

Long-range chained epistasis in influenza viruses may not be
physically- but functionally-mediated

Jean Claude Nshogozabahizi

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of Master of Science in
Biology with specialization in Bioinformatics ¹

Department of Biology
Faculty of Science
University of Ottawa

© Jean Claude Nshogozabahizi, Ottawa, Canada, 2015

¹The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Biology with specialization in Bioinformatics

Abstract

In systems biology and genomics, epistasis characterizes the impact that a substitution at a particular location in a genome can have on a substitution at another location. This phenomenon is often implicated in the evolution of drug resistance or to explain why particular ‘disease-causing’ mutations do not have the same outcome in all individuals. Hence, uncovering these mutations and their locations in a genome is a central question in biology. However, epistasis is notoriously difficult to uncover, especially in fast-evolving organisms. Here, we present a novel statistical approach that takes inspiration from a model developed in ecology and that we adapt to analyze genetic data in a typically fast-evolving system: the influenza A virus. We validate the approach using experimentally-validated data: known interactions are recovered. We further evaluate the ability of our approach to detect epistasis during antigenic shifts or at the emergence of drug resistance. We show that in all cases, epistasis is prevalent in influenza A viruses, involving many pairs of sites linked together in chains, a hallmark of historical contingency. Strikingly, interacting sites are separated by large physical distances, which entail either long-range structural effects or functional tradeoffs, for which we find support with the emergence of drug resistance. Our work paves a new way for the unbiased detection of epistasis in a wide range of organisms by performing whole-genome scans.

Résumé

En biologie des systèmes aussi bien qu'en génomique, l'épistasie caractérise l'impact qu'une substitution à un endroit particulier dans un génome peut avoir sur une substitution à un autre endroit. Ce phénomène est souvent impliqué dans l'évolution de la résistance aux médicaments ou afin d'expliquer pourquoi des mutations impliquées dans certaines maladies génétiques n'ont pas les mêmes effets chez tous les individus. Par conséquent, découvrir ces mutations et leurs emplacements dans un génome est donc une question centrale en biologie. Toutefois, l'épistasie est notoirement difficile à découvrir, en particulier dans les organismes en évolution rapide: le virus de la grippe A. Ici, nous présentons une nouvelle approche statistique qui prend l'inspiration d'un modèle développé en écologie et que nous l'adaptions pour analyser les données génétiques dans le virus de la grippe A. Nous validons l'approche à l'aide de données validées expérimentalement: les interactions connues sont récupérées. De plus, nous évaluons la capacité de notre approche pour détecter l'épistasie durant les changements antigéniques ainsi que lors de l'émergence de la résistance aux médicaments. Nous montrons que, dans tous les cas, l'épistasie est très fréquente dans le virus de la grippe, impliquant plusieurs paires de sites reliés les uns aux autres, formant des chaînes, une caractéristique de la contingence historique. Étonnamment, les sites interactifs sont séparés par des grandes distances physiques, qui entraînent des effets structurels à longue intervalle ou compromis fonctionnels. Notre travail ouvre une nouvelle façon pour la détection non-biaisée d'épistasie du génome entier.

Acknowledgements

I would like to thank my supervisor, Dr Stéphane Aris-Brosou, for the patient guidance, encouragement, and advice that he has provided throughout my time as his student. I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly. I would also like to thank all the members of my committee, Dr Alex Wong and Dr Nicolas Rodrigue, for all their time and advice over the course of this project. I would also like to express my appreciation to Dr Xuhua Xia for all his time and effort.

I am thankful to have been a part of Dr Stéphane Aris-Brosou's laboratory and would like to acknowledge the laboratory members, Jonathan Dench and Neke Ibeh, for their support and suggestions during my studies.

Of course, I thank my family for their love and encouragement throughout all of my endeavors. Great thanks goes to my wife Claudine Niyokwizera, who experienced all of the ups and downs of my research. I deeply appreciate all of her continued support and encouragement.

Finally, I would like to thank the Natural Sciences Research Council of Canada and University of Ottawa for providing the funding which allowed me to undertake this research.

Contents

List of Figures	vii
List of Tables	x
1 Introduction	1
1.1 Influenza A virus	2
1.2 Correlated evolution	9
1.3 Epistasis and historical contingency in influenza A evolution . . .	13
1.4 Evolution models to detect coevolved sites	15
1.5 Objectives	18
2 Materials and Methods	20
2.1 General approach to detect epistasis	20
(a) Repurposing of BayesTraits	20
(b) Data recoding	23
(c) Code optimization	23
2.2 Validation based on previous evidence	25
(a) Previous computational analyses	26
(b) Experimental evidence	26
(c) Nature of epistasis in influenza evolution	27
(d) Phylogenetic analyses	27

3	Results and Discussion	29
3.1	Little overlap with previous computational results	29
3.2	Our approach detects experimentally-validated pairs, and more .	35
3.3	Chained long-range interactions occur sequentially	42
3.4	Long-range interactions are likely to be functionally-mediated . .	45
4	Conclusions and Future Directions	47
4.1	Conclusions	47
4.2	Future directions	50
	Bibliography	69
5	Appendices	70
5.1	Supplementary Figures	70
5.2	Supplementary Tables	79

List of Figures

1.1	The influenza A virus structure.	4
2.1	Transitions among the four combinations of states resulting from two binary variables.	21
2.2	The amino acids properties and their categories (states) used to recode protein alignments.	22
2.3	Speed-up as a function of number of cores requested.	24
3.1	Comparison with the method by Kryazhimskiy <i>et al.</i> (2011): after FDR.	30
3.2	Comparison with the method by Kryazhimskiy <i>et al.</i> (2011): before FDR.	31
3.3	Epistatic pairs of amino acids detected in the KDBP11 H3 data set of H3N2 viruses.	32
3.4	The epistatic sites detected in human influenza hemagglutinin and the physical distance among them.	33
3.5	Epistatic pair of amino acids detected in the KDBP11 NA data set of H1N1 viruses.	34
3.6	Epistatic pair of amino acids detected in the KDBP11 NA data set of H3N2 viruses.	36

3.7 Epistatic pairs of amino acids detected in the Gong13NP data set with the ancestor / derived recoding.	37
3.8 Strength of evidence in the face of phylogenetic uncertainty.	39
3.9 The epistatic sites detected in human influenza nucleoprotein and the physical distance among them.	40
3.10 Epistatic pairs of amino acids detected in the Duan14NA data set. .	41
3.11 Epistatic pairs of amino acids detected in the Koel13HA data set. .	43
3.12 The epistatic sites detected in human influenza hemagglutinin and the physical distance among them.	44
3.13 Epistatic pairs of amino acids detected in the Adam03M2 data set. .	46
4.1 Hypothetical model to detect epistatic amino acids within and among RNA segments within the influenza A virus	53
5.1 Node support values for the phylogenetic analysis of the Gong13NP data set.	71
5.2 Node support values for the phylogenetic analysis of the Gong13NP data set.	72
5.3 Node support values for the phylogenetic analysis of the Gong13NP data set.	73
5.4 Node support values for the phylogenetic analysis of the KDBP11 H1 data set.	74
5.5 Node support values for the phylogenetic analysis of the KDBP11 N1 data set.	75
5.6 Node support values for the phylogenetic analysis of the KDBP11 H3 data set.	76
5.7 Node support values for the phylogenetic analysis of the KDBP11 N2 data set.	77

5.8 Node support values for the phylogenetic analysis of the Gong13NP data set.	78
--	----

List of Tables

5.1 Detailed results generated using the KDBP11 data set for HA in H1N1 viruses. Site-1 and site-2 show the AA positions along the alignment.	79
5.2 Detailed results generated using the KDBP11 data set for NA in H1N1 viruses. Site-1 and site-2 show the AA positions along the alignment.	80
5.3 Detailed results generated using the KDBP11 data set for HA in H3N2 viruses. Site-1 and site-2 show the AA positions along the alignment.	82
5.4 Detailed results generated using the KDBP11 data set for NA in H3N2 viruses. Site-1 and site-2 show the AA positions along the alignment.	86
5.5 Epistatic pairs of sites of amino acid detected before FDR in the Gong13NP data set of NP proteins in H3N2 human influenza A virus.	92
5.6 Detailed results generated using the Gong13NP data set. Site-1 and site-2 show the AA positions along the alignment	93
5.7 Detailed results generated using the Duan14NA data set. Site-1 and site-2 show the AA positions along the alignment.	94
5.8 Detailed results generated using the Koel13HA data set. Site-1 and site-2 show the AA positions along the alignment.	95

5.9 Detailed results generated using the Adam03M2 data set. Site-1 and site-2 show the AA positions along the alignment.	104
---	-----

Chapter 1

Introduction

Influenza A virus is one of the most important respiratory infections of humans with a rapid continuing evolution that is paramount for annual epidemics and occasional pandemics disease, responsible for the death of thousands [1]. For example, the 1918 pandemic killed 50 million people [2] and recent pandemic influenza H1N1 2009 caused the deaths of more than 14,000 people [3]. Although vaccination is the most effective way to prevent the disease, the viral proteins evolve quickly to escape the human immune system and/or antiviral drugs [1]. However, little is known about the dynamics of amino acid (AA) replacement in viral proteins. The identification of AAs that are responsible for antigenic drift, mechanism for virus variation by accumulation of mutations within antigenic or other viral proteins so that they can not be recognized by host cell antibodies, is important to understand the influenza A virus evolution, hence a way to improve the effectiveness of the influenza vaccines [4]. It has been proven experimentally that influenza A virus evolved through correlated evolution called also epistasis where the fitness effect of one mutation at a particular position is modulated by its interactions with one or many mutations at other positions in viral genome [5, 6]. However, it takes a long time and is very tedious to uncover experimentally the epistatic sites, hence a statical method is needed to detect the

interacting sites, especially in fast-evolving organism such as viruses. In this first chapter, we will review statical methods that have been developed to identify the interacting AAs in influenza A virus.

1.1 Influenza A virus

A virus is an independent evolutionary unit in continuous interaction with its host cells. The influenza A virus parasitizes many living organisms such as human, pig, bird and others. The influenza disease has received attention since the beginning of the 20th century after the influenza pandemic of 1918 [2], known as “Spanish Flu” or “La Grippe” caused by the H5N1 virus that was confused first to a common cold. Influenza viruses have been isolated since the beginning of the 20th century from chickens (A/Brescia/1902[H7N7]), from pigs in the late 1920s, from humans in the early 1930s, and from horses and domestic ducks in the 1950s [7]. It is after the Hong Kong/68 human pandemic influenza viruses that the World Health Organization (WHO) encouraged researchers to study flu disease in order to eradicate it [8].

The influenza A virus is a single-stranded, negative-sense RNA virus of the family *orthomyxoviridae* that causes seasonal outbreaks and occasional epidemics [9] in humans. The influenza A virus is distinguished from influenza B and C viruses based on antigenic differences [10]. Influenza B has four antigenic proteins (HA, NA, NB, and BM2) [11] while influenza A virus has three antigenic proteins (HA, NA and M2). Influenza C has seven segments, not eight RNAs like influenza A and B viruses. Influenza C has only one glycoprotein, hemagglutinin-esterase fusion [12], while both influenza A and B have two glycoproteins (NA and HA). Even though all these three influenza viruses can naturally infect humans, only influenza A and B viruses have been responsible for all influenza pandemics, hence the influenza A vaccines recommendations are composed with quadrivalent vaccines [13]. The quadrivalent vaccines provide coverage against two influenza A virus strains (H1N1 and H3N2) and two in-

fluenza B virus strains (from two B lineages: B/Yamagata and B/Victoria [14]) that allow evasion of immunity. Influenza A is the most virulent human pathogen [15] and accounts for about three quarter of total influenza avidity each season [13]. Influenza A virus is a good candidate to understand RNA viruses evolution due to the large availability of its genomic data and different databases that contain all influenza sequences such as NCBI-Influenza Virus Resource [4], Global Initiative on Sharing All Influenza Data [16] and Influenza Research Database [17]. The last database includes viruses circulating from 1918 up to present viruses in different areas of the world, as well as strains recommended by WHO, by predicting which strains of the virus are most likely to be circulating in the next year, for vaccine development in different seasons.

The genome of influenza A is composed of eight segments (Fig. 1.1) enveloped by the lipid bilayer. Nowadays, fourteen proteins [18] (p.58) are known to be encoded by these eight RNA strands: PB2, PB1, N40, PB1-F2, PA, PA-X, HA, NP, NA, M1, M2, NS1 and N1. To be infectious, a single influenza A virus must contain all eight unique RNA segments that play different roles in viral genome [15]. These eight RNA segments and their functions are described below.

- Polymerase subunit 2 (PB2): PB2 protein (759 AAs) is encoded by RNA segment 1 and plays a role during initiation of viral mRNA transcription and replication. PB2 recognizes and binds the 5' cap structures of host cell mRNAs for use as viral mRNA transcription primers [19]. PB2 is one of the proteins that form the influenza RNA polymerase complex (PB1, PB2 and PA)
- Polymerase subunit 1 (PB1): PB1 protein (757 AAs) encoded by RNA segment 2 functions as the protein responsible for elongation of the primed nascent viral mRNA and also as elongation protein for template RNA and vRNA synthesis. PB1 plays a role in both the assembly of polymerase protein complex (PB, PB2 and PA) and the catalytic function of RNA polymerization [15]. This

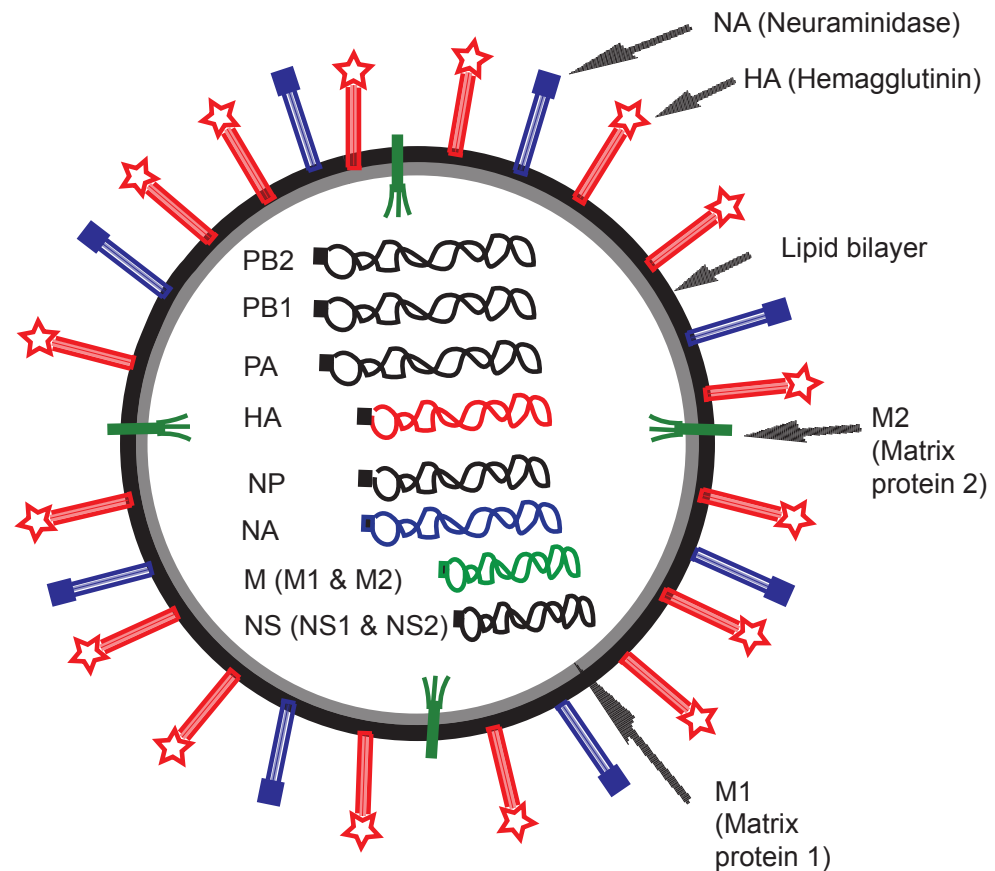


Figure 1.1: **The influenza A virus structure.** The virus is composed with eight segments, listed from largest to smallest: Polymerase subunit 2 (PB2), Polymerase subunit 1(PB1), Acidic Polymerase Protein (PA), Hemagglutinin (HA), Nucleoprotein (NP), Neuraminidase (NA), Matrix protein (M), and Nonstructural protein (NS). Figure is adapted from [1].

gene encodes also N40 protein (718 AAs) whose function is not known yet and PB1-F2 protein (87 AAs) that functions as the apoptosis regulator/virulence

factor [18].

- Acidic Polymerase Protein (PA): PA protein (716 AAs) encoded by RNA segment 3 is the smallest subunit of the influenza RNA polymerase complex. PA is essential for viral transcription and replication [20]. This gene encodes PA-X protein (252) that modulates host response [18].
- Hemagglutinin (HA): HA protein (566 AAs) is the integral membrane protein and the major surface antigen of the virus which is encoded by RNA segment 4. HA is responsible for attaching to sialic acid-containing receptors on the cell surface [15]. HA is composed of HA1 and HA2 that are connected by disulphide linkage. In nature, there are 18 recognized subtypes [21] of HA (H1-H18), that differ by at least 30% in the amino acid sequence of HA1 and are not serologically cross-reactive [7].
- Nucleoprotein (NP): NP protein (498 AAs) encoded by RNA segment 5 plays a role in the assembly and regulation of viral ribonucleoprotein complexes [22]. NP is abundantly synthesized in infected cells and is the second most abundant protein in the influenza virus. NP is also a major target of the host cytotoxic T-cell immune response[7]. NP plays also a role as a scaffold for viral RNA during genome packaging within infected cells [23].
- Neuraminidase (NA): NA protein (454 AAs) encoded by RNA segment 6 is also an integral membrane glycoprotein and a second major surface antigen of the virus. NA cleaves terminal sialic acid of host cell from glycoproteins to free virus particles and facilitate virus spread [24]. NA is highly mutable with variant selection partly in response to host immune pressure [7]. There are eleven subtypes [21] of NA (called N1-N11) have been identified in nature and they are not serologically cross-reactive.

The influenza A viruses are divided into subtypes on the basis of two viral

antigens (HA and NA) of the virus. For example, an “H3N2 virus” designates an influenza A virus subtype that has an HA 3 protein and an NA 2 protein. Similarly an “H1N1” virus has an HA 1 protein and an NA 1 protein. These influenza A subtypes infect different animal species, H17N10 and H18N1 are found only in bats while all other subtypes can infect birds [21]. Currently, only H1N1 and H3N2 are circulating among humans [13].

- Matrix protein (M): influenza virus RNA segment 7 encodes both matrix protein 1 (M1) and matrix protein 2 (M2) using different reading frames. M1 (252 AAs) functions in initiating of virus assembly while M2 (97 AAs), integral membrane protein, serves as a signal for transport to the cell surface [7]. M2 acts as a proton channel to control the pH of the Golgi during HA synthesis and to allow acidification of the interior of the virion during virus uncoating [23].
- Nonstructural protein (NS): RNA segment 8 encodes the two nonstructural proteins NS1 and NS2, using different reading frames, that play roles in virus replication. NS1 is believed to regulate cellular and viral protein expression by binding to different RNA molecules [25]. NS2 is believed to play role as the nuclear export protein [23].

In this project, we used NP, NA, HA, and M2 because they are the ones that have been targeted for the development of antiviral drugs. Gong and coworkers have shown how natural mutations in nucleoprotein interacted gradually along the evolution to stabilize the transcription activity of NP [5]. NP is a good candidate to understand well the influenza A virus evolution as it accumulates almost one mutation per year [5] which makes easier to track its evolution. Influenza antigens NA and HA are recognized by human antibodies [1], hence they are the most studied viral proteins in order to understand how they evolve quickly to escape human immune system and/or antiviral drugs [26, 6, 27]. The other important protein in influenza A evolution is M2 that has been targeted to develop antiviral drugs. For example, amantadine and

rimantadine have been developed to block M2 channel because of its crucial role in viral infection and replication into the host cell cytoplasm [28].

The continuing viral evolution is the most prominent in the influenza antigenic proteins, NA and HA, but also occurs in each of the eight gene segments of the virus [29]. The mutations in the proteins encoded by these eight RNA segments can occur by these following evolutionary mechanisms [7]: (i) antigenic drift (point mutations), (ii) genetic shift (gene reassortment), (iii) defective-interfering particles and (iv) recombination. Each of these mechanisms may contribute to the evolution of influenza viruses. However, recent studies have shown that RNA recombination is very rare or absent in human influenza A virus [30, 31]. Thus, homologous recombination is unlikely to play a major role in influenza A virus evolution. Fodor and colleagues have shown that defective interfering RNAs, incomplete RNAs are due to a single mutation in the polymerase subunit A [32]. However, this mutation responsible for promoting the synthesis of defective interfering RNAs is compensated by other mutations to correct this defect [32]. This may explain that the defective-interfering in RNA is not that important in influenza A virus evolution to produce new viruses. Steel and Lowen have explained the importance of reassortment, a process by which influenza A viruses swap gene segments to produce the new influenza A viruses [33]. For the reassortment, two viruses are of different subtypes (for example H3N1 and H7N7) while for the recombination, the exchange of genetic material occurs between two related viruses during coinfection of a host cell. The antigenic drift consists of small changes in influenza A virus genome that happen continually over time as the virus replicates. The genetic shift in influenza A virus happens only occasionally while antigenic drift occurs all the time in the virus [34](p.28).

Influenza A virus evolves very quickly due to its three characteristics [7]: large populations, short generation times and high mutation rates. The complete cycle of viral replication within a host cells often takes only a few hours [35] and generates many thousands of new viral particles. RNA polymerase [7] has poorer proofread-

ing activity than DNA polymerase. RNA polymerase makes one error per replicated genome, thus each host cell can produce 10,000 new viral mutants to infect neighboring cells [36]. This high mutation rate explains why RNA viruses evolve so readily and hence the driving force behind influenza viral antigenic drift [7]. The mutation in the primary viral protein can be base alterations, insertions and deletions. The influenza A virus accumulates these point mutations over time which sometimes lead to new viruses with different antigenic properties [27], hence they escape from immune system.

The important biological problem in investigating and controlling influenza A virus outbreaks is to show which evolutionary mechanisms interplay between antigenic properties and dynamics of influenza A subtype evolution. Each year, viruses belonging to a single antigenic cluster are responsible for almost all the infections and different antigenic clusters replace each other every 2-5 years in H3N2 influenza A viruses [37]. This antigenic cluster can be observed on a phylogenetic tree where the change of cluster is characterized by a long trunk and short side branches representing closely related sequences that co-circulate every year [38]. Recent studies have shown that double mutations are sufficient to change antigenic clusters in one influenza A virus genome [39, 40]. Koel and colleagues have identified seven positions in influenza A virus H3N2 genome responsible for clade transitions [27]. For example, they have shown that only double mutations of K156Q and E158K in HA protein are responsible for transition of 1995 to 1997 H3N2 cluster. This finding shows that not all substitutions in influenza A virus are responsible for clade transitions but few are key determinants to produce new viruses with different antigenic properties. These findings strengthened the results of the previous studies [40, 26, 41] suggesting that genetic and antigenic distance do not evolve at the same rhythm: few mutations are responsible for antigenic shifts while many genetic mutations occur within antigenic cluster. For instance, even though the viral hemagglutinin has the highest mutation rate, only 10 clusters in HA occurred in H3N2 viruses from 1968 up to 2003 [27].

If double mutations are sufficient to cause antigenic cluster transitions, what is the biological mechanism responsible to slow down the antigenic cluster changes in high mutable viruses? One hypothesis is that co-mutations may be crucial to retain viral fitness[27]. In this study, we identify the epistatic sites and investigate the role of epistasis in influenza A evolution.

1.2 Correlated evolution

Bateson used for the first time the term “epistasis” to describe the masking of one mutation by another [42]. Since then, different scientific disciplines have adopted slightly different meanings to refer to the non-additive effect of two alleles at two loci [43], depending on whether these effects are measured with respect to fitness, as in genetics, or with respect to a particular phenotypic trait, as in systems biology or genomics where epistasis is synonymous with “genetic interactions” [44]. The phenomenon of epistasis in all biological cases is similar; the effect of one biological unit depends on the state of another biological unit [45]. One of the simplest types of epistasis is that which occurs between mutations within a protein. In this thesis, epistasis is defined as the fitness effect of one mutation at a particular position that is modulated by its interactions with one or many mutations at other positions in protein. There are three major categories of epistasis [45]:

- Functional epistasis: is a strictly functional interaction without a direct genetic link within proteins to strengthen the same biological pathway. This category of epistasis has been proven experimentally to exist [46, 47, 48, 49]. For example, Wang and colleagues have shown that a secondary substitution (M182T) far from active site restores thermodynamic stability lost of extended-spectrum beta-lactamase mutant enzyme that was caused by previous mutation (G238S) [49].

- Compositional epistasis: describes the way that a specific genotype is put together, or composed, and the influence that this specific genetic background has on the effects of a given set of alleles [45]. In other words, this epistasis is defined as the masking effect of one locus over another locus. The only way that this effect can be manifested, however, is by combinatorially substituting one mutation for another against a standard background [45]. For example, two loci can be correlated to offer insights on the complex biological pathways underlying human diseases [50].
- Statistical epistasis: means any statistical deviation from the additive combination. For example, the deviation induced by simultaneously substituting two alleles at different loci within a randomly selected individual within the population, after taking into account what would expect the effect of substituting each allele separately would be [45]. In mathematical model, the relationship between multi locus genotypes and phenotypic variation in a population is not predictable based solely on the actions of the genes considered singly [51].

In this project, we focus on *functional epistasis* in influenza A virus protein. We show that epistatic sites identified are not physically linked.

At the same time that influenza A virus evolves quickly to evade the immune system of host cell, the compensatory and permissive mutations aid the virus to maintain its stability and functions. The compensatory mutations are mutations that correct viral fitness lost due to previous mutations while permissive mutations are initial mutations that facilitate the occurring subsequent mutation occur to increase the viral fitness [52]. It has been shown that some mutations compensate the negative effects on HA protein stability caused by previous mutations [53]. The negative effects caused by certain mutations can be alleviated by making certain combinations to one or more mutations [39]. For example, Lunzer and coworkers have shown that 63 of 168 single mutations chosen with the same function in wild-type *Escherchia*

coli isopropymalate dehydrogenase were deleterious when substituted alone. Hereby, epistatic interactions are necessary to preserve protein functions [54]. It has been hypothesized that epistasis is a common occurrence in the evolution of viruses generating numerous compensatory mutations [5, 55]. However, epistasis can be negative if a result of correlated mutations is smaller than additive effects of the individual mutations. In case of deleterious mutations, epistasis is negative (synergistic epistasis) when the interaction of deleterious mutations decreases the fitness of the virus while positive epistasis is when an interaction between deleterious mutations increases genetic robustness by reducing their combined effect on fitness [56]. The reduced effect of combined mutations on fitness (antagonistic epistasis) appears to be more common in RNA viruses than the opposite phenomenon (synergistic epistasis) [57]. As RNA viruses have the highest mutation rate, a fundamental and potential detectable consequence of positive epistasis is the evolution of compensatory changes in gene sequences [56].

The importance of epistasis is evident for influenza A virus that accumulates mutations at a high rate by allowing the virus to adapt rapidly in response to immunological and drug pressure [39, 5]. However, the high mutation can be a way to accumulate many deleterious mutations and hence the virus loses its fitness [57]. Influenza A viruses have evolved to live up with mutations close to their maximum tolerable mutation rates [58]. This tolerable mutation rate is the accumulation of mutations below the mutational meltdown threshold to allow adaptation through the fixation of beneficial mutations. It is believed that the factors that might prevent the mutational meltdown of influenza A viruses populations are high number generations and large population sizes of viruses produced in one cell host infected by many viruses [57]. Consequently, subsequent mutations contribute significantly additional effects, so that multiple fixations occur rapidly and simultaneously. The epistasis compensates sufficiently the accumulation of deleterious mutations [59].

Epistasis has become an interesting topic in complex genetic diseases such as diabetes, asthma and in virus disease such as HIV and flu [60]. Recent studies have focused on epistasis at the molecular level with the aim to identify the interacting sites that play a role in protein functions, structures, and evolutionary dynamics of complex genetic systems [61, 39]. The typical example of intramolecular epistasis is compensatory evolution in RNA secondary structure [62]. The double stranded regions of RNA secondary structures are formed by Watson-Crick pairing of complementary bases. The double stranded stem of RNA is essential for RNA structure, hence the conservation of its secondary structure is very important [61]. If any mutation occurs at a site involved in Watson-Crick pair within stems, it is expected to be deleterious if the secondary structure breaks up. However, RNA secondary structure can be restored when a second compensatory mutation occurs at the appropriate site on the opposite strand of the stem to establish the pairing [62]. The compensatory mutations can be observed in DNA sequence alignments, and served for predictions of RNA evolution [61]. However, the methods developed to predict the coevolution based on RNA secondary structure ignore mutations that occur in other regions of secondary structure like loop region (single stranded). Hence, these predictions on RNA secondary structure may ignore some important biological information. To identify the interacting sites without ignoring any mutated site, an ideal method should use the protein sequences to identify all possible coevolving AAs. However, it is hard to predict the correlated evolution based on just protein sequence alignment as each site has up to twenty possibilities. In this thesis, the chosen method to detect the interacting sites compares all possible mutant sites in protein sequences based on their phylogenetic relationships.

1.3 Epistasis and historical contingency in influenza A evolution

Organisms can be grown under different conditions and sequenced at regular time-points to find all the substitutions that occurred and, among these, the so-called permissive substitutions that made possible the emergence of a particular phenotype or adaptation. Gong and coworkers have shown that V280A mutation was preceded by permissive mutations including M136I that occurred prior to its own fixation [5]. These permissive mutations that allow others to be fixed in virus form the chain of events, that altogether lead to the idea of historical contingency [63]. In any cases, uncovering historical contingency in fast-evolving organisms, such as viruses, can be challenging.

Gould has described the historical contingency as the paths that life can evolve on that are constrained by historical events that are often random [64]. Gould proposed the thought experiment of “rewinding the tape of evolution and replaying it” to test the hypothesis whether life would evolve the same or different ways. The author argued that it is difficult to construct each step of life as the evolutionary paths are many but the destination possibilities are limited. Hence, an integrated approach is required to comprehensively understand protein evolution.

Thinking about different historical events that can constrain evolutionary pathways, they are many and varied. At the molecular level, the evolution of the virus can be constrained by genetic drift (accumulation of mutation in gene) where the frequency of alleles changes randomly over time, natural selection whereby organisms better adapted to their environment tend to survive and produce more offspring, and the environment such as temperature and pH. Some proteins such as enzymes require particular structures to accomplish their functions. To rewind the evolutionary history of enzymes from the ancestral starting point demands to know what is the

starting point and also all structural protein constraints. Some of these constraints are less understood such as those arise from DNA packaging, gene splicing, and well-coordinated gene expression [65]. In addition, protein folding must be considered in order to regain the same protein conformation in order to avoid the interactions with substrate occurring by chance. All these biological constraints show how difficult it could be to replay the protein evolutionary history.

Recent work by Blount and colleagues [66] has showed that historical contingency has occurred through permissive mutations. These permissive mutations do not affect the protein themselves but they facilitate subsequent mutations to produce a beneficial effect on protein. These interactions between permissive and beneficial mutations show how important the epistasis is implicated in the evolutionary history. The experimental determination of the chained events of permissive and beneficial mutations take some time. Blount and coworkers have conducted the long-term evolution experiment, over twenty years, with *Escherichia coli* [66] to characterize permissive (potentiating) mutations facilitating the fixation of beneficial mutations in evolutionary history of that bacteria. The alternative way to determine rapidly the evolutionary events has been suggested by Harms and coworkers where the uncovering historical contingency would be to bypass the culturing and growing phase and directly perform ancestral protein reconstructions based on a computational model [67]. This can unravel the chain of substitutions that happened in the deep past, and their phenotypic effect can be assessed through site-directed mutagenesis experiment [67]. However, if the proteins are very divergent, many fixed differences may have no phenotypic effect, so that it becomes difficult to identify target mutagenesis. A similar difficulty is expected to be found in fast-evolving organisms, such as viruses. In these conditions of high divergence, a purely computational approach may be the only way to uncover historical contingency.

The study of Dunham and coworkers has analyzed the evolutionary pathways of two distinct lineages of swine-adapted A/H1N1 (European Avian-Like and Classical

Swine H1N1) that evolved from avian-like A/H1N1 ancestors [68]. To do so, they compared the genetic changes occurred in these two different lineages to adapt to their species host. Their results showed that the historical contingency played a role in the emergence of viruses from an avian reservoir into a new mammalian host [68]. Dunham and colleagues have shown the impact of the historical contingency in influenza A evolution, however they did not show how mutations were linked in both A /H1N1viruses. In this thesis, the chained interacting mutations in viral proteins are shown.

