HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

MASTER THESIS

Mr. B.Sc. Ing.
Andre Marquardt

Identification and Characterization of Mammalian Signatures of Viral Adaptation: A Computational Approach

2015

Department
Mathematics / Science / Computer Science

MASTER THESIS

Identification and Characterization of Mammalian Signatures of Viral Adaptation: A Computational Approach

Author:
Andre Marquardt

Degree Course:
Molekularbiologie/Bioinformatik

Tutorial Group:
MO13w1-M

First Examiner:
Prof. Dr. rer. nat. Röbbe Wünschiers

Second Examiner:
Prof. Dr. rer. nat. Alice Carolyn McHardy

Mittweida, August 2015

Bibliographic Description

Marquardt, Andre: Identification and Characterization of Mammalian Signatures of Viral Adaptation: A Computational Approach, 99 pages, 21 figures, 32 tables, Hochschule Mittweida, University of Applied Sciences, Department
Mathematics / Science / Computer Science

Master Thesis, 2015

Satz: $\LaTeX$

Referat

I. Contents

II. List of Figures

III. List of Tables

Acronyms

ssRNA single-stranded RNA

HA Haemagglutinin

NA Neuraminidase

MSA Multiple Sequence Alignment

PB1 Polymerase basic protein 1

PB2 Polymerase basic protein 2

M1 Matrix protein 1

M2 Matrix protein 2

M42 M2-like protein

PA Polymerase acidic protein

NP Nucleoprotein

NS1 Nonstructural protein 1

NEP Nuclear export protein

TP True Positive

FP False Positive

TN True Negative

FN False Negative

FTP File Transport Protocol

NCBI National Center for Biotechnology Information

MUSCLE multiple sequence comparison by log-expectation

IPoSuS Identify Patches of Sites under Selection

AP AdaPatch

OR OmegaRatios

OV OmegaValues

NG+RF Nei Gojoborj + random Fitch

NG+NRF Nei Gojoborj + nonrandom Fitch

PDB Protein Data Bank

Abstract

Influenza viruses are single-stranded RNA (ssRNA) viruses, which are divided into three distinct genera: A,B and C. Their genome is divided into eight segments. Whilst the influenza types B and C evolve slowly, viruses of the type A evolve very fast, causing mild to severe infections, and are a constant harm for the human race. Beside the usual genetic mutations for altering the genetic information, the genomes content can randomly be altered by reassortment events. In this case a host cell has to be co-infected by two (or even more) influenza viruses, which emerge as a new virus containing segments of both (all co-infecting) ancestors.

Beside this special case of reassortment events the influenza A virus already has a very high mutation rate. The reason of the fast genetic alteration and the resulting evasion of the host immune system, is a proofreading-lacking polymerase. Especially genetic alterations in one of the two major surface glycoproteins - Haemagglutinin (HA) - can have massive influence considering the ability of the virus to infect people. This protein shows preferred amino acids that are under extreme selective pressure. Additionally HA is of substantial importance for infecting host cells. Genetic alterations in this protein is one reason influenza A viruses are constantly able to evade the host immune system, because they are targeted by antibodies. Exogenous materials are specifically recognized by the host immune system and is very specific for some surface amino acids or their properties. Already little changes in sites known to be important for evading host responses can cause the evasion of the virus, because the binding and therefore the inactivation through antibodies is affected. The high mutation rate of the influenza A virus, especially in the HA protein causes the need for almost annual vaccinations.

Changes in these preferred amino acids are involved in adaption to an increasingly immune population and are of major interest, because they provide the ability to reinfect the population.

We want to establish a fully automated framework for data download and determining sites of proteins under selection, utilizing a user-friendly and user-individual input. Combining existing tools into a user-friendly pipeline will make it more easy for biologists to find sites under selection.

In this work we introduce such a pipeline, called IPoSuS (**I**dentify **P**atches **o**f **S**ites **u**nder **S**election) and use it for analysis of the influenza A virus protein HA. Furthermore, IPoSuS can be applied onto every single dataset and protein with given sequences and according background data. Based on already existing datasets for evaluation we additionally tested new statistical approaches to find sites under positive selection, which makes it possible to not only use the gold standard, but also ω-values, including more information than only counts of synonymous and non-synonymous mutations, to make the results more convincing and more factful.

Using IPoSuS for the protein HA of different influenza A virus subtypes results into some new findings regarding host and subtype specificities. The obtained results favor one of

the five approaches tested, namely the already established AdaPatch approach using a newly introduced counting scheme. But the results also confirm the possible usage of the approaches using ω-ratios and ω-values and the newly introduced statistical test. The only downside of these new approaches are the fewer amount of results, compared to the established and favored one.

Zusammenfassung

Influenza Viren sind einzelsträngige RNA Viren, welche in drei unterschiedliche Typen unterteilt werden: A, B und C. Sie alle besitzen ein in acht Segmente unterteiltes Genom. Während Viren vom Typ B und C eher langsam im Menschen evolvieren, evolvieren Viren vom Typ A sehr schnell und verursachen milde bis schwere Erkrankungen. Sie sind daher eine ständige Gefahr für den Menschen. Neben den üblichen genetischen Mutationen zur Veränderung des Erbguts hat Influenza die Fähigkeit zur Reassortierung. In diesem Fall muss eine Wirtszelle gleichzeitig von zwei (oder mehr) unterschiedlichen Influenza Viren befallen sein, welche als neuer Virus von der Wirtszelle hervorgebracht werden und Segmente von beiden (allen) gleichzeitig infizierenden Vieren enthalten.

