))	2.50	2.51
53	0.060	0.155	V(I)	I((S))	2.12	2.42
59	2.08e-03	0.018	N((K))	S((N,R))	2.65	2.19
64	0.067	0.163	Q	H(N)	3.34	3.04
66	0.085	0.182	Y((F))	Y	3.38	3.50
67	6.62e-03	0.043	(L,I,V)	V	1.63	2.83
69	0.025	0.099	I((V))	I	2.58	2.66
70	0.022	0.093	I(S)((R,N))	N((S))	1.90	2.44
71	0.069	0.163	N(V)	N	2.41	2.79
72	0.069	0.163	T	T	2.68	2.73
74	3.23e-04	6.33e-03	(L,F,S,V)((I))	V	1.28	2.81
75	0.032	0.109	P(A)((L,V,I))	V(I)	1.98	2.10
76	0.074	0.170	T(A,G)((N,V))	A((V,T))	1.42	2.65
78	0.069	0.163	(K,Q,N)	K((E,N,Q))	1.94	2.51
79	0.012	0.063	A(T,D)	D((G))	2.00	2.98

Location	P value	FDR cutoff	Residues		Sel. constraint (d)	
			Avian	Human	Avian	Human
80	7.93e-04	9.32e-03	V((R,A,M))	(I,V,K)((T,S))	2.21	1.32
81	0.023	0.093	(A,D,I,T,V)	T	1.28	2.71
82	0.043	0.125	P(S)	S	2.20	2.47
83	0.057	0.151	(K,A,V,M)((I))	V(M)	1.48	2.44
85	0.034	0.112	L(I)	L	2.14	2.63
93	0.100	0.195	P	S(P)	3.17	2.12
95	0.070	0.163	(N,R,S)	S(R)	1.63	2.05
99	0.018	0.083	I(V)	I	2.05	2.66
101	0.016	0.079	S	T	2.50	2.73
105	0.011	0.060	S(G)	S((G))	2.01	2.44
111	0.036	0.113	K(R)	K	2.18	2.89
114	0.055	0.148	V	V((I))	2.83	2.77
116	0.078	0.171	V	V	2.78	2.83
136	0.062	0.156	Q	Q(K)	3.30	2.50
149	0.030	0.108	V((A))	V(F,I)	2.70	2.05
157	2.11e-04	5.47e-03	T	A	2.71	2.87
189	3.08e-08	6.04e-06	S((G))	G	2.40	2.60
195	0.086	0.182	I(V)	I	2.28	2.66
200	0.041	0.120	N	N(D)	2.77	2.25
206	0.086	0.182	L(V)	L	2.21	2.63
210	0.078	0.171	G	G	2.56	2.60
211	0.032	0.109	I((M))	I	2.53	2.66
214	4.73e-06	3.09e-04	D	E(G)	3.07	2.15
220	4.22e-04	6.82e-03	R((G))	K(R)	2.57	2.37
221	4.52e-04	6.82e-03	N(G)	K	2.41	2.87
222	0.022	0.093	N(D,S)	Q(R)((E,K))	2.08	2.29
223	0.090	0.187	I((T))	I	2.51	2.66
232	0.020	0.091	A	V((T,A))	2.86	2.48
241	0.011	0.060	V(I)	I	2.17	2.66
250	0.021	0.093	Q	(L,A)((Q,T,P))	3.34	1.44
257	0.034	0.112	K(R)	K	2.37	2.89
258	0.038	0.118	I(M)	I	2.34	2.66
263	0.030	0.108	V	V(I)	2.83	2.12
264	8.09e-04	9.32e-03	(I,A,V)	T	1.71	2.68
267	0.035	0.113	V((A))	I((M,V))	2.74	2.40
273	0.079	0.171	N	N	2.74	2.79
274	2.80e-05	9.13e-04	Y	(S,F)((Y))	3.50	2.12
285	0.019	0.088	A(S)	T	2.49	2.67
287	0.031	0.108	E(K)	T(I)	2.13	2.28
288	5.84e-04	8.18e-03	I(V)	V	2.08	2.83
289	8.64e-04	9.41e-03	(I,T,M)	M	1.82	3.47
309	6.95e-03	0.043	N(D)	N	2.37	2.79
311	2.12e-05	8.30e-04	E((D))	D	2.59	3.07

Location	P value	FDR cutoff	Residues		Sel. constraint (d)	
			Avian	Human	Avian	Human
329	7.22e-03	0.043	N	K(E)((R))	2.79	2.43
339	1.27e-03	0.012	S((L))	T(Y)((N))	2.40	2.11
340	1.18e-05	5.80e-04	(L,S,P)((H))	V((A,H,P))	1.55	2.26
341	4.13e-04	6.82e-03	N	D	2.76	3.05
351	6.82e-04	8.91e-03	F((Y))	Y	3.01	3.49
354	0.028	0.108	G	D	2.60	3.07
355	0.092	0.188	N((K,S))	N((D))	2.66	2.73
365	1.20e-03	0.012	T(I,P)	N((S,T))	2.05	2.59
367	0.062	0.156	S	L((I,F))	2.50	2.31
369	0.048	0.130	S	K(H)((Q,R))	2.50	2.43
382	8.60e-08	8.42e-06	E((G,D))	D(N)	2.35	2.30
386	8.80e-03	0.051	S(N,E)((R,D))	(R,K,S)((D,N))	1.52	1.38
388	0.030	0.108	S((L))	S(L)((F))	2.42	1.95
390	0.059	0.155	K	K((R))	2.84	2.61
393	4.35e-03	0.031	I	V(I)	2.63	2.30
394	0.031	0.108	V((I))	V	2.76	2.83
396	0.094	0.189	I((T))	I(T,V,M)	2.58	1.60
427	4.44e-03	0.031	I	V(I)	2.66	2.46
430	2.23e-04	5.47e-03	R((L))	L((Q,R))	2.59	2.30
432	0.042	0.124	K(Q)((E))	(R,E,K)	2.32	1.78
434	0.040	0.120	(G,N,S,E)((K))	N((G,E,R,T))	1.16	1.80
451	0.095	0.191	S(G)	G(S)	2.16	2.11
454	0.070	0.163	V((I))	A(V)	2.53	2.51
455	2.60e-04	5.66e-03	G(S,D)	N(D)	1.65	2.47
N2						
7	0.052	0.125	I((T))	I	2.59	2.66
9	0.024	0.087	T(A)	T	2.12	2.73
19	0.080	0.159	I(T)((A,S))	T((I,A))	1.77	2.41
22	0.023	0.086	L(F)	F	2.26	3.27
24	0.073	0.150	M	M((T))	3.45	3.29
26	0.048	0.125	I	T	2.66	2.73
28	9.10e-03	0.055	L(I)	I((V))	1.96	2.35
31	0.012	0.057	T(M)	T((I))	2.37	2.63
33	0.012	0.057	V(M,I)	V	2.08	2.83
38	0.084	0.163	K((R,N))	K	2.35	2.86
39	0.048	0.125	Q(P,H)	Q	2.41	3.32
40	0.032	0.099	(N,T,S)((H))	Y(C,H)	1.55	2.70
41	1.58e-03	0.025	E((G))	E	2.43	2.74
42	0.045	0.124	C	F(Y)	3.66	2.90
44	8.26e-03	0.055	(I,P,T,N)((K,V,S))	(P,V,I)	1.19	1.81
45	0.043	0.124	(P,L,S)((T))	P((L))	1.60	3.05
48	0.072	0.150	N(S)((D))	N	1.87	2.78
50	4.12e-03	0.039	V(A)((T,I))	V((A))	1.98	2.69