1.4 Evolution models to detect coevolved sites

Influenza A viruses are good candidates for exploring the nature of epistatic interactions among mutations as they have compact genomes, less than 15 kb in length, with overlapping reading frames and functional secondary structure, both are predicted to lead to strong epistasis [56]. In order to detect correlating sites, hundred of billions of SNP pairs in protein must be considered which requires an exhaustive work and significant computational cost to conduct a full investigation in viral proteins. Fast, sensitive and specific methods are needed to identify and predict the coevolving AAs that are susceptible to be mutated before an outbreak of influenza A virus. To detect experimentally the coevolved sites in influenza A virus is very tedious as it requires the analysis of each possible pair or more than a pair of mutations. Therefore, it takes a long time to detect experimentally the interacting sites in protein as it requires to construct natural evolution and to assay for the functions that contribute to biological fitness [5].

Traditional methods were based on the idea that coevolving pairs of sites in a protein should leave a typical signature in a sequence alignment, which can be detected using quantities such as mutual information [39]. However, the epistatic sites may be identified by these methods only if the divergence times between samples are very

large [39]. Most of models developed to detect interacting sites have met with only limited success in identifying pairs of sites basing on alignments. These challenges are: (i) distant sites may be as important as adjacent sites in compensatory process; (ii) number of sites in the compensatory process may be large so that the pairwise correlation signal is too weak to be detected; (iii) epistatic signal changes with evolutionary time, and (iv) methods chosen may be insufficient for separating the evolutionary process from background noise. Therefore, an ideal method should take account of phylogenetic relationship among sequences to identify substitutions at pairs of sites on phylogenetic branches [69].

A number of studies have used likelihood methods to identify coevolved sites in the specific case of the evolution of sites involved in RNA structures [70, 71]. These RNA methods benefit from the precise knowledge of RNA secondary structure, and strict and concise rules which govern the RNA structures. For example, in the stem of RNA structure, the pairwise relationships between sites are pre-defined, and the likelihood of correlated relationships between all pairs of this stem can be tested simultaneously [72]. However, a major requirement of these RNA methods is the availability of multiple sequences from divergent viruses [73]. In addition to that, the RNA secondary structure remains constant over the phylogenetic tree. Despite of the restrictions of these methods, they have contributed to understand better the coevolution and the dynamics of adaptation. Shapiro and colleagues have developed a phylogenetic method to detect positive epistasis in gene sequences of RNA viruses [56]. Their model calculates the number of sites involved in compensatory mutations and infers the phylogenetic distribution of the observed compensatory mutations. They have found that the majority of epistatic sites occurred along internal phylogenetic branch which indicate they are beneficial. Nasrallah and Huelsenbeck have proposed the other phylogenetic model that detects epistatic sites [61] in simulated 5S rRNA alignments. However, this model implies the computability of the posterior probability which may be complex with large data sets.

The prediction of protein structure from alignment is considerably less precise than RNA structure and the rules which govern its formation are extremely complicated [72]. An ideal method to identify the interacting AAs should be the one that maps sites on phylogenetic tree so that their correlated evolution can be tested. However, proteins are built up from a pool of twenty AAs that can be in any combination among the sequences at each sites, permitting a complexity of coevolutionary relations between sites that defies analysis within the limits of current sequence availability and computational power. Pollock and coworkers have developed a phylogenetic model to detect interactions AAs [72] in protein. However, their approach recodes the data sets as positively and negatively charged residues and ignores neutral AAs. The epistatic AAs detected may be false as some residues are ignored.

Poon and coworkers [74] has developed the method implemented in a program called Spidermonkey that detects the interacting sites using Bayesian graphic models (BGMs). A BGM is a representation of a joint probability distribution in which each node represents a distinct random variable [75]. This method is based on ancestral reconstructions to compare mutational paths by using the logical rule of “if A then B” which means that wherever phylogenetic node “A” is seen, then the phylogenetic node “B” will be added regardless how node “A” was derived. This dependence may be problematic when there are many phylogenetic nodes to analyze as many representations of the same data are possible. Moreover, the analyses with Spidermonkey are limited to small datasets [74]: 150 sequences and 1000 nodes if $k=1$ or 75 nodes if $k=2$, where k is the maximum number of parents per node.

Recently, Kryazhimskiy and colleagues developed a phylogenetic method to detect consecutive non-synonymous substitutions, one at leading (first) and other at consecutive trailing (second) site, that coevolved [39]. This method requires a large number of substitutions per sites, especially at the trailing site. Moreover, this method requires the parsimony reconstruction which may be problematic when there are so many internal phylogenetic branches. Their method performs badly when for proteins

that evolve slowly [39], thus this approach could not detect the interacting sites in influenza A virus nucleoprotein which accumulates few mutations per year [5]. Despite of these shortcomings of this method, it has been used to empirical data and identified some coevolving mutations proven experimentally, hence it is compared to our approach.

1.5 Objectives

In this thesis, we develop the novel statistical method, to detect the epistatic AAs in influenza A virus, that alleviates the restrictions of above phylogenetic methods. We take inspiration from statistical method developed by Pagel and Meade [76, 69] in ecology with the aim at detecting correlated evolution among phenotypic traits. As the original method analyzes only binary data, we adapted this statistical method to protein level (twenty AAs) to test if AAs have coevolved. Our maximum likelihood method circumvents the need for explicit ancestral reconstructions while maintaining reasonable computational costs.

To validate our approach, we applied our method to the same data sets used with previous computational method [39] and data sets that have been experimentally proven to contain epistatic sites [5]. We compared epistatic sites detected by our approach to what had been identified by previous studies [39, 5, 6]. As we have a large data of protein alignments to analyze, we reduce the computational burden and speedup the analysis with the parallel code and analyzing only one of possible identical sites along protein sequences (see Chapter 2). The analysis of large data sets lead us to revisit the nature of epistasis in influenza viruses.

We extend our analysis by applying our approach to different data sets to test different hypotheses on the role of epistasis in influenza A virus evolution: (i) The important problem for studies of influenza A virus evolution is to show what mechanism between the antigenic properties and the genetic profile of virus. It has been

hypothesized that virus cluster changes are due to the co-mutations [27]. To test this hypothesis, we applied our approach to data sets [27] tested for the co-mutations responsible for clade changes with the aim to test whether epistasis is responsible for clade changes in influenza A virus. (ii) The majority of current influenza A viruses that are circulating, they are resistant to antiviral drugs due to the certain mutations in influenza A virus genome. There is a hypothesis that virus resistance to drugs is not caused by the drug pressure but instead to co-mutations in the influenza proteins [77]. To test this hypothesis, we applied our approach to M2 data sets with the aim to analyze whether the mutation (S31N) known to confer the resistance to M2 inhibitors [78] has coevolved with any other mutation in M2 protein.

Chapter 2

Materials and Methods

2.1 General approach to detect epistasis

(a) Repurposing of BayesTraits

In order to model correlated evolution at the molecular level, we employed the model developed by Pagel and Meade [69] and implemented in BayesTraits. This model was originally developed as a Markov process to investigate the coevolution of discrete binary traits within phylogenetic trees, with the aim of testing whether a dependent model of trait evolution fits the data better than an independent model. The dependent model allows two traits to coevolve as the rate of change at one trait depends on the state at the other trait, while the independent model does not place any restriction on rates of changes [69]. As these two models are nested, a likelihood ratio test can be employed for model selection.

In its original formulation, binary trait states are represented by zeros and ones, a dash representing missing values. When two traits are simultaneously considered, they can be in four different configurations. As it is assumed that only one trait can change in an infinitesimal amount of time, eight transitions can occur, each of them taking place with a time-homogeneous rate $q_{i,j}$ (see Fig. 2.1). Dual transitions,

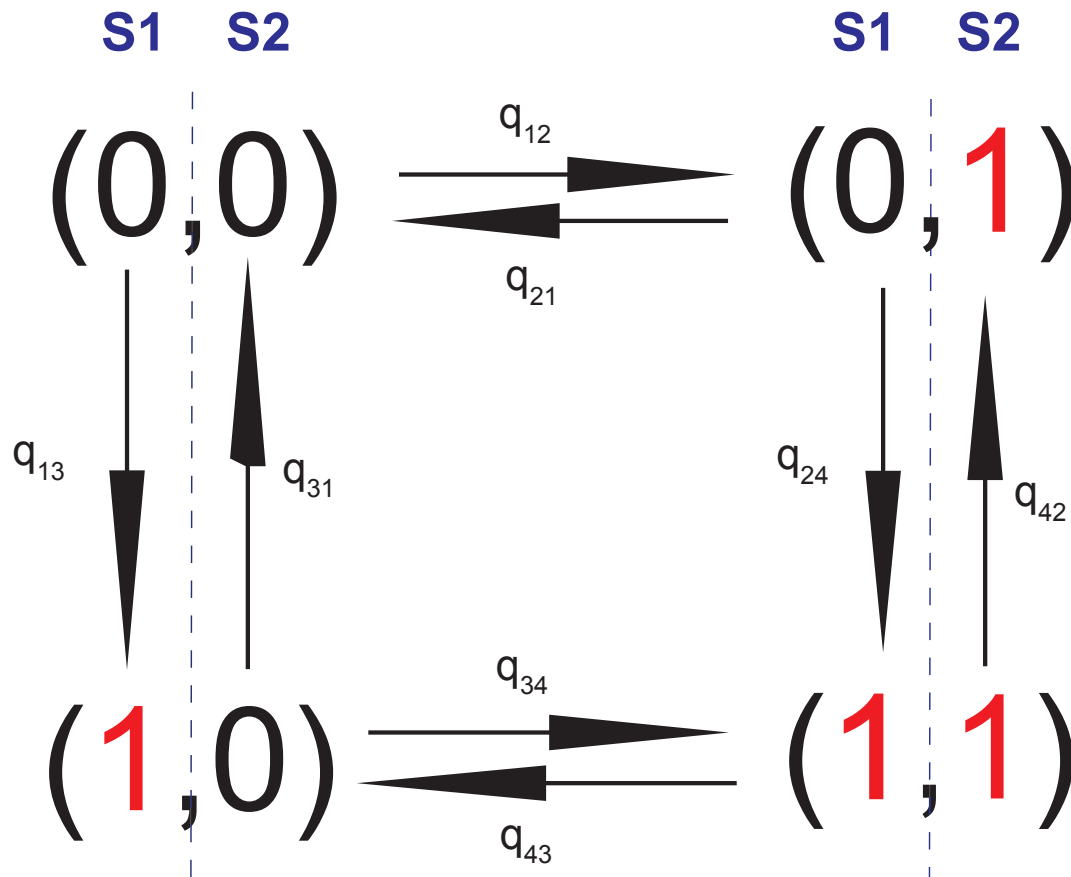


Figure 2.1: **Transitions among the four combinations of states resulting from two binary variables.** Pair (i, j) identifies the states of a particular pair of traits, that here are amino acid positions S1 and S2. Transitions are represented by arrows. Instantaneous rates of change are denoted $q_{i,j}$. Following the original model, we assume that there are no self and no double transitions.

in which the states of both variables instantaneously change, are not assumed to occur [69]. The state of one trait changes while the state of the other trait is held constant in each transition [69]. For example, in both transitions of rates q_{12} and q_{34} (Fig. 2.1), the second variable changes from “0” to “1” states where the first variable held constant “0” or “1”. If both transitions q_{12} and q_{34} are not equal, it means that the change of states in the second variable depends on the background of the first

variable, thus the two traits are coevolved. Otherwise, if q_{12} is equal to q_{34} , it signifies that the change of the sates in the second variable does not depend on the sate of the first variable, hence both traits are not coevolved. The model compares all eight transitions $[(q_{12}, q_{34}); (q_{21}, q_{43}); (q_{13}, q_{24}) \text{ and } (q_{31}, q_{42})]$ to test whether two traits are or not correlated. With the same logic, we test whether two sites (S1 and S2) along protein sequences are coevolved.

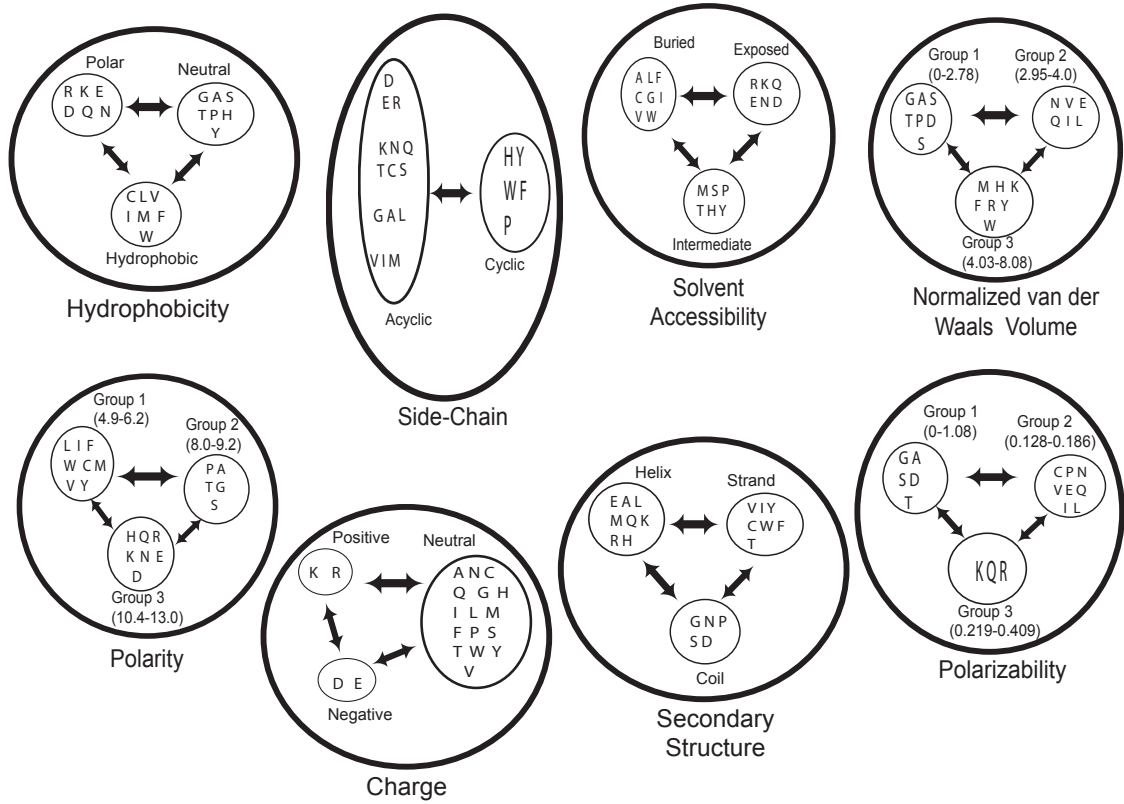


Figure 2.2: **The amino acids properties and their categories (states) used to recode protein alignments.** For each category, the arrows show the different transitions among physiochemical properties. For continuous categories, the numbers between bracket indicate the range of values employed in the discretization process as per the `protr` R package.

(b) Data recoding

Here however, our goal is not to study binary traits, but to analyze the coevolution of pairs of sites along a sequence alignment. In the case of proteins, each site has twenty AAs (states) possible. Data recoding is therefore necessary to reconcile the data with the approach, and in this optic we evaluated two different strategies. First, AAs were partitioned according to their physicochemical properties (Fig. 2.2). Binary properties (side-chain group types) were naturally recoded “0”/“1” where “0” and “1” are assigned respectively to acyclic and cyclic AAs. For those with $k > 2$ states (the seven other properties), we compared each of the k states against the other ones. For instance, in the case of charge, we first assigned state “0” to negative and state “1” for non-negative (positive and neutral), and circled through the two other states. Alignments were recoded based on the physiochemical properties recorded in the R package `protr` ver. 0.2-1 [79]. Second, AAs were classified as either ancestral or derived: at each position of an alignment, the ancestral state was defined as the consensus AA present in a clade used to root the tree. In the case of viral sequences, this rooting clade was defined either as that containing the oldest sequences and this clade was removed for downstream analysis, or as the basal clade in a relaxed molecular clock analysis (see below).

(c) Code optimization

To discover pairs of AAs that are potentially interacting, we need to run the above model on all $\frac{n(n-1)}{2}$ pairs of sites in an alignment of length n . This computation can be prohibitively long even with an alignment of modest size. For instance, the influenza H3N2 nucleoprotein has 498 AAs, which lead to analyzing 123,753 pairs of sites under both the dependent and the independent model (*i.e.*, 247,506 models need to be run). To alleviate the computational cost, we first compressed the alignment into site patterns [80] (p.105). The compression depends on the sequence alignment

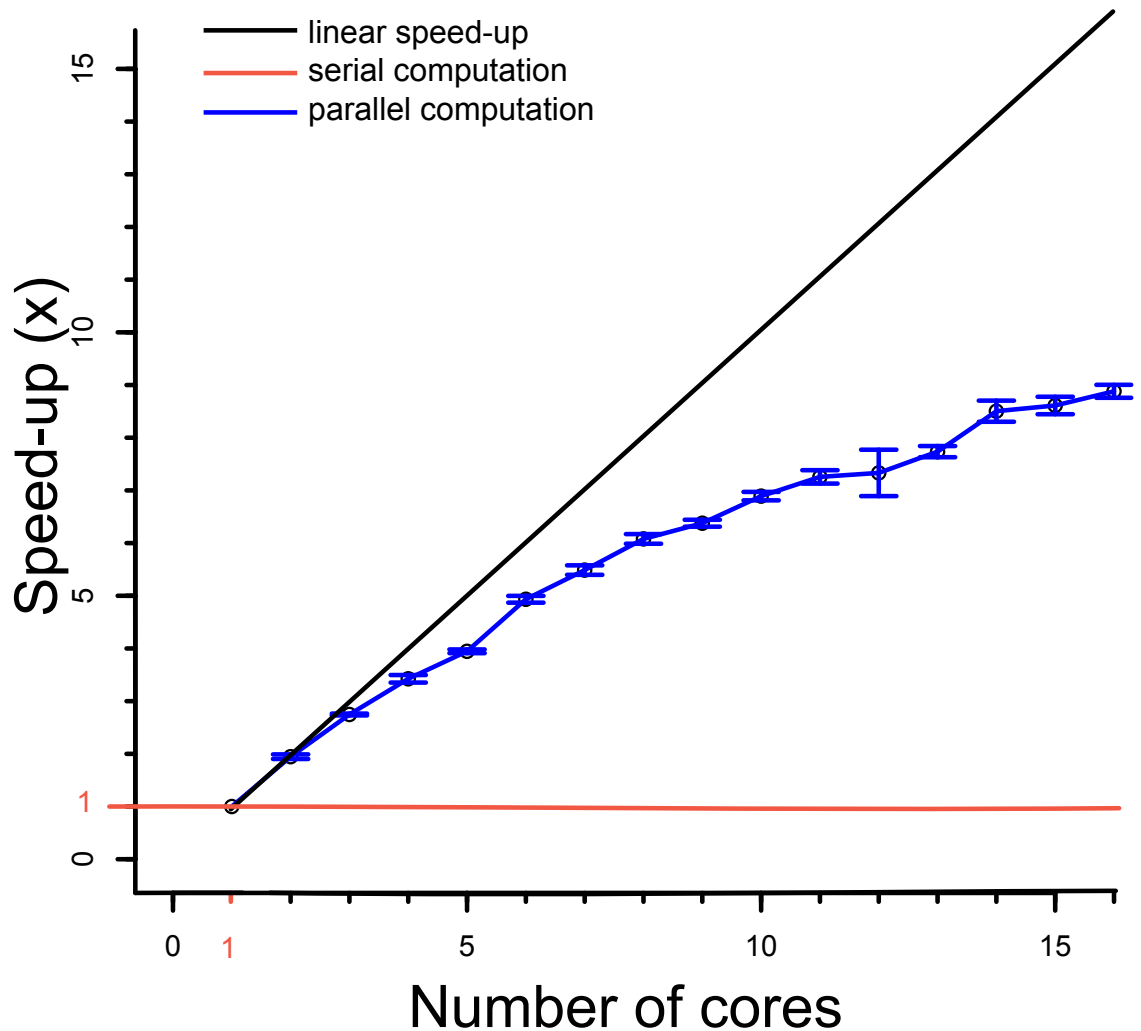


Figure 2.3: **Speed-up as a function of number of cores requested.** Speed-up is a measure of “acceleration” (how many times faster is the computation with n cores) compared to a serial computation (on a single core; in red). The black line indicates the ideal situation of linear scaling of performance with resource. The actual speed-up curve is shown in blue. Error bars show one standard error, based on four replicates. Overheads due to latency (among others) lead to a non-linear speed-up curve.

divergence. Second, as pairs of sites can be independently compared, we parallelized the code using R's `foreach` ver.1.4.2 and `doMC` (ver.1.3.3) [81] packages to take advantage of multicore/multiprocessor architectures (Fig. 2.3); when performing site comparisons, we parallelized the outer loop. According to Amdahl's Law [82], if a fraction p of a computation can be run in parallel and the rest must run serially, then the speedup will be $1/(1-p)$. Theoretically, the maximum number of processors that can be used effectively is the inverse of the fraction of time the task must proceed on a single thread [83] (Fig. 2.3: linear speed-up). However, there is a limit in performance depending on the fraction of time spent waiting for access to shared storage elements [82]. With the sequential serial code, the increase of the number of processors does not speed up the code as there is overhead (Fig. 2.3: serial computation).

To account for multiple testing, we computed the false discovery rate (FDR) according to the Benjamini-Hochberg (BH) procedure [84]. Briefly, the FDR was controlled at the nominal level where p -values of null hypotheses were ordered from smallest to largest. The smallest p -value was ranked of $i = 1$, the next had $i = 2$, etc. Then each individual p -value was compared to $(i/m)Q$, where m is the total number of tests and Q is the chosen false discovery rate. The largest p -value that had $P < (i/m)Q$ was significant, and all p -value smaller than it was also significant. We used R (v3.0.2) for all analyses [85]. In the analyses presented below, the pairs of AAs identified to be interacting were subsequently mapped on three-dimensional protein models predicted by homology modeling with SWISS-MODEL [86] and plotted using KiNG [87]. The physical (Euclidian) distance between two residues on the three-dimensional protein structure was determined between their alpha carbons (C_α).

2.2 Validation based on previous evidence

In order to validate our method to detect epistatic pairs of sites (AA), we undertook two approaches: the analysis of data sets that were analyzed by previous computa-

tional epistasis method and data sets in which epistasis was experimentally confirmed.

(a) Previous computational analyses

We reanalyzed the data sets analyzed in a previous computational study [39]. The four data sets obtained from the original authors consist of: 1,219 HA and 1,836 NA sequences from H1N1 viruses, as well as 2,149 HA and 2,339 NA sequences from H3N2 viruses. These data sets are here denoted KDBP11-H1 (HA in H1N1), KDBP11-N1 (NA in H1N1), KDBP11-H3 (HA in H3N2) and KDBP11-N2 (NA in H3N2), respectively.

(b) Experimental evidence

As computational studies make predictions that are not always tested or validated, we reanalyzed data sets in which epistasis was experimentally confirmed. Focusing our work on influenza A viruses, a number of recent studies reported evidence for epistasis and we present results on two of these viruses (H1N1 and H3N2). First, we reanalyzed data published by Gong and coworkers [5] which comprised 424 H3N2 human influenza A nucleoprotein (NP) sequences spanning 42 years between 1968-2010. This gene was originally chosen because it evolves relatively slowly and is hence amenable to experimental validation as all substitutions can be easily tested by site-directed mutagenesis [5]. The viral sequences that they used were downloaded from the IVR database [4]. This data set is here denoted Gong13NP.

We also retrieved a second data set as analyzed by Duan and coworkers [88], who used an alignment that contained 1366 human influenza A H1N1 neuraminidase (NA) collected between 1999 and 2009; as in their analysis, the 2009 H1N1 pandemic sequences were excluded. This data set is denoted Duan14NA.

(c) Nature of epistasis in influenza evolution

After performing these validations on experimentally-validated data, we set out to investigate the nature of epistasis by testing two hypotheses. First, we revisited the work by Koel and coworkers [27] that experimentally validated the existence of AA substitutions involved in changes of antigenic clusters. For this, we used 877 H3N2 human influenza hemagglutinin (HA) sequences as used in [27], collected between 1968 and 2003, to test if these substitutions responsible for antigenic changes also showed signal of epistasis. This data set is denoted Koel13HA.

Second, we tested if our statistical approach could detect some evidence for epistasis at a pair of sites involving the S31N substitution, which is responsible for conferring resistance to the anti-influenza drug Adamantane [78]. For this, we retrieved 668 H3N2 human influenza A matrix protein 2 (M2) sequences that were collected between 1968 and 2003. This data set is denoted Adam03M2.

(d) Phylogenetic analyses

Maximum likelihood was employed to estimate phylogenetic tree for the Gong13NP, Duan14NA, Adam03M2, KDBP11-H1, KDBP11-N1, KDBP11-H3 and KDBP11-N2 alignments with FastTree ver. 2.1.7 [89] under WAG + Γ_4 model to account for among-site rate variation. Trees were rooted using the earliest sequences, which were Aichi68-NPA/Aichi/2/1968, A/Victoria/JY2/1968, A/Albany/17/1968, AAD17229/USA/1918, CAA24269/Japan/1968, AAF77036/USA/1918 and ABI92283/Australia/1968, respectively. The R package APE was used to visualize the trees [90]. For the Koel13HA data set, a rooted tree was reconstructed using BEAST ver. 1.8.0 under a relaxed molecular clock assuming an uncorrelated lognormal prior [91] and a constant-size coalescent prior under the FLU + Γ_4 substitution model. Analyses were run in duplicate to check convergence, for a total of 100 million steps with a thinning of 5,000; log files were combined with LogCombiner after conservatively removing

the first 10% of each chain as a burn-in period, as checked with Tracer ver.1.5 (tree.bio.ed.ac.uk/software/tracer).

Phylogenetic uncertainty was taken into account by running the algorithm on bootstrapped trees (or trees sampled from the posterior distribution). Because only the Gong13NP data set showed a large proportion of SH-like aLRT [92] node support values in the low $[0.0, 0.8]$ range (Fig. 18-25), the results of the bootstrap analyses are only shown for this case.

Chapter 3

Results and Discussion

3.1 Little overlap with previous computational results

As a first evaluation of our algorithm, we reanalyzed four large data sets (1–2,000 sequences) previously analyzed with another method designed to detect epistasis [39]. The data were recoded AAs as ancestral/derived states (Chapter 2).

Comparing the coevolved sites detected by both methods (the previous method and ours) in each data set (Fig. 3.1, Tables 1-4), our algorithm detects fewer pairs of sites than the previous method. Our approach failed to detect any epistatic pair of sites in KDBP11-HA (H1N1) while a single pair of sites in both KDBP11-NA (H1N1) and KDBP11-NA (H3N2), and sixteen pairs of sites in KDBP11-HA (H3N2) were identified by our approach. The previous study has detected more than two hundred pairs of interacting sites in each data set (they also applied an FDR correction to their tests). Overall, there is barely an overlap between two methods. Is our method weak to identify the possible epistatic sites? It looks like the first impression that the results give when we compare both methods, but the results are different of each data set. The following paragraphs are the discussion on each data set.

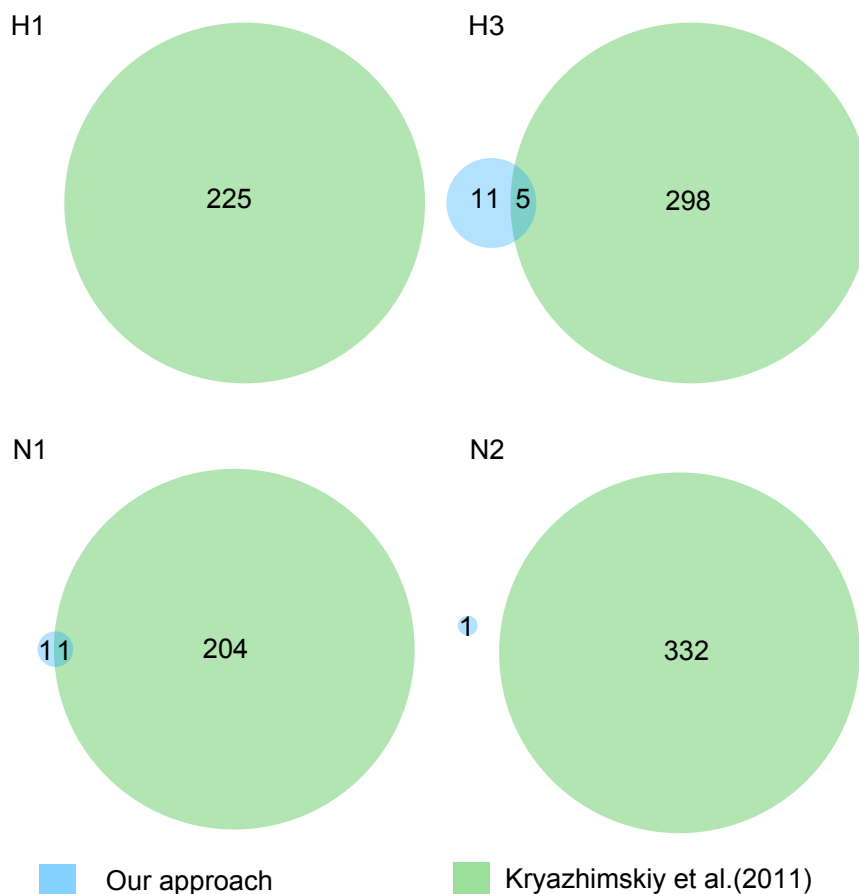


Figure 3.1: **Comparison with the method by Kryazhimskiy *et al.* (2011): after FDR.** The numbers in the Venn diagrams show the total of pairs of epistatic sites detected by our method (in blue) and the previous method (in green) using the four large KDBP11 data sets. The diameter of each circle is proportional to the number of pairs of epistatic sites detected by each method.

Among these data sets, it is only in hemagglutinin (H3N2) we detect many coevolved sites that form the epistatic network (Fig. 3.3.B). This network indicates that the mutations identified are not random. The central mutation in this network at 172 position has experienced four mutations (Fig. 3.3. A): K172E, E172K, K172Q and Q172H, that coevolved with other fifteen mutations in HA protein (H3N2) evolution. When these detected mutations are mapped on three dimensional protein structure

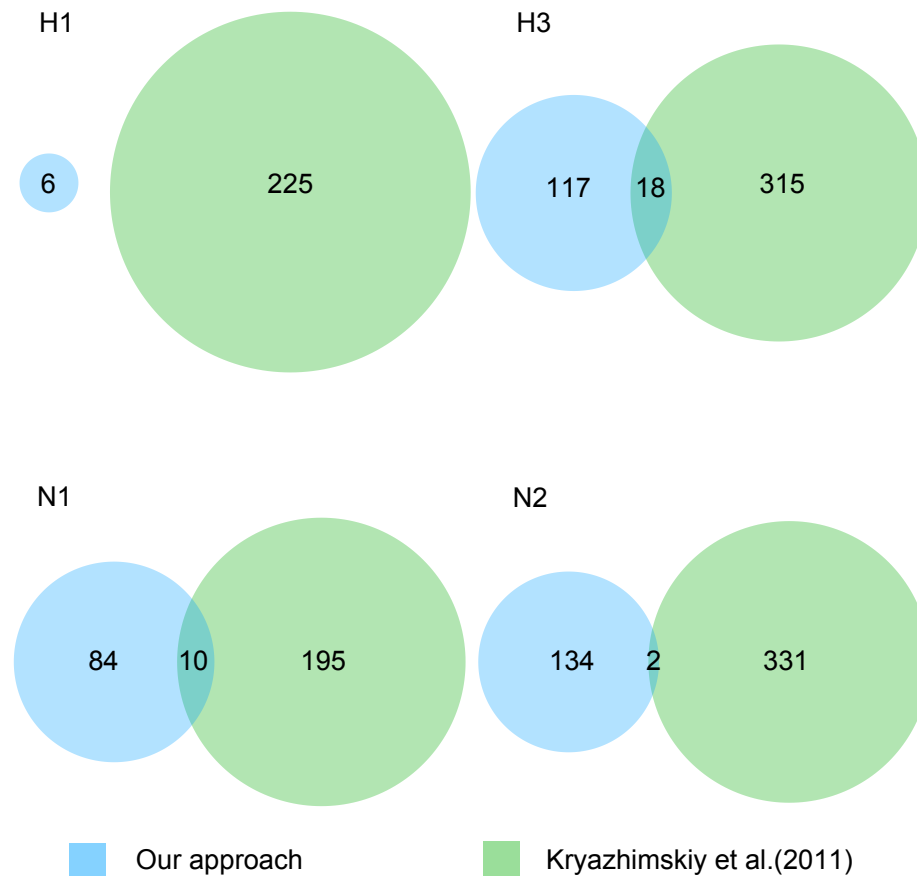


Figure 3.2: **Comparison with the method by Kryazhimskiy *et al.* (2011): before FDR.** The numbers in the Venn diagrams show the total of pairs of epistatic sites detected by our method (in blue) and the previous method (in green) using the four large KDBP11 data sets. The diameter of each circle is proportional to the number of pairs of epistatic sites detected by each method.