Neben diesem speziellen Fall von Reassortierungen haben Influenza A Viren bereits eine hohe Mutationsrate. Ursache für die schnelle genetische Veränderung und die damit verbundene Umgehung des Immunsystems des Wirts, ist eine Polymerase der eine *proofreading*-Untereinheit fehlt. Speziell genetische Veränderungen in einem der beiden Oberflächenglykoproteinen - Hämagglutinin (HA) - können massive Auswirkungen auf die Fähigkeit des Virus haben, Menschen zu infizieren. Dieses Protein zeigt bevorzugte Aminosäuren auf, welche unter extremen Selektionsdruck stehen. Genetische Veränderungen in diesem Proteinen ist einer der Hauptgründe dafür, dass Influenza A Viren das Immunsystems des Wirts umgehen können, da dieses bevorzugt von Antikörpern erkannt wird. Exogenes Material wird sehr spezifisch durch das Immunsystem erkannt und ist sehr speziell auf die Oberläche einiger Aminosäuren und deren Eigenschaften angepasst. Schon kleine Veränderungen an bekannten, wichtigen Stellen können zum Umgehen der Immunantwort des Wirts führen, da die Bindung und damit die Markierung durch Antikörper beeinträchtig ist. Die hohe Mutationsrate von Influenza A Viren, speziell im HA protein, ist der Grund für die Notwendigkeit von jährlichen Impfungen. Da Veränderungen in diesen bevorzugten Aminosäuren an der Adaption und Reinfektion an die zunehmend immunisierten Gesellschaft beteiligt sind, sind sie von größtem Interesse.

Wir wollen nun einen vollautomatischen Arbeitsablauf für den Datendownload und weitere Analysen etablieren, welcher Bereiche des Proteins bestimmt, die sich unter Selektionsdruck befinden und dabei eine benutzerfreundliche und individuelle Eingabe ermöglicht. Die Kombination aus schon existierenden Tools und einem automatischen, benutzerfreundlichen Arbeitsablauf, wird es jedem Biologen ermöglichen Bereiche unter Selektionsdruck zu finden.

In dieser Arbeit führen wir genau einen solchen Arbeitabslauf ein, genannt IPoSuS (**I**dentify **P**atches **o**f **S**ites **u**nder **S**election), und verwenden ihn für Analysen des HA Proteins des Influenza A Viruses. Daneben bietet IPoSuS auch die Möglichkeit für jeden möglichen Datensatz und jedes Protein verwendet werden zu können, sofern die Sequenzen gegeben und dazugehörende Hintergrunddaten vorhanden sind. Auf Grundlage von schon bestehenden Datensätzen zur Evaluation wurden zusätzlich neue sta-

tistische Ansätze zur Auffindung von Positionen unter selektiven Druck getestet, welche es möglich machen, nicht die Standardmethode verwenden zu müssen, sondern auch ω-Werte, welche mehr Informationen beinhalten als die Anzahl von synonymen und nicht synonymen Mutationen, um das Ergebnis aussagekräftiger und überzeugender zu machen.

Die Verwendung von IPoSuS für das HA protein von verschiedenen Subtypen des Influenza A viruses resultierte in einigen neuen Erkenntnissen bezüglich der Wirts und Subtypspezifität. Die erhaltenen Resultate bevorzugen einen der fünf getesteten Ansätze, nämlich den bereits etablierten AdaPatch ansatz mit einer neu eingeführten Zählmethode. Aber die Ergebnisse bestätigt gleichzeitig die Verwendbarkeit der Ansätze, die ω-ratios und ω-Werte und den neu eingeführten statistischen Test verwenden. Der einzige Nachteil dieser neuen Ansätze ist die verhältnismäßig kleinere Anzahl an Ergebnissen, verglichen zum etablierten und favorisierten Ansatz.

1 Introduction

This chapter serves the purpose of understanding of the following work and is needed to understand the thesis as a whole, containing crucial information about the virus and all used bioinformatic and mathematic tools or methods.

1.1 Motivation

Among diseases caused by viral infection one of the most important is caused by influenza A viruses. Causing a respiratory disease this virus gets transmitted by droplets of body fluids, e.g. tears and salivary [56]. Infecting healthy individuals the influenza A virus causes a mild to severe disease resulting in around 5 million infections each year. Infections especially among the young and elderly can result in serious illness, causing around 500.000 deaths per year [47]. By recommending a vaccine strain every year the World Health Organization (WHO) reduces the health and the resulting worldwide economical burden. The vaccine grants a temporary immunity, which humans could not achieve in the last hundreds of years, or at least reduces the risk of getting infected by the current dominant influenza A virus strain, resulting in fewer infected individuals.

The single-stranded RNA (ssRNA) genome of the influenza A virus is divided into eight segments, which can be exchanged by reassortment. These reassortment events can result in huge antigenic differences and therefore can lead to fitness advantages compared to the previous dominant strain. In some cases the reassorted segments are from different influenza A viruses, which occur when a host cell is co-infected by more than one influenza A virus. Reassortment events are the main evolutionary mechanism beside the usual genetic mutations [47, 48, 56].

Reassortment events are quite rare compared to genetic mutations. The frequently occurring genetic alterations in the coding sequences are caused by the proofreading-lacking RNA-polymerase [44] of the influenza A viruses.