Location	P value	FDR cutoff	Residues		Sel. constraint (d)	
			Avian	Human	Avian	Human
51	5.36e-04	0.012	V((M,T))	M	2.39	3.43
52	0.071	0.150	P	L(P)	3.21	2.19
57	0.012	0.057	I(T)((V))	I((T))	2.13	2.56
58	0.082	0.161	I((V))	I((L))	2.45	2.59
59	0.049	0.125	E((K))	E	2.67	2.74
60	5.79e-03	0.047	R(K)	R	2.30	2.83
62	4.28e-03	0.039	(I,T,M)((V))	I((T))	1.59	2.59
66	0.046	0.124	V((I))	V	2.76	2.83
69	9.91e-03	0.055	N	T(A)	2.79	2.33
70	1.82e-04	6.14e-03	S(H,N)	N	1.96	2.78
72	2.10e-03	0.027	T(I)((V))	T	2.00	2.70
73	0.046	0.124	(L,T,I)	I	1.59	2.66
77	2.05e-03	0.027	(I,K,T)((V,L))	I((K))	1.36	2.32
79	0.012	0.057	(P,L)((S))	P	1.96	3.21
81	2.83e-04	8.22e-03	A((V,M,L,I,T))	(P,L,A,T)	1.92	1.62
83	7.97e-05	3.24e-03	G(E,D)((K))	E	1.61	2.70
85	0.052	0.125	R(K)	R	2.40	2.83
86	0.010	0.055	(N,D,S,T)	N	1.53	2.70
93	0.072	0.150	Q	(N,K,D)	3.30	1.81
95	0.058	0.133	T(A)	T	2.41	2.73
100	0.054	0.125	F((L))	F	3.23	3.30
113	0.027	0.087	D(G)((N))	D	2.49	3.07
116	0.015	0.065	V(I)	V	2.41	2.83
125	7.62e-04	0.014	(G,S,D)	D	1.63	3.07
126	3.57e-03	0.038	(L,P,T)((H,S))	P((S))	1.68	3.08
143	0.066	0.144	E(R,T)((K))	(G,V)((R,T))	1.61	1.77
147	7.60e-04	0.014	G	D((N))	2.60	2.94
149	0.012	0.057	I(S)((A,T))	(S,A,V)((I,T))	1.73	1.45
150	0.032	0.099	H	R	4.01	2.83
155	4.04e-03	0.039	H	Y(H)	3.96	3.19
187	9.78e-03	0.055	K(R)	K	2.26	2.89
192	3.44e-04	8.74e-03	V(I)	V	2.48	2.83
199	0.033	0.099	(G,K)((N,R))	K(E)	1.58	2.26
206	0.053	0.125	I((V))	I	2.59	2.66
210	0.021	0.079	(K,M,I,R,V)((L))	I(V)((K,R))	1.37	1.84
212	9.32e-03	0.055	V(A)((T,I))	V	2.08	2.83
216	6.64e-03	0.048	(G,V,A)((S))	V(G,S)	1.45	1.92
220	0.052	0.125	K(Q)	K(R)	2.41	2.40
221	0.082	0.161	N	(N,E,K,D)	2.78	1.48
234	0.059	0.133	N(S)	N	2.39	2.79
238	0.050	0.125	T(A)	T((A))	2.40	2.64
257	0.064	0.141	I(V)	I	2.13	2.66
267	0.010	0.055	P	(T,K,L,Q)((S))	3.17	1.37

Location	P value	FDR cutoff	Residues		Sel. constraint (d)	
			Avian	Human	Avian	Human
275	0.015	0.065	V(I)	V((I))	2.30	2.77
283	8.85e-04	0.015	R(Q)	R	2.36	2.83
284	0.017	0.068	Y((H))	Y	3.29	3.50
286	2.03e-03	0.027	(I,E,N)((D))	G((D))	1.51	2.45
290	0.092	0.176	V	V((I))	2.79	2.52
296	0.075	0.152	K(R)	K	2.50	2.86
305	0.054	0.125	I((V))	I	2.59	2.66
308	0.027	0.087	(E,V,T,A)	K(T)	1.57	2.26
310	0.016	0.065	Y	H	3.48	4.01
311	0.027	0.087	S((N))	S	2.38	2.50
312	0.017	0.068	I((V))	I	2.55	2.66
313	0.025	0.087	D(G)((N,E))	V((A))	1.84	2.67
315	6.05e-03	0.047	G(S,R)((N))	S(R)	1.53	2.14
328	4.65e-05	2.36e-03	N	K((R))	2.79	2.70
331	2.49e-03	0.028	(I,R,G,S)	S(R)((N))	1.38	1.86
332	0.046	0.124	S	S(F)((Y))	2.48	2.07
336	0.025	0.087	N	H(Y,N)	2.78	2.72
338	4.47e-03	0.039	R(K)	L(Q,W)((K,R))	2.34	1.79
342	0.073	0.150	N(E)((D))	N	2.14	2.78
347	0.044	0.124	P	H(R)((Q))	3.21	3.26
356	0.047	0.124	(I,Y,V,D,S)((N))	(I,V,D)	1.21	1.75
360	0.079	0.159	(L,V,I)	V	1.65	2.81
367	0.018	0.069	(S,K,E)((N))	S((G,N))	1.42	2.34
368	0.064	0.141	(K,D,S,R)((E))	E(D)	1.43	2.41
369	5.46e-03	0.046	D	K(E)	3.07	2.18
370	0.027	0.087	S	L(S)((F))	2.45	1.75
378	2.43e-03	0.028	R(K)	K	2.29	2.89
380	0.071	0.150	I(T)((V))	I	1.97	2.66
381	1.46e-05	9.88e-04	G((D,N))	E(D)	2.36	2.42
384	4.38e-06	4.45e-04	(A,T,I)((V,N,S))	V(I)	1.18	2.06
385	0.021	0.079	(T,I,N)	K(N,R)	1.73	1.86
386	2.74e-06	4.45e-04	A((P))	P((S))	2.68	3.09
390	0.028	0.087	S	L(S)	2.50	2.24
393	0.050	0.125	N(S)	N	2.25	2.79
396	6.80e-03	0.048	V(I)	V	2.21	2.83
399	6.69e-03	0.048	D	E	3.06	2.74
400	0.036	0.106	(N,S,R,G)	R(S)((K))	1.35	2.00
401	8.91e-03	0.055	(S,N,D,E)	(D,G,K)((S,N))	1.41	1.65
403	0.016	0.065	W(S)	R(M)((S))	2.63	2.43
415	0.054	0.125	K((R))	K	2.82	2.89
431	0.042	0.122	P	K(N)((S,E,Q))	3.21	1.96
435	0.015	0.065	R	E((K))	2.83	2.48
437	0.027	0.087	W	L	3.95	2.63

Location	P value	FDR cutoff	Residues		Sel. constraint (d)	
			Avian	Human	Avian	Human
445	0.011	0.055	V((A))	V	2.65	2.81
466	8.91e-03	0.055	F	L	3.30	2.63
NP						
16	0.047	0.185	(S,G,D)	D((G))	1.61	2.65
31	0.014	0.112	R	R(K)	2.83	2.17
34	0.042	0.180	G(S,V)((D))	G(D)((S,N))	1.55	1.83
61	5.10e-03	0.055	I((M))	L((I,M))	2.52	2.23
67	0.037	0.176	V((I))	V	2.73	2.83
77	1.37e-04	8.30e-03	R(K)	K(R)	2.44	2.52
84	0.042	0.180	S(N)	S	2.18	2.50
100	0.031	0.167	R((I))	V((I))	2.71	2.56
101	1.88e-07	3.45e-05	(E,D,N)	G(N)((D))	1.90	1.96
102	3.60e-04	0.013	G	G((R))	2.60	2.44
103	0.017	0.131	K	K(R)	2.89	2.42
105	0.045	0.182	V(M)((I,T))	M(V,I,T)	2.03	1.98
131	2.71e-04	0.012	A	A(R)	2.90	2.51
136	6.30e-04	0.019	(L,M,I)	I(M)	1.91	2.32
214	0.029	0.167	R(K)	K((R))	2.43	2.69
217	0.041	0.180	I((T,V))	(S,I,G)((N))	2.40	1.49
236	9.90e-03	0.090	R((K))	K(R)	2.78	2.45
283	3.50e-03	0.049	L	P	2.63	3.21
285	0.027	0.164	V	V((I))	2.83	2.75
286	6.10e-03	0.062	A	S	2.90	2.50
290	0.025	0.161	D((N))	D((A,G,N))	2.99	2.35
305	1.40e-03	0.032	R(K)	K((R))	2.44	2.80
313	0.011	0.098	F(L,S)	Y((F))	2.00	3.21
329	0.013	0.110	V(I)	V	2.22	2.83
335	9.62e-04	0.025	S	S((F))	2.50	2.38
343	4.30e-03	0.053	V	L	2.83	2.63
353	3.10e-03	0.047	(V,I,L)((A))	(C,S,F,L)((I,V))	1.45	1.48
357	2.00e-03	0.036	Q	K((R))	3.29	2.65
373	0.032	0.169	A(T)	A(N)((S,T))	2.14	1.93
375	1.26e-04	8.30e-03	(V,D,E,G)((S,N))	G(V)((E))	1.30	1.97
400	0.038	0.176	R((K))	K(R)	2.61	2.27
411	3.80e-03	0.050	T	T(A)	2.73	2.27
421	0.041	0.180	E	E(D)	2.74	2.21
422	0.037	0.176	R	K((R))	2.81	2.57
423	0.027	0.164	A((S,T))	(P,S,T,A)	2.59	1.69
425	1.90e-03	0.036	I	I(V)	2.66	2.11
430	0.030	0.167	T(A,K)((I))	(N,A,I,S)((T))	1.70	1.26
442	0.048	0.187	T((A))	A((T))	2.48	2.66
447	5.10e-03	0.055	M	(M,I,L)	3.50	2.02
455	0.024	0.161	D((N))	E(D)	3.01	2.38