(Fig. 3.3.C), they link with long-range interactions and these coevolved sites are not linked to their physical distance (Fig. 3.4). Thereby, the long range-interactions that link epistatic sites in pairs may be functional epistasis types as these sites are not physically close. These long range-interactions may raise a question whether the epistatic sites we identified are real. Jonathan Dench, my laboratory partner, has shown with simulation data (manuscript submitted) that our approach distinguishes

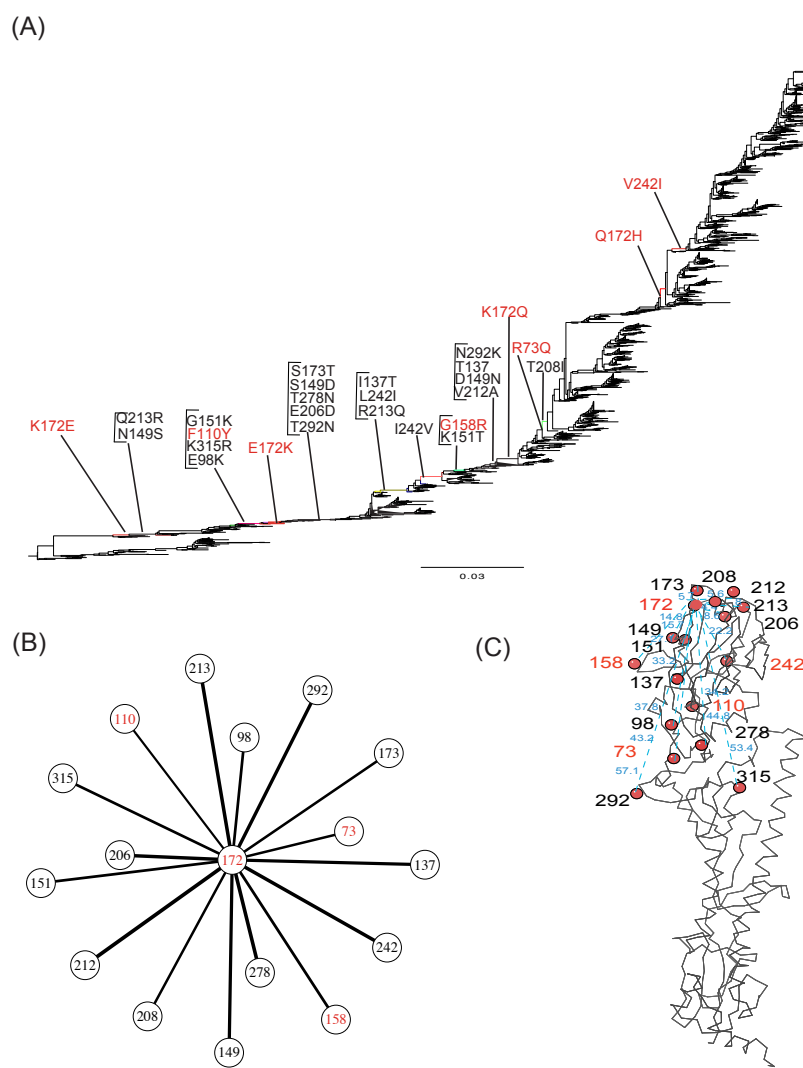


Figure 3.3: **Epistatic pairs of amino acids detected in the KDBP11 H3 data set of H3N2 viruses.** (A) The epistatic mutations that we detected are plotted on the HA phylogenetic tree. The substitutions in red are the ones detected by both Kryazhimskiy *et al.* (2011) method and our approach. (B) Chained epistasis of interacting AAs. The thickness of the links is proportional to $-\log_{10} P$ -value, the strength of evidence supporting the interaction. (C) The epistatic sites are mapped on three dimensional H3 protein structure (2YP2). The numbers show the AA positions detected by both methods (in red) are those detected only in this study (in black). The numbers in blue show the physical distance between epistatic sites (in Å).

between false positive and true epistatic sites, thereby these correlated sites may be real. If the epistatic sites detected are real, why have we identified less interacting sites than the previous computational method [39]? Again basing on simulation results (manuscript submitted), our approach has an excellent specificity but with moderate sensitivity. This average sensitivity may be one of the reasons why we detected fewer epistatic sites than the previous study and fail to detect any interacting AA in hemagglutinin of H1N1.

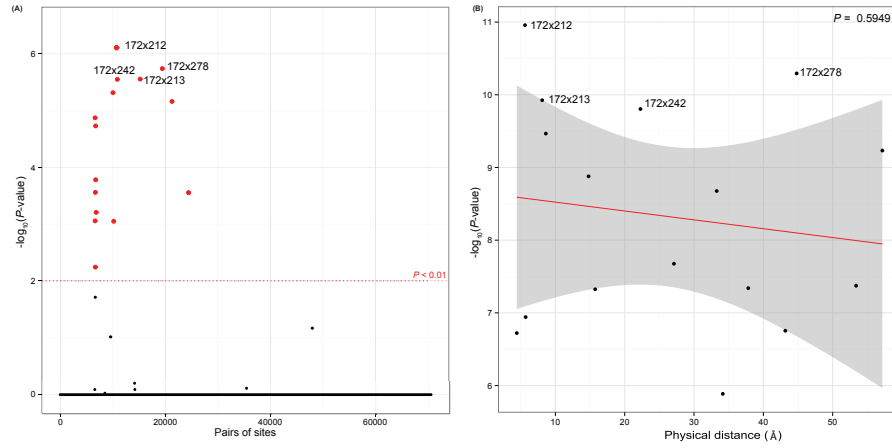


Figure 3.4: The epistatic sites detected in human influenza hemagglutinin and the physical distance among them. (A) Pairs of AA sites analyzed using the KDBP11 H3 data set. Each pair is indexed using an arbitrary ordering. Pair of sites detected as epistatic with a $P < 0.01$ after FDR are shown in red. (B) The strength of the association is not related to physical distance (in Å) between sites of each epistatic pair. The significance of the regression (slope) is shown in the top right corner of the panel.

For neuraminidase data sets in both H1N1 (Fig. 3.5) and H3N2 (Fig. 3.6), we detected a single pair of coevolved sites for each data. The pair of sites (H275Y×D354G) that we detected in H1N1 was also identified by the previous computational method while the pair of sites (N41I×Q5K) in H3N2 was identified only by our approach. Why does our approach detect only single pair of epistatic site in neuraminidase while this protein evolves quickly? Analyzing our results in detail for both methods, the re-

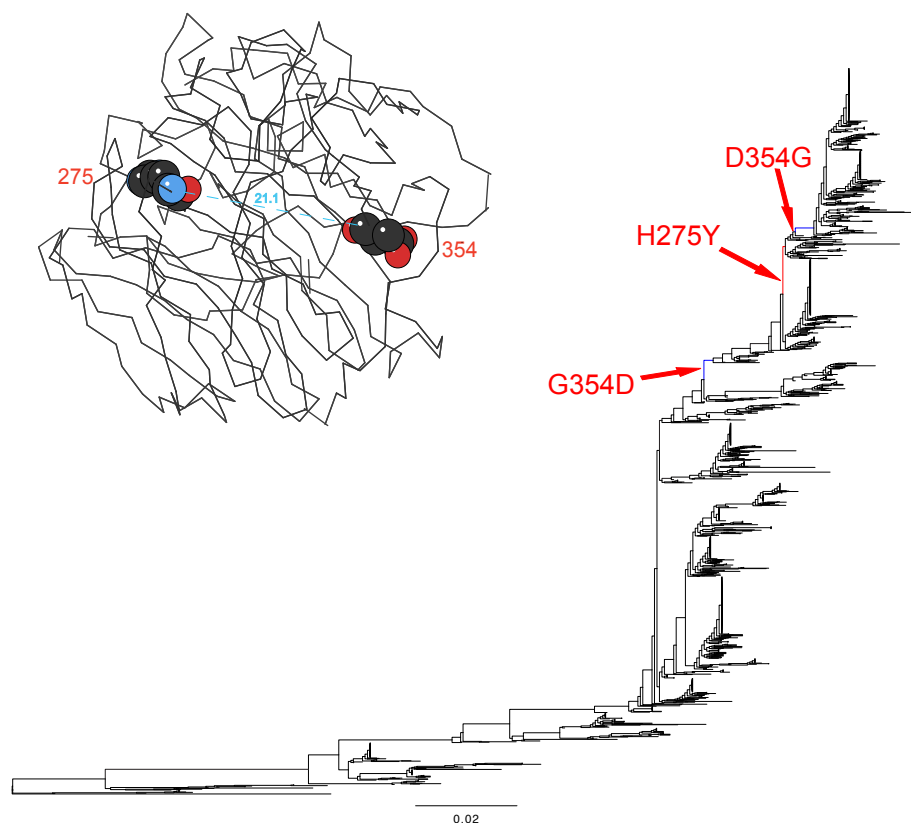


Figure 3.5: **Epistatic pair of amino acids detected in the KDBP11 NA data set of H1N1 viruses.** The epistatic mutations that we detected are plotted on the NA phylogenetic tree. Inset: the epistatic sites are mapped on three dimensional NA protein structure (2BAT). The numbers show the AA positions detected by both methods (in red). The numbers in blue show the physical distance between epistatic sites (in Å).

sults of the previous computational study have been presented with P -value cut-off 0.0004. Despite of this small P -value cut-off, the previous method [39] identified more epistatic pairs of AAs, 2 to 30 times, than our approach whose P -value cut-off is 0.01 after FDR. The study of Kryazhimskiy and colleagues [39] compared both their method and the Spidermonkey method developed by Poon and his colleagues. Similarly, Spidermonkey detected almost 10 times fewer epistatically interacting sites

than the method of Kryazhimskiy and coworkers when it was applied to the same data sets [39]. The results of Spidermonkey strengthen the hypothesis that the approach of Kryazhimskiy and coworkers may not be that specific. In other words, our approach may be too rigorous to select the true interacting sites among true epistatic pairs of sites. To test this hypothesis, we compared our results before FDR to theirs. The comparison of both methods shows a bigger overlap (Fig. 3.2) than the previous comparison with our results after FDR. The number of epistatic pair of sites identified before FDR are six times more than the ones detected after FDR. Arguably, the increasing of overlap between their approach and our method with our results before FDR may indicate that the previous study [39] might have identified false positive coevolved sites.

3.2 Our approach detects experimentally-validated pairs, and more

As the previous data sets were only examined from a computational point of view, we turned to additional data that are experimentally validated. At least two such data sets exist in the literature, the Gong13NP [5] and Duan14NA [88] data.

First, we analyzed the Gong13NP data set of 424 influenza NP protein sequences [5]. As this is the smallest data set in our study, we evaluated our two ways of recoding the data into binary character states, first by partitioning AAs into their physiochemical properties. With this first approach, four pairs of sites were detected before FDR correction (Table 5), but none of them matching the pairs detected by Gong and coworkers. Note however that after FDR, no interactions were significant. This suggests that even if the physicochemical recoding seemed a priori as a good idea, capitalizing on the chemistry of life, it wastes statistical power on multiple three-way tests (Chapter 2; Fig. 2.2). Otherwise, the lack of the identifica-

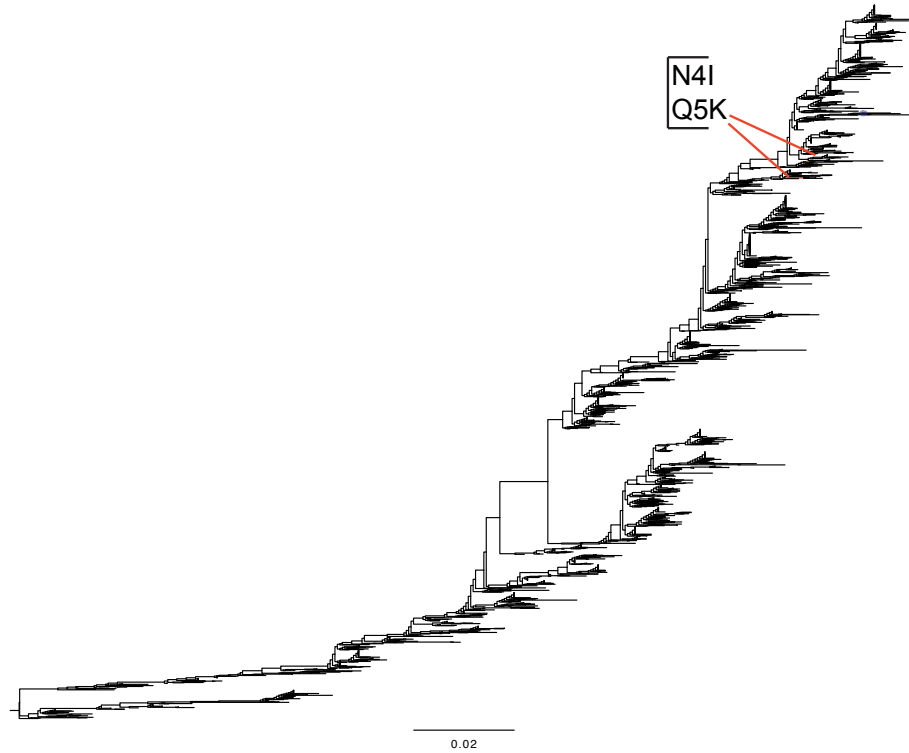


Figure 3.6: **Epistatic pair of amino acids detected in the KDBP11 NA data set of H3N2 viruses.** The epistatic mutations that we detected are plotted on the NA phylogenetic tree. The numbers show the AA positions detected by our method

tion of epistatic pairs of sites by recoding data into their physiochemical properties may be caused by some inaccessible AAs due to their aggregation. For example, the protein regions rich in hydrophobic residues are responsible of aggregation process of polypeptide chains [93, 94] (p. 46). The hydrophobic residues are hidden in aggregated regions and hence they may not be exposed to interact with hydrophilic AAs.

Second, we recoded AAs as ancestral/derived states (Chapter 2). Under this recoding, seven pairs of sites were detected: L259S×N334H, D421E×I425V, R246G×R470K, G217S×N334H, G217S×V343L, L259S×D421E and V186I×L259S (Fig. 3.7A, Table 6). Among these, the L259S and N334H substitutions both occurred in 1973 and,

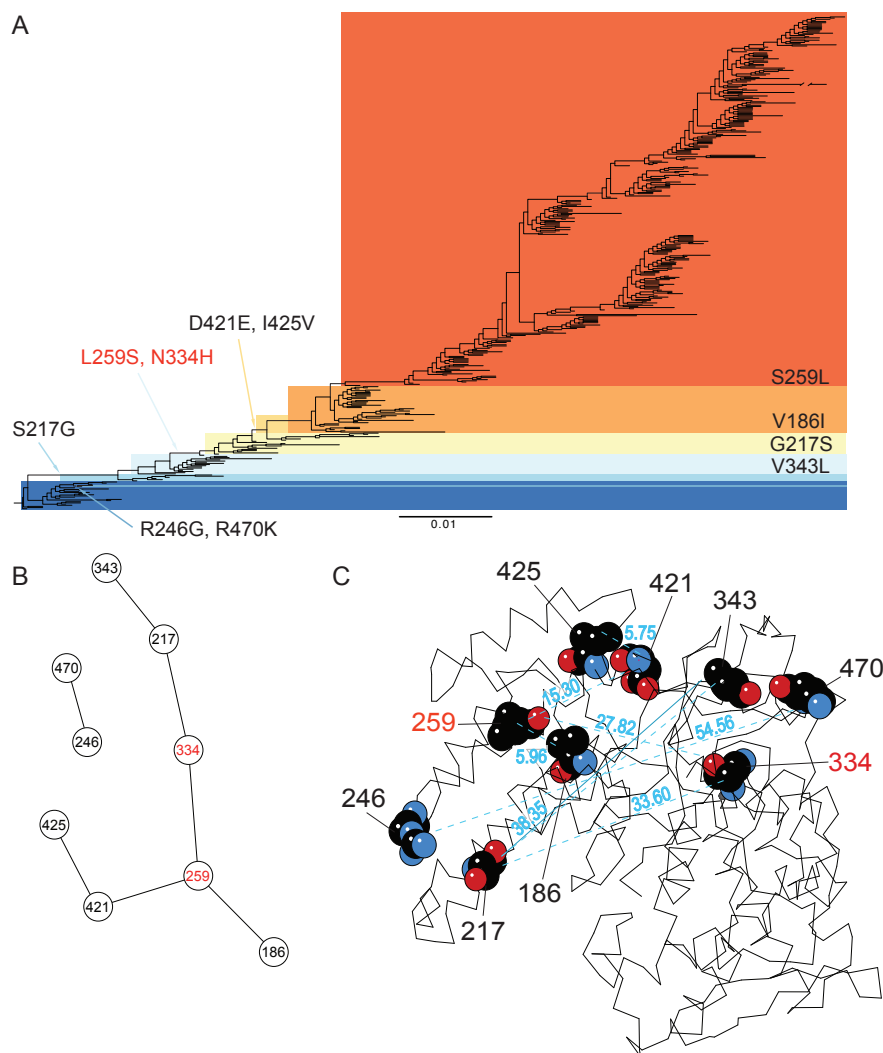


Figure 3.7: **Epistatic pairs of amino acids detected in the Gong13NP data set with the ancestor / derived recoding.** (A) The epistatic mutations that we detected are plotted on the NP phylogenetic tree. The substitutions in red were experimentally validated [5]. (B) Chained epistasis of interacting AAs. The thickness of the links is proportional to $-\log_{10} P$ -value, the strength of evidence supporting the interaction. (C) The epistatic sites are mapped on three dimensional NP protein structure (3ZDP). The numbers show the AA positions experimentally validated (in red) and those detected only in this study (in black). The numbers in blue show the physical distance between epistatic sites (in Å).

critically, were experimentally shown to coevolve [5]. However, we detected six more pairs of sites that were never shown to be epistatic. These could be false positives, but again our simulations suggest that our approach is extremely specific (manuscript submitted).

The experimental design employed by Gong and coworkers essentially permits the discovery of hierarchical epistasis, where the effect of a mutation masks completely the effect of another one. As such, it can not detect “gradual epistasis”, where the non-additive effects are due to > 1 substitution [95]. Supporting this hypothesis is our unexpected finding of not only > 1 epistatic pairs, but that almost all these pairs are all linked to the experimentally confirmed L259S and N334H substitutions (Fig. 3.7B). Indeed, apart from R246G $\times$ R470K, all the five other AAs interact directly or indirectly with the experimentally confirmed pair. If the gradual hypothesis is correct, it would be expected that the chained epistatic AAs would undergo substitutions along the same phylogenetic lineage. Here again, this prediction is not validated by the phylogenetic distribution of these substitutions: (i) R246G $\times$ R470K occur along its own lineage; (ii) in the trunk lineage, we first find S217G, then V343L, then back to G217S, then the famous L259S/N334H (ordering cannot be told from the tree), followed by V186I and finishing with D421E/I425V (ordering cannot be told from the tree), exactly following the network reconstructed in Fig. 3.7B. In the face of phylogenetic uncertainty (Fig. 3.8A), the distributions of P -values measuring the strength of evidence for epistasis (Fig. 3.8B) show that, if detection of R246G $\times$ R470K is severely affected by phylogenetic noise, the six other pairs remain largely significant. This robustness to phylogenetic uncertainty strongly supports the hypothesis that such a network of epistatic sites, which we term “chained epistasis”, undergoes substitutions along the same phylogenetic lineage. However in this case, experimental evidence is not available to validate the gradual nature of this chained pattern of epistasis.

Can we say anything about the nature of these epistatic interactions? If they were physical, we would expect to observe that interacting sites would be in close

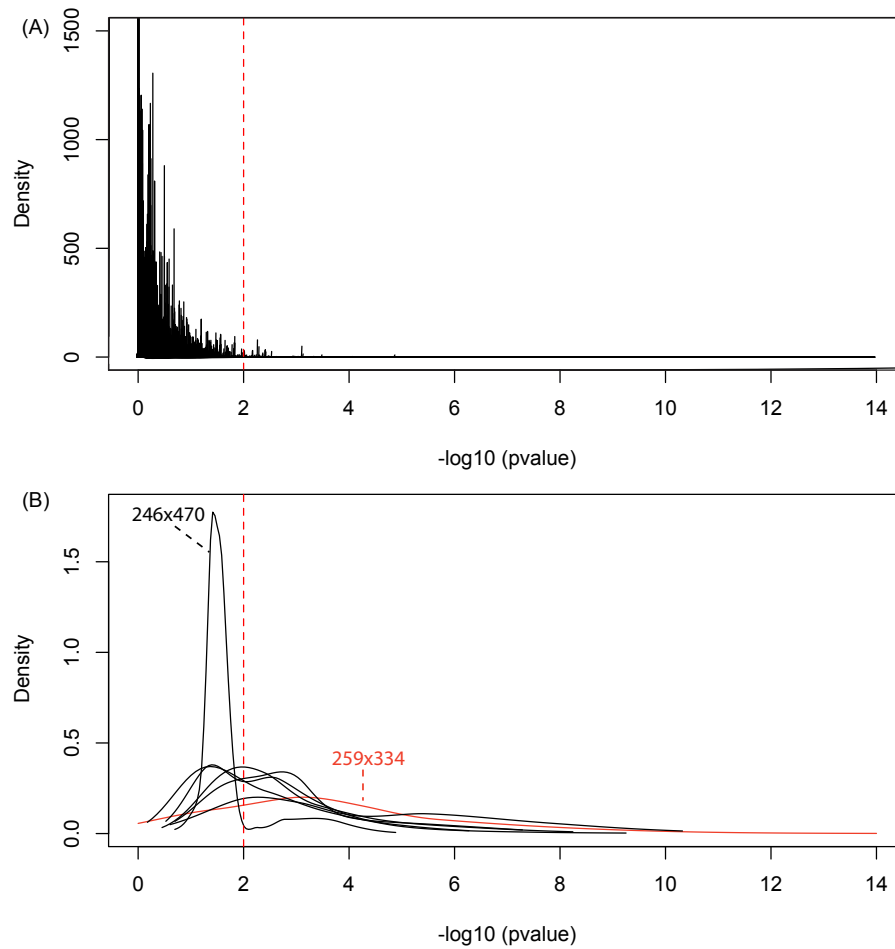


Figure 3.8: **Strength of evidence in the face of phylogenetic uncertainty.** Both panels show the strength of evidence on a $-\log_{10}$ scale when testing for epistasis in the case of the Gong13NP data set for all $\frac{424(424-1)}{2} = 89,676$ pairs of sites. Each density represents the distribution of the P -value for one particular pair of sites in the face of phylogenetic uncertainty, over 100 bootstrapped trees (A) for pairs of sites where no significant interaction was detected by using the maximum likelihood tree and (B) for the seven significant pairs. The dashed vertical red line shows the significance cutoff ($P < 0.01$, *before FDR*). In (A), the plot is censored at 1500 in order to visualize all distributions.

proximity in the folded protein. However, the spatial distribution of these epistatic pairs of sites on the protein structure does not conform to this prediction: while two

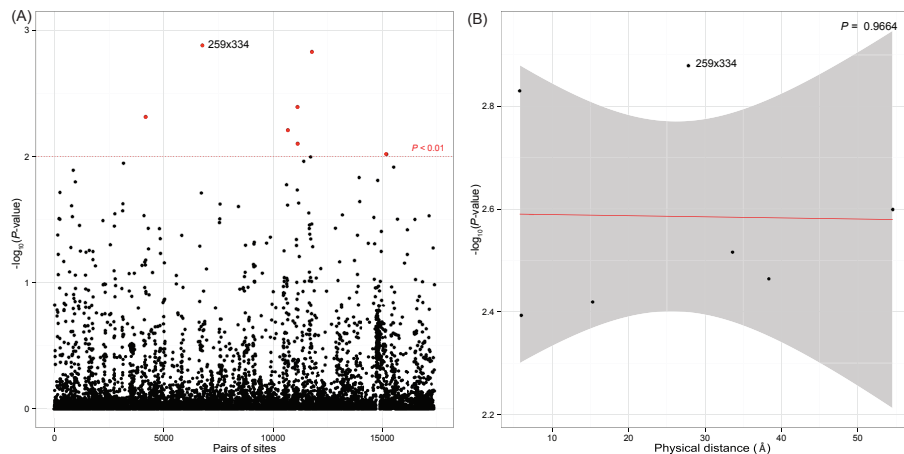


Figure 3.9: **The epistatic sites detected in human influenza nucleoprotein and the physical distance among them.** (A) Pairs of AA sites analyzed using the Gong13NP data set. Each pair is indexed using an arbitrary ordering. Pair of sites detected as epistatic with a $P < 0.01$ *before* FDR are shown in red. (B) The strength of the association is not related to physical distance (in Å) between sites of each epistatic pair. The significance of the regression (slope) is shown in the top right corner of the panel.

residues are considered physically linked when the distance of their alpha carbons is ≤ 8.5 Å [96], we find that the average distance between interacting AAs is 25.9 Å ($sd = 18.1$), three times as much (Fig. 3.7C). This distribution strongly suggests that chained epistasis is not linked by physical interactions (Fig. 3.9).

To further validate our approach, we also analyzed the Duan14NA data set, comprising 1366 NA sequences of pre-pandemic H1N1 viruses [88]. We identify the epistatic pair H275Y×D354G (Fig. 3.10, Table 7), which was experimentally proven to confer oseltamivir resistance that dominated the population in 2008-2009 [88]. In their study, these authors showed that D354G was the main mutation responsible for maintaining the function of NA after alteration of enzyme activity by H275Y, so that this is potentially the strongest existing interaction. However, our approach failed to identify the five other mutations (V234M, R222Q, K329E, D344N and D354G) that were further identified by Duan and coworkers to be interacting with H275Y. This

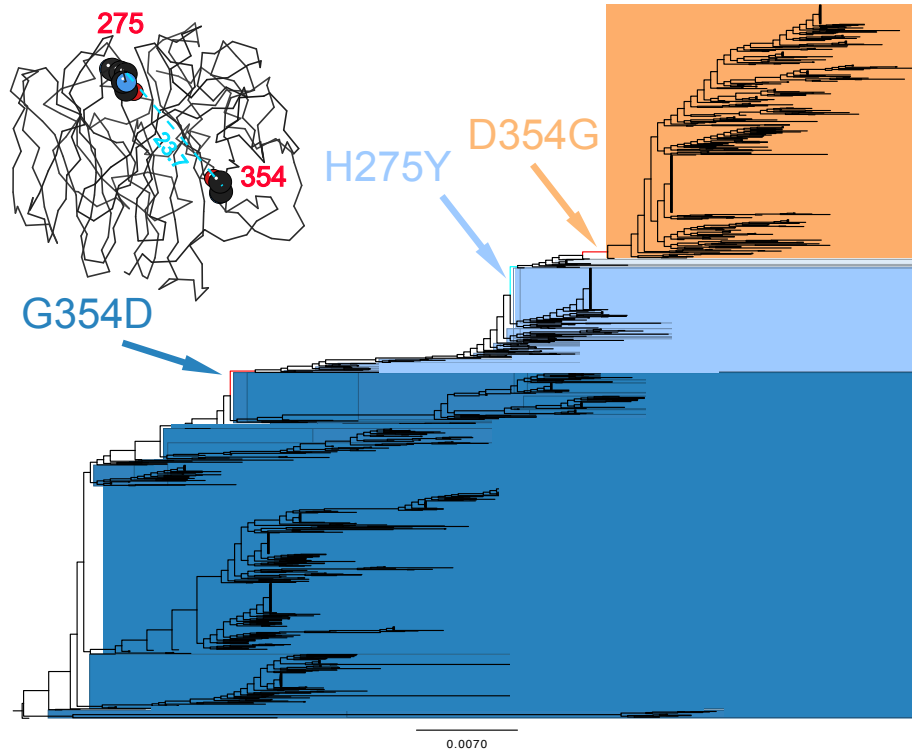


Figure 3.10: **Epistatic pairs of amino acids detected in the Duan14NA data set.** The epistatic mutations that we detected are plotted on the NA phylogenetic tree. Inset: the epistatic sites are mapped on three dimensional NA protein structure (1HA0). The numbers show the AA positions experimentally validated (in red). The numbers in blue show the physical distance between epistatic sites (in Å).

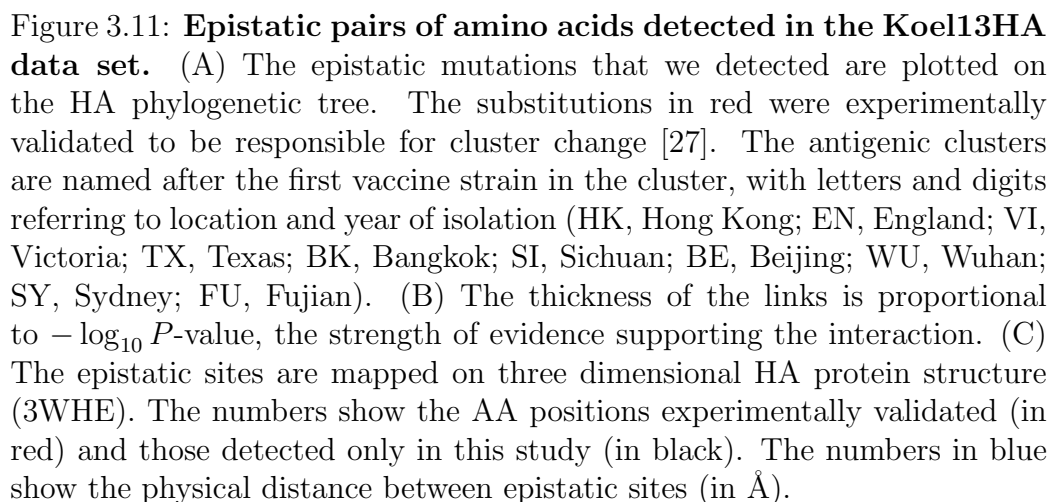
estimated H1N1 tree is more symmetrical in shape than the one estimated for the Gong13NP and has shorter average branch lengths (see scale bar in Fig.3.10). Our simulation results suggest that in this case, the sensitivity of our approach can be very small (paper submitted). This expected low sensitivity might explain why we fail to detect the five additional sites interacting with 275. In spite of this negative result, we find that the interacting pair is, again, a long-range interaction, weighting in at 23.7 Å (Fig. 3.10, inset). The objective of the next two subsections is to explore more systematically the nature of epistasis in influenza viruses, focusing more specifically on (i) the existence of chained long-range epistasis and (ii) the nature of these long

range interactions.

3.3 Chained long-range interactions occur sequentially

In a second study, the original authors investigated the mutations involved in changes of antigenic clusters, and found that double mutations could suffice to explain such cluster changes [27]. The high mutation rate of this virus' antigene however does not explain why these cluster changes do not occur more often than the observed 3.3 years, leading these authors to postulate that “co-mutations” may be required to maintain viral fitness. In light of our study, the immediate interpretation of these results would be that epistasis is involved during cluster change. We re-analyzed these data in order to test this hypothesis.

By doing so, we find that some of the substitutions previously (and experimentally) implicated in cluster change are indeed involved in epistasis (Fig. 3.11A,B, Table 8). Among all ten influenza clusters, we did not detect any epistatic sites in two cluster transitions: EN72/VI75 and SI87/BE89 because only single mutations were responsible for clade changes [27]. In particular, as in the Gong13NP and KDBP11-H3 data sets, we find evidence for chained epistasis, either as short chains such as 146×155×158, involved in two consecutive cluster changes, but we also find a much larger network of interactions involving 20 sites, some of which being also involved in the last four cluster changes (Fig. 3.11B). Again, as in the Gong13NP data set, a temporal sequence of substitutions along this network can be found: G124D, found at the SI87/BE89 transition, interacts with K299R, G172D and E82K; G172D interacts with sites involved in the next transition, BE89/BE92, such as position 135, which is again involved in the next transition, BE92/WU95, where 172 is interacting with 172, 262 and 226, which is interacting with 156 which is involved in the WU95/SY97



transition; finally, 156 is interacting with 192, involved in the SY97/FU02 transition. It is tentative to propose that such chained epistasis may provide an explanation, as an evolutionary constraint, to the paradox of high mutation rate and slow antigenic evolution. Yet, can we delve further into the nature of these constraints?