Proteins consist of two different main parts: A part that is highly conserved and plays an important role in protein stability or function and a second part that is highly variable. Genetic alterations in areas of the conserved part may significantly influence the stability of the protein and could lead to a misfolded and malfunctioned protein, because of altered binding sites or binding recognition sites. Minding this fact one may conclude that mutations, especially mutations getting fixed or become predominant, are not arbitrary. Furthermore it is noticeable that relevant mutations to influenza A viruses are often occurring in predefined areas, proven by wet lab experiments and are often adjacent to highly conserved regions [34]. In terms of influenza A viruses HA is an example. HA is one of the two major surface glycoproteins of influenza A viruses and plays a pivotal role in host cell invasion, further explained in Section 1.2.2. Alterations in this protein can lead to host immune system evasion and altered pathogenicity and mainly appear near the conserved receptor binding site of the protein. However, mutations in other proteins or segments can yield altered pathogenicity too [70], but detecting the amino acids that are altered more favored could yield better understanding of the evolution of influenza A viruses and ideally better vaccine-strain selection. This evolutionary aspect becomes important considering the two parts - conserved and variable - of a protein, because the preferred changed amino acids are under constant selective pressure. Detecting such sites under selection are of major interest, because they are responsible for the constant evasion of the hosts immune system.

HA is not the only protein exposed on the outer surface of the membrane of the influenza virus and is therefore not the only protein of interest. Additionally to the factor of exposure all proteins have to act jointly to yield optimal results and alterations in every protein are therefore of interest. The knowledge about sites under selection in different proteins of the virus could lead to new insights about the needed adaptive changes for host switches and host adaption. Having detailed information about the sites predominantly involved in immune system evasion or host specific adaption could advance treatment as much as further the general understanding of influenza A infections, because of better vaccine strain selection or the possibility for a longer lasting vaccine, but also an improved understanding in cross species infections.

There are several approaches to determine sites under selection, most mentionable are the ones of Nei and Gojoborj [55], Bush [9] and Suzuki [75]. All these approaches have in common that there is no automated framework, usable for every biologist, which makes it hard to use the approaches on own data.

Establishing an automated framework can be used as a quick and "easy-to-use" tool for detecting sites under selection in every protein of influenza A viruses. It could be utilized to gain a more detailed view on sites under selection, according to different years and strains, but it could also improve the understanding of the diversification of sites under selection. An automated framework would make it is easy to get comparable but also a high amount of different results and data for further investigations and comparisons.

In general IPoSuS identifies patches of sites under selection for input sequences and 3D-structure, where a patch is a pooling of different amino acid positions. In this work we focus on the HA protein of influenza A viruses of different subtypes to get a more detailed view of the changes of this protein. Nevertheless the aim is to introduce a tool that can be applied on every protein.

1.2 Influenza A Viruses

Influenza viruses belong to the family Orthomyxoviridae and can be divided into 3 distinct genera - A, B and C. All these viruses cause influenza, which is commonly known as "the flu" [83]. Infections mainly affect the upper respiratory tract, but can also influence the bronchi, depending on the moieties recognized by the virus (see Chapter 1.3). Symptoms can be either mild or severe [14] and commonly include high fever, a runny nose or headache. Many other symptoms can occur beside these mentioned. However, 33 percent of influenza infections show no symptoms at all [12].

Since influenza is a droplet infection, there are three main ways how influenza viruses can be transmitted. The first way is by direct transmission, such that an infected person directly sneezes into someone eyes, nose or mouth. The second way is the air borne route, which means that the produced aerosol, that gets exhaled of an infected person through coughing or sneezing, is inhaled. The third and last possible way of getting infected is through hand-to-eye, hand-to-nose or hand-to-mouth contact with contaminated surfaces.

Influenza A viruses are able to infect humans, not only by direct contact, but also by using the airborne route of infection. Because of this ability and their fast genetic alteration, influenza A viruses are much more threatening to human health than influenza viruses of type B and C. Nevertheless, humans are not the only species that can be infected by influenza A viruses. The natural reservoir of influenza A viruses are birds and waterfowls, but swines, cats, dogs and horses can be infected by influenza A viruses as well [81].

Influenza A viruses can be of either spheric or of filamentous appearance [17, 39] and are about 120 nm in diameter [15, 39]. The virus contains a negative-sense ssRNA genome, which is divided into eight segments with a total genome length of around 14,000 nucleotides [5]. Each segment encodes for at least one protein. The segments are numbered ascending and starting with the longest [46] due to the nomenclature conventions. Some segments are encoding for more than one protein, due to splicing, resulting in 14 different proteins for each influenza A virus. All proteins with their function are combined in Table 1.1.

Table 1.1: Detailed view of the viral segments and their encoded proteins and their function. Modified from [7] with added information from [48, 53, 85].

Segment	Encoded protein(s)	Protein function
1	Polymerase basic protein 2 (PB2)	mRNA cap recognition
2	Polymerase basic protein 1 (PB1) PB1-F2 PB1-N40	RNA elongation, endonuclease activity Pro-apoptotic activity Unknown [53]
3	Polymerase acidic protein (PA) PA-X	Protease activity Modulates host response [53]
4	Haemagglutinin (HA)	Major antigen, receptor binding and fusion activities
5	Nucleoprotein (NP)	Nuclear import regulation
6	Neuraminidase (NA)	Sialidase activity, virus release
7	Matrix protein 1 (M1) Matrix protein 2 (M2) M2-like protein (M42)	vRNP interaction, RNA nuclear export, viral budding Virus uncoating and assembly Functional complementary to M2 [85]
8	Nonstructural protein 1 (NS1) Nuclear export protein (NEP)	Regulation of host gene expression Nuclear export of RNA