Location	P value	FDR cutoff	Residues		Sel. constraint (d)	
			Avian	Human	Avian	Human
459	0.034	0.173	Q	R	3.34	2.83
470	7.80e-03	0.075	K	R	2.89	2.83
472	2.50e-03	0.042	T	T(A)	2.73	2.31
480	0.025	0.161	D	D(E)	3.04	2.54
483	0.044	0.182	N(K)	N	2.25	2.79
496	0.030	0.167	Y(F)	Y	2.85	3.50
497	0.018	0.131	D(E)((A))	D	2.40	3.01
NS1						
4	6.85e-03	0.110	N	I(H)((N))	2.76	2.16
22	3.62e-03	0.079	(L,F,I,V)	V	1.51	2.78
41	3.09e-03	0.079	K((R))	K(R)	2.82	2.20
53	3.40e-03	0.079	D	D(N)	3.02	2.24
67	5.30e-03	0.103	(D,R,G)((Q,P,E,W,N))	(N,D,K)	1.34	1.85
70	0.013	0.131	K(E,R)((D))	K	1.70	2.89
81	5.08e-04	0.029	I(T)((V,M))	M((V))	1.93	3.39
82	0.011	0.116	A((T))	A(V,T)	2.79	1.92
84	4.84e-04	0.029	(V,M,G,S)((L,A,I,T))	T(A)((V))	1.17	1.97
104	9.40e-03	0.110	M	M((I))	3.50	3.32
105	9.44e-03	0.110	L	L((V))	2.63	2.48
112	1.49e-03	0.052	(T,I,A)((V))	(E,S,V,I)	1.52	1.33
114	0.015	0.142	P(S)((G))	P	2.08	3.16
115	7.06e-03	0.110	L	L((F))	2.63	2.51
127	7.87e-03	0.110	(T,N)((K,A,R,D,S))	N((K,S,D))	1.32	2.35
129	0.018	0.157	T(I)((M,V))	(M,I,T)	1.87	2.00
196	0.019	0.157	E	K(E)	2.70	2.13
202	8.66e-03	0.110	A(T)	A((S))	2.30	2.80
211	0.018	0.157	R(G)	G((R))	2.20	2.31
215	8.84e-06	1.55e-03	(P,S,L)((T,A))	T((P))	1.65	2.60
227	6.65e-04	0.029	E((G,K))	R(G)((E))	2.43	2.19
PA						
55	1.20e-03	0.082	D	N((D))	3.02	2.70
337	9.89e-04	0.082	(A,T,V)	S((T))	1.76	2.44
356	2.66e-04	0.035	K((R))	R((K))	2.70	2.68
397	3.00e-03	0.157	E	E((G))	2.74	2.60
552	1.94e-04	0.035	T	S	2.73	2.50
PB1						
52	4.10e-04	0.032	K((R))	R(K)	2.74	2.17
171	4.63e-03	0.168	M((V))	M(I)((L))	3.25	2.49
213	6.98e-03	0.172	N((S,K,T))	N(D)((K))	2.48	2.13
215	5.68e-03	0.168	R(K)	R((K))	2.17	2.70
298	1.46e-03	0.086	L((V))	I(L)	2.54	2.14
327	5.26e-03	0.168	R	K(R)	2.81	2.20
469	5.66e-03	0.168	T	T(I)	2.71	2.11

Location	P value	FDR cutoff	Residues		Sel. constraint (d)	
			Avian	Human	Avian	Human
517	3.63e-04	0.032	I((V))	V(I)	2.58	2.06
584	7.67e-07	1.81e-04	R((H))	Q(H)	2.73	2.93
667	7.27e-03	0.172	I((T,V))	(I,T,V)	2.30	1.67
PB2						
44	6.18e-04	0.023	A(S)	L(S)	2.53	2.15
81	3.10e-03	0.058	T(A)((I))	V(M)((I))	2.09	2.26
105	9.68e-05	0.013	T(A)((I,M))	V(M)((I))	2.01	2.42
111	8.90e-03	0.130	Y	H((Y))	3.50	3.81
143	7.60e-03	0.118	R	Q	2.83	3.34
199	2.78e-04	0.023	A	S	2.88	2.50
290	0.012	0.153	G	G((R))	2.60	2.48
395	4.00e-03	0.071	A	V	2.90	2.83
453	0.012	0.153	P((S,T))	H(P,S)((Q))	2.71	2.65
475	5.46e-04	0.023	L((M))	M	2.51	3.50
493	1.80e-03	0.039	R((K))	K((R))	2.53	2.70
522	0.015	0.177	Q	Q((H))	3.34	3.25
537	0.012	0.153	W	W((R))	3.95	3.79
569	2.40e-03	0.049	T((A))	A((S))	2.51	2.69
613	1.10e-03	0.035	V(A)((I))	T(I,A)	2.33	1.82
627	1.20e-03	0.035	E(K)	K	2.20	2.89
655	4.30e-03	0.071	V(A)	I(V)	2.27	2.24
661	5.91e-04	0.023	A(T)((V))	T((V))	2.28	2.51
682	1.50e-03	0.038	G	S(N)	2.60	1.94
684	7.63e-05	0.013	A((T))	S(T)	2.69	1.95
702	1.60e-03	0.038	K(R)	R	2.42	2.78
740	3.83e-04	0.023	D	D(N)	3.03	2.24

D. Examples of sites identified in chapter 3 in structural context

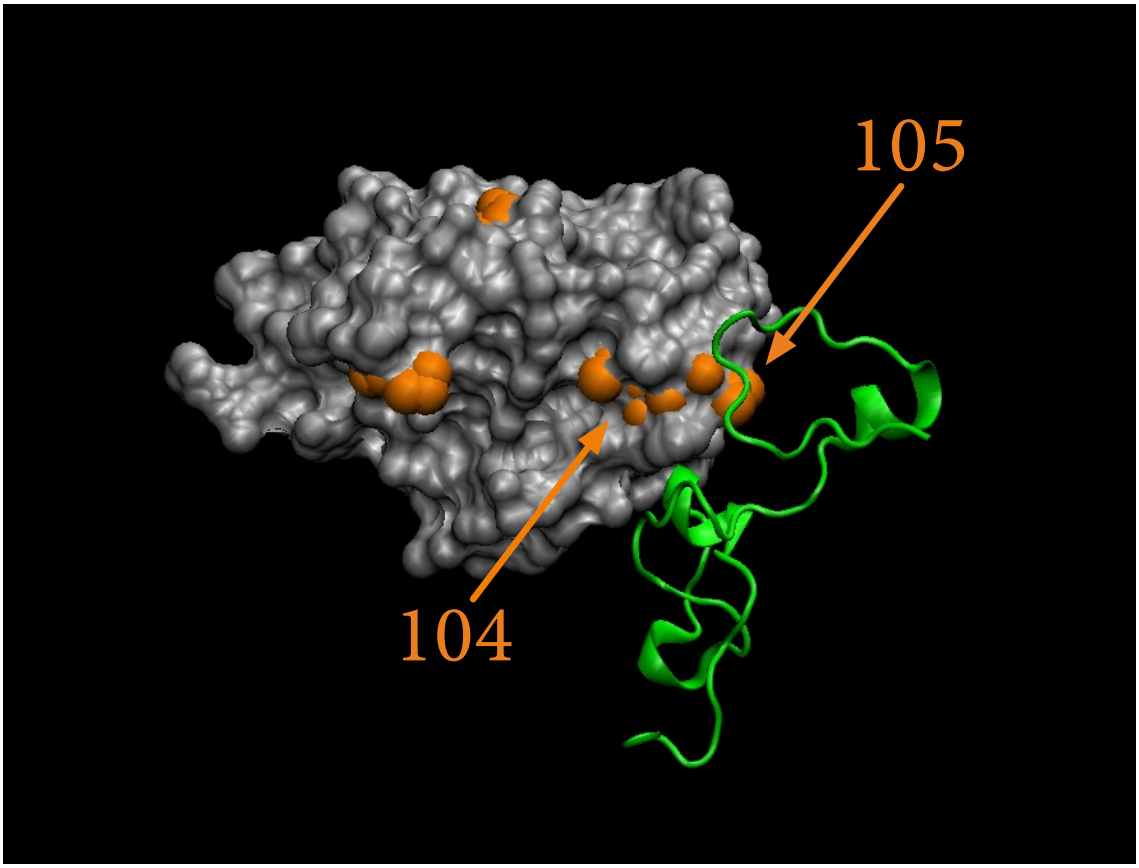


Figure D.1.: Structure of NS1 complex (residues 85-215) with human cellular factor CPSF30 (PDB: 2RHK). Sites identified as having changing selective constraints are highlighted in orange.