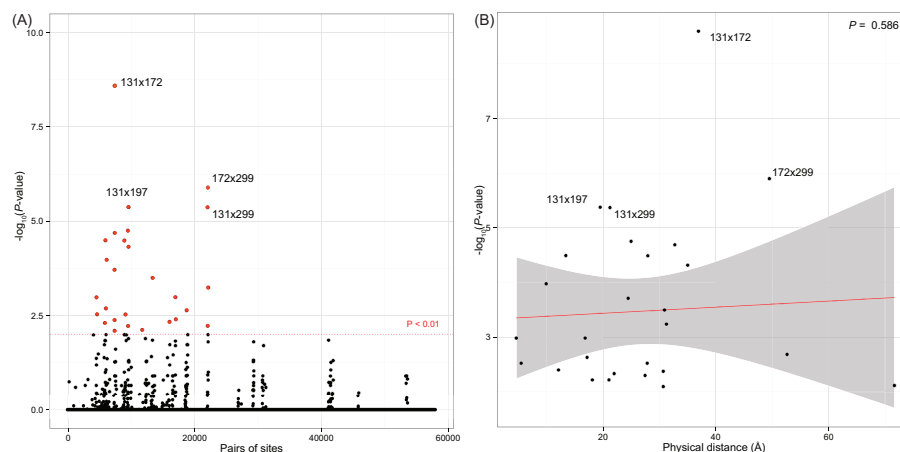


Figure 3.12: The epistatic sites detected in human influenza hemagglutinin and the physical distance among them. (A) Pairs of AA sites analyzed using the Koel13HA data set. Each pair is indexed using an arbitrary ordering. Pair of sites detected as epistatic with a $P < 0.01$ *after* FDR are shown in red. (B) The strength of the association is not related to physical distance (in Å) between sites of each epistatic pair. The significance of the regression (slope) is shown in the top right corner of the panel.

At first inspection, Fig. 3.11C suggests that all these interactions are located in the head of the HA protein and hence might respond to steric constraints, physically-mediated. (Fig. 3.12A,B, Table 8) confirms that the strength of the association is not related to physical distance between sites of each epistatic pair. However again, the average distance between pairs of interacting sites is $23.7 \text{ Å} (sd = 10.4)$, which is much larger than 8.5 Å . Can we obtain some evidence about the nature of such long-range interactions?

3.4 Long-range interactions are likely to be functionally-mediated

The analysis of a fourth data set, Adam03M2, sheds some unexpected light on this question. Influenza viruses are resistant to M2 inhibitors, such as adamantane, and this resistance is associated with the S31N amino acid substitution, which is found in more than 95% of the currently circulating viruses [97, 98]. There is evidence supporting that the spread of S31N may be unrelated to drug selection pressure and instead results from its interaction with advantageous mutations located elsewhere in the viral genome [77] or maybe merely in the M2 gene. To test this epistatic hypothesis, we used our approach to analyze a data set of M2 sequences. The results show that only one pair of epistatic sites (S31N $\times$ V51I) is detected (Fig. 3.13, Table 9). Again, it is a long-range interaction (35.96 Å), but the reason why this data set is illuminating is that the mutation at position 51 has been shown to play a role in virus replication by stabilizing the amphipathic helices of the M2 protein [99]. Thus, V51I may enhance the fitness of M2 protein to increase the frequency of adamantane resistance associated with S31N mutation. The reversion I51V that appeared in few sequences in 2000 (red clade in Fig. 3.13) was apparently quickly lost, which supports the hypothesis that V51I mutation is (i) compensating S31N and (ii) has permitted the emergence of S31N.

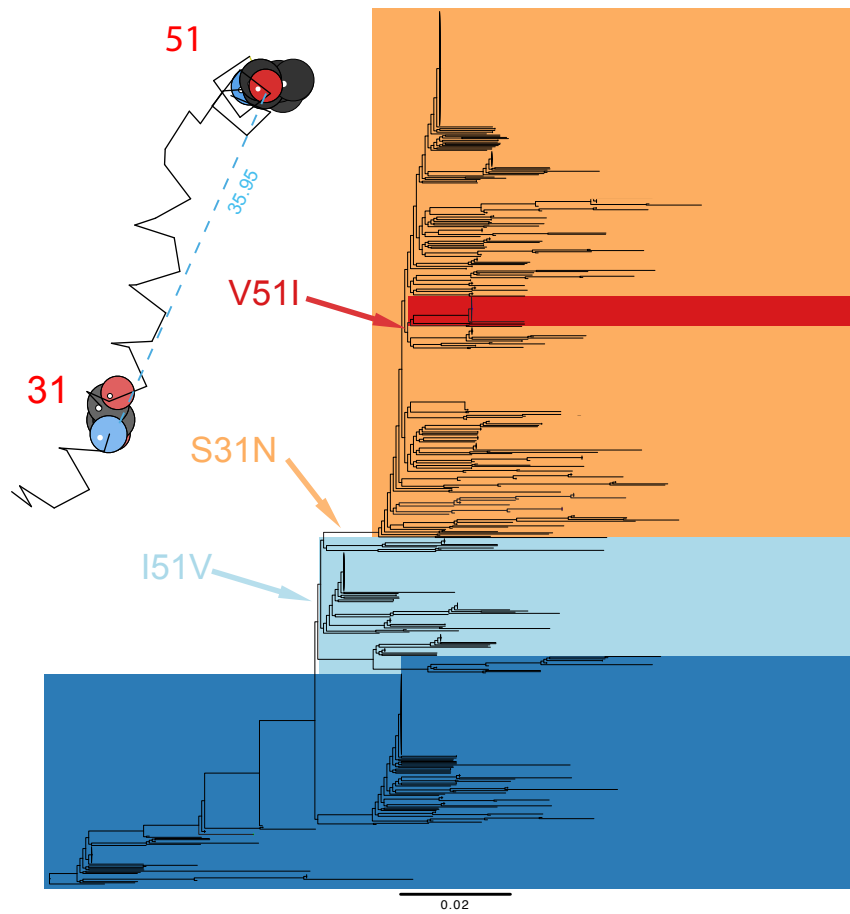


Figure 3.13: **Epistatic pairs of amino acids detected in the Adam03M2 data set.** The epistatic mutations that we detected are plotted on the M2 phylogenetic tree. Inset: the epistatic sites are mapped on three dimensional M2 protein structure (2KIH). The numbers show the AA positions detected (in red). The numbers in blue show the physical distance between epistatic sites (in Å).

Chapter 4

Conclusions and Future Directions

4.1 Conclusions

Epistasis has been studied at molecular level to understand the protein structure, function and evolutionary dynamics of complex genetic systems. Since Fisher described almost a century ago epistasis as the non-additive effect of two alleles at two loci [43], there has been different interpretations ranging from the phenotypic impact of specific alleles in the context of a specific genetic background [45], to a strictly functional interaction without a direct genetic link within proteins to strengthen the same biological pathway. While this category of epistasis is supported by ample experimental evidence [46, 47, 48, 49], functional epistasis differs from what is commonly understood in molecular biology, where mutated AAs need to be physically close to one another on the three dimensional protein structure to be effectively interacting [96].

It is this idea of interaction, which entails coevolution, that we harnessed here by co-opting a method developed in ecology to test for the correlated evolution of phenotypic traits [69]: instead of treating pairs of traits as such, we repurposed the method to deal with pairs of AA sites. Because the original method was developed

to handle binary traits, we explored data recoding and showed that treating AAs as ancestral/derived consensus states was a more sensible (albeit less intuitive) option than using physicochemical properties. We then validated our approach against other computational results, showing little overlap, and against experimentally-validated results, showing wide overlap, hereby suggesting that detected pairs of AAs are indeed genuine.

With this good statistical behavior, the further analyses of many data sets showed a consistent pattern: that of finding (i) many pairs of sites involved in epistatic interactions, (ii), that these pairs of sites are not independent but “chained” along a network of sites where substitutions occur sequentially (hereby supporting the idea of gradual epistasis [95], probably as some of these substitutions are permissive to others that showing evidence for historical contingency in fast-evolving organisms, such as in the case of our M2 data, where I51V seems to be required for the resistance mutation S31N to occur without fitness loss thanks to a functional interaction [99]) and (iii), more intriguingly, that these epistatic pairs of sites form long-range interactions. Hereby precluding the idea of a close physical link as in the case of tRNA molecules [100, 62], so that these long-range interactions must bring about stability [101] and/or conformational [102, 103, 67] and/or functional changes, as in the case of the M2 data set. While there is evidence that epistasis can be prevalent in RNA viruses (at least 31% in [56]) and in bacteria (15% in [48]), it is not impossible that functional epistasis reflects an evolutionary constrain stronger in RNA viruses than in larger and more redundant genomes: because these viruses have a small size, mutations are expected to have large fitness effects, which can be alleviated by compensatory mutations [104]. The accumulation of permissive substitutions can be seen as a coevolutionary process, where what happens at one AA site depends on what happened at another site. This result is not absolutely novel, as it was already found in [39], but these authors never clearly pointed this out in their work, maybe because its significance was unclear.

Many mutations have more impact than one mutation on a protein evolution [105], each mutation in epistatic network contributes to the protein function and/or structure during the influenza A virus evolution. As these chained mutations lead to the historical contingency, it is problematic to trace the most probably evolution trajectory as many pathways are possible among possible mutations. Most of empirical studies have conducted the analysis of correlated evolution by jumping from one mutation to the target mutation without considering beneficial mutational neighbor [106, 48]. Obviously, it is very tedious and takes a long time to uncover experimentally correlated mutations as it requires the knowledge of the fitness function of the protein and map each mutation to its fitness. To make it more complicated, the effect of each mutation varies with genetic background [106]; one mutation may be beneficial or deleterious depending on the genetic background. Here, our approach alleviates all these restrictions to detect all possible interacting mutations by scanning all mutations in protein and mapping them on phylogenetic tree to reveal the evolutionary tape of protein.

Our choice of focusing on a segmented RNA virus such as influenza allowed us to limit our analysis to searching for evidence of epistasis within genes, *contra* among genes. This way of analyzing data intra-genically might make sense in the case of such viruses, as different segments code for proteins involved in relatively different functions, but experimental evidence in other organisms, such as yeasts, shows that epistatic interactions can involve multiple genes and hence be *inter-genic* [107]. Although the computational costs of whole-genome scans can seem prohibitive, assessing the genomic prevalence of epistasis remains an unexplored area – but one that my laboratory is currently investigating.

4.2 Future directions

Should chained epistasis be a common feature of evolution, a number of paradigms ranging from our understanding of the nature of interactions ([39] and this work) to the missing heredity in genome-wide association studies [51], may have to be revisited. One additional example is the idea of diminishing return, where consecutive adaptive substitutions increase the fitness of organisms by smaller and smaller amounts, as observed in experimental evolution studies [108, 109]. In this context, how can we explain the existence of large networks of chained epistatic sites? One possibility would be that a changing environment creates new adaptive landscapes, and that a natural population (*contra* those from *in vitro* studies) do not climb peaks on the landscape but rather chase moving targets [110] (p. 36). While it is not clear how such landscapes are robust to changing environments [111], they are certainly a reality in the world of viruses, where vaccination regularly alters the adaptive landscape, hereby leading to chained networks of epistatic interactions as a result – a mere by-product of evolution. But then, one can wonder if the metaphor of a landscape itself is appropriate when all mutational trajectories are not accessible from particular genomic backgrounds [106, 107]. This may be one of the reasons why evolution is so difficult to predict [106, 112]. Here however, the analysis of the M2 data set provides us with the evidence that these long-range interactions are functionally mediated. The generality of these results remain to be explored.

The influenza A virus is highly mutated to keep pace with rapidly changing environments. It is expected that natural selection promotes the mutations in viral proteins without causing the loss of proteins functions. Although it is well accepted that adaptive evolution occurs in influenza, the importance of episodic and positive selection to lead to influenza epidemiological dynamics have not been systematically investigated. Some epidemiological models have assumed continuous antigenic change [113, 114], while others have assumed episodic antigenic change [115]. These

models have tried to understand the mechanisms that limit the genetic diversity of the virus despite its high transmissibility and mutation rate. Recently, Trevor and colleagues have shown that both continuous and episodic selection exist in influenza A virus [116]. Their results showed that episodic selection can play a role in the low diversity of influenza A virus. However, they did not clarify which force might be responsible for maintaining low diversity in viral HA. One member of our group, Neke, has shown that maintenance of multiple lineages (through positive selection) in Ebola virus enables the proliferation of epistasis by facilitating the most recent diversification event (Neke et al., manuscript in preparation). Some mutations that have evolved through episodic selection were implicated in correlated evolution in Ebola viruses. We propose that future work on the dynamics of the influenza virus should investigate the role of epistasis in fixation of episodic mutations. Moreover, as our approach analyzes only the dependence of two mutations in protein, models of complete dependence should be employed on our datasets to validate the ability of our approach to detect all possible dependent sites [117, 118, 119].

In this thesis, we have focused only on intramolecular interactions. However, it has been shown that intermolecular interactions have impacts on influenza A virus evolution [120, 121, 122]. For example, Mintae and coworkers have identified six pairs of correlated positions [122] among HA, NA, and M2 proteins that form the viral envelope. To detect epistatic sites among and within all eight RNA segments, all genome virus should be scanned. The idea will be to align AA sequences of all RNA segments (Fig. 4.1) and code to which segment the sequence sites belong. The sites detected would be plotted on three dimensional proteins encoded by all eight segments to predict viral protein-protein interactions. To detect epistatic sites among proteins would help to understand how amino acid residues located at the interface of two different proteins alleviate physical constraints by probably coevolving. However, it would be challenging to scan the whole genome in case of influenza A virus that has the segmented genome. Two types of influenza A virus that parasitize one host cell,

they can exchange their segments during their genomes assembly after their replication, the phenomena called reassortment [33]. As our approach consists of analyzing how AAs interact on phylogenetic tree, the genome scanning analysis will require the intermediate step to check whether all RNA genes belong to the same influenza A virus types. Westgeest and colleagues have conducted genome wide analysis of reassortment [123] in influenza A virus. They used GiRaF [124] to detect reassortment events between all segments. Hereby, this step can be realized. Detected reassortment events would help to reconstructed the viral genome of one influenza subtype such as H3N2 by aligning all RNA segmented genes. The detected AAs between two proteins can be visualized on 3D structures of the protein complex using EVcomplex method(<http://evfold.org>).

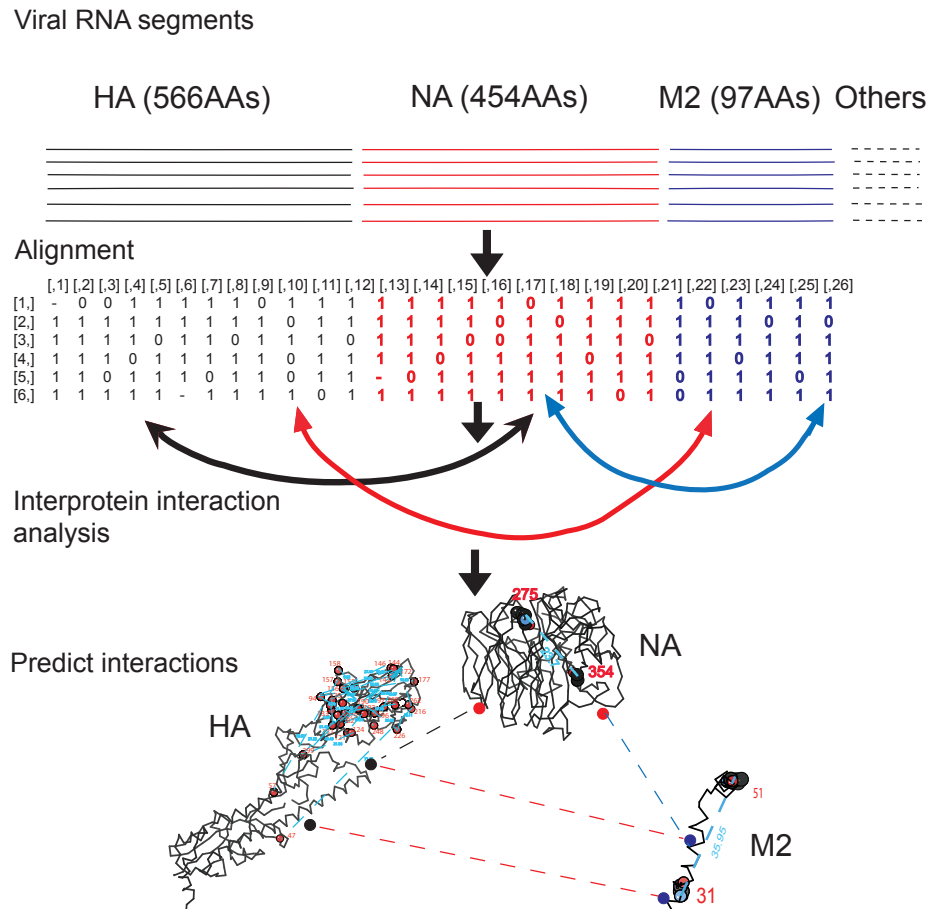


Figure 4.1: **Hypothetical model to detect epistatic amino acids within and among RNA segments within the influenza A virus.** The viral RNA segments are aligned together and then our approach detects the interacting AAs. Coevolving AAs are mapped on three dimensional protein structures.

Bibliography

- [1] Nelson MI, Holmes EC. The evolution of epidemic influenza. *Nature reviews genetics*. 2007;8(3):196–205.
- [2] Taubenberger JK, Morens DM. 1918 Influenza: the mother of all pandemics. *Rev Biomed*. 2006;17:69–79.
- [3] Hanafi S, Hayatshahi A, Torkamandi H, Javadi MR, et al. Evaluation of treatment with Oseltamivir during the 2009 H1N1 (swine flu) pandemic: the problem of incomplete clinical information. *Health Information Management Journal*. 2012;41(1):31.
- [4] Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, et al. The influenza virus resource at the National Center for Biotechnology Information. *Journal of virology*. 2008;82(2):596–601.
- [5] Gong LI, Suchard MA, Bloom JD. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife*. 2013;2.
- [6] Tria F, Pompei S, Loreto V. Dynamically correlated mutations drive human Influenza A evolution. *Scientific reports*. 2013;3.
- [7] Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. Evolution and ecology of influenza A viruses. *Microbiological reviews*. 1992;56(1):152–179.

- [8] Potter CW. A history of influenza. *Journal of applied microbiology*. 2001;91(4):572–579.
- [9] Chen LM, Rivaller P, Hossain J, Carney P, Balish A, Perry I, et al. Receptor specificity of subtype H1 influenza A viruses isolated from swine and humans in the United States. *Virology*. 2011;412(2):401–410.
- [10] Lamb RA, Krug R, Knipe D. *Fields virology*. Fields Virology. 2001;1.
- [11] Hatta M, Kawaoka Y. The NB protein of influenza B virus is not necessary for virus replication in vitro. *Journal of virology*. 2003;77(10):6050–6054.
- [12] Matsuzaki Y, Sato K, Sugawara K, Takashita E, Muraki Y, Morishita T, et al. Isolation of an influenza C virus introduced into Japan by a traveler from Malaysia. *Journal of clinical microbiology*. 2005;43(2):993–995.
- [13] Ambrose CS, Levin MJ. The rationale for quadrivalent influenza vaccines. *Human vaccines & immunotherapeutics*. 2012;8(1):81–88.
- [14] Biere B, Bauer B, Schweiger B. Differentiation of influenza B virus lineages Yamagata and Victoria by real-time PCR. *Journal of clinical microbiology*. 2010;48(4):1425–1427.
- [15] Cheung TK, Poon LL. Biology of influenza a virus. *Annals of the New York Academy of Sciences*. 2007;1102(1):1–25.
- [16] Bogner P, Capua I, Lipman DJ, Cox NJ, et al. A global initiative on sharing avian flu data. *Nature*. 2006;442(7106):981–981.
- [17] Squires RB, Noronha J, Hunt V, García-Sastre A, Macken C, Baumgarth N, et al. Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza and other respiratory viruses*. 2012;6(6):404–416.

- [18] Krug RM, Fodor E. The virus genome and its replication. Textbook of Influenza, 2nd Edition. 2013;p. 57–66.
- [19] Perales B, Ortín J. The influenza A virus PB2 polymerase subunit is required for the replication of viral RNA. *Journal of virology*. 1997;71(2):1381–1385.
- [20] Maier HJ, Kashiwagi T, Hara K, Brownlee GG. Differential role of the influenza A virus polymerase PA subunit for vRNA and cRNA promoter binding. *Virology*. 2008;370(1):194–204.
- [21] Tong S, Zhu X, Li Y, Shi M, Zhang J, Bourgeois M, et al. New world bats harbor diverse influenza A viruses. *PLoS Pathog*. 2013;9(10):e1003657.
- [22] Turrell L, Lyall JW, Tiley LS, Fodor E, Vreede FT. The role and assembly mechanism of nucleoprotein in influenza A virus ribonucleoprotein complexes. *Nature communications*. 2013;4:1591.
- [23] Ye Q, Krug RM, Tao YJ. The mechanism by which influenza A virus nucleoprotein forms oligomers and binds RNA. *Nature*. 2006;444(7122):1078–1082.
- [24] Shtyrya Y, Mochalova L, Bovin N. Influenza virus neuraminidase: structure and function. *Acta naturae*. 2009;1(2):26.
- [25] Hale BG, Randall RE, Ortín J, Jackson D. The multifunctional NS1 protein of influenza A viruses. *Journal of General Virology*. 2008;89(10):2359–2376.
- [26] Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, et al. Mapping the antigenic and genetic evolution of influenza virus. *Science*. 2004;305(5682):371–376.
- [27] Koel BF, Burke DF, Bestebroer TM, van der Vliet S, Zondag GC, Vervaet G, et al. Substitutions Near the Receptor Binding Site Determine Major Antigenic Change During Influenza Virus Evolution. *Science*. 2013;342(6161):976–979.

- [28] Dai JP, Zhao XF, Zeng J, Wan QY, Yang JC, Li WZ, et al. Drug Screening for Autophagy Inhibitors Based on the Dissociation of Beclin1-Bcl2 Complex Using BiFC Technique and Mechanism of Eugenol on Anti-Influenza A Virus Activity. *PloS one*. 2013;8(4):e61026.
- [29] Lamb RA, Choppin PW. The gene structure and replication of influenza virus. *Annual review of biochemistry*. 1983;52(1):467–506.
- [30] Boni MF, Zhou Y, Taubenberger JK, Holmes EC. Homologous recombination is very rare or absent in human influenza A virus. *Journal of virology*. 2008;82(10):4807–4811.
- [31] Boni MF, de Jong MD, van Doorn HR, Holmes EC. Guidelines for identifying homologous recombination events in influenza A virus. *PLoS One*. 2010;.
- [32] Fodor E, Mingay LJ, Crow M, Deng T, Brownlee GG. A single amino acid mutation in the PA subunit of the influenza virus RNA polymerase promotes the generation of defective interfering RNAs. *Journal of virology*. 2003;77(8):5017–5020.
- [33] Steel J, Lowen AC. Influenza A virus reassortment. In: *Influenza Pathogenesis and Control-Volume I*. Springer; 2014. p. 377–401.
- [34] Leung PC, Ooi EE. SARS war: Combating the disease. World Scientific; 2003.
- [35] Zhu L, Li Y, Li S, Li H, Qiu Z, Lee C, et al. Inhibition of influenza A virus (H1N1) fusion by benzenesulfonamide derivatives targeting viral hemagglutinin. *PLoS One*. 2011;6(12):e29120.
- [36] Boivin S, Cusack S, Ruigrok RW, Hart DJ. Influenza A virus polymerase: structural insights into replication and host adaptation mechanisms. *Journal of Biological Chemistry*. 2010;285(37):28411–28417.

- [37] Plotkin JB, Dushoff J, Levin SA. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proceedings of the National Academy of Sciences*. 2002;99(9):6263–6268.
- [38] Fitch WM, Bush RM, Bender CA, Cox NJ. Long term trends in the evolution of H (3) HA1 human influenza type A. *Proceedings of the National Academy of Sciences*. 1997;94(15):7712–7718.
- [39] Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB. Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS genetics*. 2011;7(2):e1001301.
- [40] Shih ACC, Hsiao TC, Ho MS, Li WH. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proceedings of the National Academy of Sciences*. 2007;104(15):6283–6288.
- [41] Lin JH, Chiu SC, Lin YC, Cheng JC, Wu HS, Salemi M, et al. Exploring the molecular epidemiology and evolutionary dynamics of influenza A virus in Taiwan. *PloS one*. 2013;8(4).
- [42] Bateson W, Mendel G. *Mendel's principles of heredity*. 448. Cambridge [Eng.]University Press.; 1909.
- [43] Fisher RA. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the royal society of Edinburgh*. 1919;52(02):399–433.
- [44] Lehner B. Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*. 2011;27(8):323–331.
- [45] Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*. 2008;9(11):855–867.

- [46] Gregersen JW, Kranc KR, Ke X, Svendsen P, Madsen LS, Thomsen AR, et al. Functional epistasis on a common MHC haplotype associated with multiple sclerosis. *Nature*. 2006;443(7111):574–577.
- [47] Buckholtz J, Sust S, Tan H, Mattay V, Straub R, Meyer-Lindenberg A, et al. fMRI evidence for functional epistasis between COMT and RGS4. *Molecular psychiatry*. 2007;12(10):893–895.
- [48] Weinreich DM, Delaney NF, DePristo MA, Hartl DL. Darwinian evolution can follow only very few mutational paths to fitter proteins. *science*. 2006;312(5770):111–114.
- [49] Wang X, Minasov G, Shoichet BK. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *Journal of molecular biology*. 2002;320(1):85–95.
- [50] Carlborg Ö, Haley CS. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*. 2004;5(8):618–625.
- [51] Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays*. 2005;27(6):637–646.
- [52] Gong LI, Bloom JD. Epistatically interacting substitutions are enriched during adaptive protein evolution. *PLoS Genet*. 2014;10(5):e1004328.
- [53] Byrd-Leotis L, Galloway SE, Agbogu E, Steinhauer DA. Influenza Hemagglutinin (HA) Stem Region Mutations That Stabilize or Destabilize the Structure of Multiple HA Subtypes. *Journal of virology*. 2015;89(8):4504–4516.
- [54] Lunzer M, Golding GB, Dean AM. Pervasive cryptic epistasis in molecular evolution. *PLoS Genet*. 2010;6(10):e1001162.

- [55] Myers JL, Wetzel KS, Linderman SL, Li Y, Sullivan CB, Hensley SE. Compensatory hemagglutinin mutations alter antigenic properties of influenza viruses. *Journal of virology*. 2013;87(20):11168–11172.
- [56] Shapiro B, Rambaut A, Pybus OG, Holmes EC. A phylogenetic method for detecting positive epistasis in gene sequences and its application to RNA virus evolution. *Molecular biology and evolution*. 2006;23(9):1724–1730.
- [57] Pereira F, Amorim A. Evolution: Viruses. *Brenners Encyclopedia of Genetics* Eds MALOY S, HUGHES K New York: Elsevier p. 2013;566.
- [58] Elena SF, Sanjuán R. Adaptive value of high mutation rates of RNA viruses: separating causes from consequences. *Journal of virology*. 2005;79(18):11555–11558.
- [59] Bhattacharjee A, Mallik S, Kundu S. Compensatory Mutations Occur Within the Electrostatic Interaction Range of Deleterious Mutations in Protein Structure. *Journal of molecular evolution*. 2015;80(1):10–12.
- [60] Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*. 2002;11(20):2463–2468.
- [61] Nasrallah CA, Huelsenbeck JP. A phylogenetic model for the detection of epistatic interactions. *Molecular biology and evolution*. 2013;30(9):2197–2208.
- [62] Chen Y, Carlini DB, Baines JF, Parsch J, Braverman JM, Tanda S, et al. RNA secondary structure and compensatory evolution. *Genes & genetic systems*. 1999;74(6):271–286.
- [63] Mohrig JR, Moerke KA, Cloutier DL, Lane BD, et al. Importance of historical contingency in the stereochemistry of hydratase-dehydratase enzymes. *Science*. 1995;269(5223):527.

- [64] Gould SJ. Wonderful life: the Burgess Shale and the nature of history. WW Norton & Company; 1990.
- [65] Worth CL, Gong S, Blundell TL. Structural and functional constraints in the evolution of protein families. *Nature Reviews Molecular Cell Biology*. 2009;10(10):709–720.
- [66] Blount ZD, Borland CZ, Lenski RE. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proceedings of the National Academy of Sciences*. 2008;105(23):7899–7906.
- [67] Harms MJ, Thornton JW. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature*. 2014;512(7513):203–207.
- [68] Dunham EJ, Dugan VG, Kaser EK, Perkins SE, Brown IH, Holmes EC, et al. Different evolutionary trajectories of European avian-like and classical swine H1N1 influenza A viruses. *Journal of virology*. 2009;83(11):5485–5494.
- [69] Pagel M, Meade A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *The American Naturalist*. 2006;167(6):808–825.
- [70] Schöniger M, Von Haeseler A. A stochastic model for the evolution of autocorrelated DNA sequences. *Molecular phylogenetics and evolution*. 1994;3(3):240–247.
- [71] Rzhetsky A. Estimating substitution rates in ribosomal RNA genes. *Genetics*. 1995;141(2):771–783.
- [72] Pollock DD, Taylor WR, Goldman N. Coevolving protein residues: maximum likelihood identification and relationship to structure. *Journal of molecular biology*. 1999;287(1):187–198.

- [73] Turner TF. Evolutionary Genetics: Concepts and Case Studies. *BioScience*. 2007;57(4):375–376.
- [74] Poon AF, Lewis FI, Frost SD, Pond SLK. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. *Bioinformatics*. 2008;24(17):1949–1950.
- [75] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann; 2014.
- [76] Pagel M. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London B: Biological Sciences*. 1994;255(1342):37–45.
- [77] Simonsen L, Viboud C, Grenfell BT, Dushoff J, Jennings L, Smit M, et al. The genesis and spread of reassortment human influenza A/H3N2 viruses conferring adamantane resistance. *Molecular biology and evolution*. 2007;24(8):1811–1820.
- [78] Holsinger LJ, Nichani D, Pinto LH, Lamb RA. Influenza A virus M2 ion channel protein: a structure-function analysis. *Journal of virology*. 1994;68(3):1551–1563.
- [79] Xiao N, Xu Q, Cao D. *protr: Protein Sequence Descriptor Calculation and Similarity Computation with R*. R package version 02-1, URL [http://CRAN.R-project.org/package= protr](http://CRAN.R-project.org/package=protr). 2014;.
- [80] Yang Z. Computational molecular evolution. vol. 21. Oxford University Press Oxford; 2006.
- [81] Analytics R, Weston S. *foreach: Foreach looping construct for R*; 2013.

- [82] Rodgers DP. Improvements in multiprocessor system design. In: ACM SIGARCH Computer Architecture News. vol. 13. IEEE Computer Society Press; 1985. p. 225–231.
- [83] Amdahl GM. Validity of the single processor approach to achieving large scale computing capabilities. In: Proceedings of the April 18-20, 1967, spring joint computer conference. ACM; 1967. p. 483–485.
- [84] Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. *Behavioural brain research*. 2001;125(1):279–284.
- [85] Team RDC, et al.. R: A language and environment for statistical computing. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013. url: <http://www.R-project.org>; 2005.
- [86] Guex N, Peitsch MC. SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling. *electrophoresis*. 1997;18(15):2714–2723.
- [87] Chen VB, Davis IW, Richardson DC. KING (Kinemage, Next Generation): a versatile interactive molecular and scientific visualization program. *Protein Science*. 2009;18(11):2403–2409.
- [88] Duan S, Govorkova EA, Bahl J, Zaraket H, Baranovich T, Seiler P, et al. Epistatic interactions between neuraminidase mutations facilitated the emergence of the oseltamivir-resistant H1N1 influenza viruses. *Nature communications*. 2014;5.
- [89] Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PloS one*. 2010;5(3):e9490.
- [90] Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*. 2004;20(2):289–290.