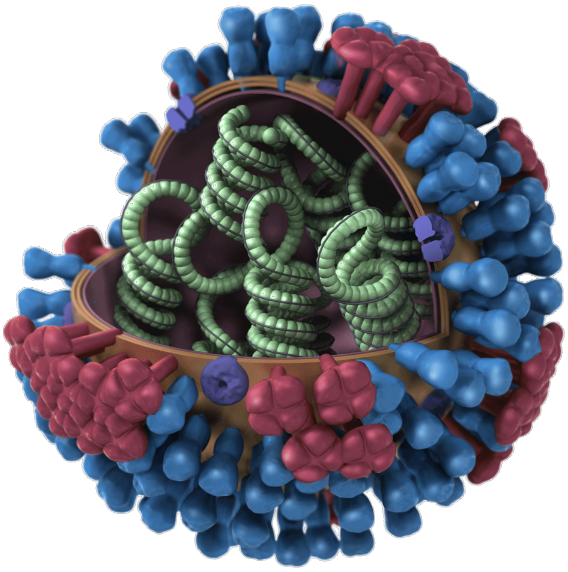

Figure 1.1: Figure showing the whole influenza A virus including the ratio of HA (in blue) to NA (in red) and the inner content (green) of the virus. The M2 ion-channel is shown in purple. Picture taken from http://www.cdc.gov/flu/images.htm accessed: 16.08.2015 at 20:55.

The genome of influenza A viruses is inside of a viral envelope. This viral envelope contains a lipid-bilayer obtained from the host and covers the capsid containing the virus genome. On the surface of influenza A viruses there are two major surface glycoproteins - HA and NA - encoded by segments 4 and 6, respectively. HA and NA are present in an approximated ratio of four to one on the membrane of influenza A [7]. Figure 1.1 shows the structure of an influenza A virus.

There are 18 known subtypes of HA and eleven of NA [48,74,77,80]. Important for individual classification of each influenza A virus is the combination of HA and NA subtypes it encodes for [49]. Generally an influenza A virus is classified as HxNy, where x and y stand for the number of the subtype of HA and NA obtained of the virus. Viruses of type H3N2, as an example, encode for subtype three of HA and subtype two of NA. The whole nomenclature of isolated and sequenced viruses contains more information, than just the subtypes of HA and NA, also considering the year and origin of isolation. The whole annotation of a influenza A virus is described as: influenza virus strain / location of isolation / consecutive number by the WHO / year of isolation (subtype classification). As example: A / Sydney / 5 / 97 (H3N2). There are also strains that are categorized as pandemic by the WHO and get classified with a "p" in front of the usual classification, e.g. the previous pandemic pH1N1 strain.
Additionally HA and NA play important roles in the virus replication cycle (see Chapter 1.2.2).

1.2.1 Evolutionary Mechanisms

HA and NA are, as already mentioned before, located on the outer surface of the virus and can therefore be targeted by host antibodies and are both under extreme selective pressure. The constant evasion of the host immune system is facilitated through two mechanisms in influenza A viruses, namely genetic drift for minor changes and antigenic shift for greater changes [63].

The occurrence of minimal antigenic changes, due to mutations in HA and NA, is called antigenic drift. This drift can refer to either genetic or antigenic drift and denotes mutations in the RNA sequence of the virus. Influenza A viruses have a high mutation rate, leading to a fast evolution of the virus. This rapid evolution is caused by the prone RNA polymerase that misses a proofreading unit. Slight changes in the genes encoding for HA and NA can already have a high impact for the virus, because it can change the infectious capabilities. This evolution is occurring under the pressure of the host immune system and its antibodies [48, 63]. Because of the high amount of different possibilities to form new viral strains [58] that can be derived by drift, the virus is capable of infecting humans that already received a vaccine [27] or have been infected previously.

Greater changes can be achieved by antigenic shift, which describes the reassortment of segments between different and distinct influenza A subtypes, e.g. different serotypes. Reassortments can occur when a host cell is co-infected by two, or more, different and distinct influenza A viruses. Segments reassortments can also lead to viruses with segments from different parental viruses [48, 58]. This event can lead to viruses with mixed genes from strains that infect different species [63]. Influenza A viruses generated by a reassortment event are often able for cross species infections and host switches, which means the virus has the possibility to infect other hosts than the foregone, due to the new combination of segments.

1.2.2 Replication Cycle

A host is mandatory for the replication of influenza A viruses. The first step for invading a host cell is the recognition of sialic acid (SA) moieties on the surface of host cell membranes and the subsequent binding to it [7]. Responsible for this initial binding to the host cell is the viral HA.

The main activity of HA is to bind sialic acids and to induce the fusion between virus and host cell membrane. The initial binding is mediated by a subunit of HA. This subunit, the head of HA, is also called HA1 and induces a change in conformation which triggers the fusion of the virus membrane with the host membrane [33]. But before this fusion can occur, the pH-value has to be lowered, to around 6.0. Because of the lowered pH-value the HA1 subunit gets protonated, causing a positive charge. Ultimately, this results in HA1 subunits repelling each other and detach from the stem. The stem is also called HA2 and is the second subunit of HA. By detaching the HA1 subunits from the HA2 subunits, the HA2 subunits get activated [32] and triggers the fusion of the two membranes, by partially unfolding and releasing a hidden hydrophobic portion, which functions as a hook. This activation is unique and the virus looses its pathogenicity after the first activation.