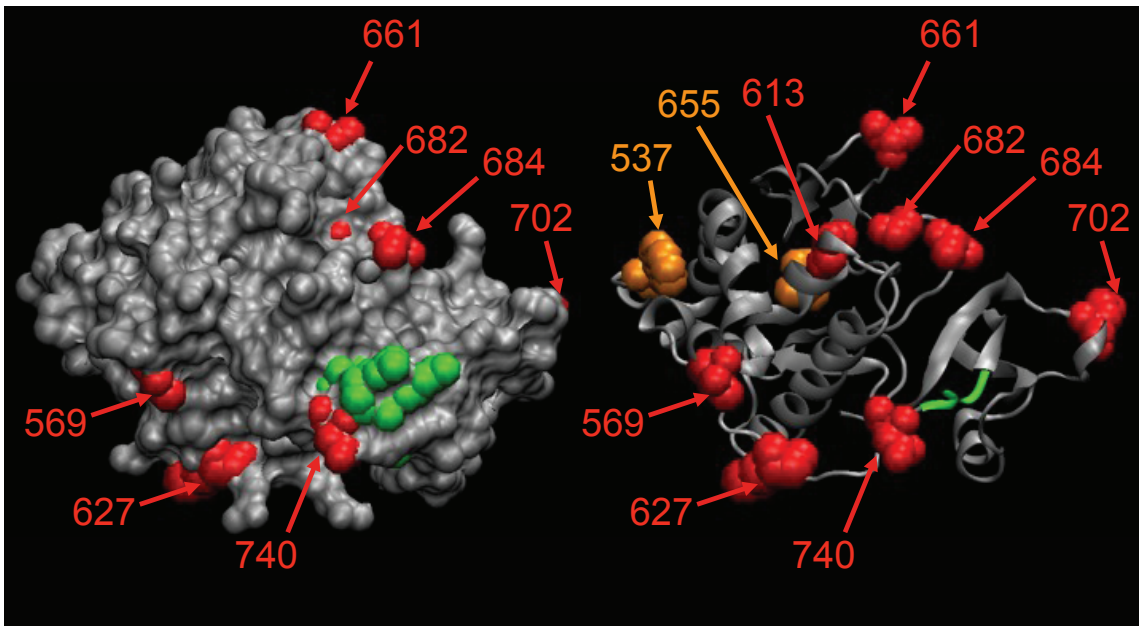


Figure D.2.: Structure of PB2 C-terminal domain (residues 535-742) (PDB: 3CW4). Selected sites identified as having changing selective constraints are highlighted in red (at FDR < 0.05) and orange (FDR < 0.20).

Site	Residue (Av → Hu)	Relative ASA
537	W → W((R))	0.05
569	T((A)) → A((S))	0.56
613	V(A)((I)) → T(I,A)	0.13
627	E(K) → K	0.50
655	V(A) → I(V)	0.00
661	A(T)((V)) → T((V))	0.73
682	G → S(N)	0.37
684	A((T)) → S(T)	0.69
702	K(R) → R	0.73
740	D → D(N)	0.82

Table D.1.: Sites identified on PB2 C-terminal domain (see figure above) listed with avian and human residues, and relative values of absolute surface area (ASA) calculated using the ASAView program (Ahmad *et al.*, 2004).

E. Accession numbers of sequences used in chapter 4

The following sequences were analysed in addition to those from chapter 3 (listed in appendix A).

H1

AAD25309 ACO24983 ACO25089 AAD25312 AAD25311 AAD25310 CAP49183 ACO25026 ABV60698 ACO25122 ACO25100 ACO25133 ACO25016 ACO25036 ACO24994 ACO25069 ACO25005 ACO25058 ACO25047 CAC86606 CAC86616 CAC86624 CAC86608 AAC57169 ACO25155 ACN67524 ABE12634 ABE27153 ACN72617 CAC86610 ACJ06667 ABV60697 ABD78104 CAP49192 AAD05218 AAD05215 AAD05216 AAD05217 ABS53372 ABS53353 ABS53363 AAD25303 ABD79255 ABV25635 ABW38010 ABW71481 ABV82573 AAD25302 ABV25636 ABV82595 ABV82584 ABR28724 ABR28702 ABV25637 ABR28603 AAB39851 ABR15852 ABR15863 ABS49921 ABU80287 ABQ45533 ABQ45458 ABD95712 ABU80210 ABR15874 ABR28691 BAH02160 AAC57168 AAC57166 BAG49742 BAH02180 BAF47397 BAG49619 ABQ45436 ABX58657 ABU80232 BAH02170 AAD25301 ABR28614 ABU80410 ABW86585 ABW71503 ABW86574 BAH02090 ABY40407 BAH02050 ABY40408 ABU80276 ABR28636 ABR28713 ABR28669 ABR28658 ABX58646 AAA72339 ABB86937 ABB86907 ABB86946 AAA19934 AAR90881 ABD85123 ABB86877 ABB86887 ABY81426 ABR29605 ABS50111 BAH02030 AAF87276 AAL29715 ABF71860 BAH02040 ABS50121 ABR29595 AAY56898 ACA25337 ABR29565 AAF87280 AAF87275 AAF87284 AAF87283 AAZ79392 ABQ42448 ACI48760 ABV25638 AAL87868 AAL87866 ABV25640 AAL29709 AAL87869 AAN46827 AAL29712 ABV25643 ACE77927 ACE77928 AAL29714 AAL87867 ABV25639 AAL87870 ABQ42444 ABQ42446 AAF75994 AAL29710 AAL87871 AAL87872 AAL87865 ACH69547 ABV25641 AAL29713 ABV25642 ABG34254 ACE77931 ACE77930 ACE77929 ACE77933

A/California/06/2009(H1N1)

N1

AAF77044 ABW71484 ABV82576 ACD85157 ABR28705 ABV82587 ABV25646 ABR28727 ABR28606 ABR15822 ABQ45417 ABD95715 ABU80290 ABW36325 AAF77043 ABR28540 BAG49744 ABY81429 ABS50114 ABD85121 ABR29608 AAD00584 ABB86936 ABB86876 ABB86906 ABB86947 ACE77987 ABA46958 ACA25342 ABI54390 ABB86957 ABR29588 ACH69548 ABB86886 ABA27442 ABV25650 ABA27434 ACD65205 CAC86317 ABS49946 CAC86315 CAC86314 CAC85490 ACO25113 ACO25060 AAF77046 ACO25071 ACO25038 ACO25049 CAC86316 CAC85492 BAH02042 BAH02032 BAH02072 BAH02062 ABY40396 CAP49184 ACO25157 ABD78107 CAC85488 ACJ06668 ACN67525 ABE12637 ABE27156 ACN72621

A/California/06/2009(H1N1)

A/Anhui/1/2005(H5N1)

A/Bangladesh/207095/2008(H5N1)

A/Hong Kong/156/97(H5N1)

A/Hong Kong/378.1/2001(H5N1)

NP

AAA43453 AAA43676 AAA52255 AAA52256 AAA52260 ABR28706 AAA43670 ABS49925 BAG49743 BAH02151 BAH02171 BAG49623 BAH02181 AAA43455 ABY81430 AAA73110 AAA74749 BAH02101 BAH02121 ACL11953 BAH02141 AAV36516 AAA51481 ABB86885 AAL26994 AAZ79397 ACA25341 AAF73886 AAF73880 AAF73884 AAL87893 ABB86875 ABB86945 ABB86935 ABB86895 ABB86925 AAL87890 ACE78019 AAL87892 AAF75997 AAG01771 AAG01762 AAG01789 ABA27433 ABY40429 AAN46830 ACE78009 ABA27441 ACH69553 ABG34249 ABI54394 ACO24984 ACO25145 ACO25101 AAA43456 ACO25027 ACO25112 AAA52271 CAC85241 ACO25006 ACO25070 ACO25017 ACO25059 BAH02091 BAH02071 BAH02031 BAH02051 ACO25123 ACO25156 CAA81461 ACF94710 CAP49194 AAK69308 ACN67528 CAP49180 CAN89845 ABO44039 ACJ06676 CAC85236 CAC85229 ABE12638 ABD62837 ABD78108

A/California/06/2009(H1N1)

A/Anhui/T2/2006(H5N1)

A/Bangladesh/207095/2008(H5N1)

A/Hong Kong/156/97(H5N1)

A/Hong Kong/378.1/2001(H5N1)

NS1

AAC36141 CAC86627 AAA43499 AAC36142 CAC40061 CAC40059 AAC36139 ACO25031 CAN89846 ACO25020 ACO24999 ACO25074 CAC87414 BAH02046 CAP49187 CAC86635 BAH02126 CAC86631 AAR12442 AAR12454 BAH02076 BAH02036 BAH02116 CAC86641 ACF94712 ABE12639 ABD62838 ABS53375 ABU62959 ABD62799 CAP49201 ABD78109 CAP49196 ABE27169 ABE27158 ACN72634 AAA43684 AAA43495 AAB50995 ABW71486 AAC35570 AAA43497 ABR28542 AAC36137 BAH02176 BAG49625 BAH02186 BAH02096 BAH02056 ABR28663 AAR90878 ABB86883 ABB86893 ABB86943 ABB86873 ABB86933 ACE78057 ABX89000 BAH02146 BAH02106 BAH02136 AAZ79396 AAB51007 ABL75553 ABY81725 AAL29804 AAL29803 AAL29801 ABR87891 ACE78061 AAL29789 AAL29798 AAL29805 ACE78049 ABY40427

A/California/06/2009(H1N1)

A/Anhui/T2/2006(H5N1)

A/Bangladesh/207095/2008(H5N1)

A/Hong Kong/156/97(H5N1)

A/Hong Kong/378.1/2001(H5N1)