- [91] Drummond AJ, Ho SY, Phillips MJ, Rambaut A, et al. Relaxed phylogenetics and dating with confidence. *PLoS biology*. 2006;4(5):699.
- [92] Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic biology*. 2006;55(4):539–552.
- [93] Uversky VN, Fink A. Protein Misfolding, Aggregation and Conformational Diseases: Part B: Molecular Mechanisms of Conformational Diseases. vol. 6. Springer Science & Business Media; 2007.
- [94] Alberts B, Bray D, Hopkin K, Johnson A, Lewis J, Raff M, et al. Essential cell biology. Garland Science; 2013.
- [95] Aylor DL, Zeng ZB. From classical genetics to quantitative genetics to systems biology: modeling epistasis. *PLoS Genet*. 2008;4(3):e1000029.
- [96] Atilgan AR, Akan P, Baysal C. Small-world communication of residues and significance for protein dynamics. *Biophysical journal*. 2004;86(1):85–91.
- [97] Wang J, Wu Y, Ma C, Fiorin G, Wang J, Pinto LH, et al. Structure and inhibition of the drug-resistant S31N mutant of the M2 ion channel of influenza A virus. *Proceedings of the National Academy of Sciences*. 2013;110(4):1315–1320.
- [98] Garcia V, Aris-Brosou S. Comparative dynamics and distribution of influenza drug resistance acquisition to protein m2 and neuraminidase inhibitors. *Molecular biology and evolution*. 2014;31(2):355–363.
- [99] Stewart SM, Pekosz A. Mutations in the membrane-proximal region of the influenza A virus M2 protein cytoplasmic tail have modest effects on virus replication. *Journal of virology*. 2011;85(23):12179–12187.

- [100] Kimura M. The role of compensatory neutral mutations in molecular evolution. *Journal of Genetics*. 1985;64(1):7–19.
- [101] Thomas VL, McReynolds AC, Shoichet BK. Structural bases for stability–function tradeoffs in antibiotic resistance. *Journal of molecular biology*. 2010;396(1):47–59.
- [102] Mitraki A, Fane B, et al. Global suppression of protein folding defects and inclusion body formation. *Science*. 1991;253(5015):54.
- [103] Newcomb RD, Campbell P, Ollis D, Cheah E, Russell R, Oakeshott J. A single amino acid substitution converts a carboxylesterase to an organophosphorus hydrolase and confers insecticide resistance on a blowfly. *Proceedings of the National Academy of Sciences*. 1997;94(14):7464–7468.
- [104] Sanjuán R, Elena SF. Epistasis correlates to genomic complexity. *Proceedings of the National Academy of Sciences*. 2006;103(39):14402–14405.
- [105] Tomatis PE, Fabiane SM, Simona F, Carloni P, Sutton BJ, Vila AJ. Adaptive protein evolution grants organismal fitness by improving catalysis and flexibility. *Proceedings of the National Academy of Sciences*. 2008;105(52):20605–20610.
- [106] Weinreich DM. *Predicting Molecular Evolutionary Trajectories in Principle and in Practice*. eLS. 2010;p. 1–9.
- [107] Sorrells TR, Booth LN, Tuch BB, Johnson AD. Intersecting transcription networks constrain gene regulatory evolution. *Nature*. 2015;523(7560):361–365.
- [108] Schoustra SE, Bataillon T, Gifford DR, Kassen R, et al. The properties of adaptive walks in evolving populations of fungus. *PLoS biology*. 2009;7(11):2474.
- [109] de Visser JAG, Krug J. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*. 2014;15:480–490.

- [110] Gavrillets S. Fitness landscapes and the origin of species (MPB-41). Princeton University Press Princeton, NJ; 2004.
- [111] Hartl DL. What can we learn from fitness landscapes? Current opinion in microbiology. 2014;21:51–57.
- [112] Sandie R, Aris-Brosou S. Predicting the emergence of H3N2 influenza viruses reveals contrasted modes of evolution of HA and NA antigens. Journal of molecular evolution. 2014;78(1):1–12.
- [113] Ferguson NM, Galvani AP, Bush RM. Ecological and immunological determinants of influenza evolution. Nature. 2003;422(6930):428–433.
- [114] Recker M, Pybus OG, Nee S, Gupta S. The generation of influenza outbreaks by a network of host immune responses against a limited set of antigenic types. Proceedings of the National Academy of Sciences. 2007;104(18):7711–7716.
- [115] Koelle K, Cobey S, Grenfell B, Pascual M. Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. Science. 2006;314(5807):1898–1903.
- [116] Kim K, Kim Y. Episodic nucleotide substitutions in seasonal influenza virus H3N2 can be explained by stochastic genealogical process without positive selection. Molecular biology and evolution. 2014;p. msu332.
- [117] Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. Protein evolution with dependence among codons due to tertiary structure. Molecular Biology and Evolution. 2003;20(10):1692–1704.
- [118] Rodrigue N, Lartillot N, Bryant D, Philippe H. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene. 2005;347(2):207–217.

- [119] Rodrigue N, Philippe H, Lartillot N. Assessing site-interdependent phylogenetic models of sequence evolution. *Molecular biology and evolution*. 2006;23(9):1762–1775.
- [120] Chen BJ, Leser GP, Jackson D, Lamb RA. The influenza virus M2 protein cytoplasmic tail interacts with the M1 protein and influences virus assembly at the site of virus budding. *Journal of virology*. 2008;82(20):10059–10070.
- [121] Thaa B, Herrmann A, Veit M. Intrinsic cytoskeleton-dependent clustering of influenza virus M2 protein with hemagglutinin assessed by FLIM-FRET. *Journal of virology*. 2010;84(23):12445–12449.
- [122] Mintaev RR, Alexeevski AV, Kordyukova LV. Co-evolution analysis to predict protein–protein interactions within influenza virus envelope. *Journal of bioinformatics and computational biology*. 2014;12(02):1441008.
- [123] Westgeest KB, Russell CA, Lin X, Spronken MI, Bestebroer TM, Bahl J, et al. Genomewide analysis of reassortment and evolution of human influenza A (H3N2) viruses circulating between 1968 and 2011. *Journal of virology*. 2014;88(5):2844–2857.
- [124] Nagarajan N, Kingsford C. GiRaF: robust, computational identification of influenza reassortments via graph mining. *Nucleic acids research*. 2010;p. gkq1232.
- [125] Rambaut A. FigTree v1. 4. Molecular evolution, phylogenetics and epidemiology Edinburgh, UK: University of Edinburgh, Institute of Evolutionary Biology. 2012;.
- [126] Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, et al. The immune epitope database 2.0. *Nucleic acids research*. 2010;38(suppl 1):D854–D862.

- [127] Nagai K. RNA—protein complexes. *Current opinion in structural biology*. 1996;6(1):53–61.
- [128] Charles W Fox JBW. *Evolutionary Genetics: Concepts and Case Studies*. 1-608. Oxford University Press; 2006.
- [129] Wolf JB, Brodie ED, Wade MJ. *Epistasis and the evolutionary process*. Oxford University Press; 2000.
- [130] RAARMK L. *Orthomyxoviridae: The Viruses and Their Replication*. Fields Virology Philadelphia, PA: Lippincott Williams & Wilkins. 2001;.
- [131] Webster RG, Monto AS, Braciale TJ, Lamb RA. *Textbook of influenza*. John Wiley & Sons; 2014.
- [132] Garcia V, Aris-Brosou S. Comparative dynamics and distribution of influenza drug resistance acquisition to protein M2 and neuraminidase inhibitors. *Molecular biology and evolution*. 2013;p. mst204.
- [133] ;.
- [134] Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution*. 2012;29(8):1969–1973.
- [135] Pielak RM, Chou JJ. Solution NMR structure of the V27A drug resistant mutant of influenza A M2 channel. *Biochemical and biophysical research communications*. 2010;401(1):58–63.
- [136] Belshe R, Smith MH, Hall C, Betts R, Hay A. Genetic basis of resistance to rimantadine emerging during treatment of influenza virus infection. *Journal of virology*. 1988;62(5):1508–1512.

- [137] Sugrue R, Hay A. Structural characteristics of the M2 protein of influenza A viruses: Evidence that it forms a tetrameric channel. *Virology*. 1991;180(2):617–624.
- [138] Bjellqvist B, Hughes GJ, Pasquali C, Paquet N, Ravier F, Sanchez JC, et al. The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis*. 1993;14(1):1023–1031.
- [139] Berx GW, van Hengel JW. *Cell Biology* (C003220);.
- [140] Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences*. 1996;93(23):13429–13429.
- [141] Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*. 2005;21(9):2104–2105.
- [142] Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*. 2009;25(9):1189–1191.
- [143] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*. 2004;32(5):1792–1797.
- [144] Furuse Y, Suzuki A, Oshitani H. Large-scale sequence analysis of M gene of influenza A viruses from different species: mechanisms for emergence and spread of amantadine resistance. *Antimicrobial agents and chemotherapy*. 2009;53(10):4457–4463.
- [145] Kryazhimskiy S, Bazykin GA, Plotkin J, Dushoff J. Directionality in the evolution of influenza A haemagglutinin. *Proceedings of the Royal Society B: Biological Sciences*. 2008;275(1650):2455–2464.

Chapter 5

Appendices

5.1 Supplementary Figures

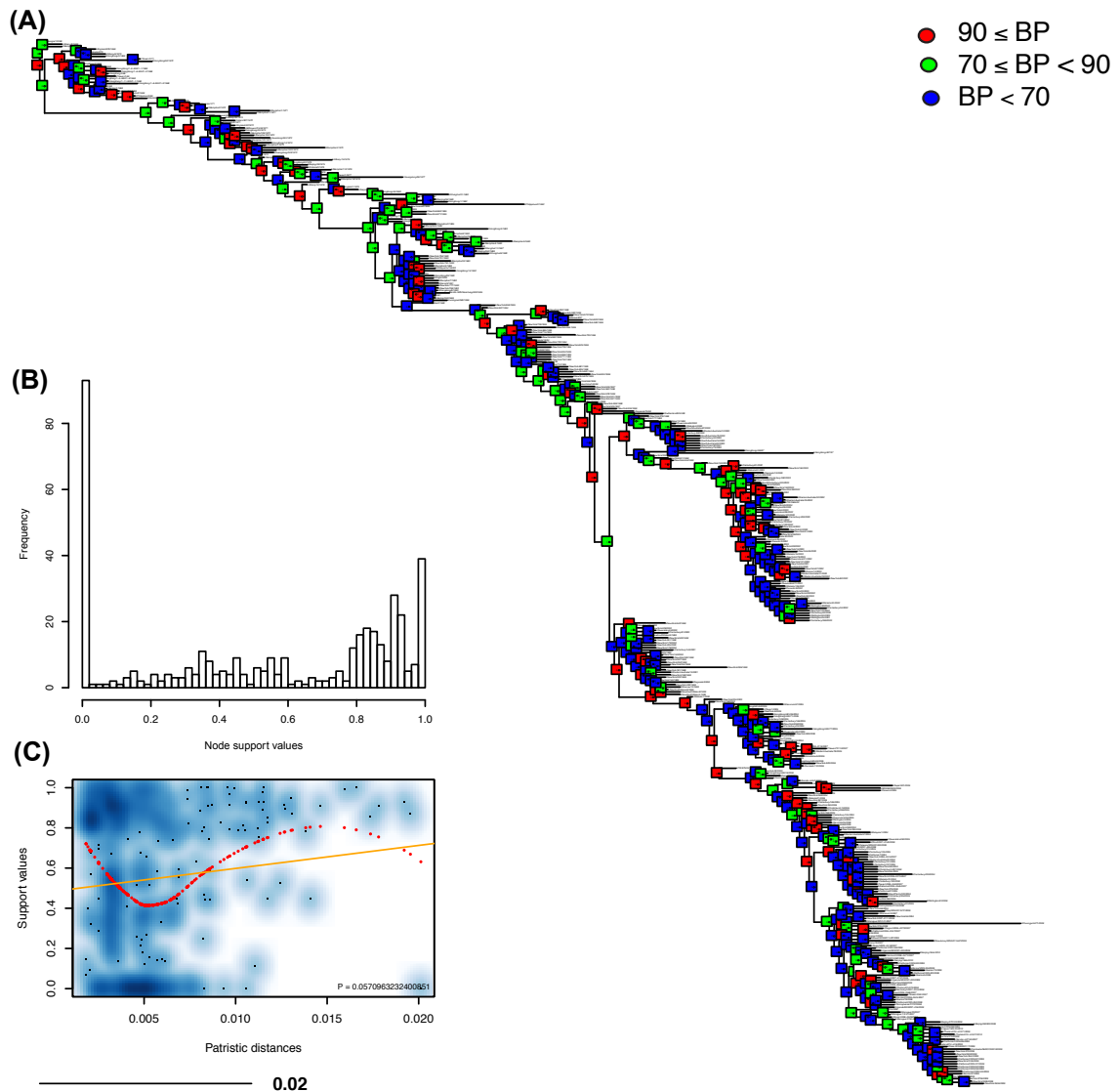


Figure 5.1: **Node support values for the phylogenetic analysis of the Gong13NP data set.** (A) The estimated maximum likelihood phylogenetic tree. Support values (SH-like aLRT) are color-coded as shown in the top right corner of the panel. (B) The distribution of node support values over the entire tree. (C) Relationship between support values and patristic distance for each pair of sequence. The significance of fitted linear regression (orange line) is shown in the bottom right corner of the panel. The loess (locally estimated scatterplot smoothing) is shown in red.

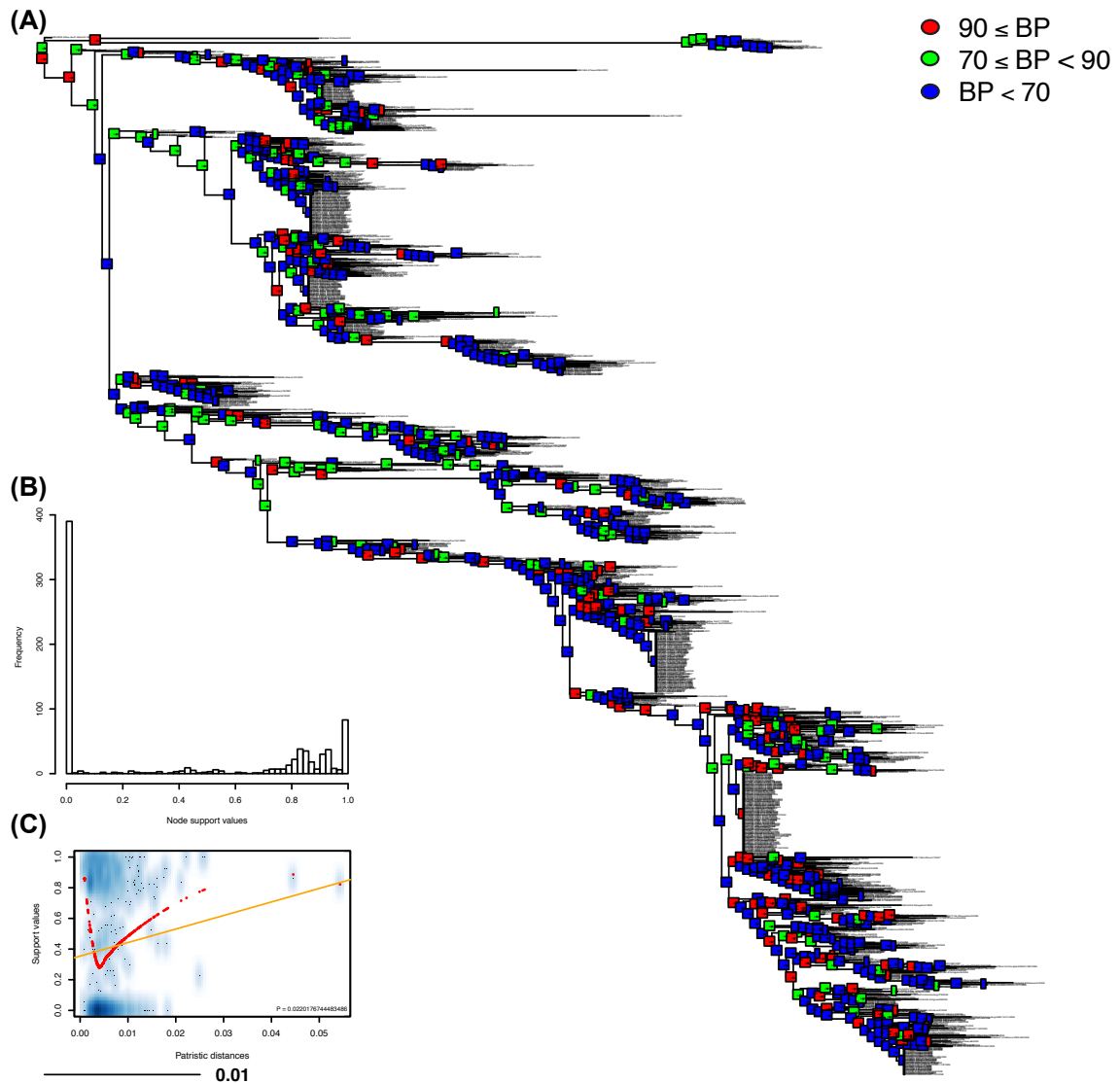


Figure 5.2: **Node support values for the phylogenetic analysis of the Gong13NP data set.** (A) The estimated maximum likelihood phylogenetic tree. Support values (SH-like aLRT) are color-coded as shown in the top right corner of the panel. (B) The distribution of node support values over the entire tree. (C) Relationship between support values and patristic distance for each pair of sequence. The significance of fitted linear regression (orange line) is shown in the bottom right corner of the panel. The loess (locally estimated scatterplot smoothing) is shown in red.

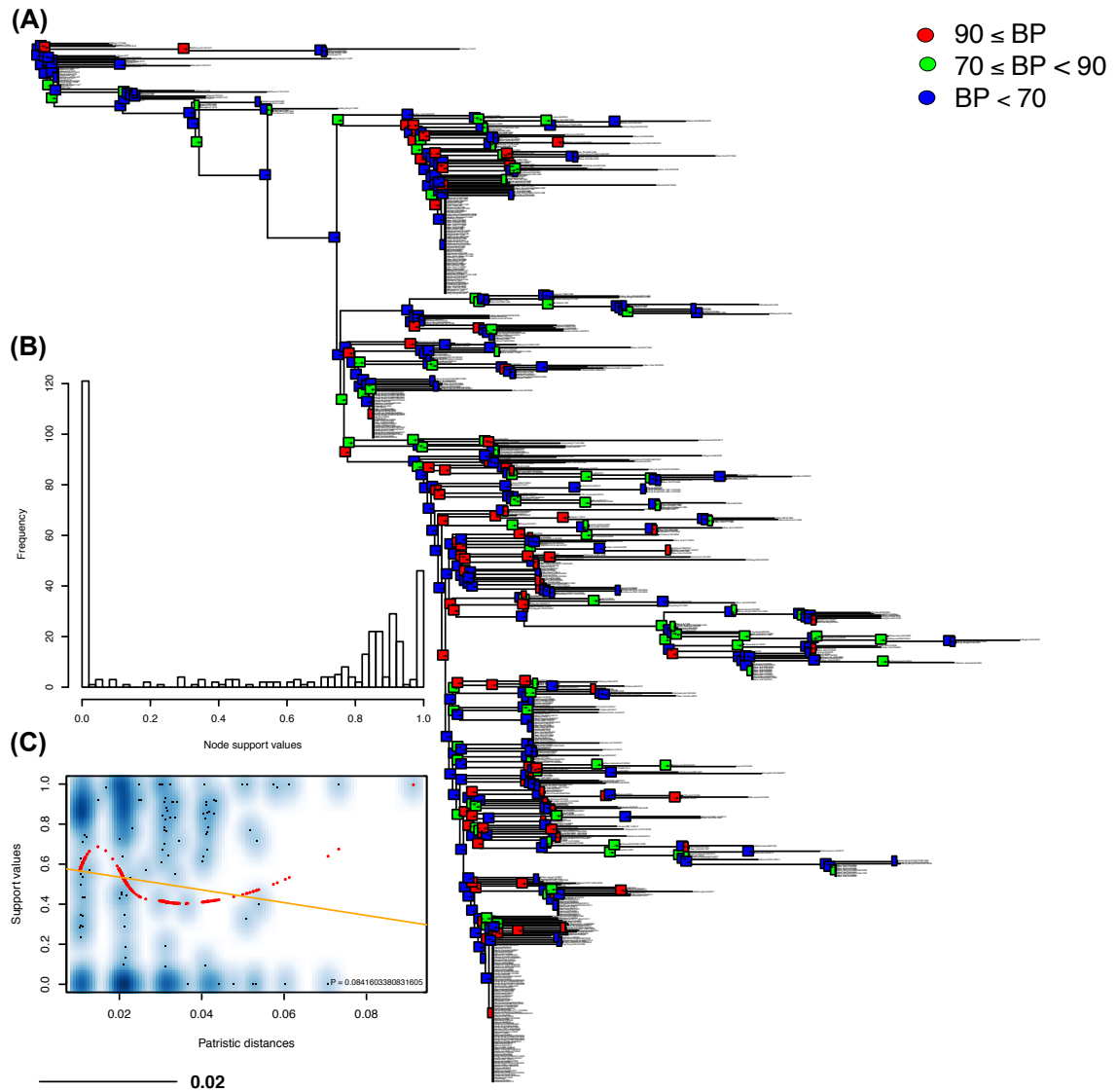


Figure 5.3: **Node support values for the phylogenetic analysis of the Gong13NP data set.** (A) The estimated maximum likelihood phylogenetic tree. Support values (SH-like aLRT) are color-coded as shown in the top right corner of the panel. (B) The distribution of node support values over the entire tree. (C) Relationship between support values and patristic distance for each pair of sequence. The significance of fitted linear regression (orange line) is shown in the bottom right corner of the panel. The loess (locally estimated scatterplot smoothing) is shown in red.

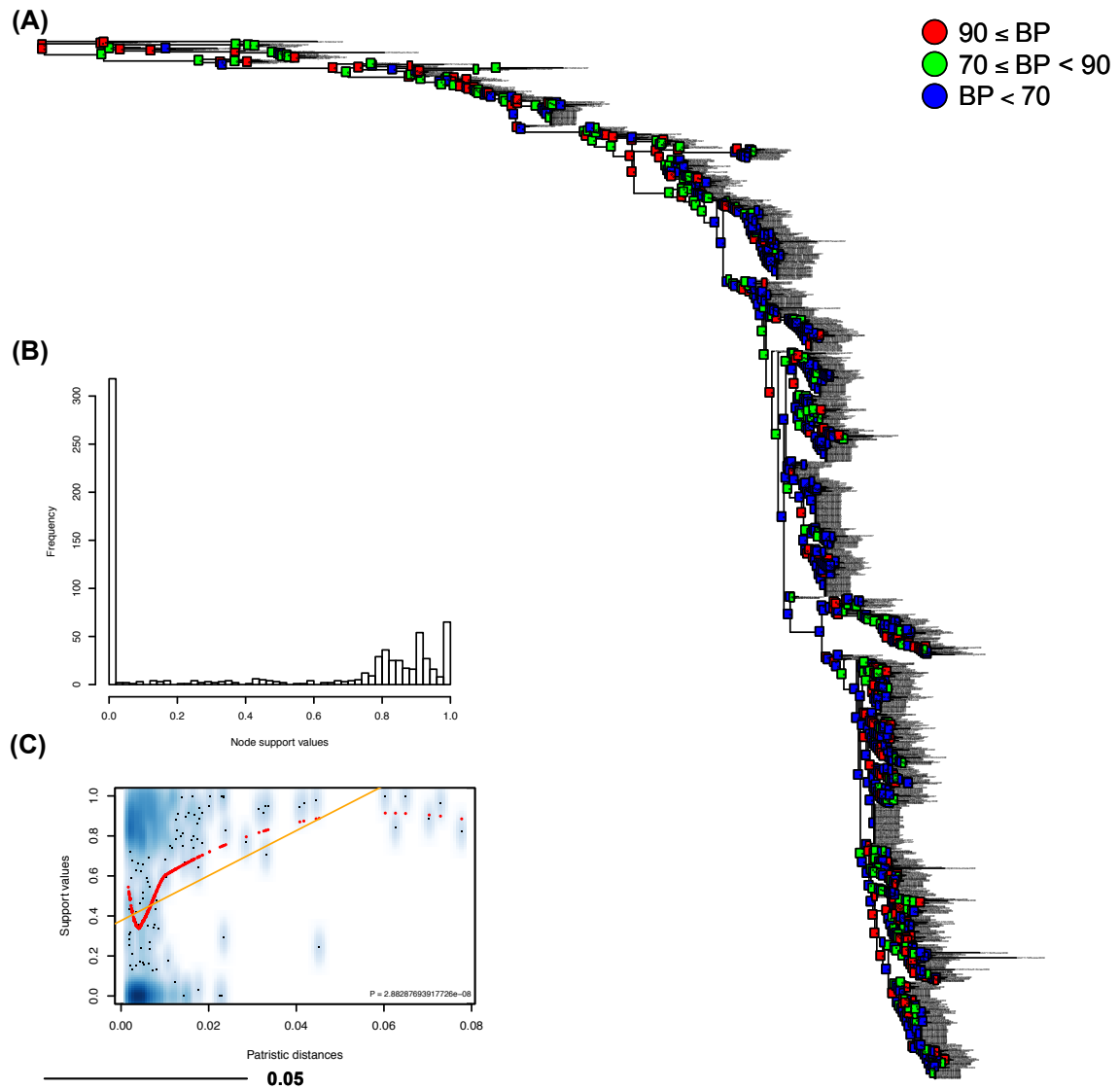


Figure 5.4: **Node support values for the phylogenetic analysis of the KDBP11 H1 data set.** (A) The estimated maximum likelihood phylogenetic tree. Support values (SH-like aLRT) are color-coded as shown in the top right corner of the panel. (B) The distribution of node support values over the entire tree. (C) Relationship between support values and patristic distance for each pair of sequence. The significance of fitted linear regression (orange line) is shown in the bottom right corner of the panel. The loess (locally estimated scatterplot smoothing) is shown in red.

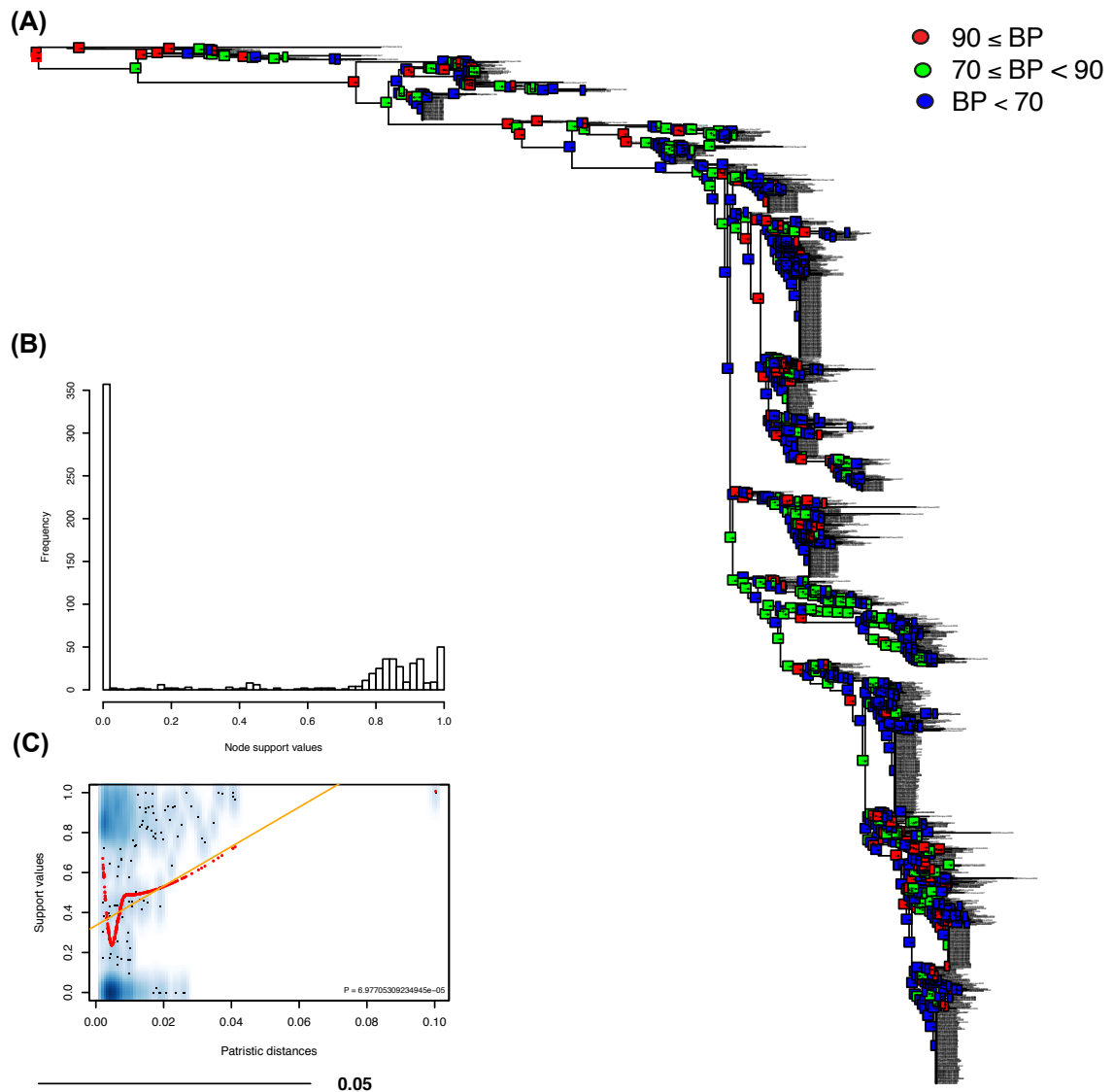


Figure 5.5: **Node support values for the phylogenetic analysis of the KDBP11 N1 data set.** (A) The estimated maximum likelihood phylogenetic tree. Support values (SH-like aLRT) are color-coded as shown in the top right corner of the panel. (B) The distribution of node support values over the entire tree. (C) Relationship between support values and patristic distance for each pair of sequence. The significance of fitted linear regression (orange line) is shown in the bottom right corner of the panel. The loess (locally estimated scatterplot smoothing) is shown in red.

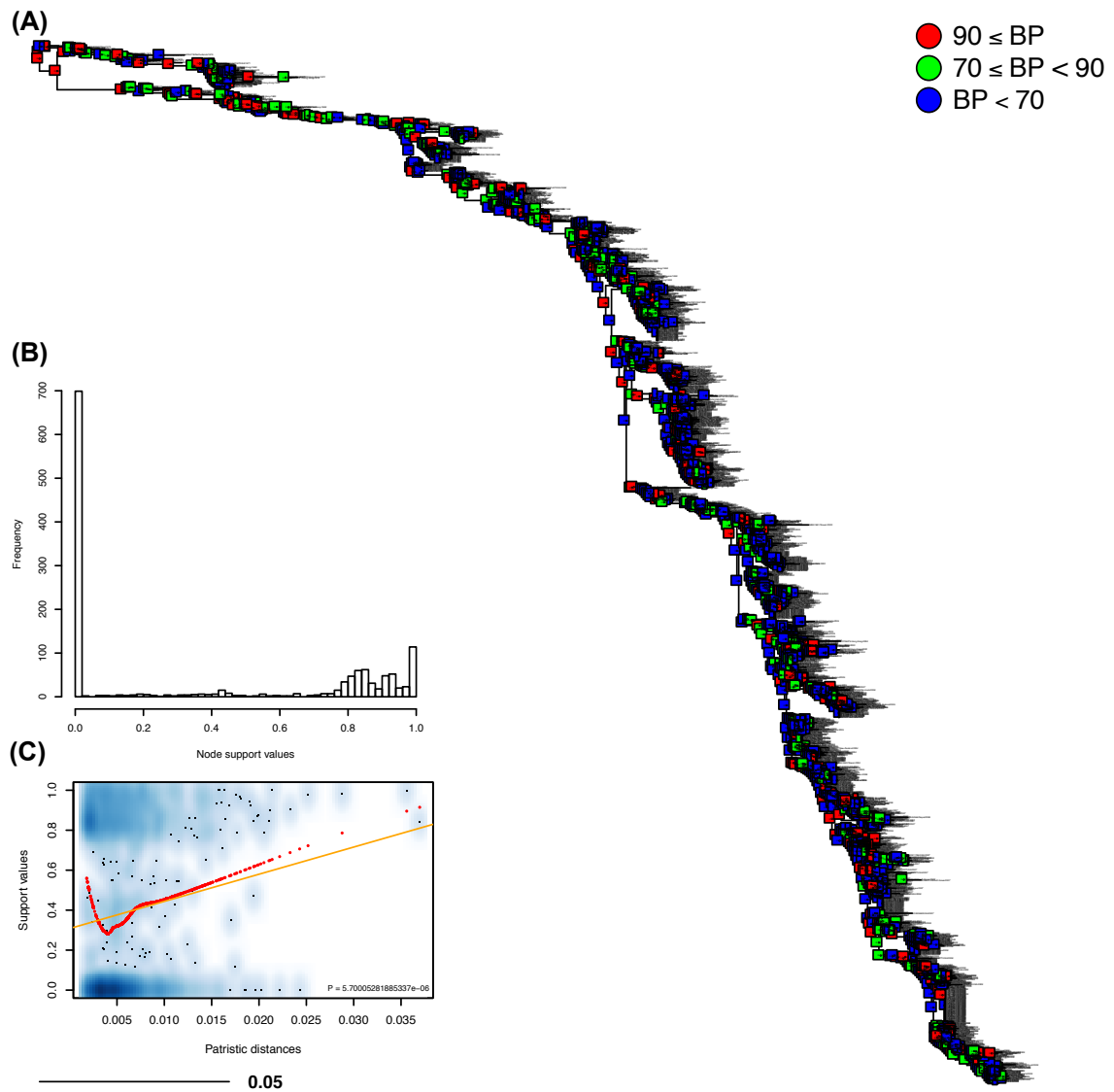


Figure 5.6: **Node support values for the phylogenetic analysis of the KDBP11 H3 data set.** (A) The estimated maximum likelihood phylogenetic tree. Support values (SH-like aLRT) are color-coded as shown in the top right corner of the panel. (B) The distribution of node support values over the entire tree. (C) Relationship between support values and patristic distance for each pair of sequence. The significance of fitted linear regression (orange line) is shown in the bottom right corner of the panel. The loess (locally estimated scatterplot smoothing) is shown in red.