The second step after the HA1 and HA2 mediated fusion is the entry into the host cell. Through receptor-mediated endocytosis [48] the virus enters the host cell. After the fusion of viral envelope and host membrane the content of the virus is released and translocated to the nucleus of the host cell. The infiltrated virus' segments and mRNA are then translated and replicated by the host. Subsequently the newly produced proteins are released into the cytoplasm. The trans-golgi secretory pathway then transports the new translated proteins to the membrane. The formation of the new virus is induced, after the mature HA, NA and M2 proteins arrive there, assisted by M1 [48]. The second major surface glycoprotein -NA - is now required for releasing the progeny viruses from the host cell [7]. Responsible for this release is the cleavage of the glycosidic linkages of neuraminic acids [66] initially formed by HA. Before the NA activity cuts the sialic acid binding and therefore releases the new virus, it gets newly arranged and packed into a part of the host cell membrane. For illustration see Figure 1.2.

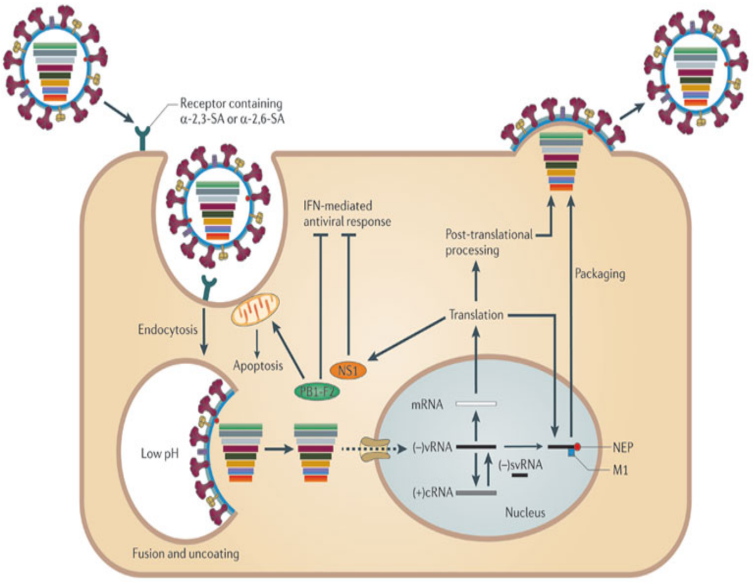

Figure 1.2: Figure showing the replication cycle from influenza A viruses. Starting from the attachment onto the membrane through Haemagglutinin (HA), over the endocytosis and the virus protein replication, ending with the release of new viruses. Taken from [48]

1.2.3 Tropism

Furthermore HA is important for host tropism [7]. This means that every host has its special fundamentals. This is of importance because, as already mentioned in Chapter

1.2, influenza A viruses may infect a broad variety of different hosts. This variety can be explained by the occurrence of special sialic acid moieties on the surface of host cells, recognized by HA. There are two possible ways of linkage between galactose, located on the outer surface of the membrane, and the sialic acid. It is either a α-2,3 or a α-2,6 linkage. A α-2,3 linkage refers to a linked carbon atom at position two in the sialic acid and the carbon atom at position three in the hexose of galactose. The α-2,6 linkage refers to a binding between the second carbon atom of SA and the sixth carbon atom of galactose.

On the one hand there are waterfowl's, which are the natural reservoirs for influenza A viruses [72], but their respiratory tract does only contain α-2,3-SA receptors, which is very specific. On the other hand, swines are mentioned as mixing vessels [28] for influenza A viruses, because their respiratory tract contains another receptor beside the α-2,3-SA aforementioned α-2,6-SA. Humans also have both receptors, but in different areas. The α-2,6-SA receptor is located in nose and throat whereas the α-2,3-SA can be found in the bronchi. Another interesting fact considering host tropism is the prevalent temperature in this mentioned areas. In waterfowl's the temperature in the receptor containing area is around $40°$C whereas the temperature in swines is around $39°$C - for both receptors - and in humans they are around $33°$C for the α-2,6-SA receptor area and around $37°$C for the α-2,3-SA receptor area, respectively (Figure 1.3) [48].

Because of these host dependencies the infections with influenza A viruses cause different symptoms. As already mentioned in Chapter 1.2 influenza A viruses have the possibility to infect the upper respiratory tract and the bronchi of humans. The binding of SA moieties could also play an important role in host switches and could therefore be of further interest.

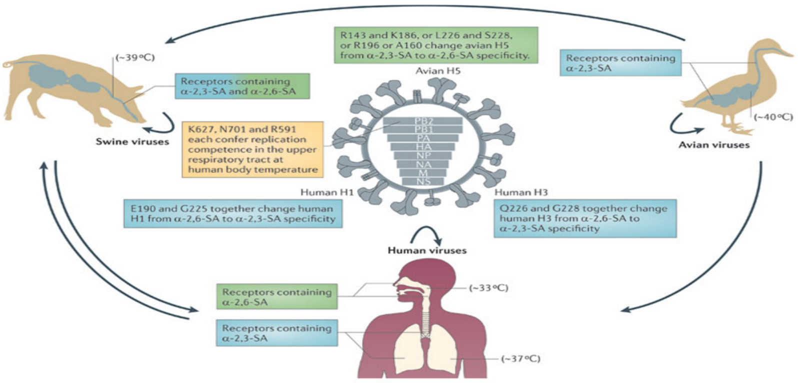
R143 and K186, or L226 and S228, or R196 or A160 change avian H5 from α-2,3-SA to α-2,6-SA specificity.
Avian H5
(~39 °C)
Receptors containing α-2,3-SA and α-2,6-SA
Receptors containing α-2,3-SA
(~40 °C)
Swine viruses
Avian viruses
K627, N701 and R591 each confer replication competence in the upper respiratory tract at human body temperature
PB2
PB1
PA
HA
NP
NA
M
NS
Human H1
Human H3
E190 and G225 together change human H1 from α-2,6-SA to α-2,3-SA specificity
Q226 and G228 together change human H3 from α-2,6-SA to α-2,3-SA specificity
Human viruses
Receptors containing α-2,6-SA
(~33 °C)
Receptors containing α-2,3-SA
(~37 °C)

Figure 1.3: Figure showing the different host specific conditions for influenza A viruses. Also showing the different receptor types the hosts got and the location of them. Taken from [48].