PA

AAA43617 AAA43675 ABW71488 ABB86882 ABV82580 ABR28731 ABD95719 AAA43681 ABB86872 ABB86892 ABB86942 ABR15826 BAH02169 BAG49626 BAH02179 ABR28665 AAZ79399 ACA25345 ABB86951 ABB86956 ACO24982 CAC37005 ABS49950 ACO25025 BAH02049 BAH02119 ACL11951 BAH02139 BAH02089 BAH02099 ACF94708 ABS53371 ABE12641 ACJ06675 CAC85222 CAC37006 BAH02029 BAH02039 BAH02109 BAH02129 ACO25015 ACO24993 ACO25004 ACO25068 CAC84685 ABD78111 CAC85217 ABD62840 ABD62801 ABE27171 ACE78077 ACE78084 ABQ41897 ACE78078 AAL87913 AAL87919 ABI54398 ACE78069 AAL87917 AAL87914 AAG01765 AAL87916 ABR87895 ACI48768 AAN46832 ACE78082 AAF76000 AAL87918 ABA46959 ABA27431

A/California/06/2009(H1N1)

A/Anhui/T2/2006(H5N1)

A/Bangladesh/207095/2008(H5N1)

A/Hong Kong/156/97(H5N1)

A/Hong Kong/378.1/2001(H5N1)

PB2

AAG01766 AAL87934 AAF76002 AAL87932 ACE78141 ABR87897 ACI48761 ACE78138 ABQ41895 ABY81853 AAL87930 ACE78137 AAL87929 AAL87935 ABI54402 ACJ53898 AAN46834 ABB86900 ABA46960 ACE78126 ABY40436 ACD65210 ABA27437 AAA43126 ABW71491 ABV82583 ABR28712 ABR15829 BAH02147 BAG49628 BAH02177 ABU80220 AAA43125 ABB86890 ABY81436 ABB86870 ABB86880 ABB86930 ABB86940 ABR29574 AAZ79398 ACA25347 AAA43652 ACO25022 ACO25107 ABS49953 ACO25054 ACO24990 ACO25065 CAC37000 ABS53370 ACJ06674 ABD62842 ABD62803 ABE27174 ABO44045 ABD78114 ACF94705 ABE12644 ABS53350 BAH02087 BAH02067 BAH02117 ACL11949 BAH02137 BAH02047 BAH02037 BAH02097 BAH02107 BAH02027 BAH02127

A/California/06/2009(H1N1)

A/Anhui/T2/2006(H5N1)

A/Bangladesh/207095/2008(H5N1)

A/Hong Kong/156/97(H5N1)

A/Hong Kong/378.1/2001(H5N1)

F. Plots of host adaptedness using per-gene FDRs

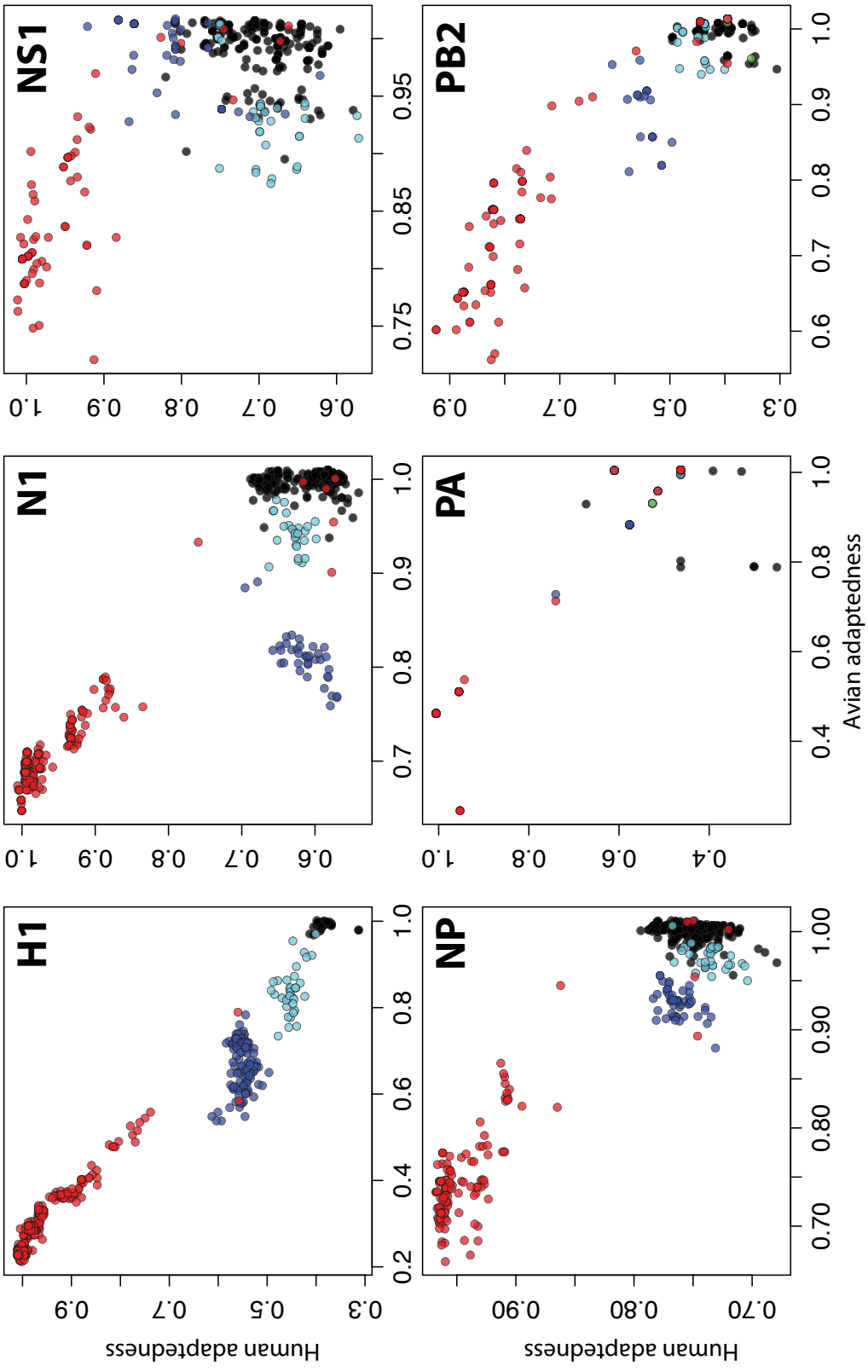


Figure F.1.: Host adaptedness values for a series of different virus sequences (black: avian, red: human, blue: classical swine, cyan: Eurasian swine, green: triple-reassortant). The multiple hypothesis correction in this figure was applied per-gene rather than genome-wide. The most striking difference is the reduced resolution of PA adaptedness when using 5 identified sites rather than 27 sites. Qualitative results are similar to the results presented in Chapter 4.

G. Plots of adaptedness of NP using sites selected with FDR < 0.2 or 0.05

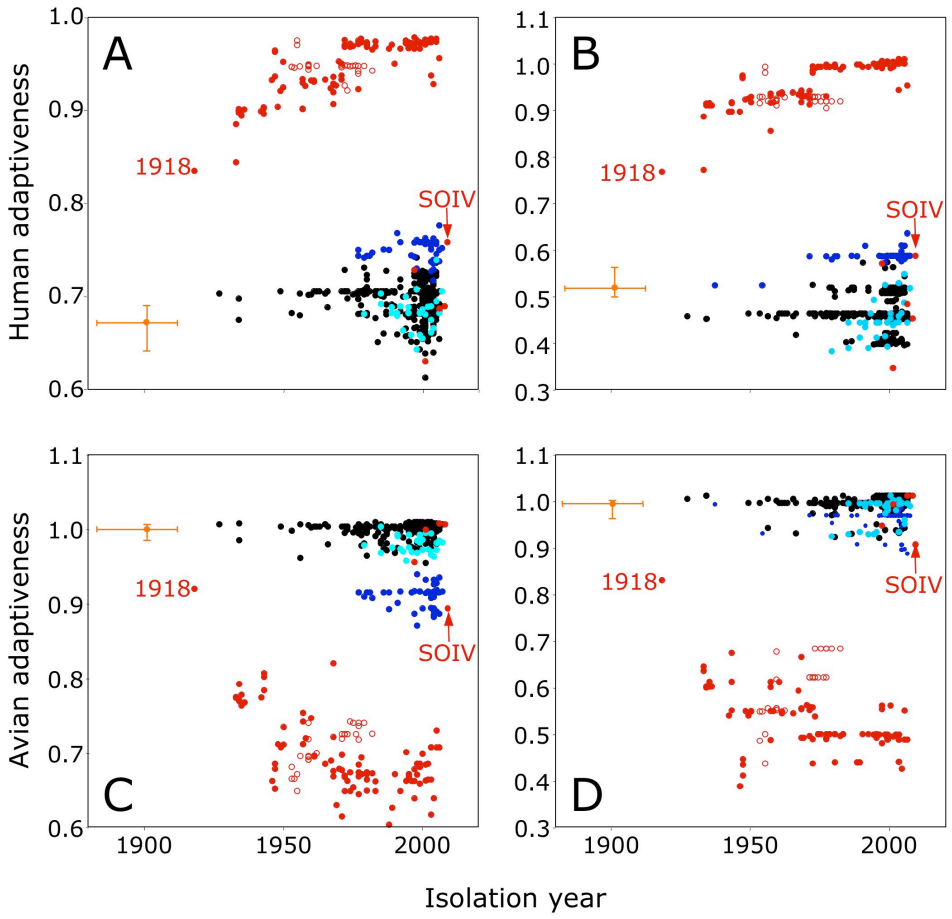


Figure G.1.: Plots of human (A,B) and avian (C,D) adaptedness values for NP as a function of isolation year, computed either with sites selected based on false discovery rate (FDR) < 0.20 (A, C) or FDR < 0.05 (B, D). Qualitative results are relatively insensitive to the choice of FDR.