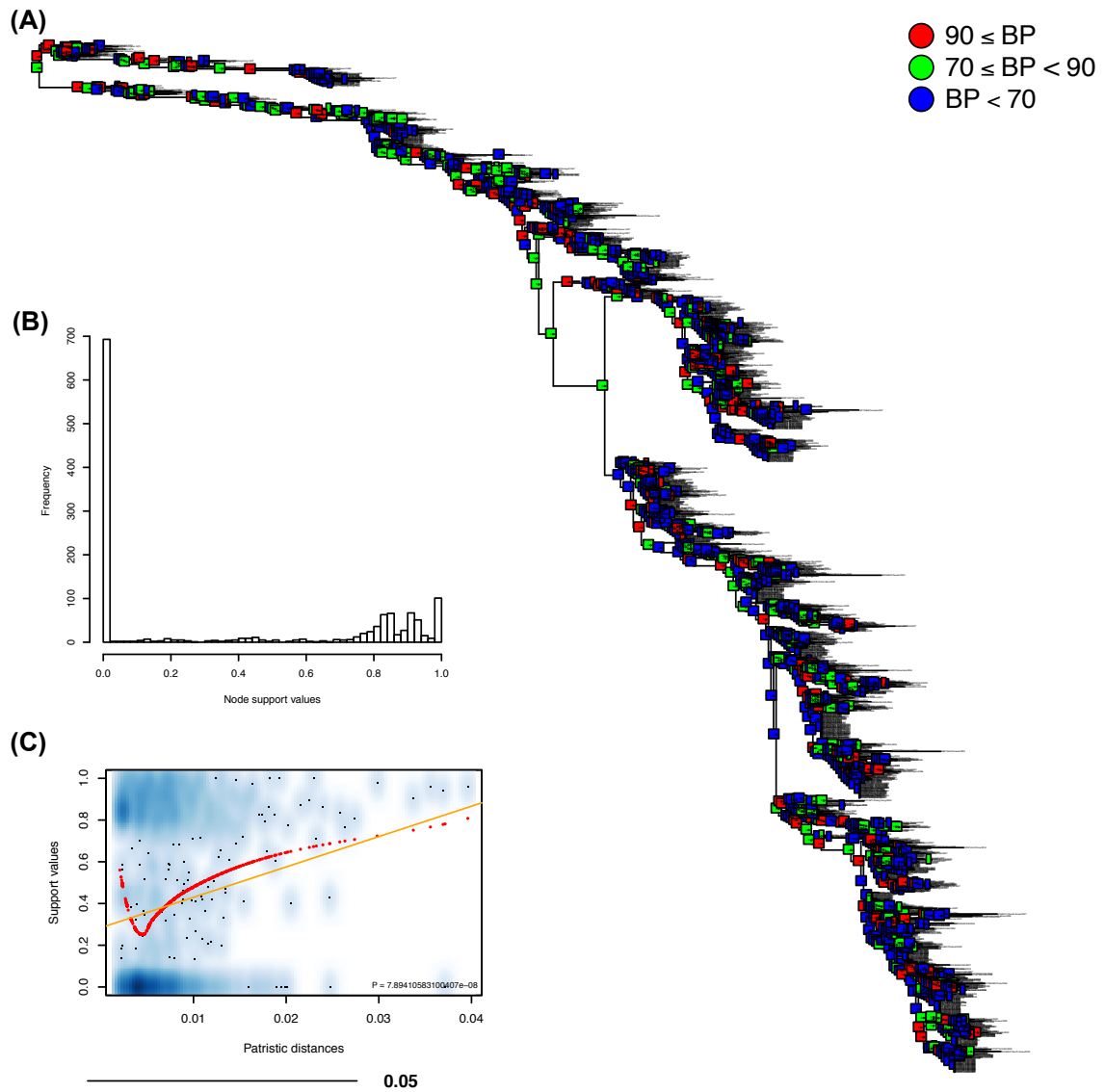


Figure 5.7: **Node support values for the phylogenetic analysis of the KDBP11 N2 data set.** (A) The estimated maximum likelihood phylogenetic tree. Support values (SH-like aLRT) are color-coded as shown in the top right corner of the panel. (B) The distribution of node support values over the entire tree. (C) Relationship between support values and patristic distance for each pair of sequence. The significance of fitted linear regression (orange line) is shown in the bottom right corner of the panel. The loess (locally estimated scatterplot smoothing) is shown in red.

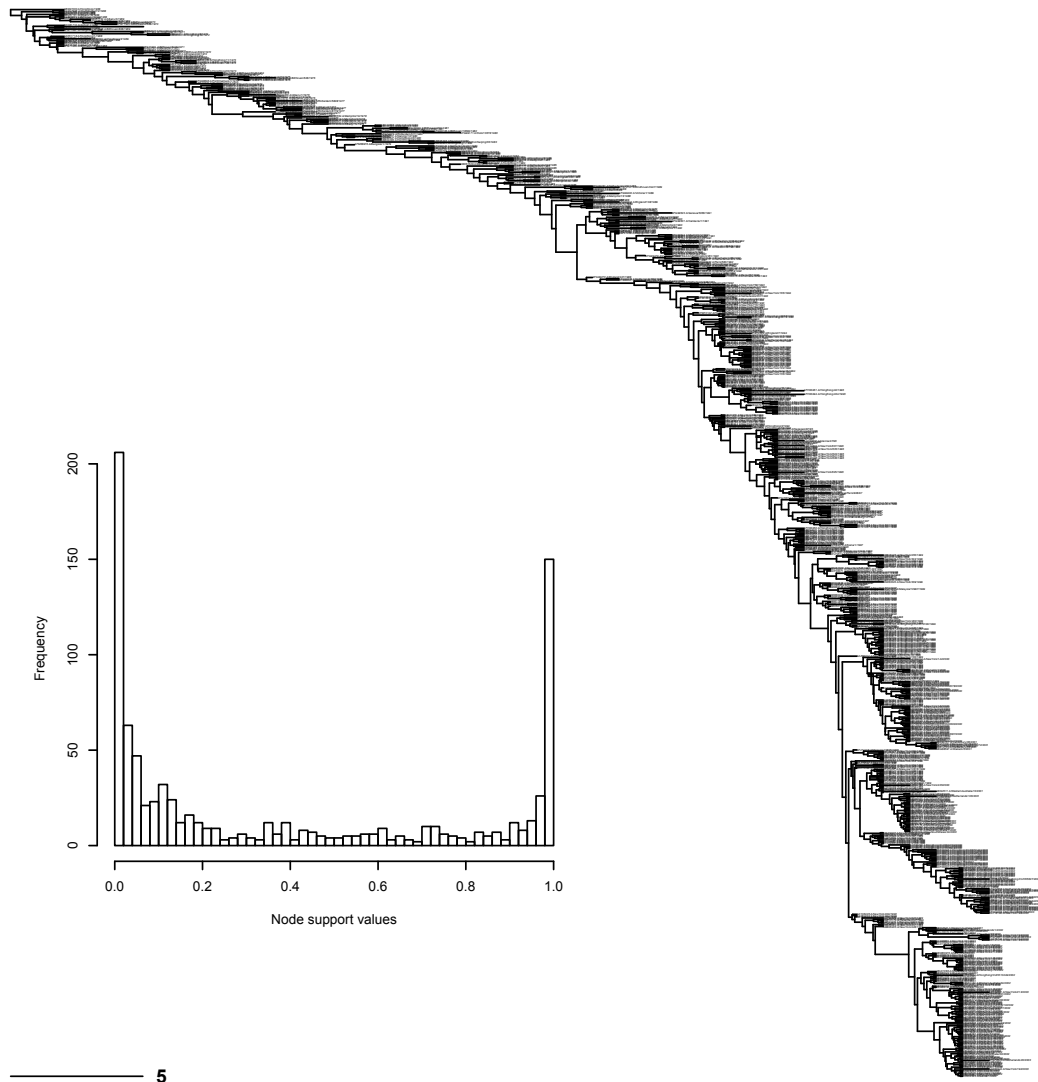


Figure 5.8: **Node support values for the phylogenetic analysis of the Gong13NP data set.** (A) The estimated maximum likelihood phylogenetic tree. Support values (SH-like aLRT) are color-coded as shown in the top right corner of the panel. (B) The distribution of node support values over the entire tree.

5.2 Supplementary Tables

Table 5.1: Detailed results generated using the KDBP11 data set for HA in H1N1 viruses. Site-1 and site-2 show the AA positions along the alignment. The lkl terms show the log-likelihood values after optimization under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The P -values are computed using the standard χ^2 approximation. FDR-corrected P -values follow BH. Only pairs significant at the nominal 1% level (P -value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	site-2	lkl-indep	lkl-dep	test-stat	P -value	P -value (FDR)
71	73	-38.485092	-29.055355	18.859474	0.000837511178384709	0.984913145780418
12	73	-58.238767	-49.80791	16.861714	0.00205625270630705	0.99999998318065
73	74	-56.071941	-48.042412	16.059058	0.00294093184608624	0.99999998318065
11	73	-35.737211	-28.307707	14.859008	0.00500275725669519	0.99999998318065
41&45&54&84	65	-23.662771	-16.698284	13.928974	0.00752522577487336	0.99999998318065
73	88	-29.379655	-22.670017	13.419276	0.00939886101619647	0.99999998318065

Table 5.2: Detailed results generated using the KDBP11 data set for NA in H1N1 viruses. Site-1 and site-2 show the AA positions along the alignment. The lkl terms show the log-likelihood values after optimization under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The P -values are computed using the standard χ^2 approximation. FDR-corrected P -values follow BH. Only pairs significant at the nominal 1% level (P -value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	site-2	lkl-indep	lkl-dep	test-stat	P -value	P -value (FDR)
275	354	-151.206402	-122.116619	58.179566	7.00E-12	2.34E-07
461	466	-80.750412	-63.80729	33.886244	7.86E-07	0.0131356962707871
234	275	-140.234912	-128.581571	23.306682	0.00010995716148976	0.9999999999955
50	286	-56.127815	-45.461299	21.333032	0.000271976976207289	0.9999999999955
200	388	-45.296802	-35.134716	20.324172	0.000430930754599101	0.9999999999955
254	270	-64.855428	-54.755728	20.1994	0.000456107550876772	0.9999999999955
249	275	-155.496669	-145.664956	19.663426	0.000581886713368229	0.9999999999955
336	427	-55.900157	-46.137891	19.524532	0.000619734205504541	0.9999999999955
70	466	-51.543637	-41.860559	19.366156	0.000665869398006613	0.9999999999955
435	461	-85.07601	-75.485225	19.18157	0.000723940959468261	0.9999999999955
136	270	-65.039998	-55.451621	19.176754	0.000725521312396338	0.9999999999955
23	274	-71.48771	-61.947853	19.079714	0.000758101690989998	0.9999999999955
34	336	-41.971628	-32.49925	18.944756	0.00080583070915563	0.9999999999955
200	336	-32.754213	-23.344507	18.819412	0.000852815473945556	0.9999999999955
52	466	-49.232332	-40.012794	18.439076	0.00101259734365267	0.9999999999955
105	466	-41.824315	-32.72081	18.20701	0.00112426857912984	0.9999999999955
263	336	-61.402675	-52.397998	18.009354	0.00122891413529491	0.9999999999955
357	383	-36.893395	-27.896693	17.993404	0.00123776646505713	0.9999999999955
200	274	-26.322141	-17.541908	17.560466	0.00150363511573715	0.9999999999955
336	388	-58.414566	-49.684818	17.459496	0.00157333138969318	0.9999999999955
45	275	-140.885071	-132.221152	17.327838	0.00166901818191334	0.9999999999955
220	336	-55.603238	-46.94017	17.326136	0.00167029202285462	0.9999999999955
200	466	-45.927375	-37.268853	17.317044	0.00167711308491059	0.9999999999955
270	396	-86.099665	-77.448379	17.302572	0.00168802709857463	0.9999999999955
23	452	-111.291565	-102.649438	17.284254	0.00170194209259467	0.9999999999955
248	466	-118.425817	-109.808757	17.23412	0.00174060696636791	0.9999999999955
383	466	-46.488327	-37.899067	17.17852	0.00178450131560515	0.9999999999955
93	275	-131.342994	-122.778536	17.128916	0.00182458253941953	0.9999999999955
275	396	-150.982066	-142.567216	16.8297	0.00208588559096545	0.9999999999955
23	173	-72.887506	-64.473463	16.828086	0.00208739064264929	0.9999999999955
334	466	-53.692311	-45.366264	16.652094	0.00225809043765124	0.9999999999955
143	467	-22.125427	-13.864434	16.521986	0.00239306161668063	0.9999999999955
382	466	-40.933061	-32.681373	16.503376	0.00241300606556916	0.9999999999955
332	467	-13.418112	-5.178711	16.478802	0.00243959327709231	0.9999999999955
173	275	-121.008318	-112.792149	16.432338	0.00249065498094114	0.9999999999955
52	461	-57.912071	-49.789049	16.246044	0.00270617009931429	0.9999999999955
390	467	-27.201104	-19.132944	16.13632	0.00284159256104222	0.9999999999955
454	466	-58.950808	-50.92717	16.047276	0.00295637796235626	0.9999999999955
34	287	-36.310054	-28.313782	15.992544	0.00302918482382142	0.9999999999955

Continued on next page

Table 2 -Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
86	467	-27.507537	-19.537999	15.939076	0.00310201359905926	0.9999999999955
105	461	-50.425315	-42.465231	15.920168	0.00312817848327285	0.9999999999955
23	220	-87.652607	-79.760612	15.78399	0.00332314529606204	0.9999999999955
15	200	-40.968346	-33.088335	15.760022	0.00335867675914581	0.9999999999955
130	331	-68.45982	-60.657796	15.604048	0.00359920495061494	0.9999999999955
383	461	-55.168084	-47.384994	15.56618	0.0036601113151743	0.9999999999955
388	466	-71.514138	-63.817153	15.39397	0.00395012290198515	0.9999999999955
15	274	-47.6502	-39.963419	15.373562	0.00398595416950454	0.9999999999955
17&163	466	-39.08149	-31.412476	15.338028	0.00404910689811722	0.9999999999955
336	366&367	-27.081351	-19.422294	15.318114	0.00408492800602434	0.9999999999955
232	467	-32.514391	-24.881063	15.266656	0.00417893881066467	0.9999999999955
16	466	-58.934043	-51.335155	15.197776	0.00430811378552565	0.9999999999955
334	461	-62.37204	-54.783778	15.176524	0.00434875529997469	0.9999999999955
254	336	-49.567648	-41.996368	15.14256	0.0044144902055856	0.9999999999955
101	336	-35.215942	-27.696958	15.037968	0.00462311335217314	0.9999999999955
57	466	-66.016384	-58.50624	15.020288	0.0046593229258225	0.9999999999955
82	130	-71.473515	-63.99239	14.96225	0.00478015697097867	0.9999999999955
262	274	-49.90798	-42.455595	14.90477	0.00490286108222038	0.9999999999955
307	466	-69.509585	-62.078536	14.862098	0.00499594947834114	0.9999999999955
23	275	-158.091078	-150.690456	14.801244	0.00513171047589567	0.9999999999955
275	352	-137.045128	-129.652756	14.784744	0.00516914075797925	0.9999999999955
14	336	-35.207757	-27.822389	14.770736	0.00520112818749763	0.9999999999955
200	250	-31.51041	-24.131537	14.757746	0.0052309646028319	0.9999999999955
84	287	-51.4864	-44.112225	14.74835	0.00525265069975567	0.9999999999955
95	396	-86.471666	-79.157217	14.628898	0.00553616002222657	0.9999999999955
106	466	-54.834282	-47.543352	14.58186	0.00565187543337498	0.9999999999955
70	461	-60.223376	-52.946299	14.554154	0.00572114045480998	0.9999999999955
15	388	-66.626424	-59.386456	14.479936	0.00591081509238112	0.9999999999955
34	386	-34.010318	-26.778058	14.46452	0.00595097969768998	0.9999999999955
211	274	-28.759868	-21.562656	14.394424	0.00613700829742769	0.9999999999955
270	434	-62.234339	-55.133788	14.201102	0.00668012274939944	0.9999999999955
336	418	-47.091265	-40.052959	14.076612	0.00705451363202247	0.9999999999955
275	418	-127.272988	-120.235474	14.075028	0.00705940740052502	0.9999999999955
15	262	-64.553026	-57.518177	14.069698	0.00707589887701154	0.9999999999955
15	250	-52.840018	-45.970497	13.739042	0.00817623254556632	0.9999999999955
427	466	-68.999729	-62.132703	13.734052	0.0081940589981907	0.9999999999955
52	105	-18.865615	-12.01964	13.69195	0.00834598796723063	0.9999999999955
23	336	-77.920897	-71.082516	13.676762	0.00840146912942918	0.9999999999955
382	461	-49.517328	-42.720275	13.594106	0.00870979383594128	0.9999999999955
173	274	-34.404951	-27.611158	13.587586	0.00873458058217291	0.9999999999955
220	466	-68.70281	-61.910932	13.583756	0.00874917305685707	0.9999999999955
79	201	-56.900371	-50.132123	13.536496	0.00893121045800649	0.9999999999955
27	200	-17.748473	-10.982961	13.531024	0.00895252549151837	0.9999999999955
95	275	-141.70149	-134.940735	13.52151	0.00898970381832009	0.9999999999955
95	270	-76.81909	-70.063706	13.510768	0.00903186238930109	0.9999999999955
287	466	-53.307035	-46.564777	13.484516	0.00913570767452088	0.9999999999955

Continued on next page

Table 2 -Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	<i>P</i> -value	<i>P</i> -value (FDR)
34	250	-40.728247	-33.988081	13.480332	0.00915236589278601	0.9999999999955
220	386	-47.628595	-40.892513	13.472164	0.00918497169373855	0.9999999999955
64	275	-111.303341	-104.594838	13.417006	0.0094081495237065	0.9999999999955
6	45	-54.820441	-48.117229	13.406424	0.00945156809500369	0.9999999999955
93	95	-66.832594	-60.144784	13.37562	0.00957907695899363	0.9999999999955
23	254	-81.617017	-74.937806	13.358422	0.00965099565284477	0.9999999999955
130	344	-88.20565	-81.54131	13.32868	0.00977661863990886	0.9999999999955
101	466	-48.315515	-41.66894	13.29315	0.0099287850007318	0.9999999999955
14	466	-48.307329	-41.663856	13.286946	0.00995559156861858	0.9999999999955

Table 5.3: Detailed results generated using the KDBP11 data set for HA in H3N2 viruses. Site-1 and site-2 show the AA positions along the alignment. The lkl terms show the log-likelihood values after optimization under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The *P*-values are computed using the standard χ^2 approximation. FDR-corrected *P*-values follow BH. Only pairs significant at the nominal 1% level (*P*-value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	site-2	lkl-indep	lkl-dep	test-stat	<i>P</i> -value	<i>P</i> -value (FDR)
172	212	-95.260701	-66.640486	57.24043	1.10E-11	7.77E-07
172	278	-86.257381	-59.220203	54.074356	5.08E-11	1.79E-06
172	213	-102.53157	-76.376144	52.310852	1.19E-10	2.77E-06
172	242	-90.576939	-64.712052	51.729774	1.57E-10	2.77E-06
172	206	-77.13814	-52.08129	50.1137	3.42E-10	4.82E-06
172	292	-79.869861	-55.371555	48.996612	5.85E-10	6.87E-06
149	172	-142.326561	-118.68048	47.292162	1.33E-09	1.33E-05
137	172	-98.903071	-75.740267	46.325608	2.11E-09	1.86E-05
158	172	-126.00687	-105.252318	41.509104	2.11E-08	0.000165161699520559
172	315	-89.372521	-69.349836	40.04537	4.24E-08	0.00027882614746233
98	172	-82.689173	-62.746316	39.885714	4.57E-08	0.00027882614746233
151	172	-87.87812	-67.97483	39.80658	4.75E-08	0.00027882614746233
172	173	-140.454483	-121.475613	37.95774	1.14E-07	0.000620009254642246
73	172	-103.779205	-85.255579	37.047252	1.76E-07	0.000886978319149443
172	208	-135.618671	-117.172459	36.892424	1.90E-07	0.000890932476549366
110	172	-75.266596	-58.856524	32.820144	1.30E-06	0.00572883065011465
140	172	-107.194851	-92.150417	30.088868	4.69E-06	0.0194692816187225
172	468	-106.731919	-93.090927	27.281984	1.74E-05	0.0682778067287818
172	202	-142.37568	-129.171471	26.408418	2.62E-05	0.0971264383410068
147	236	-133.393355	-122.281267	22.224176	0.000180842506048995	0.637469833822707
236	402	-157.755002	-146.913319	21.683366	0.000231702291790925	0.777857693869534
11	172	-85.524571	-74.819616	21.40991	0.000262583155412055	0.817110082323141
218	236	-131.09927	-120.410813	21.376914	0.000266574920474216	0.817110082323141
149	188	-99.854232	-89.369955	20.968554	0.00032124768836983	0.943665084586376

Continued on next page

Table 3-Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
172	241	-86.603477	-76.275782	20.65539	0.000370578823137624	0.999999999755
236	241	-123.959968	-113.711205	20.497526	0.00039822044147797	0.999999999755
171	236	-118.15418	-107.907915	20.49253	0.000399127786084086	0.999999999755
171	245	-157.753608	-147.576446	20.354324	0.000425056094516707	0.999999999755
154	213	-197.645557	-187.505926	20.279262	0.000439830166879762	0.999999999755
171	209	-80.154649	-70.265916	19.777466	0.000552529122753942	0.999999999755
2	377	-132.82316	-123.057211	19.531898	0.000617667224629415	0.999999999755
236	238	-118.153036	-108.440276	19.42552	0.000648191457816139	0.999999999755
208	262	-151.535949	-141.930562	19.210774	0.000714430439784697	0.999999999755
2	91	-138.037954	-128.438014	19.19988	0.000717963619603745	0.999999999755
243	245	-160.870729	-151.288714	19.16403	0.000729713083036265	0.999999999755
236	243	-121.367667	-111.865979	19.003376	0.000784744795860992	0.999999999755
172	402	-120.398298	-110.939689	18.917218	0.000815928741833449	0.999999999755
153	205	-24.643147	-15.201595	18.883104	0.000828611814868485	0.999999999755
172	466	-157.397841	-147.982333	18.831016	0.00084835434424424	0.999999999755
172	238	-80.796828	-71.453865	18.685926	0.000905832076696589	0.999999999755
15	233	-70.853696	-61.536033	18.635326	0.000926769459038801	0.999999999755
172	546	-189.734083	-180.459916	18.548334	0.000963889098073767	0.999999999755
147	209	-95.395213	-86.201372	18.387682	0.00103634281682119	0.999999999755
91	236	-145.73871	-136.637576	18.202268	0.00112667292298252	0.999999999755
69	205	-121.873963	-112.777017	18.193892	0.00113093223673189	0.999999999755
236	546	-227.090789	-218.053187	18.075204	0.00119302430998547	0.999999999755
41	236	-125.973027	-116.954557	18.03694	0.00121375112605182	0.999999999755
172	175	-88.807657	-79.810745	17.993824	0.00123753255760162	0.999999999755
79	153	-33.498754	-24.656956	17.683596	0.00142275580259887	0.999999999755
15	159	-37.517793	-28.759554	17.516478	0.00153361582691103	0.999999999755
41	172	-88.616345	-79.881762	17.469166	0.0015665207568073	0.999999999755
65	122	-102.978119	-94.30307	17.350098	0.00165244594718095	0.999999999755
172	262	-153.063464	-144.390213	17.346502	0.00165511207631241	0.999999999755
79	99	-25.345169	-16.69659	17.297158	0.00169212803581087	0.999999999755
149	154	-237.440547	-228.843112	17.19487	0.00177148141516981	0.999999999755
73	108	-115.35578	-106.801142	17.109276	0.00184069638482554	0.999999999755
171	172	-80.797011	-72.251506	17.09101	0.00185580885609293	0.999999999755
91	209	-107.740568	-99.199889	17.081358	0.00186384383996996	0.999999999755
15	16&294	-36.476618	-27.980262	16.992712	0.0019392618772972	0.999999999755
69	79	-130.729512	-122.265429	16.928166	0.00199606187258083	0.999999999755
172	218	-93.743876	-85.28066	16.926432	0.00199761019179456	0.999999999755
172	391	-128.171689	-119.717041	16.909296	0.00201297504854159	0.999999999755
41	209	-87.974885	-79.53564	16.87849	0.00204089059165746	0.999999999755
16&294	153	-23.713307	-15.294666	16.837282	0.00207882969689865	0.999999999755
91	245	-184.795639	-176.381198	16.828882	0.0020866482405304	0.999999999755
241	245	-163.739881	-155.336381	16.807	0.00210715192978539	0.999999999755
16&294	69	-120.944089	-112.559717	16.768744	0.00214347647400648	0.999999999755
154	212	-190.374688	-182.016027	16.717322	0.00219327641052169	0.999999999755
2	209	-166.160652	-157.850034	16.621236	0.00228940680823264	0.999999999755
41	245	-165.029956	-156.72176	16.616392	0.00229436147719242	0.999999999755
172	188	-94.673858	-86.429946	16.487824	0.00242979879049932	0.999999999755

Continued on next page

Table 3-Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
245	402	-196.811909	-188.679286	16.265246	0.00268313322906089	0.999999999755
236	391	-165.528372	-157.46435	16.128044	0.0028520733753099	0.999999999755
238	245	-158.160035	-150.114325	16.09142	0.00289891315078628	0.999999999755
236	377	-140.523917	-132.487123	16.073588	0.00292199249379554	0.999999999755
16&294	79	-24.885007	-16.863386	16.043242	0.00296168483869053	0.999999999755
2	243	-115.11334	-107.124119	15.978442	0.00304822810272354	0.999999999755
69	99	-121.404162	-113.41741	15.973504	0.00305492417351449	0.999999999755
15	79	-46.26147	-38.306894	15.909152	0.00314352243424154	0.999999999755
91	172	-108.382028	-100.430462	15.903132	0.0031519388392941	0.999999999755
99	153	-24.173396	-16.232303	15.882186	0.00318139608280876	0.999999999755
202	209	-141.73422	-133.800234	15.867972	0.0032015399706149	0.999999999755
15	153	-45.089704	-37.168746	15.841916	0.00323879264171545	0.999999999755
218	245	-170.86688	-163.033675	15.66641	0.00350106845546394	0.999999999755
15	363	-58.456346	-50.681616	15.54946	0.00368732505795455	0.999999999755
158	191	-89.091102	-81.393364	15.395476	0.00394749133007244	0.999999999755
73	391	-94.804708	-87.148318	15.31278	0.00409457552700854	0.999999999755
237	546	-195.14541	-187.5139	15.26302	0.00418566143234478	0.999999999755
147	245	-172.450284	-164.83437	15.231828	0.00424377090065453	0.999999999755
154	242	-186.387587	-178.823874	15.127426	0.00444409450263472	0.999999999755
69	142	-122.232786	-114.672781	15.12001	0.00445867239861419	0.999999999755
65	545	-130.41267	-122.858469	15.108402	0.00448158503491347	0.999999999755
209	546	-189.092623	-181.581983	15.02128	0.00465728390114206	0.999999999755
128	313	-82.258413	-74.770147	14.976532	0.00475013965260518	0.999999999755
142	153	-25.001957	-17.553857	14.8962	0.00492141870197038	0.999999999755
73	173	-107.087502	-99.705605	14.763794	0.00521705225442526	0.999999999755
15	205	-37.405868	-30.030795	14.750146	0.00524849869616506	0.999999999755
172	235	-141.653506	-134.310796	14.68542	0.00540018592006486	0.999999999755
79	205	-25.814931	-18.491374	14.647114	0.00549197516844835	0.999999999755
15	264	-36.099708	-28.795231	14.608954	0.00558493640637592	0.999999999755
15	69	-142.320462	-135.018307	14.60431	0.00559635447213735	0.999999999755
15	99	-36.936173	-29.656677	14.558992	0.00570898572169087	0.999999999755
73	188	-61.306877	-54.083756	14.446242	0.0059989478750111	0.999999999755
188	213	-60.059242	-52.839282	14.43992	0.0060156272313775	0.999999999755
172	179	-78.89473	-71.708123	14.373214	0.00619441330064496	0.999999999755
154	217	-231.698276	-224.538646	14.31926	0.00634282405485831	0.999999999755
212	262	-111.177979	-104.037812	14.280334	0.00645205687136396	0.999999999755
91	402	-91.63414	-84.524912	14.218456	0.00662950374489812	0.999999999755
209	218	-93.1015	-86.026803	14.149394	0.00683320368361695	0.999999999755
151	213	-53.263503	-46.203451	14.120104	0.0069214361698553	0.999999999755
138	223	-33.911451	-26.881617	14.059668	0.00710703496939846	0.999999999755
15	229	-43.520539	-36.496346	14.048386	0.00714221807017035	0.999999999755
360	366	-23.046363	-16.058858	13.97501	0.00737524206945139	0.999999999755
209	402	-119.756839	-112.786258	13.941162	0.00748522941595586	0.999999999755
2	238	-112.256468	-105.303071	13.906794	0.00759854979001506	0.999999999755

Continued on next page

Table 3-Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
15	176	-38.773334	-31.830015	13.886638	0.0076657892646832	0.999999999755
15	179	-39.249057	-32.318012	13.86209	0.00774846751649405	0.999999999755
554	562&565	-35.843729	-28.949767	13.787924	0.00800360001575418	0.999999999755
262	468	-122.649197	-115.787454	13.723486	0.00823193099993291	0.999999999755
160	400	-34.253878	-27.405005	13.697746	0.00832491007154013	0.999999999755
312	493	-22.524909	-15.678997	13.691824	0.00834644676191187	0.999999999755
2	171	-112.289281	-105.456631	13.6653	0.00844357817895147	0.999999999755
209	243	-83.369645	-76.565585	13.60812	0.00865675023049517	0.999999999755
79	162	-36.750957	-29.966928	13.568058	0.00880923276612045	0.999999999755
209	495	-101.320696	-94.54879	13.543812	0.00890279004421735	0.999999999755
79	176	-27.181719	-20.431972	13.499494	0.00907631684629118	0.999999999755
12	557	-72.257759	-65.510564	13.49439	0.00909651270958733	0.999999999755
18	79	-34.203613	-27.511071	13.385084	0.00953972412230608	0.999999999755
391	466	-148.423343	-141.739556	13.367574	0.0096126583070183	0.999999999755
433&449	514	-22.064224	-15.401217	13.326014	0.00978795692183365	0.999999999755
79	276	-67.062134	-60.400447	13.323374	0.00979919729328571	0.999999999755
154	158	-221.120857	-214.461438	13.318838	0.00981853976584857	0.999999999755
16&294	205	-16.029383	-9.371419	13.315928	0.00983096827222008	0.999999999755
209	238	-80.155108	-73.505347	13.299522	0.00990132596967297	0.999999999755
306	403	-22.358672	-15.713962	13.28942	0.00994489331108972	0.999999999755