1.3 Sites Under Selection

1.3.1 Selection

Selection of traits, alleles or genes is a main part of the evolutionary process. In a population an individual with advantages, and therefore the best adapted traits, is more likely to be successful and to reproduce. Influenza A viruses are under a permanent selection pressure, due to the constant evasion of the host immune system. Alterations that get fixed in a certain population are denoted as evolution. The ones not getting fixed usually get lost, because they were not of major advantage.

There are three possibilities of selection - neutral, purifying and positive selection.

In case of neutral selection there is no favored trait and therefore the alterations are not caused by any selective pressure but by random genetic drift [31], which is completely undirected.

Compared to this first case purifying and positive selection are driven by selective pressure. A trait under purifying selection is prevented from being altered. This type of natural selection is also called negative selection. This is because this type selectively removes alleles, mainly deleterious ones [45]. Positive selection, also called directional selection, terms the extreme favoring of a special trait over all others. An individual with a trait under positive selection has a high fitness advantage over all other individuals. A trait under positive selection has the possibility to become fixed and therefore be persistent in the resulting population [51].

1.3.2 Measures of Selection

The most commonly used method to evaluate the pressure of selection for proteins is the counting of synonymous and non-synonymous mutations for each site (described in detail in Chapter 2.5.1). There are two more possible methods, namely a maximum-likelihood [90] and an approximative [90], which can be used instead of counting synonymous and non-synonymous mutations. In comparison, when the dataset is big enough, all three methods tend to the same results [36] and it is more important which assumptions are implicit in the used method [90]. Possible in this coherence are assumptions on the mutation rate, a correcting for multiple substitutions or the divergence between the considered sequences. Because the method of counting synonymous and non-synonymous mutations does not need any further assumptions, the ratio is therefore the more powerful model of evolution compared to the other two [90]. All three methods are based on a multiple sequence alignment (MSA) (see Chapter 2.2) and a preceding ancestral state reconstruction (see Chapter 2.3.2) of protein-coding gene sequences.

In the end the ratio test, which are introduced in Section 2.5.1, results in counts of synonymous and non-synonymous mutations, whose ratio constitute in a value called ω. This $omega$-ratio can be used for interpreting the results, even without further assumptions [9] but gets more viable with additional assumptions, e.g. the codon-frequencies

[75]. In this context three co-domains are of interest, which all three refer to one of the in Section 1.3.1 introduced cases. The first case is represented by $\omega \approx 1$, similar to the absence of selective pressure. In the second case, described by $\omega < 1$ the site is under purifying selection, meaning natural selection prevents the replacement with other amino acids. The last case $\omega > 1$ describes the favoring of amino acid changes, namely positive selection. The procedure of calculating and using ω-values is also known as Bayes prediction.

1.3.3 Known Sites under Selection

Some parts of a protein are under constant selective pressure. In case of influenza A viruses HA has such fast evolving parts, compared to other parts or other proteins, mostly located near the receptor binding site (RBS). With the fast evolution in this parts, HA permanently avoids the host immune system, because already slight changes can lead to the possibility to escape the immune system. The alterations usually happen in the epitope parts of HA. Epitopes are part of the antigen that is recognized by the antibodies and where the antibodies bind, which is the reason why changes in this part can have such capital effects.

The epitopes for HA of influenza A viruses are known because of experimental studies. The references [13] and [84] provide information on the already known epitopes. Table 1.2 summarizes them for the influenza A virus subtype H3.

Table 1.2: Table showing the known epitopes with residues under selection for the protein HA of influenza A virus subtype H3. Values taken from [84].

Epitopes	Residues
A	122, 124, 126, 130-133, 135, 137, 138, 140, 142-146, 150, 152, 168
B	128, 129, 155-160, 163-165, 186-190, 192-194, 196-198
C	44-48, 50, 51, 53, 54, 273, 275, 276, 278-280, 294, 297, 299, 300, 304, 305, 307-312
D	96, 102, 103, 117, 121, 167, 170-177, 179, 182, 201, 203, 207-209, 212-219, 226-230, 238, 240, 242, 244, 246-248
E	57, 59, 62, 63, 67, 75, 78, 80-83, 86-88, 91, 92, 94, 109, 260-262, 265

The subtypes H3 and H1 of HA are the most important for humans, because the combinations H3N2 and H1N1 are the actual circulating subtypes in humans.
Table 1.3 shows the known epitopes for influenza A viruses from subtype H1.

Table 1.3: Table showing the known epitopes under selection with corresponding residues for the protein HA of influenza A virus subtype H1. Values taken from [13].

Epitopes	Residues
Sa	124, 125, 154, 156, 158, 159, 161-163
Sb	152, 155, 188, 189, 192, 194
Ca1	165, 169, 178, 203, 236, 269
Ca2	136, 139, 141, 220, 221
Cb	74, 75, 77-85, 118

This two tables, containing all the positions of amino acids in patches for subtypes H3 and H1, are furthermore used as main evaluation data and are both given in H3 numbering. Numbering in this case means, that the residue positions are matched from one subtype to another, regarding to a Multiple Sequence Alignment (MSA). Some other publications [11,24,26,29,41,42,60,71] are added, such that there is a broad variety additional to the patch data.