H. Expected distributions of S

Figure H.1.: Distribution of selection coefficients showing mutations and substitutions expected under neutral, nearly neutral and adaptive models of molecular evolution. From Akashi (1999) and Yang (2006).

I. Distributions of S for simulations

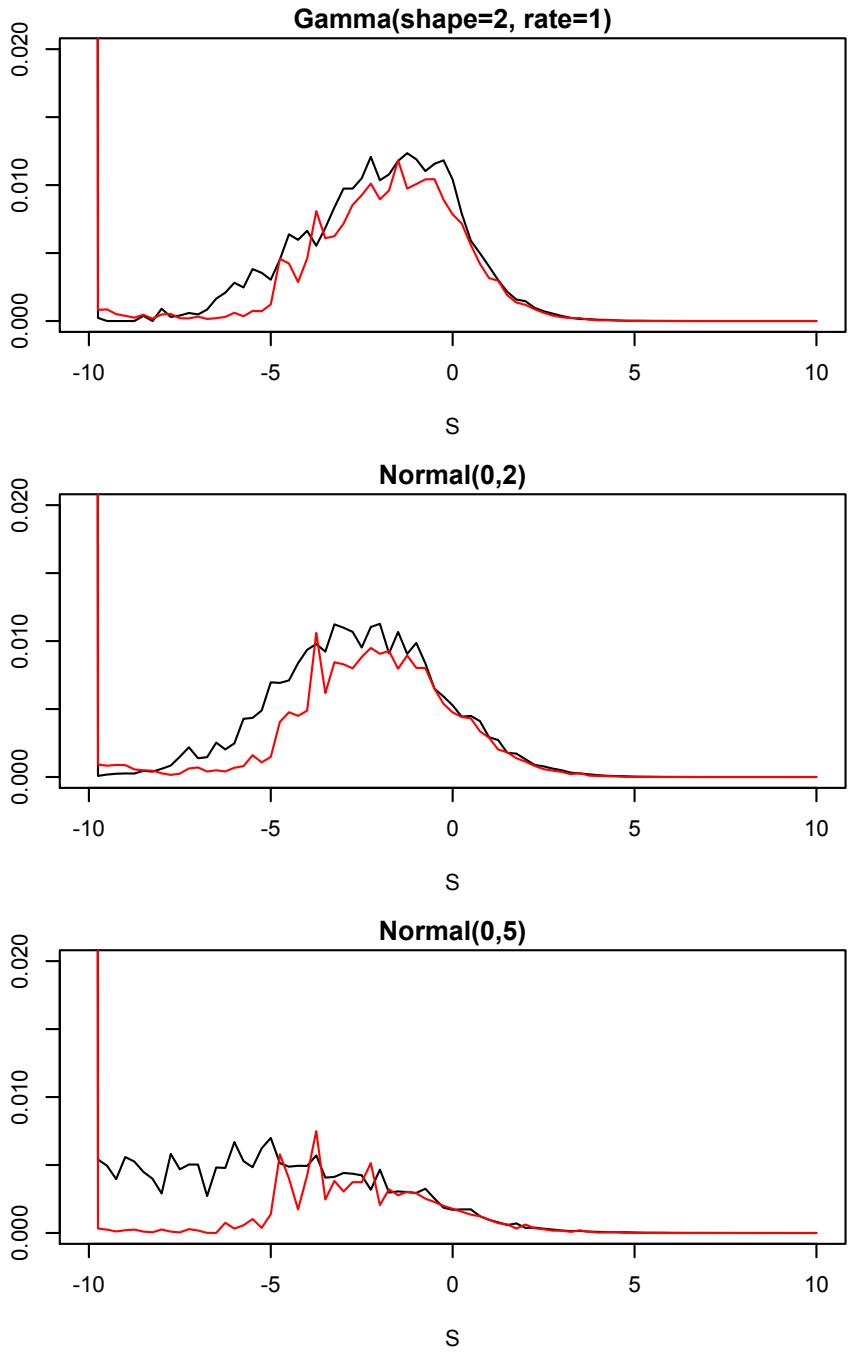


Figure I.1.: Distribution of S for simulated data with fitness drawn from gamma distribution, normal ($\sigma = 2$) and ($\sigma = 5$) (top to bottom) on 256 taxa tree. The known (true) distribution is shown in black, while the estimated distribution is shown in red.

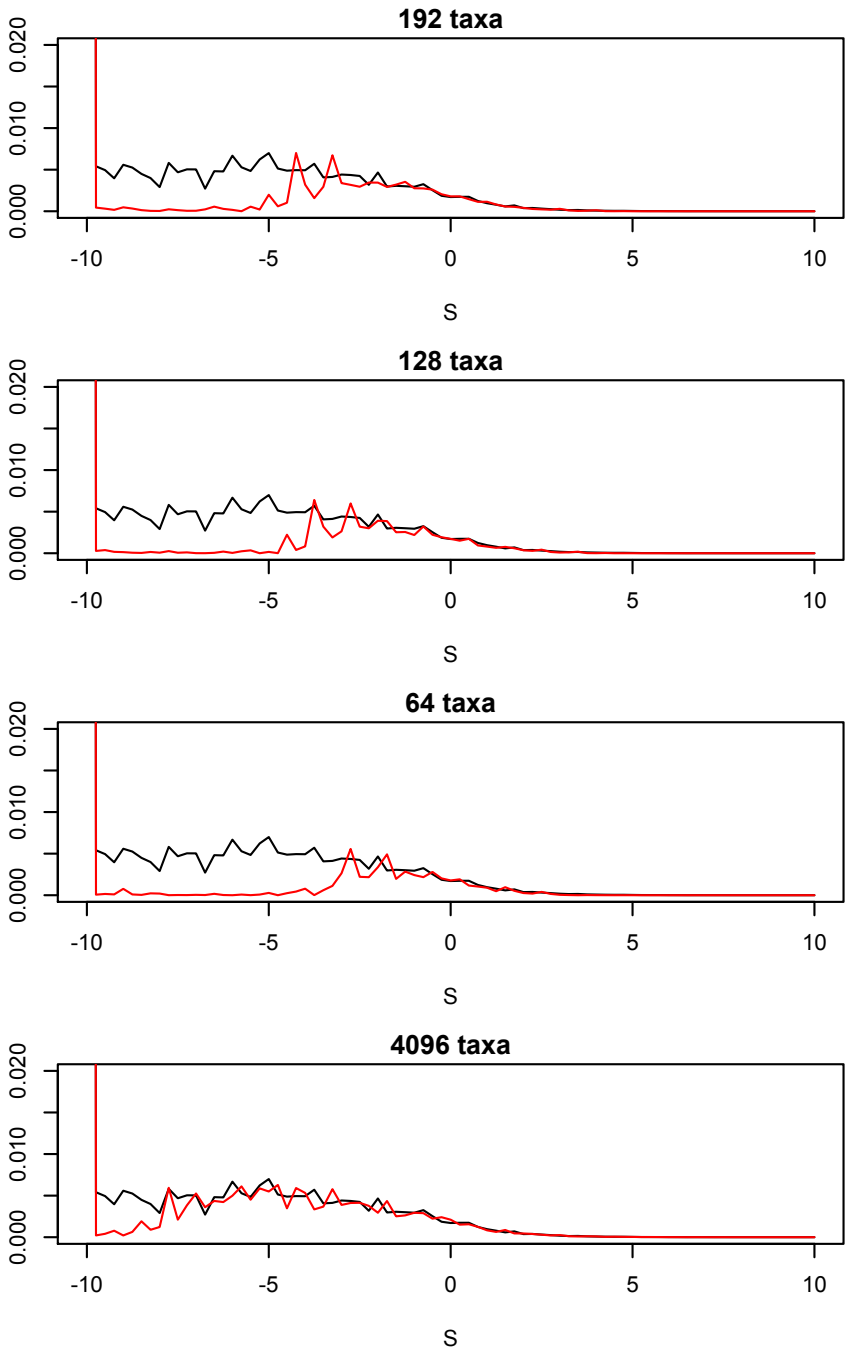


Figure I.2.: Distribution of S for simulated data with fitness drawn from normal ($\sigma = 5$) distribution. The known (true) distribution is shown in black, while the estimated distribution is shown in red. Top to bottom: 192, 128, 64 and 4096 taxa trees.