Table 5.4: **Detailed results generated using the KDBP11 data set for NA in H3N2 viruses. Site-1 and site-2 show the AA positions along the alignment.** The lkl terms show the log-likelihood values after optimization under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The P -values are computed using the standard χ^2 approximation. FDR-corrected P -values follow BH. Only pairs significant at the nominal 1% level (P -value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	site-2	lkl-indep	lkl-dep	test-stat	P -value	P -value (FDR)
4	5	-140.551535	-115.819373	49.464324	4.67E-10	2.52E-05
2	4	-113.963582	-97.896387	32.13439	1.80E-06	0.0348153544795633
30	215	-191.722818	-175.885065	31.675506	2.23E-06	0.0348153544795633
18	215	-141.968795	-126.28716	31.36327	2.58E-06	0.0348153544795633
42	215	-142.317227	-126.887929	30.858596	3.27E-06	0.0353044758154209
215	216	-155.86052	-140.880835	29.95937	4.99E-06	0.0448602999915356
23	215	-153.032804	-138.311892	29.441824	6.36E-06	0.0490014117590225
2	5	-141.640317	-128.095522	27.08959	1.91E-05	0.128596052265248
267	338	-111.370643	-97.965384	26.810518	2.17E-05	0.13016322352628
215	387	-150.643803	-137.556013	26.17558	2.92E-05	0.157372062936907
215	372	-175.876948	-162.935756	25.882384	3.34E-05	0.16323158527533
215	437	-159.054349	-146.202314	25.70407	3.63E-05	0.16323158527533

Continued on next page

Table 4-Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
215	310	-181.656122	-169.582247	24.14775	7.46E-05	0.306265388827206
215	265	-170.896434	-158.890888	24.011092	7.95E-05	0.306265388827206
147	310	-168.960636	-157.463052	22.995168	0.000126907935161791	0.456496303305972
23	30	-114.499431	-103.25457	22.489722	0.000160093160150421	0.539874159317256
194	215	-159.70242	-148.579431	22.245978	0.000179042843320709	0.568260920836011
151	208	-298.020159	-286.967044	22.10623	0.000190894074553305	0.572215593699896
147	387	-137.948316	-127.159345	21.577942	0.000243156189340787	0.689103997241034
150	215	-153.709274	-142.974131	21.470286	0.000255431832323016	0.689103997241034
50	208	-128.316646	-117.840993	20.951306	0.000323786781901036	0.803369469548565
221	332	-97.455311	-87.072967	20.764688	0.000352563215472612	0.803369469548565
172	221	-81.930036	-71.559024	20.742024	0.000356226180617258	0.803369469548565
221	399	-106.187484	-95.819908	20.735152	0.000357344266979864	0.803369469548565
267	339	-131.381651	-121.137016	20.48927	0.000399720952128035	0.862693747720811
147	372	-163.181462	-153.039367	20.28419	0.000438844807024497	0.910704246454374
147	194	-147.006933	-136.934545	20.144776	0.000467583059149623	0.934404131091743
208	215	-181.839584	-171.88551	19.908148	0.000520686430556361	0.9999999999875
155	390	-49.067913	-39.428541	19.278744	0.000692771360469524	0.9999999999875
151	437	-275.23493	-265.88771	18.69444	0.000902355388088139	0.9999999999875
223	234	-44.100347	-34.755491	18.689712	0.000904284433107305	0.9999999999875
47	267	-99.701921	-90.428847	18.546148	0.000964840550960244	0.9999999999875
151	216	-272.031728	-262.847282	18.368892	0.00104516099261065	0.9999999999875
208	338	-112.919637	-103.829411	18.180452	0.00113779992965746	0.9999999999875
120	267	-163.302895	-154.224436	18.156918	0.00114992491961707	0.9999999999875
147	215	-217.560704	-208.519712	18.081984	0.00118938836160598	0.9999999999875
82	370	-127.044629	-118.03821	18.012838	0.00122698883474071	0.9999999999875
248	249	-105.019997	-96.015487	18.00902	0.00122909886439504	0.9999999999875
458	459	-50.477343	-41.617396	17.719894	0.00139974359347339	0.9999999999875
2	208	-124.23767	-115.415743	17.643854	0.00144837988963775	0.9999999999875
267	400	-82.484163	-73.810887	17.346552	0.00165507497646944	0.9999999999875
221	338	-96.858553	-88.214593	17.28792	0.00169914824242634	0.9999999999875
42	151	-258.496695	-250.021015	16.95136	0.00197546496346734	0.9999999999875
4	30	-133.032124	-124.591141	16.881966	0.00203772173065686	0.9999999999875
460	462	-57.896555	-49.533401	16.726308	0.00218449237417284	0.9999999999875
151	265	-287.076993	-278.747965	16.658056	0.00225208872620708	0.9999999999875
2	30	-134.120903	-125.81302	16.615766	0.00229500254974246	0.9999999999875
221	267	-115.812898	-107.51043	16.604936	0.00230612129517527	0.9999999999875
367	426	-85.840734	-77.551632	16.578204	0.00233379375398302	0.9999999999875
3	5	-114.796378	-106.517702	16.557352	0.00235560629205034	0.9999999999875
18	151	-258.149354	-249.890235	16.5182380000001	0.00239706526466199	0.9999999999875
30	151	-307.903387	-299.654305	16.498164	0.00241862109504687	0.9999999999875
47	338	-80.747576	-72.549737	16.395678	0.00253168438019746	0.9999999999875
150	151	-269.889838	-261.71149	16.3566959999999	0.00257604302664682	0.9999999999875
328	334	-79.102208	-70.930845	16.342726	0.00259212548843057	0.9999999999875
47	208	-101.250916	-93.09991	16.302012	0.00263956227260953	0.9999999999875
151	338	-277.516803	-269.38117	16.271266	0.00267595081040717	0.9999999999875
467	469	-50.458942	-42.381348	16.155188	0.00281783950547188	0.9999999999875
82	267	-112.040302	-103.962751	16.155102	0.00281794732667073	0.9999999999875
52	370	-112.243004	-104.167152	16.151704	0.0028222107706144	0.9999999999875
4	42	-83.625444	-75.583108	16.084672	0.00290762566973246	0.9999999999875
147	150	-141.013787	-132.990365	16.046844	0.00295694582647521	0.9999999999875

Continued on next page

Table 4-Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
2	42	-84.714222	-76.731945	15.964554	0.00306709755514389	0.999999999875
148	265	-154.588796	-146.624072	15.929448	0.00311530979260433	0.999999999875
47	339	-100.758584	-92.848993	15.819182	0.00327164401657698	0.999999999875
82	172	-78.15744	-70.259604	15.795672	0.00330596158678165	0.999999999875
468	469	-78.861775	-70.974619	15.774312	0.00333744774713474	0.999999999875
52	267	-97.238677	-89.396752	15.68385	0.00347409805255838	0.999999999875
23	151	-269.213363	-261.383175	15.660376	0.0035104477183352	0.999999999875
4	23	-94.34211	-86.512211	15.659798	0.00351134745896731	0.999999999875
4	216	-97.160477	-89.335684	15.649586	0.00352728137476244	0.999999999875
127	221	-94.72738	-86.927875	15.59901	0.00360725014316809	0.999999999875
82	151	-278.186472	-270.400877	15.57119	0.00365199548860917	0.999999999875
241	287	-70.044727	-62.261923	15.565608	0.00366103903998682	0.999999999875
331	435	-52.400018	-44.629912	15.540212	0.00370246262268259	0.999999999875
2	216	-98.249254	-90.486048	15.526412	0.00372516482327112	0.999999999875
2	23	-95.430889	-87.692935	15.475908	0.0038094215182668	0.999999999875
155	328	-56.759264	-49.047166	15.424196	0.00389763482251559	0.999999999875
248	267	-135.975774	-128.313444	15.32466	0.00407311889187079	0.999999999875
47	249	-68.746144	-61.193065	15.106158	0.00448602771443796	0.999999999875
127	285	-65.277967	-57.73946	15.077014	0.00454412162507856	0.999999999875
4	18	-83.278102	-75.767891	15.020422	0.00465904744183665	0.999999999875
359	419&420	-22.759934	-15.258199	15.00347	0.0046940255441954	0.999999999875
199	267	-166.807698	-159.307124	15.001148	0.00469883671018834	0.999999999875
221	437	-94.576657	-87.11821	14.916894	0.00487672489268509	0.999999999875
2	18	-84.366883	-76.913846	14.906074	0.0049000434090325	0.999999999875
267	385	-74.1345	-66.768057	14.732886	0.00528853411749663	0.999999999875
18	30	-103.435422	-96.078339	14.714166	0.00533229497135612	0.999999999875
334	384	-72.339251	-64.983297	14.711908	0.00533759735197847	0.999999999875
89	322	-22.493935	-15.217846	14.552178	0.00572611215121643	0.999999999875
149	390	-39.239054	-32.011958	14.454192	0.00597803773405736	0.999999999875
194	335	-94.541538	-87.341945	14.399186	0.00612419180716595	0.999999999875
199	215	-216.7733	-209.575485	14.39563	0.00613375997292154	0.999999999875
65	381	-85.121412	-77.963863	14.315098	0.00635441619646426	0.999999999875
267	381	-85.467318	-78.315731	14.303174	0.00638774244647655	0.999999999875
151	172	-262.588286	-255.441413	14.2937459999999	0.0064142139127632	0.999999999875
51	431	-150.81514	-143.689073	14.252134	0.0065323406301101	0.999999999875
159	322	-22.114499	-14.992086	14.244826	0.006553305353589	0.999999999875
52	199	-133.721387	-126.615921	14.210932	0.00665140378507478	0.999999999875
199	331	-144.286475	-137.186706	14.199538	0.00668470326733372	0.999999999875
467	468	-79.489709	-72.425517	14.128384	0.00689638112617863	0.999999999875
260&452	261	-21.77714	-14.727302	14.099676	0.00698363284295656	0.999999999875
267	342	-135.50191	-128.473591	14.056638	0.00711646736040761	0.999999999875
257	282&352	-22.199851	-15.184335	14.031032	0.00719667016713788	0.999999999875
329	356	-32.774535	-25.765332	14.018406	0.00723654248121752	0.999999999875
149	370	-105.774035	-98.7817	13.98467	0.00734414381931736	0.999999999875
52	62	-101.5899	-94.629657	13.920486	0.00755320371228818	0.999999999875
268	282&352	-21.514353	-14.561359	13.905988	0.00760122744767489	0.999999999875
267	462	-110.321548	-103.372697	13.897702	0.00762880846715241	0.999999999875
338	370	-126.37497	-119.429498	13.890944	0.00765137588288978	0.999999999875
16	21	-65.851742	-58.90748	13.888524	0.00765947302770498	0.999999999875
52	151	-263.384837	-256.46375	13.842174	0.00781618608652146	0.999999999875

Continued on next page

Table 4-Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
215	249	-149.334812	-142.465949	13.737726	0.00818093018169386	0.9999999999875
153	368	-19.522221	-12.667663	13.709116	0.00828371284576168	0.9999999999875
155	248	-106.249358	-99.395435	13.707846	0.00828830455355267	0.9999999999875
227	228	-21.372586	-14.519539	13.706094	0.0082946430297719	0.9999999999875
47	344	-55.414002	-48.597998	13.632008	0.00856706076175984	0.9999999999875
197	462	-94.856671	-88.054269	13.604804	0.00866927281861274	0.9999999999875
69	308	-21.946527	-15.156018	13.581018	0.00875961957864446	0.9999999999875
165	274	-28.831479	-22.046053	13.570852	0.00879851364628936	0.9999999999875
11	355	-29.487763	-22.70613	13.563266	0.00882764692170956	0.9999999999875
43	53	-62.18546	-55.418248	13.534424	0.00893927565536456	0.9999999999875
346	390	-34.096673	-27.337333	13.51868	0.00900079186093183	0.9999999999875
47	127	-78.616404	-71.858896	13.515016	0.00901516741845965	0.9999999999875
267	331	-107.803764	-101.056404	13.49472	0.00909520561465604	0.9999999999875
141	151	-254.977731	-248.235514	13.484434	0.00913603386506068	0.9999999999875
106	368	-23.105528	-16.365453	13.48015	0.00915309118315888	0.9999999999875
56	267	-91.562243	-84.838056	13.448374	0.00928058750487426	0.9999999999875
71	127	-86.15186	-79.442996	13.417728	0.00940519423249664	0.9999999999875
356	466	-34.056448	-27.348043	13.41681	0.00940895194781366	0.9999999999875
432	462	-109.814891	-103.10911	13.411562	0.00943046216358712	0.9999999999875
153	344	-33.436465	-26.766585	13.33976	0.00972963369836866	0.9999999999875
151	385	-240.280661	-233.617236	13.32685	0.00978440009898351	0.9999999999875
18	23	-64.745408	-58.082295	13.326226	0.00978705483280917	0.9999999999875

Table 4-Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
2	42	-84.714222	-76.731945	15.964554	0.00306709755514389	0.999999999875
148	265	-154.588796	-146.624072	15.929448	0.00311530979260433	0.999999999875
47	339	-100.758584	-92.848993	15.819182	0.00327164401657698	0.999999999875
82	172	-78.15744	-70.259604	15.795672	0.00330596158678165	0.999999999875
468	469	-78.861775	-70.974619	15.774312	0.00333744774713474	0.999999999875
52	267	-97.238677	-89.396752	15.68385	0.00347409805255838	0.999999999875
23	151	-269.213363	-261.383175	15.660376	0.0035104477183352	0.999999999875
4	23	-94.34211	-86.512211	15.659798	0.00351134745896731	0.999999999875
4	216	-97.160477	-89.335684	15.649586	0.00352728137476244	0.999999999875
127	221	-94.72738	-86.927875	15.59901	0.00360725014316809	0.999999999875
82	151	-278.186472	-270.400877	15.57119	0.00365199548860917	0.999999999875
241	287	-70.044727	-62.261923	15.565608	0.00366103903998682	0.999999999875
331	435	-52.400018	-44.629912	15.540212	0.00370246262268259	0.999999999875
2	216	-98.249254	-90.486048	15.526412	0.00372516482327112	0.999999999875
2	23	-95.430889	-87.692935	15.475908	0.0038094215182668	0.999999999875
155	328	-56.759264	-49.047166	15.424196	0.00389763482251559	0.999999999875
248	267	-135.975774	-128.313444	15.32466	0.00407311889187079	0.999999999875
47	249	-68.746144	-61.193065	15.106158	0.00448602771443796	0.999999999875
127	285	-65.277967	-57.73946	15.077014	0.00454412162507856	0.999999999875
4	18	-83.278102	-75.767891	15.020422	0.00465904744183665	0.999999999875
359	419&420	-22.759934	-15.258199	15.00347	0.0046940255441954	0.999999999875
199	267	-166.807698	-159.307124	15.001148	0.00469883671018834	0.999999999875
221	437	-94.576657	-87.11821	14.916894	0.00487672489268509	0.999999999875
2	18	-84.366883	-76.913846	14.906074	0.0049000434090325	0.999999999875
267	385	-74.1345	-66.768057	14.732886	0.00528853411749663	0.999999999875
18	30	-103.435422	-96.078339	14.714166	0.00533229497135612	0.999999999875
334	384	-72.339251	-64.983297	14.711908	0.00533759735197847	0.999999999875
89	322	-22.493935	-15.217846	14.552178	0.00572611215121643	0.999999999875
149	390	-39.239054	-32.011958	14.454192	0.00597803773405736	0.999999999875
194	335	-94.541538	-87.341945	14.399186	0.00612419180716595	0.999999999875
199	215	-216.7733	-209.575485	14.39563	0.00613375997292154	0.999999999875
65	381	-85.121412	-77.963863	14.315098	0.00635441619646426	0.999999999875
267	381	-85.467318	-78.315731	14.303174	0.00638774244647655	0.999999999875
151	172	-262.588286	-255.441413	14.2937459999999	0.0064142139127632	0.999999999875
51	431	-150.81514	-143.689073	14.252134	0.0065323406301101	0.999999999875
159	322	-22.114499	-14.992086	14.244826	0.006553305353589	0.999999999875
52	199	-133.721387	-126.615921	14.210932	0.00665140378507478	0.999999999875
199	331	-144.286475	-137.186706	14.199538	0.00668470326733372	0.999999999875
467	468	-79.489709	-72.425517	14.128384	0.00689638112617863	0.999999999875
260&452	261	-21.77714	-14.727302	14.099676	0.00698363284295656	0.999999999875
267	342	-135.50191	-128.473591	14.056638	0.00711646736040761	0.999999999875
257	282&352	-22.199851	-15.184335	14.031032	0.00719667016713788	0.999999999875
329	356	-32.774535	-25.765332	14.018406	0.00723654248121752	0.999999999875
149	370	-105.774035	-98.7817	13.98467	0.00734414381931736	0.999999999875
52	62	-101.5899	-94.629657	13.920486	0.00755320371228818	0.999999999875
268	282&352	-21.514353	-14.561359	13.905988	0.00760122744767489	0.999999999875
267	462	-110.321548	-103.372697	13.897702	0.00762880846715241	0.999999999875
338	370	-126.37497	-119.429498	13.890944	0.00765137588288978	0.999999999875
16	21	-65.851742	-58.90748	13.888524	0.00765947302770498	0.999999999875
52	151	-263.384837	-256.46375	13.842174	0.00781618608652146	0.999999999875

Continued on next page

Table 4-Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
215	249	-149.334812	-142.465949	13.737726	0.00818093018169386	0.9999999999875
153	368	-19.522221	-12.667663	13.709116	0.00828371284576168	0.9999999999875
155	248	-106.249358	-99.395435	13.707846	0.00828830455355267	0.9999999999875
227	228	-21.372586	-14.519539	13.706094	0.0082946430297719	0.9999999999875
47	344	-55.414002	-48.597998	13.632008	0.00856706076175984	0.9999999999875
197	462	-94.856671	-88.054269	13.604804	0.00866927281861274	0.9999999999875
69	308	-21.946527	-15.156018	13.581018	0.00875961957864446	0.9999999999875
165	274	-28.831479	-22.046053	13.570852	0.00879851364628936	0.9999999999875
11	355	-29.487763	-22.70613	13.563266	0.00882764692170956	0.9999999999875
43	53	-62.18546	-55.418248	13.534424	0.00893927565536456	0.9999999999875
346	390	-34.096673	-27.337333	13.51868	0.00900079186093183	0.9999999999875
47	127	-78.616404	-71.858896	13.515016	0.00901516741845965	0.9999999999875
267	331	-107.803764	-101.056404	13.49472	0.00909520561465604	0.9999999999875
141	151	-254.977731	-248.235514	13.484434	0.00913603386506068	0.9999999999875
106	368	-23.105528	-16.365453	13.48015	0.00915309118315888	0.9999999999875
56	267	-91.562243	-84.838056	13.448374	0.00928058750487426	0.9999999999875
71	127	-86.15186	-79.442996	13.417728	0.00940519423249664	0.9999999999875
356	466	-34.056448	-27.348043	13.41681	0.00940895194781366	0.9999999999875
432	462	-109.814891	-103.10911	13.411562	0.00943046216358712	0.9999999999875
153	344	-33.436465	-26.766585	13.33976	0.00972963369836866	0.9999999999875
151	385	-240.280661	-233.617236	13.32685	0.00978440009898351	0.9999999999875
18	23	-64.745408	-58.082295	13.326226	0.00978705483280917	0.9999999999875

Table 5.5: **Epistatic pairs of sites of amino acid detected before FDR in the Gong13NP data set of NP proteins in H3N2 human influenza A virus.** The residues were recoded according to their physicochemical properties as in Fig 3

Epistatic pair of Amino acids (years)	Amino acid physicochemical properties groups	Detected by Gong and coworkers	<i>P</i> -value after FDR **
A131S×T373N (2007–1969)	- Hydrophobicity - Normalized van der Waals Volume - Polarity - Solvent Accessibility	No	1
A131S×E375G (2007–1991)	- Hydrophobicity - Normalized van der Waals Volume - Polarity - Solvent Accessibility - Charge - Polarizability	No	1
A131S×R384G* (2007–1991)	- Polarity - Solvent Accessibility - Charge - Polarizability	No	1
A373S×L259S* (1969–1973)	- Polarizability - Normalized van der Waals Volume - Polarity - Secondary structure	No	1
* Deleterious mutation **FDR: False Discovery Rate			

Table 5.6: **Detailed results generated using the Gong13NP data set. Site-1 and site-2 show the AA positions along the alignment.** The lkl terms show the log-likelihood values after optimization under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The P -values are computed using the standard χ^2 approximation. FDR-corrected P -values follow BH. Only pairs significant at the nominal 1% level (P -value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	site-2	lkl-indep	lkl-dep	test-stat	P -value	P -value (FDR)
259	334	-24.112557	-15.110641	18.003832	0.00123197177337009	1
421	425	-41.33507	-32.643448	17.383244	0.00162806970000984	1
259	421	-28.433097	-20.76289	15.340414	0.00404483575372983	1
186	259	-52.01484	-44.549463	14.930754	0.0048470139256368	1
259	411	-43.108741	-35.931834	14.353814	0.00624738040528983	1
286	421	-21.478869	-14.567582	13.822574	0.00788339582234365	1
246	470	-11.518628	-4.755542	13.526172	0.00897146705998841	1

Table 5.7: **Detailed results generated using the Duan14NA data set.** **Site-1** and **site-2** show the AA positions along the alignment. The lkl terms show the log-likelihood values after optimization under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The P -values are computed using the standard χ^2 approximation. FDR-corrected P -values follow BH. Only pairs significant at the nominal 1% level (P -value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	site-2	lkl-indep	lkl-dep	test-stat	P -value	P -value (FDR)
275	354	-136.541395	-116.877463	39.327864	5.96E-08	0.0020535384925
151	450	-139.615983	-128.751821	21.728324	0.000226982022307198	0.999999999955
151	332	-159.091073	-148.686575	20.808996	0.000345509589050774	0.999999999955
69	352	-55.379769	-44.989787	20.779964	0.000350115379225246	0.999999999955
287	354	-63.891067	-54.806851	18.168432	0.00114397690881296	0.999999999955
151	173	-137.888164	-128.945172	17.885984	0.00129904744386045	0.999999999955
82	151	-146.103838	-137.666486	16.874704	0.00204434758921312	0.999999999955
151	453	-144.681804	-136.271468	16.820672	0.00209431794817483	0.999999999955
68	332	-95.891365	-87.504276	16.774178	0.00213827945675782	0.999999999955
23	151	-160.568287	-152.31876	16.499054	0.0024176613631598	0.999999999955
93	95	-53.766448	-45.618077	16.296742	0.00264576463273458	0.999999999955
23	452	-93.45056	-85.319271	16.262578	0.00268632246945488	0.999999999955
130	267	-121.808649	-113.710559	16.19618	0.00276690614373509	0.999999999955
151	262	-136.41958	-128.732775	15.37361	0.00398586951995639	0.999999999955
83	151	-154.666562	-147.216912	14.8993	0.00491469796176947	0.999999999955
151	367	-151.047796	-143.685161	14.72527	0.00530629497982016	0.999999999955
354	454	-78.108496	-70.771995	14.673002	0.00542977590314364	0.999999999955
100	151	-130.504031	-123.195851	14.61636	0.00556677477166456	0.999999999955
68	78	-106.513245	-99.36027	14.30595	0.00637996856547018	0.999999999955
6	329	-100.48581	-93.444414	14.082792	0.00703545228325753	0.999999999955
149	151	-156.535571	-149.509767	14.051608	0.00713215283964763	0.999999999955
151	452	-164.589539	-157.573367	14.032344	0.00719253929803354	0.999999999955
151	393	-145.958782	-138.989157	13.93925	0.0074914900717713	0.999999999955
130	239	-53.068511	-46.106562	13.923898	0.00754194493815419	0.999999999955
6	45	-49.538259	-42.597146	13.882226	0.00768058503537805	0.999999999955
234	382	-52.245962	-45.500083	13.491758	0.00910694436005499	0.999999999955

Table 5.8: **Detailed results generated using the Koel13HA data set. Site-1 and site-2 show the AA positions along the alignment.** The lkl terms show the log-likelihood values after optimization under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The P -values are computed using the standard χ^2 approximation. FDR-corrected P -values follow BH. Only pairs significant at the nominal 1% level (P -value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	site-2	lkl-indep	lkl-dep	test-stat	P -value	P -value (FDR)
131	172	-56.325232	-90.641292	68.63212	4.41E-14	2.56E-09
172	299	-63.869614	-91.063367	54.387506	4.37E-11	1.27E-06
131	197	-54.723437	-80.250126	51.053378	2.18E-10	4.20E-06
131	299	-52.796784	-27.588776	50.416016	2.96E-10	4.28E-06
156	196	-149.015007	-125.514329	47.001356	1.52E-09	1.77E-05
82	172	-54.520069	-77.674156	46.308174	2.12E-09	2.05E-05
131	190	-42.767375	-65.287127	45.039504	3.90E-09	3.23E-05
142	156	-144.5229	-122.150763	44.744274	4.49E-09	3.26E-05
94	197	-44.415095	-66.252821	43.675452	7.49E-09	4.83E-05
155	158	-23.289118	-44.194232	41.810228	1.83E-08	0.000105870061375635
121	172	-81.928955	-102.095109	40.332308	3.69E-08	0.000194716683130647
156	226	-182.029994	-162.474344	39.1113	6.61E-08	0.000319184934128395
275	299	-72.503845	-91.350619	37.693548	1.30E-07	0.000577962847914969
172	262	-76.631249	-94.724662	36.186826	2.65E-07	0.00104000518981192
131	133	-89.996826	-108.07345	36.153248	2.69E-07	0.00104000518981192
57	157	-71.313012	-54.028379	34.569266	5.69E-07	0.00206323564386095
260	275	-84.652373	-101.740308	34.17587	6.86E-07	0.00233855447015707
156	192	-159.698581	-142.985686	33.42579	9.77E-07	0.0029905995524521
133	135	-92.222911	-108.932543	33.419264	9.80E-07	0.0029905995524521
226	262	-121.085486	-104.732525	32.705922	1.37E-06	0.00397691373256576
135	172	-56.241943	-72.48434	32.484794	1.52E-06	0.00420337615276976
94	248	-30.942845	-14.856788	32.172114	1.76E-06	0.00464875522733443
146	155	-44.326771	-60.286742	31.919942	1.99E-06	0.00500694975261313
172	196	-85.20475	-100.881214	31.352928	2.59E-06	0.00604869206333314
124	299	-58.586432	-74.256782	31.3407	2.61E-06	0.00604869206333314
47	216	-57.331129	-41.954512	30.753234	3.44E-06	0.00766433062728581
94	172	-46.654893	-61.943841	30.577896	3.73E-06	0.00801358453394216
226	276	-110.892793	-95.935017	29.915552	5.09E-06	0.0103080793574517
124	190	-55.926525	-70.837009	29.820968	5.32E-06	0.0103080793574517
133	157	-101.076796	-115.984826	29.81606	5.34E-06	0.0103080793574517
197	299	-60.894938	-75.768088	29.7463	5.51E-06	0.0103080793574517
190	193	-67.907935	-53.10205	29.61177	5.87E-06	0.0104351974002796
82	124	-52.888221	-67.681553	29.586664	5.94E-06	0.0104351974002796
146	219	-93.45458	-107.94919	28.98922	7.86E-06	0.0130615143536782
82	197	-56.427429	-70.918084	28.98131	7.89E-06	0.0130615143536782
157	226	-110.350463	-95.997502	28.705922	8.97E-06	0.0144137159332662
124	450	-84.747798	-99.073679	28.651762	9.20E-06	0.0144137159332662

Continued on next page

Table 8 -Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
133	157	-101.076796	-115.984826	29.81606	5.34E-06	0.0103080793574517
197	299	-60.894938	-75.768088	29.7463	5.51E-06	0.0103080793574517
190	193	-67.907935	-53.10205	29.61177	5.87E-06	0.0104351974002796
82	124	-52.888221	-67.681553	29.586664	5.94E-06	0.0104351974002796
146	219	-93.45458	-107.94919	28.98922	7.86E-06	0.0130615143536782
82	197	-56.427429	-70.918084	28.98131	7.89E-06	0.0130615143536782
157	226	-110.350463	-95.997502	28.705922	8.97E-06	0.0144137159332662
124	450	-84.747798	-99.073679	28.651762	9.20E-06	0.0144137159332662
156	193	-141.906891	-127.689876	28.43403	1.02E-05	0.0147925885126242
131	156	-115.085158	-129.301989	28.433662	1.02E-05	0.0147925885126242
57	156	-132.329673	-118.114981	28.429384	1.02E-05	0.0147925885126242
131	262	-53.772767	-67.932639	28.319744	1.07E-05	0.0151901024859121
196	375	-88.241385	-102.325965	28.16916	1.15E-05	0.015908596206199
219	299	-84.39992	-98.45473	28.10962	1.19E-05	0.0159765755133827
54	155	-44.92534	-58.755838	27.660996	1.46E-05	0.0189908627484213
197	276	-64.598716	-78.409744	27.622056	1.49E-05	0.0189908627484213
124	276	-64.414062	-78.209269	27.590414	1.51E-05	0.0189908627484213
172	248	-47.379421	-61.153635	27.548428	1.54E-05	0.0189908627484213
156	262	-136.522344	-122.848716	27.347256	1.69E-05	0.0200587546743246
347	384	-45.407296	-59.078106	27.34162	1.70E-05	0.0200587546743246
155	173&323	-21.437778	-34.852155	26.828754	2.15E-05	0.0249596943834507
133	138	-233.09576	-220.018628	26.154264	2.95E-05	0.0334827438576916
146	260	-68.844774	-55.848314	25.99292	3.17E-05	0.0353938089077092
219	276	-90.87916	-103.849445	25.94057	3.25E-05	0.0355802907411882
155	225	-35.075461	-47.921275	25.691628	3.65E-05	0.0391978185272695
156	246	-221.264234	-208.527253	25.473962	4.04E-05	0.0418124876197024
94	133	-81.269174	-93.938962	25.339576	4.30E-05	0.0437190995686334
54	159	-50.291102	-62.915365	25.248526	4.48E-05	0.0448167713614273
156	375	-139.427401	-126.910084	25.034634	4.95E-05	0.0486452103978388
156	197	-135.830135	-123.368725	24.92282	5.21E-05	0.0500389568285764
375	453	-75.964483	-63.513684	24.901598	5.27E-05	0.0500389568285764
196	220	-190.197331	-177.794496	24.80567	5.50E-05	0.0514666003308989
124	197	-74.605806	-86.927028	24.642444	5.94E-05	0.0534856947229171
135	173&323	-32.68001	-20.37102	24.61798	6.00E-05	0.0534856947229171
142	220	-188.033821	-175.725054	24.617534	6.00E-05	0.0534856947229171
124	375	-77.707519	-90.001173	24.587308	6.09E-05	0.0534856947229171
163	172	-48.55901	-60.814908	24.511796	6.31E-05	0.054559298682351
190	197	-59.452012	-71.688215	24.472406	6.42E-05	0.054744735840454
229	452	-226.236426	-214.036553	24.399746	6.64E-05	0.0557938638811007
146	275	-78.195493	-90.366425	24.341864	6.82E-05	0.0564876048789984
159	248	-28.837736	-40.832342	23.989212	8.03E-05	0.0654365820654832
121	133	-118.085097	-130.066287	23.96238	8.13E-05	0.0654365820654832
124	201	-89.206528	-101.002581	23.592106	9.64E-05	0.0755854137686689
156	276	-126.358671	-114.563495	23.590352	9.65E-05	0.0755854137686689
50	275	-110.836472	-122.471673	23.270402	0.000111809463839929	0.0864212615840093
124	157	-66.812128	-78.413177	23.202098	0.000115381222295596	0.0879987775898886