1.4 Goal

The first goal is to introduce an automated framework for detecting patches of sites under selection,which has a broad range of possible appliance. Regarding this goal the first aim is to complete an automated framework for the detection of sites under selection comparable to [79] and make it usable for every interested person, not only for those with a background in bioinformatics. Providing an automated tool including the graph-cut algorithm (see Chapter 2.7) makes it easy for biologists to find sites under selection in supported proteins using IPoSuS. The automated graphical output of the results additionally makes it easy to understand and interpret them. Furthermore the automated data download makes it comfortable in use because no data preprocessing is needed to fit the programs need.
For a broader application range of the framework, the second main goal is to update the used statistical test, since the current approach uses a test on the counts of synonymous and non-synonymous mutations the aim is to establish a statistical test working on ω-values instead. Because ω-values contain more information than the pure counts of synonymous and non-synonymous mutations. The hope is to get better and more differentiated results, based on comparison with recent results [79].
The third goal is to evaluate the functionality of IPoSuS on the basis of the HA protein for different influenza A subtypes. On basis of the obtained results it also should be possible to decide which of the five tested approaches is the best, regarding the task.
Since the algorithm in [79] was only applied to the proteins HA of subtype H1N1, HA of

subtype H3N2 and the protein PB2 one further goal is to make the algorithm applicable for all possible proteins, not only of influenza A viruses, although the focus for this study lies on the HA protein of different subtypes of influenza A viruses. Being able to analyze and compare results of every influenza A virus protein could ideally lead to further understandings of host switches or adaption needed to maintain in a host. Beside this special analysis the possibility to consider every single protein of an organism in case of selection can lead to further understanding which proteins evolve at which rate.

2 Methods

In this chapter of the work all used methods, which are needed for the algorithm to work, are introduced and explained. Furthermore all needed basic methods are explained.

2.1 Database

The decision on the used database is of paramount importance. Since we want to create and establish an automated workflow the database has to be machine accessible, for example through File Transport Protocol (FTP). Even though the publication, on which this thesis is based [79], uses data from the Global Initiative on Sharing Avian Influenza Data (GISAID)(www.gisaid.org), it was not practicable for us to use the same database, because it does not provide a needed machine accessibility. We decided to use the Influenza Database hosted by the National Center for Biotechnology Information (NCBI) [4]
(http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html) for this thesis, because it provides a FTP connection. We receive coding sequences from the database, which are needed to perform a MSA. The received sequences can be assumed to be of high quality, because they are from the NIAID Influenza Genome Sequencing Project as well as from GenBank.

2.2 Multiple Sequence Alignment (MSA)

In contrast to pairwise alignment algorithms such as Smith-Waterman [68] or Needleman-Wunsch [54], a MSA uses more than two biological sequences as input. They should be of similar length and should also consist of the same elements, therefore be composed of either amino (protein) or nucleic acids (DNA/RNA). MSA's are usually used to infer homology from the input data, or at least the evolutionary relationship between the sequences. There are several different tools to complete this task, e.g. multiple sequence comparison by log-expectation (MUSCLE) [19], ClustalOmega [67] or CLUSTALW2 [40]. We decided to use MUSCLE, because it outperforms the other applications.

2.2.1 MUSCLE

MUSCLE is a computer program for creating MSA's using distance estimations, log-expectation scores and tree-dependent restricted partitioning. As shown in [19] MUSCLE outperforms other established MSA tools, such as T-Coffee and CLUSTALW in speed and accuracy.

MUSCLE makes use of two different distance measurements of a pair of sequences: the $k$mer distance for unaligned pairs and the Kimura distance for aligned pairs. The $k$mer distance is a contiguous subsequence with length k. Generating the distance matrices MUSCLE uses unweighted pair group method with arithmetic mean(UPGMA) [69]. After this it computes a progressive alignment and refines the results with the Kimura distance followed by another UPGMA generated tree. Further progressive alignments and repeatedly computing subtree profiles with re-aligning and score determination finally result in an optimal tree for the given input sequences. The workflow of the multiple sequence comparison by log-expectation (MUSCLE) algorithm is shown in Figure 2.1.