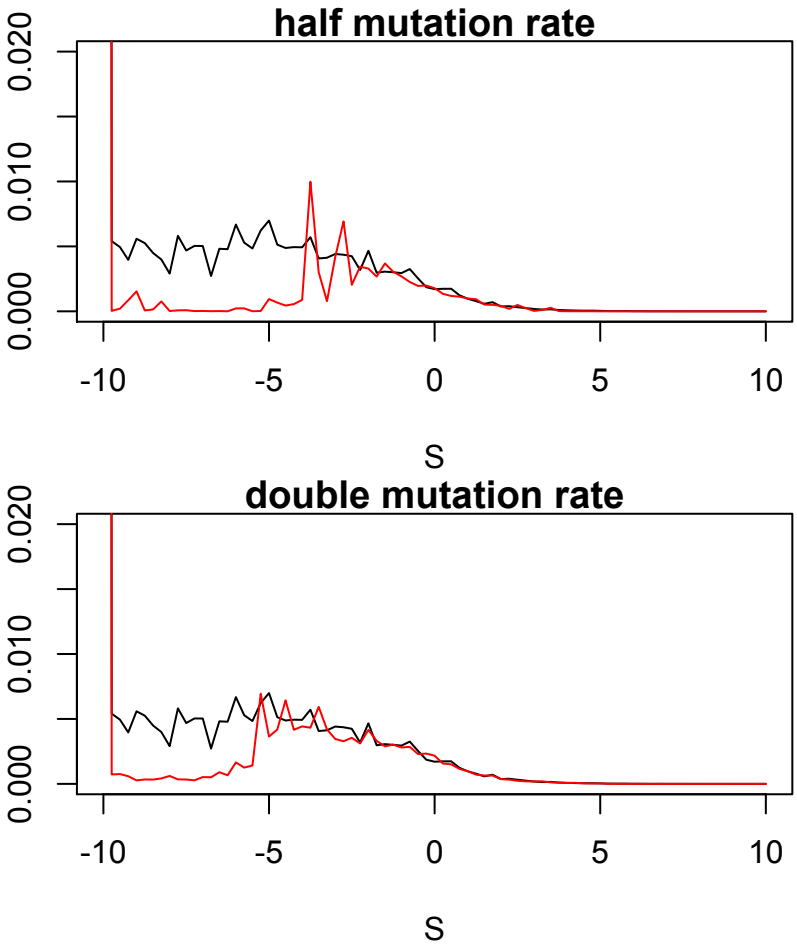


Figure I.3.: Distribution of S for simulated data with fitness drawn from normal ($\sigma = 5$) distribution. The known (true) distribution is shown in black, while the estimated distribution is shown in red. Top to bottom: half mutation rate and double mutation rate on 256 taxa tree.

J. Placental mammal species used in analysis in chapter 5

Acinonyx jubatus, Ailuropoda melanoleuca, Ailurus fulgens, Ammotragus lervia, Anomalurus sp., Antilope cervicapra, Arctocephalus forsteri, Arctocephalus pusillus, Arctocephalus townsendi, Arctodus simus, Artibeus jamaicensis, Balaena mysticetus, Balaenoptera acutorostrata, Balaenoptera bonaerensis, Balaenoptera borealis, Balaenoptera brydei, Balaenoptera edeni, Balaenoptera musculus, Balaenoptera omurai, Balaenoptera physalus, Berardius bairdii, Bison bison, Bison bonasus, Bos grunniens, Bos indicus, Bos javanicus, Bos primigenius, Bos taurus, Bradypus tridactylus, Bubalus bubalis, Budorcas taxicolor, Callorhinus ursinus, Camelus bactrianus, Camelus dromedarius, Canis latrans, Canis lupus, Caperea marginata, Capra hircus, Capricornis crispus, Cavia porcellus, Cebus albifrons, Ceratotherium simum, Cervus elaphus, Cervus nippon, Cervus unicolor, Chalinolobus tuberculatus, Chlorocebus aethiops, Chlorocebus pygerythrus, Chlorocebus sabaeus, Chlorocebus tantalus, Choloepus didactylus, Chrysochloris asiatica, Coelodonta antiquitatis, Colobus guereza, Cricetulus griseus, Crocidura russula, Cuon alpinus, Cynocephalus variegatus, Cystophora cristata, Dasypus novemcinctus, Daubentonia madagascariensis, Delphinus capensis, Dendrohyrax dorsalis, Dicerorhinus sumatrensis, Diceros bicornis, Dugong dugon, Echinops telfairi, Echinosorex gymnura, Elaphodus cephalophus, Elephantulus sp., Elephas maximus, Enhydra lutris, Eothenomys chinensis, Episoriculus fumidus, Equus asinus, Equus caballus, Eremitalpa granti, Erignathus barbatus, Erinaceus europaeus, Eschrichtius robustus, Eubalaena australis, Eubalaena japonica, Eulemur fulvus, Eulemur macaco, Eulemur mongoz, Eumetopias jubatus, Felis catus, Galago senegalensis, Galemys pyrenaicus, Giraffa camelopardalis, Gorilla gorilla, Grampus griseus, Gulo gulo, Halichoerus grypus, Helarctos malayanus, Hemiechinus auritus, Herpestes javanicus, Hippopotamus amphibius, Homo sapiens, Hydropotes inermis, Hydrurga leptonyx, Hylobates agilis, Hylobates lar, Hylobates pileatus, Hylomys suillus, Hyperoodon ampullatus, Inia geoffrensis, Jaculus jaculus, Kogia breviceps, Lagenorhynchus albirostris, Lama glama, Lama guanicoe, Lama pacos, Lemur catta, Leptonychotes weddellii, Lepus europaeus, Lipotes vexillifer, Lobodon carcinophaga, Loris tardigradus, Loxodonta africana, Lutra lutra, Macaca fascicularis, Macaca mulatta, Macaca sylvanus, Macaca thibetana, Macroscelides proboscideus, Manis tetradactyla, Martes flavigula, Martes melampus, Martes zibellina, Megaptera novaeangliae, Meles meles, Melursus ursinus, Mesocricetus auratus, Microtus kikuchii, Microtus rossiaemeridionalis, Mirounga leonina, Mogera wogura, Monachus schauinslandi, Monodon monoceros,

Moschus berezovskii, Moschus moschiferus, Muntiacus crinifrons, Muntiacus muntjak, Muntiacus reevesi, Mus musculus, Mus terricolor, Myoxus glis, Mystacina tuberculata, Naemorhedus caudatus, Naemorhedus swinhoei, Nannospalax ehrenbergi, Nasalis larvatus, Neofelis nebulosa, Neophoca cinerea, Nomascus siki, Nyctereutes procyonoides, Nycticebus coucang, Ochotona collaris, Ochotona curzoniae, Ochotona princeps, Odobenus rosmarus, Orycteropus afer, Oryctolagus cuniculus, Otolemur crassicaudatus, Ovis aries, Pan paniscus, Pan troglodytes, Panthera pardus, Panthera tigris, Pantholops hodgsonii, Papio hamadryas, Pecari tajacu, Perodicticus potto, Phacochoerus africanus, Phoca caspica, Phoca fasciata, Phoca groenlandica, Phoca hispida, Phoca largha, Phoca sibirica, Phoca vitulina, Phocarctos hookeri, Phocoena phocoena, Physeter catodon, Pipistrellus abramus, Platanista minor, Pongo abelii, Pongo pygmaeus, Pontoporia blainvillei, Presbytis melalophos, Procavia capensis, Procolobus badius, Procyon lotor, Proedromys sp., Propithecus coquereli, Pteropus dasymallus, Pteropus scapulatus, Pygathrix nemaeus, Pygathrix roxellana, Rangifer tarandus, Rattus exulans, Rattus norvegicus, Rattus praetor, Rattus rattus, Rattus tanezumi, Rhinoceros sondaicus, Rhinoceros unicornis, Rhinolophus formosae, Rhinolophus monoceros, Rhinolophus pumilus, Rousettus aegyptiacus, Saimiri sciureus, Sciurus vulgaris, Semnopithecus entellus, Sorex unguiculatus, Sousa chinensis, Spilogale putorius, Stenella attenuata, Stenella coeruleoalba, Sus scrofa, Symphalangus syndactylus, Talpa europaea, Tamandua tetradactyla, Tarsius bancanus, Tarsius syrichta, Thryonomys swinderianus, Trachypithecus obscurus, Tremarctos ornatus, Trichechus manatus, Tscherskia triton, Tupaia belangeri, Tursiops aduncus, Tursiops truncatus, Uncia uncia, Urotrichus talpoides, Ursus americanus, Ursus arctos, Ursus maritimus, Ursus thibetanus, Varecia variegata, Vicugna vicugna, Vulpes vulpes, Zalophus californianus

K. Software tutorial for estimating the distribution of selection coefficients

K.1. Introduction

This document describes how to use the TdG12 site-wise mutation-selection model ('swMutSel0') to estimate the distribution of selection coefficients (or 'fitness effects') from an alignment of protein coding sequences.