Continued on next page

Table 8 -Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
135	197	-59.645916	-71.232886	23.17394	0.000116886421846152	0.0879987775898886
220	452	-172.908979	-161.347677	23.122604	0.000119680940907085	0.0889474890305608
229	375	-236.730719	-225.246181	22.969076	0.00012844020645264	0.0942490983298678
121	190	-62.812756	-74.191632	22.757752	0.000141546180741137	0.10174518241206
261	299	-49.605782	-38.231659	22.748246	0.000142165944029271	0.10174518241206
121	156	-142.863567	-131.514102	22.69893	0.000145424681587181	0.10280815599523
124	262	-77.39017	-88.543916	22.307492	0.000174060335175197	0.120835928031344
142	299	-76.920183	-65.772891	22.294584	0.000175094323868086	0.120835928031344
135	276	-45.598181	-56.70918	22.221998	0.000181023273510994	0.123457872534498
192	246	-181.232002	-170.157115	22.149774	0.000187120238953398	0.126132095954982
172	384	-51.709551	-62.747294	22.075486	0.000193603939211195	0.12778228350828
192	229	-254.692613	-243.665571	22.054084	0.000195512862955383	0.12778228350828
159	172	-56.782493	-67.781472	21.997958	0.000200608073363862	0.12778228350828
220	529	-182.672498	-171.678073	21.98885	0.000201447241126806	0.12778228350828
220	375	-183.67905	-172.702215	21.95367	0.000204721378502803	0.12778228350828
50	260	-94.689275	-83.712958	21.952634	0.000204818592827949	0.12778228350828
135	275	-70.650232	-81.625592	21.95072	0.000204998315788685	0.12778228350828
133	452	-111.861384	-100.91372	21.895328	0.000210267879833248	0.128754390559956
159	173&323	-24.001045	-34.937096	21.872102	0.000212517188936223	0.128754390559956
190	226	-102.042375	-91.120654	21.843442	0.000215325661845345	0.128754390559956
49	529	-83.631148	-72.719162	21.823972	0.000217254543595446	0.128754390559956
63	260	-53.36382	-42.453885	21.81987	0.000217663106345967	0.128754390559956
453	529	-75.705336	-65.007195	21.396282	0.000264224599794471	0.154718182324096
155	163	-25.280744	-35.935942	21.310396	0.000274805841863901	0.159304946528503
49	106	-57.463109	-46.821678	21.282862	0.000278286158348906	0.159725233658278
106	220	-156.810892	-146.1897	21.242384	0.000283482072796493	0.161112311372673
121	196	-91.506957	-102.10514	21.196366	0.000289505995278105	0.162938471323027
196	242	-69.271627	-58.710905	21.121444	0.000299586106181193	0.164549176608459
384	453	-41.724796	-52.280959	21.112326	0.000300836354965583	0.164549176608459
222	299	-51.144795	-40.592114	21.105362	0.000301794731485505	0.164549176608459
54	384	-45.764409	-56.310121	21.091424	0.000303721957859326	0.164549176608459
172	173&323	-46.653453	-57.104388	20.90187	0.000331175002466932	0.176130411862459
133	156	-172.823132	-162.484654	20.676956	0.00036695331471015	0.190004515112401
172	450	-88.325033	-98.655943	20.66182	0.000369494164581097	0.190004515112401
133	384	-88.779026	-99.107331	20.65661	0.000370372782606543	0.190004515112401
172	192	-97.567101	-87.261729	20.610744	0.000378197508699829	0.192316750695869
190	452	-57.611482	-47.322632	20.5777	0.000383936076347235	0.193182273008617
190	375	-64.673245	-54.391881	20.562728	0.000386564493168873	0.193182273008617
106	529	-57.730217	-47.499146	20.462142	0.000404690950940867	0.200258011824123
94	219	-71.216443	-81.439564	20.446242	0.000407632316633544	0.200258011824123
202	222	-54.493339	-44.291284	20.40411	0.000415529163839023	0.20119445225455
155	159	-32.086038	-42.28304	20.394004	0.000417445760909296	0.20119445225455
57	220	-176.494371	-166.303938	20.380866	0.000419950469601527	0.20119445225455
131	276	-42.673108	-52.800566	20.254916	0.000444730422040629	0.211319857095863
216	385	-51.12756	-41.018044	20.219032	0.000452051719206814	0.213052342783894
186	220	-232.61022	-222.575661	20.069118	0.000483950986062465	0.225227730365345
156	157	-125.887003	-115.856311	20.061384	0.000485655792576645	0.225227730365345

Continued on next page

Table 8 -Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
156	452	-128.94249	-118.936546	20.011888	0.000496707861006196	0.22852503732166
133	229	-270.049502	-260.056591	19.985822	0.000502627916866838	0.228899715244312
82	133	-89.803667	-99.772844	19.938354	0.000513588728797698	0.228899715244312
163	173&323	-16.360794	-26.326642	19.931696	0.000515144927566547	0.228899715244312
94	201	-60.913487	-70.875467	19.92396	0.000516958954721258	0.228899715244312
82	94	-31.19643	-21.238934	19.914992	0.000519069791021454	0.228899715244312
190	192	-83.70289	-73.76272	19.88034	0.000527306458202315	0.228899715244312
82	275	-66.816087	-76.753238	19.874302	0.000528754848769131	0.228899715244312
138	197	-196.176423	-186.240013	19.87282	0.000529110951228873	0.228899715244312
172	375	-80.616478	-90.527502	19.822048	0.000541455136314095	0.231027861632655
49	186	-133.668109	-123.758194	19.81983	0.000542000848405055	0.231027861632655
248	275	-60.168155	-70.032178	19.728046	0.000565066964782535	0.239101693054332
121	261	-62.405206	-72.25909	19.707768	0.000570292590757138	0.239564213668053
193	226	-125.692221	-115.873978	19.636486	0.00058904432396012	0.245661147194015
94	275	-60.533342	-70.31553	19.564376	0.000608634590062218	0.252018194185048
33	54	-54.146174	-44.396762	19.498824	0.000627001926607584	0.257782281457033
157	386	-92.544435	-82.813833	19.461204	0.000637789452684889	0.258428360570166
92	192	-114.465663	-124.189099	19.446872	0.000641947281733724	0.258428360570166
213	275	-71.171593	-80.799194	19.255202	0.000700198623615145	0.27993458076531
63	146	-45.974074	-36.384591	19.178966	0.000724795029350189	0.287783341448154
164	174	-56.108263	-46.53173	19.153066	0.000733344249998602	0.289197048791966
157	192	-90.895589	-81.382307	19.026564	0.000776555386258315	0.304168349604017
57	375	-74.320637	-64.815217	19.01084	0.000782099423448535	0.304283916626252
57	190	-56.288048	-46.791331	18.993434	0.000788282162159093	0.304644779602417
54	260	-68.566102	-59.081965	18.968274	0.000797304697197632	0.304852908491239
121	155	-58.531093	-68.012414	18.962642	0.000799338314484532	0.304852908491239
131	375	-58.396569	-67.869367	18.945596	0.000805524635998256	0.305204334306006
213	347	-51.981455	-61.436208	18.909506	0.000818779048583518	0.308211827573939
260	384	-49.10544	-58.545712	18.880544	0.000829571409825203	0.310259707274626
157	375	-68.291231	-58.86465	18.853162	0.000839904310122042	0.312110595242146
142	229	-241.330907	-231.919572	18.82267	0.000851560607108226	0.314426550280662
124	163	-51.157226	-60.491179	18.667906	0.000913234268536511	0.335064497133301
262	275	-88.295839	-97.621699	18.65172	0.000919934053339233	0.335399855799216
53	189	-86.800404	-77.48348	18.633848	0.000927388166094945	0.336004324928275
5	190	-65.780774	-56.490441	18.580666	0.000949924463000906	0.342031808199767
155	196	-62.001647	-71.268394	18.533494	0.00097036645025228	0.346443758690093
63	201	-69.798372	-60.539949	18.516846	0.000977684168894877	0.346443758690093
192	262	-100.991766	-91.739234	18.505064	0.000982895942460371	0.346443758690093
83	155	-19.979145	-29.22809	18.49789	0.000986082804620758	0.346443758690093
94	299	-34.225275	-24.98394	18.48267	0.000992877745520149	0.346729656071103
197	260	-78.239601	-69.033342	18.412518	0.00102479996611049	0.355734455301947
135	260	-57.909944	-48.723937	18.372014	0.00104369071405264	0.35971801330216
220	226	-226.004872	-216.82416	18.361424	0.0010486862902892	0.35971801330216
213	479	-67.793672	-76.960265	18.333186	0.00106212267358552	0.362183831692662
192	196	-109.842098	-100.684069	18.316058	0.00107035536081723	0.362856726705117
155	156	-112.859521	-121.987746	18.25645	0.00109950091122679	0.369693737437346
124	133	-113.111748	-122.236168	18.24884	0.00110327784330966	0.369693737437346
156	450	-146.622227	-137.504625	18.235204	0.00111007763935278	0.36983448708782
173&323	453	-35.626106	-26.537061	18.17809	0.001139011135188	0.377305574324846

Continued on next page

Table 8 -Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
124	260	-73.026933	-63.964393	18.12508	0.00116653159437441	0.384226343897071
133	450	-122.819352	-131.850334	18.061964	0.00120015628095649	0.39306813337315
133	299	-95.015347	-104.009823	17.988952	0.00124024857017624	0.402372830676089
138	196	-206.507396	-197.515936	17.98292	0.00124361941014584	0.402372830676089
131	144	-40.354927	-31.368611	17.972632	0.00124938950356557	0.402372830676089
133	246	-196.588891	-187.611804	17.954174	0.00125980814428805	0.402712445219168
53	137	-108.100392	-99.133508	17.933768	0.00127142621777521	0.402712445219168
196	276	-74.911702	-83.872319	17.921234	0.00127861478978042	0.402712445219168
160	213	-33.435388	-42.39487	17.918964	0.00127992097941532	0.402712445219168
121	220	-186.736444	-177.781519	17.90985	0.00128517858143085	0.402712445219168
63	173&323	-23.277442	-14.337132	17.88062	0.00130218482225264	0.404442993617244
142	222	-71.580004	-62.641801	17.876406	0.00130465481812014	0.404442993617244
57	155	-47.248426	-56.170924	17.844996	0.00132321182656014	0.408013774391976
75	157	-66.618	-57.708256	17.819488	0.00133847351204264	0.410536029064083
131	453	-51.268371	-42.37988	17.776982	0.00136429230267776	0.416252762032788
135	299	-42.898676	-34.04359	17.710172	0.00140587086861943	0.426692849496693
143	189	-28.579923	-19.735735	17.688376	0.00141970415627557	0.428647135100494
146	299	-53.891888	-45.069154	17.645468	0.00144733042595702	0.434724066283566
135	155	-32.919557	-41.736097	17.63308	0.00145540468664551	0.434895926210517
131	193	-60.007467	-51.202484	17.609966	0.00147058900589381	0.43505936742826
25	172	-58.717814	-49.913113	17.609402	0.00147096146310055	0.43505936742826
197	384	-51.281896	-60.064081	17.56437	0.001501002402438	0.441592177612016
143	275	-65.141732	-73.918526	17.553588	0.0015082844776122	0.441592177612016
7	479	-64.977813	-56.212322	17.530982	0.00152366577224905	0.443853793051645
92	375	-103.791523	-95.032004	17.519038	0.00153185496336772	0.444008161132133
222	225	-42.124864	-33.396397	17.456934	0.00157514071132425	0.453956964945181
260	262	-74.779231	-66.055495	17.447472	0.0015818407265642	0.453956964945181
126	155	-23.389175	-32.0967	17.41505	0.00160501264800983	0.456748998247154
57	229	-229.585554	-220.879636	17.411836	0.00160732785306916	0.456748998247154
158	275	-59.902418	-68.590982	17.377128	0.0016325406786849	0.460698883376356
155	375	-57.672848	-66.358287	17.370878	0.00163712213171519	0.460698883376356
156	220	-241.394328	-232.731327	17.326002	0.00167039235423028	0.467790554467291
229	529	-235.990236	-227.416409	17.147654	0.00180933848399223	0.504266114985719
163	201	-63.964852	-72.494195	17.058686	0.00188285276345668	0.522243898074563
186	453	-122.060406	-113.554376	17.01206	0.00192254842967854	0.528417986456766
121	197	-81.257522	-89.762545	17.010046	0.00192428154500135	0.528417986456766
33	192	-80.528773	-72.034745	16.988056	0.00194330522641417	0.528889220540984
33	197	-59.591614	-51.105486	16.972256	0.0019570882852008	0.53015143875276
192	347	-94.65143	-86.174938	16.952984	0.00197403068322821	0.532253761426695
222	452	-56.415417	-47.962534	16.905766	0.0020161546394134	0.541094835401829
82	122	-27.678802	-19.279528	16.798548	0.00211512461679475	0.56504043334374
142	529	-89.436843	-81.047975	16.777736	0.00213488335633827	0.566818104947187
157	260	-65.310951	-56.92546	16.770982	0.0021413345693192	0.566818104947187
155	452	-44.859465	-53.222155	16.72538	0.00218539790898842	0.574152125687997
63	248	-26.197581	-17.836658	16.721846	0.00218884974602462	0.574152125687997
192	220	-201.376513	-193.029727	16.693572	0.00221666057942971	0.578827990042974
201	226	-135.545613	-127.21645	16.658326	0.00225181730023905	0.582758253994899
173&323	248	-19.387097	-11.072093	16.630008	0.00228046122106784	0.587548164379124

Continued on next page

Table 8 -Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
193	452	-79.057276	-70.74903	16.616492	0.00229425908578562	0.588487607092887
159	220	-161.136734	-152.863073	16.547322	0.00236616975457205	0.604259298116923
275	450	-98.261946	-106.518177	16.512462	0.00240324816724158	0.608368105917006
63	142	-63.156418	-54.927657	16.457522	0.00246284967085508	0.620745197475953
57	159	-57.333136	-49.112254	16.441764	0.0024802118342877	0.622415065080772
7	155	-31.195166	-22.993921	16.40249	0.00252401043749195	0.630676228712968
173&323	452	-42.941814	-34.75868	16.366268	0.00256508039109282	0.636613946642521
160	163	-26.946379	-18.771702	16.349354	0.00258448294827329	0.636613946642521
121	450	-94.003495	-102.172221	16.337452	0.00259822267386667	0.636613946642521
173&323	275	-55.955738	-64.123671	16.335866	0.00260005898393745	0.636613946642521
163	192	-73.318484	-65.151683	16.333602	0.00260268251430529	0.636613946642521
124	135	-55.953698	-64.075774	16.244152	0.00270845046910417	0.659701149974659
196	246	-169.998866	-161.881753	16.234226	0.00272044501650903	0.659850199192588
163	220	-153.555007	-145.443602	16.22281	0.00273430476553704	0.660448530242427
172	217	-66.506317	-58.402986	16.206662	0.0027540282814581	0.662204794132106
190	196	-70.953145	-79.052244	16.198198	0.00276442229049456	0.662204794132106
157	347	-63.957638	-55.865698	16.18388	0.00278209317697564	0.66369523238386
452	529	-73.13894	-65.082841	16.112198	0.00287224723577861	0.679608866359536
135	242	-36.093561	-28.052785	16.081552	0.00291166269367737	0.686134497367794
106	146	-48.26613	-40.234117	16.064026	0.00293444267257603	0.688703002952359
142	172	-86.1186	-94.111434	15.985668	0.00303845546913251	0.705678222040266
131	275	-65.896907	-73.889338	15.984862	0.00303954399487061	0.705678222040266
260	347	-69.098607	-61.107562	15.98209	0.00304329059013397	0.705678222040266
124	384	-51.382023	-59.367177	15.970308	0.00305926575782245	0.706556318649272
10	275	-97.113217	-89.146958	15.932518	0.00311106406638095	0.713800871972872
163	375	-55.088634	-47.123892	15.929484	0.00311525997255713	0.713800871972872
57	226	-117.124914	-109.166056	15.917716	0.00313158743115161	0.714717021196294
209	356	-63.999716	-56.060599	15.878234	0.00318698425077535	0.722812995418078
135	160	-34.795395	-26.858048	15.874694	0.00319199804773207	0.722812995418078
124	146	-67.947634	-60.030045	15.835178	0.00324849527740056	0.732744246034672
121	157	-71.215349	-79.122179	15.81366	0.00327967285186859	0.736909438848147
142	246	-167.870281	-160.005454	15.729654	0.00340423315700045	0.761581606145444
133	190	-97.999831	-105.86085	15.722038	0.0034157532792447	0.761581606145444
173&323	197	-47.027883	-54.875769	15.695772	0.00345577878323144	0.76636948952601
33	260	-55.754481	-47.911097	15.686768	0.00346960546209574	0.76636948952601
121	124	-81.04241	-88.88343	15.68204	0.0034768876271406	0.76636948952601
155	244	-30.361867	-38.194334	15.664934	0.00350336047307531	0.769279570546121
157	248	-42.982537	-35.198709	15.567656	0.00365771846383844	0.80014316735364
82	131	-34.458742	-26.688049	15.541386	0.00370053758236977	0.800743716843867
121	131	-59.62614	-67.390754	15.529228	0.003720521168737	0.800743716843867
137	159	-47.792996	-55.556355	15.526718	0.00372465994570903	0.800743716843867
57	92	-88.344692	-80.582992	15.5234	0.00373013799529642	0.800743716843867
82	375	-57.992648	-50.234937	15.515422	0.00374334219880434	0.800743716843867
83	275	-56.443374	-64.190091	15.493434	0.00377997240351158	0.805606618498405
172	260	-75.107703	-67.368664	15.478078	0.00380576304390134	0.808132174560297
146	347	-61.728124	-53.999149	15.45795	0.00383983052194214	0.812390421010898
137	145	-72.056847	-64.335375	15.442944	0.00386542373527898	0.81483132339681

Continued on next page

Table 8 -Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
138	156	-257.771099	-250.056935	15.428328	0.0038905130048531	0.817148691635269
94	159	-30.908048	-23.204889	15.406318	0.00392859691159841	0.817506550518829
121	229	-240.06565	-232.367633	15.396034	0.00394651672317781	0.817506550518829
94	479	-55.115062	-62.812782	15.39544	0.00394755421606141	0.817506550518829
202	299	-54.730142	-47.033058	15.394168	0.00394977681998065	0.817506550518829
142	175	-74.83521	-67.141825	15.38677	0.00396272797474195	0.817506550518829
94	242	-25.239887	-17.563186	15.353402	0.00402166367405121	0.826722848172868
145	189	-50.169128	-42.522699	15.292858	0.00413080656592058	0.846158503980268
82	479	-63.230761	-70.866305	15.271088	0.00417075880973317	0.847728038740991
54	242	-45.197408	-37.563457	15.267902	0.00417663751623298	0.847728038740991
219	450	-110.984447	-118.616855	15.264816	0.00418233947006941	0.847728038740991
304	452	-79.56105	-71.934924	15.252252	0.00420563286092079	0.849479222813862
160	190	-32.237537	-39.858996	15.242918	0.00422302040104172	0.85002948836246
131	142	-62.348153	-69.959288	15.22227	0.00426173524646933	0.851919733793854
202	386	-94.210581	-86.599464	15.222234	0.00426180304985713	0.851919733793854
135	189	-31.296099	-23.698522	15.195154	0.00431310774307248	0.859212563113098
137	275	-79.92529	-87.51611	15.18164	0.00433893736023183	0.860330350979068
106	173&323	-25.962985	-18.376294	15.173382	0.00435479582101495	0.860330350979068
163	275	-61.425488	-69.009986	15.168996	0.00436324173172065	0.860330350979068
25	155	-28.866135	-21.287199	15.157872	0.00438473493320102	0.860910419108941
124	226	-120.400477	-127.976538	15.152122	0.00439588552796699	0.860910419108941
226	231	-108.30332	-100.737435	15.13177	0.00443557710712461	0.8657397383858
201	452	-87.528315	-79.967121	15.122388	0.00445399276918035	0.8657397383858
131	163	-31.985212	-24.427338	15.115748	0.00446707161298188	0.8657397383858
201	225	-79.175126	-71.623801	15.10265	0.00449298153648703	0.8657397383858
92	196	-103.242272	-110.793034	15.101524	0.00449521582291057	0.8657397383858
82	156	-114.448471	-106.905559	15.085824	0.00452648279179213	0.868874859073476
51	140	-27.99493	-20.456644	15.076572	0.00454500834659843	0.869551596872314
163	197	-51.216915	-58.746908	15.059986	0.00457840562108636	0.873059782415711
146	172	-69.695959	-62.172157	15.047604	0.00460349484717526	0.87436753389091
226	229	-279.335161	-271.816238	15.0378460000001	0.00462336226496685	0.87436753389091
193	202	-74.95527	-67.438098	15.034344	0.00463051290157857	0.87436753389091
54	450	-86.146326	-78.658884	14.974884	0.00475359387992746	0.894694276686348
83	190	-26.728921	-34.208945	14.960048	0.00478480158961803	0.897653553884003
155	453	-42.586713	-50.05972	14.946014	0.0048145074367405	0.900312890670474
275	384	-64.530687	-71.990203	14.919032	0.00487213013360921	0.908065314243136
124	219	-100.050685	-107.506677	14.911984	0.00488729304888491	0.908065314243136
57	223	-84.14084	-76.698188	14.885304	0.00494511270361342	0.912955998179841
172	276	-65.09222	-72.5239	14.86336	0.00499317170162905	0.918902106487099
63	299	-37.507042	-30.080465	14.853154	0.00501567956499505	0.920123241717604
57	186	-124.452017	-131.833345	14.762656	0.00521966724385314	0.95452400670715
82	106	-35.67608	-28.314866	14.722428	0.00531293764369445	0.968525142153985
33	271	-35.925134	-28.574097	14.702074	0.0053607505954899	0.972459371626231
220	248	-152.637438	-145.287951	14.698974	0.00536806967259607	0.972459371626231
50	172	-102.095461	-94.753884	14.683154	0.00540557358263227	0.973502643058585
146	197	-70.425473	-63.084283	14.68238	0.00540741506063247	0.973502643058585
299	452	-57.488652	-50.15792	14.661464	0.00545741157288859	0.979461761239479
173&323	213	-30.960142	-23.641103	14.638078	0.00551384952353073	0.986536595305792
157	163	-42.781998	-35.474937	14.614122	0.00557225687010077	0.993919171568435
83	122	-16.342897	-9.042956	14.599882	0.00560726279606694	0.997095166527608

Continued on next page

Table 8 -Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
157	453	-59.019873	-51.728437	14.582872	0.00564936107038549	0.998627784066229
155	160	-25.028745	-32.319985	14.58248	0.00565033488310718	0.998627784066229
33	92	-72.218032	-64.932801	14.570462	0.00568027025204132	0.9999999999955
143	248	-27.267079	-19.987242	14.559674	0.00570727434003271	0.9999999999955
220	453	-169.875305	-162.59974	14.55113	0.00572875068182932	0.9999999999955
193	219	-105.948183	-113.213606	14.530846	0.00578005494456268	0.9999999999955
196	529	-90.932077	-83.678513	14.507128	0.00584061667350455	0.9999999999955
275	278	-60.252977	-67.505748	14.505542	0.00584468850953412	0.9999999999955
164	172	-77.033793	-69.812251	14.443084	0.00600727396046807	0.9999999999955
196	347	-82.136695	-74.922972	14.427446	0.00604867095490935	0.9999999999955
135	541	-28.532892	-21.328924	14.407936	0.00610071037092097	0.9999999999955
75	146	-67.529937	-60.33342	14.393034	0.00614075431316519	0.9999999999955
124	220	-179.802458	-172.608846	14.387224	0.00615643640992691	0.9999999999955
133	163	-85.137387	-92.326747	14.37872	0.00617946086947774	0.9999999999955
219	426	-99.656838	-92.480819	14.352038	0.00625225148398223	0.9999999999955
155	172	-54.02349	-61.190159	14.333338	0.00630376701491686	0.9999999999955
54	146	-60.15157	-52.998617	14.305906	0.00638009171022702	0.9999999999955
63	160	-26.006799	-18.858222	14.297154	0.00640463271205716	0.9999999999955
135	219	-82.619265	-89.739675	14.24082	0.00656482548319859	0.9999999999955
122	384	-19.61421	-26.730605	14.23279	0.00658797731585026	0.9999999999955
220	384	-156.740571	-149.626297	14.228548	0.0066002399998899	0.9999999999955
138	142	-204.381318	-197.280048	14.20254	0.0066759139649003	0.9999999999955
197	242	-56.047905	-48.964596	14.166618	0.00678183420708611	0.9999999999955
122	160	-21.372697	-14.294871	14.155652	0.0068144956456776	0.9999999999955
172	347	-71.159305	-64.088703	14.141204	0.00685776341484157	0.9999999999955
190	248	-28.417982	-35.486246	14.136528	0.00687182432451994	0.9999999999955
244	276	-48.214431	-41.146367	14.136128	0.0068730284506795	0.9999999999955
53	278	-89.373057	-82.309062	14.12799	0.0068975713423286	0.9999999999955
196	453	-77.860993	-70.799972	14.122042	0.00691556385341208	0.9999999999955
172	226	-120.953241	-128.005738	14.104994	0.00696738871790314	0.9999999999955
131	304	-65.466472	-58.417651	14.097642	0.00698985561474619	0.9999999999955
57	121	-77.922398	-70.874583	14.09563	0.00699601642551784	0.9999999999955
106	135	-36.089592	-29.047467	14.08425	0.00703096264802816	0.9999999999955
160	219	-74.502493	-81.542656	14.080326	0.00704305225809942	0.9999999999955
155	384	-27.9318	-34.962247	14.060894	0.00710322191274571	0.9999999999955
83	145	-49.581557	-42.556249	14.050616	0.00713525027446926	0.9999999999955
172	529	-81.075068	-74.052081	14.045974	0.0071497620426999	0.9999999999955
2	155	-37.541461	-44.560506	14.03809	0.00717447520251169	0.9999999999955
242	331	-23.549119	-30.546877	13.995516	0.00730938055947039	0.9999999999955
126	193	-52.890505	-45.899921	13.981168	0.00735540286939074	0.9999999999955
106	155	-33.576133	-26.591511	13.969244	0.00739386587593449	0.9999999999955
172	219	-101.114381	-108.09349	13.958218	0.00742960749346167	0.9999999999955
331	384	-28.597205	-35.570754	13.947098	0.0074658252008365	0.9999999999955
156	384	-112.828685	-119.796767	13.936164	0.00750160569561331	0.9999999999955
160	201	-63.752392	-70.700896	13.897008	0.00763112297862256	0.9999999999955
82	242	-32.522742	-25.591032	13.86342	0.00774396575789293	0.9999999999955

Continued on next page

Table 8 -Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	P-value	P-value (FDR)
25	193	-59.454494	-52.53825	13.832488	0.0078493296237222	0.9999999999955
275	509	-57.4627	-50.551978	13.821444	0.00788728787322224	0.9999999999955
146	384	-41.72992	-48.639894	13.819948	0.0078924434293498	0.9999999999955
199	247	-65.406967	-58.507465	13.799004	0.00796496857918394	0.9999999999955
82	226	-99.63263	-92.734582	13.796096	0.00797508988390194	0.9999999999955
138	226	-242.381419	-235.484437	13.793964	0.00798251831619945	0.9999999999955
142	452	-81.076979	-74.185702	13.782554	0.00802238894678764	0.9999999999955
194	226	-145.435549	-138.544347	13.782404	0.00802291439669367	0.9999999999955
126	244	-27.433497	-34.32111	13.775226	0.00804809834599707	0.9999999999955
160	384	-25.484145	-32.36505	13.76181	0.00809537591497678	0.9999999999955
220	450	-190.460282	-183.585038	13.750488	0.0081354857282594	0.9999999999955
106	142	-65.317115	-58.454194	13.725842	0.00822347151126335	0.9999999999955
138	248	-168.733679	-161.872729	13.7219	0.00823763051129656	0.9999999999955
159	453	-46.472956	-53.324344	13.702776	0.00830666006438074	0.9999999999955
160	347	-44.814802	-37.964372	13.70086	0.00831360713550122	0.9999999999955
155	189	-21.071808	-27.91832	13.693024	0.00834207829043065	0.9999999999955
144	158	-27.55264	-34.385354	13.665428	0.00844310679467353	0.9999999999955
426	427	-61.807167	-54.976437	13.66146	0.00845773168901576	0.9999999999955
94	260	-44.074939	-37.244767	13.660344	0.00846184940515771	0.9999999999955
83	375	-51.343718	-44.527575	13.632286	0.00856602235870729	0.9999999999955
347	479	-79.971252	-86.782867	13.62323	0.00859991231801649	0.9999999999955
207	534	-49.252066	-42.451926	13.60028	0.00868638590255322	0.9999999999955
63	472	-21.409695	-14.612078	13.595234	0.00870551256512819	0.9999999999955
57	312	-68.943744	-75.736721	13.585954	0.00874079566293218	0.9999999999955
83	278	-20.432775	-13.655385	13.55478	0.00886034838818517	0.9999999999955
94	155	-25.260075	-32.034964	13.549778	0.00887967946810242	0.9999999999955
275	347	-82.205447	-88.976397	13.5419	0.00891020902684159	0.9999999999955
131	450	-65.606709	-72.374406	13.535394	0.00893549907317859	0.9999999999955
138	375	-199.776984	-193.016337	13.521294	0.00899054964298773	0.9999999999955
49	328	-57.073112	-50.313366	13.519492	0.00899760904776725	0.9999999999955
54	193	-74.387307	-81.14572	13.516826	0.00900806315130598	0.9999999999955
271	299	-46.499555	-39.741173	13.516764	0.00900830641129602	0.9999999999955
94	192	-69.05857	-62.303105	13.51093	0.00903122516243138	0.9999999999955
194	220	-204.799787	-198.048616	13.502342	0.00906506675038055	0.9999999999955
202	452	-62.267793	-55.521928	13.49173	0.00910705539793133	0.9999999999955
196	229	-243.459521	-236.717326	13.48439	0.0091362088982683	0.9999999999955
190	229	-214.765556	-208.041104	13.448904	0.00927844678736212	0.9999999999955
146	450	-78.511167	-71.791849	13.438636	0.00932000622995632	0.9999999999955
57	217	-65.101674	-58.383247	13.436854	0.00932723733819962	0.9999999999955
156	347	-130.002894	-123.288134	13.42952	0.00935705548519916	0.9999999999955
271	529	-61.85077	-55.1434	13.41474	0.00941743061537137	0.9999999999955
260	452	-69.189044	-62.482096	13.413896	0.00942088976152655	0.9999999999955

Continued on next page

Table 8 -Continued from previous page

site-1	site-2	lkl-indep	lkl-dep	test-stat	<i>P</i> -value	<i>P</i> -value (FDR)
142	201	-99.514968	-92.836301	13.357334	0.0096555631683235	0.9999999999955
157	222	-54.129263	-47.451169	13.356188	0.00966037645785334	0.9999999999955
197	386	-103.343478	-96.674705	13.337546	0.00973900452533127	0.9999999999955
53	384	-92.442923	-99.11064	13.335434	0.00974795185568822	0.9999999999955
10	34	-61.255058	-54.591963	13.32619	0.00978720801200927	0.9999999999955
131	529	-59.485435	-52.834357	13.302156	0.00988999689741887	0.9999999999955
260	299	-58.561035	-51.915353	13.291364	0.00993649480352199	0.9999999999955
63	124	-51.664336	-45.019507	13.289658	0.00994386472616582	0.9999999999955

Table 5.9: Detailed results generated using the Adam03M2 data set. Site-1 and site-2 show the AA positions along the alignment. The lkl terms show the log-likelihood values after optimization under the independent and the dependent model. The test statistic is twice the log-likelihood difference between the two models. The *P*-values are computed using the standard χ^2 approximation. FDR-corrected *P*-values follow BH. Only pairs significant at the nominal 1% level (*P*-value) are shown; those above the horizontal line inside the table are also significant after FDR.

site-1	site-2	lkl-indep	lkl-dep	test-stat	<i>P</i> -value	<i>P</i> -value (FDR)
31	51	-215.103387	-197.277061	35.652652	3.41E-07	0.00113281616036631
12	53	-77.799957	-65.876321	23.847272	8.57E-05	0.142316180090207
13	56	-46.536317	-40.493356	12.085922	0.0167236344812858	0.999999964628273
11	58	-29.888689	-23.892009	11.99336	0.0174007103450287	0.999999964628273
13	54	-71.386993	-66.032614	10.708758	0.0300397186352879	0.999999964628273
6	70	-17.476614	-12.309403	10.334422	0.0351558484101363	0.999999964628273
5	8	-12.900464	-7.964634	9.87166	0.0426459429703552	0.999999964628273
13	51	-168.765715	-164.127949	9.275532	0.0545696117650336	0.999999964628273
63	75	-16.839233	-12.370825	8.936816	0.0626983299054582	0.999999964628273
40	63	-17.108515	-12.661735	8.89356	0.0638157412109532	0.999999964628273
24	52	-58.933899	-54.596959	8.67388	0.0697883949162066	0.999999964628273
4	90	-32.302657	-27.98877	8.62777400000001	0.0711074320946067	0.999999964628273
75	96	-17.939049	-13.701758	8.474582	0.0756614405570145	0.999999964628273
5	10	-24.371314	-20.161276	8.420076	0.0773472039538335	0.999999964628273
16	17	-39.158191	-34.956543	8.403296	0.0778732712198347	0.999999964628273
75	77	-24.809162	-20.607664	8.402996	0.0778827071170962	0.999999964628273
50	88	-56.048288	-51.902973	8.290629999999999	0.0814937940632067	0.999999964628273
14	68	-47.801338	-43.671019	8.260638	0.0824839822106254	0.999999964628273
12	51	-173.532531	-169.559639	7.945784	0.0935845065252358	0.999999964628273
15	94	-27.233707	-23.320416	7.826581999999999	0.0981412984366585	0.999999964628273
9	55	-22.540834	-18.64306	7.795548	0.0993612397204912	0.999999964628273