Requirements

The steps in this tutorial have been tested on Linux and Mac OS X 10.6+. Our program does not estimate tree topology or perform branch length optimisation, therefore we recommend the use of two popular tools for this purpose: RAxML and PAML[1]. Both have Windows versions available for download, but they have not been tested by us. We assume that you already have them installed and are able to run them on your computer. Our software is written in Java and should run on all platforms that have the Java Runtime Environment (JRE) (6 or later) installed. The JRE is available for Windows, Linux and Mac OS X 10.6+.

Installation

You can download the executable files for the program from http://mathbio.nimr.mrc.ac.uk/. Download the tdg12.zip file and extract the contents. The download contains the files required for this tutorial and `tdg12.jar`, which is the Java binary file for the program.

[1] However, you can use other programs if you prefer.

K.2. An example analysis

Preparing the alignment and tree

For the purposes of this tutorial, we will be estimating the distribution of selection coefficients of a set of mammalian mitochondrial ATP8 protein coding genes. The download provides an alignment, atp8.phyl (which was built using PRANK) and a tree, atp8.tree (estimated by RAxML with branch lengths optimised using PAML's codeml). Full details are available from the program's websites. The options used for running RAxML were:

```
raxmlHPC -f a -x 12345 -p 12345 -N 100 -m GTRGAMMA -s
atp8.phy -n atp8
```

and the PAML codeml control file (codeml.ctl) is available in the download. We optimised the branch lengths using the FMutSel0 model in codeml.

Getting estimates of global parameters from codeml

There are a number of global, site-invariant, parameters that are required for the TdG12 model. They are:

1. τ (tau) - rate of multiple substitutions.
2. κ (kappa) - transition/transversion bias.
3. π (pi) - base nucleotide composition.
4. μ (mu) - branch scaling.

Each of these parameters can be estimated under the TdG12 model but this requires significant computational resources and the use of a distributed version of the software (not covered here). However, you can get obtain good estimates for branch lengths, κ, π and

μ from the faster PAML codeml analysis. The results file (named 'mlc') contains the new tree (with re-estimated branch lengths) and estimates of κ and π. An estimate of μ can be calculated from $\mu = 3 \times T_{dS}/T$, where T_{dS} is the tree length in dS (synonymous changes) units and T is the total tree length, also found in the codeml 'mlc' results file. *Remember to use the tree from the 'mlc' file for the TdG12 program.*

Therefore, the estimates for the global parameters for the ATP8 gene alignment are:

$\kappa = 4.93022$

$\pi = \{0.25299, 0.22268, 0.43860, 0.08572\}$

$\mu = 3 \times T_{dS}/T = 3 \times 70.7207/113.24814 = 1.8734.$

We choose a small number for τ, e.g. 1.0×10^{-2}.

Analysing the data using the site-wise mutation-selection model

You can now run the TdG12 program to estimate the site-wise fitness of each amino acid in our alignment. The available options for running the program are:

-t The tree file in NEWICK format (required). Remember to use the tree supplied by PAML's codeml.

-s The protein coding alignment file in PHYLIP sequential format (required).

-gc The genetic code, 'standard' or 'vertebrate_mit' (required).

-tau Rate of multiple substitutions (required).

-kappa Transition/transversion bias (required).

-pi Comma-separated (with no spaces) base nucleotide frequencies (T,C,A,G) (required).

-mu Branch scaling factor (required).

-useapprox Use the approximation for unobserved residues which allows for faster computation (optional).

-site Location of the site to analyse (optional, default = all sites in the alignment).

-optimruns The number of restarts for the optimisation routine (optional, default = 1).

-threads The number of threads to use for processing in a multicore environment (optional, default = 1).

You start the analysis of the set of ATP8 genes by running:

```
java -cp tdg12.jar tdg.Analyse -s atp8.phy -t atp8.tree -gc
vertebrate_mit -tau 1e-2 -kappa 4.93 -pi
0.2530,0.2227,0.4386,0.0857 -mu 1.8734 > tdg.out
```

You must add '`> tdg.out`' to redirect the analysis output to a file named tdg.out.

Using multicore/multiple CPUs

If you are running the program on a computer with multicore or multiple CPUs, you can specify the -**threads** option. Usually, this would be the number of available cores − 1. For example, to utilise 3 cores you would run:

```
java -cp tdg12.jar tdg.Analyse -s atp8.phy -t atp8.tree -gc
vertebrate_mit -tau 1e-2 -kappa 7.8 -pi 0.25,0.25,0.25,0.25
-mu 2.3 -threads 3 > tdg.out
```

Parsing the results and calculating the distribution

Once the program completes, the results saved in tdg.out need to be processed to calculate the distribution of selection coefficients. The output looks something like (truncated):

```
Site 1 - Residues: [1/20] { 1:(12, M), 2:(0, A), 3:(1, R), ... }
Site 1 - Optimisation run (267 evaluations). lnL = -4.939901E-5, Params = {0.0, ...}
Site 1 - Homogeneous model lnL: -4.939901011180276E-5
Site 1 - Fitness: { -13.06494, -18.20550, -16.32137, ... }
Site 1 - Pi: { 4.79931E-6, 5.79345E-8, 2.70189E-7, ... }
...
```

The output is quite verbose. Run the following command (specifying the name of the result file with '-o tdg.out'):

```
java -cp tdg12.jar tdg.results.All -o tdg.out -gc
vertebrate_mit -tau 1e-6 -kappa 7.8 -pi 0.25,0.25,0.25,0.25
-mu 2.3
```

The global parameters should be the same as were specified when you run the analysis. This command reads the results file and the global parameters from the command-line, and writes the files:

F.txt Fitness values for each amino acid at each site

Q0.txt Neutral mutation matrix for entire alignment

S.txt Selection coefficients matrix for each site

QS.txt Mutation with selection matrix for each site

PiS.txt Codon frequencies at each site

PiAA.txt Amino acid frequencies at each site

distribution.mutations.csv the distribution of selection coefficients for mutations

distribution.substitutions.csv the distribution for substitutions.

The last two are comma-separated value files that can be opened in a program like Excel to plot the distribution. The three columns in the distribution files are (i) the histogram bins for S (ii) all mutations/substitutions and (iii) nonsynonymous mutations/substitutions.

K.3. Simulating data using the mutation-selection model

Simulating a single set of fitnesses (for one site)

```
java -cp tdg12.jar tdg.results.All -tree sim.tree -output
out.phy -sites 100 -fitness 0.0,0.2,0.3,1.0,0.5,0.6,0.7,0.8
-characters A,R,N,D,C,Q,E,H -tau 0 -pi 0.25,0.25,0.25,0.25
-mu 1.0 -gc standard
```

This command will write an alignment file 'out.phy' (in PHYLIP format) simulating a site (100 times) on the tree 'sim.tree' with the specified fitnesses for the specified characters. Characters that do not appear in the -characters option are not observed at that site (i.e. have fitness of $-\infty$). Set the global parameters as required.

Simulating multiple sets of fitnesses (for multiple sites)

Create a file with 20 fitnesses values on each line, each line corresponding to a single location in the alignment. Each fitness must be separated by a space and each site should be on a new line. For example, to simulate an alignment with 5 sites:

```
-21 -21 -21 -21 -21 -21 -21 -21 -21 2.19 3.07 -21 2.39 -21 -21 -21 -21 -21 -21 -21
-21 -21 1.74 -21 -21 -21 -21 -21 -21 -21 -21 -21 -21 -21 4.23 -21 -21 -21 -21
```

```
0.93 -21 -21 -21 -21 -21 -21 -21 -21 0.31 -21 -21 2.63 -21 -21 -21 4.11 -21 -21 6.39
2.75 -21 -21 -21 -21 -21 -21 -21 -21 0.92 -21 -21 1.71 -21 2.28 0.72 1.33 -21 -21 3.51
-21 -21 -21 -21 -21 -21 -21 -21 -21 -21 -21 -21 -21 -21 -21 0.79 -21 -21 -21 -21
```

Each line specifies the fitnesses of each amino acid in the canonical IUPAC order. Residues that are unobserved at a site are given a fitness < -20. If this file is saved as 'F.txt', the alignment is generated using the command:

```
java -cp tdg12.jar tdg.results.All -tree sim.tree -output
out.phy -fitnessfile F.txt -tau 0 -pi 0.25,0.25,0.25,0.25
-mu 1.0 -gc standard
```

K.4. Colophon

TdG12 uses the following libraries:

1. Colt Project (http://acs.lbl.gov/software/colt/). For linear algebra.
2. PAL: Phylogenetic Analysis Library (http://www.cebl.auckland.ac.nz/pal-project/). Reading/traversing/writing NEWICK trees and PHYLIP alignments.
3. Apache Commons Math (http://commons.apache.org/math/). For optimisation.
4. Guava: Google Core Libraries (http://code.google.com/p/guava-libraries/).
5. JCommander (http://jcommander.org/). Parsing and managing command-line options.
6. Simple Java HTTP server (http://www.simpleframework.org/).
7. Asynchronous Http Client library for Java (https://github.com/sonatype/async-http-client). Used by tdg.distributed.Master to make asynchronous HTTP requests to the slaves.