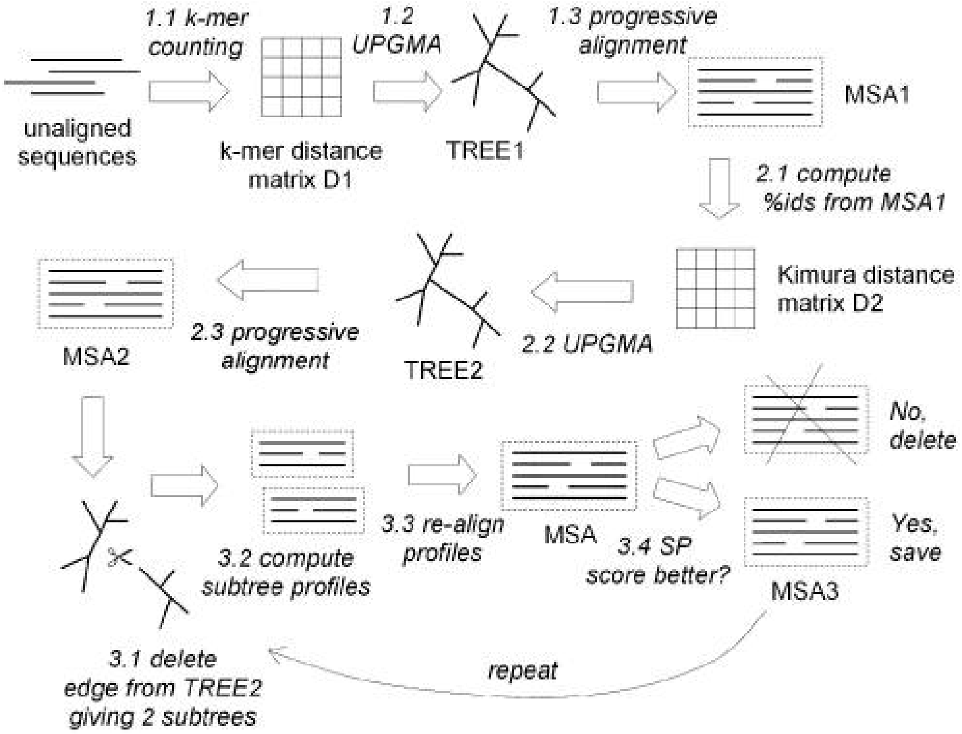
1.1 k-mer counting
1.2 UPGMA
1.3 progressive alignment
unaligned sequences
k-mer distance matrix D1
TREE1
MSA1
2.1 compute %ids from MSA1
Kimura distance matrix D2
2.2 UPGMA
TREE2
2.3 progressive alignment
MSA2
No, delete
Yes, save
3.3 re-align profiles
MSA
3.4 SP score better?
MSA3
3.2 compute subtree profiles
3.1 delete edge from TREE2 giving 2 subtrees
repeat

Figure 2.1: Diagram summarizing the MUSCLE algorithm with its three main steps. 1. draft progressive, 2. improves progressive and 3. refinement. Taken from [19].

Since the intended automated process of the framework and the fact, that the downloaded sequences are not curated manually, there is need for an additional step, that automatically curates the MSA. A tool that can perform this task is TrimAl.

2.2.2 TrimAl

Because of the possibility to self define the input parameters, it could be possible to receive sequences that are not of equal length. Moreover it could be possible to get false annotated sequences in the requested dataset. Thus we integrate TrimAl [10] in the workflow. This tool allows the automated removal of spurious sequences or of poorly aligned regions. Additionally to the web-server this tool provides a command line application. Many parameters are adjustable, allowing the user to obtain the best results for their use. For example trimAl can be used to remove all gaps from a MSA or columns that consists only of gaps. It can also be used to determine a minimum overlap of a position with other positions in the column to be labeled as a good position. Based on this labeling trimAl can be used to remove sequences below a minimum percentage of good positions.

Furthermore this program is able to process different types of files and can save the results in different types, e.g. phylip, nexus or fasta. The different output formats can be used as input for different further steps.

2.3 Phylogenetic Inference

Phylogenetic inference methods are used to reconstruct a phylogenetic tree that displays the evolutionary relationship among a set of different species or entities such as influenza viruses. Each individual entity or species is represented as taxa. In the phylogeny each taxa is represented by a given sequence. Taxa being evolutionarily closer related to each other are arranged more close together in the phylogeny than distinct ones. Using phylogenetic tree inference methods makes it possible to infer information about the evolution of specific genes, or traits in general. For the purpose of reconstructing a phylogenetic tree a MSA is needed with sequences representing the leaf nodes of the tree.

There are two different kinds of phylogenetic trees - rooted and unrooted. The unrooted tree does not visualize the most common ancestor, whereas the rooted tree contains this relation. Every unrooted tree can be converted to a rooted one by using a so called outgroup, which either can be added to the dataset or can be included. An outgroup is a species or entity that is far related to the remaining dataset. Figure 2.2 shows the difference in presentation between a rooted and an unrooted tree.

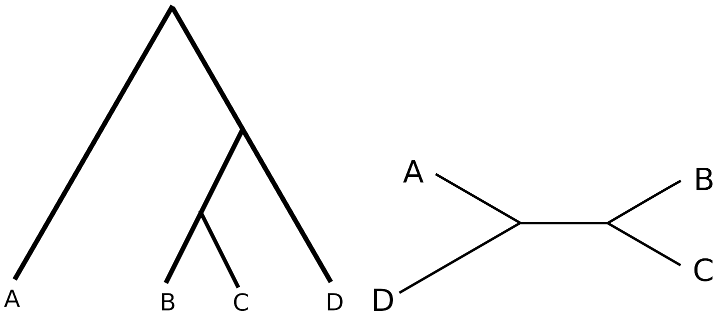

Figure 2.2: Figure showing the difference between a rooted tree on the left and an unrooted phylogenetic tree on the right. The tree on the left got rooted by defining the sequence A as an outgroup.

There are four widely used methods to solve the problem of finding the optimal tree -

distance-matrix method [52], maximum parsimony [20], maximum likelihood [21] and Bayesian inference [25]. All methods have their own specific strengths and weaknesses [91]. In this thesis we use FastTree to generate the phylogenetic tree, which is a approximately-maximum-likelihood method, based on the input sequences and the generated MSA (see Section 2.2) with MUSCLE (see Section 2.2.1), that got automatically curated by TrimAl.

2.3.1 FastTree

In general it is a heuristic problem to generate a phylogenetic tree, based on a MSA. This is due to the NP hardness of the task. Therefore every algorithm or heuristic has its right to exist. Usually the differences are determined by speed or accuracy. In this thesis we decided to use FastTree [61] for inferring phylogenetic trees, based on the MSA output by MUSCLE and TrimAl. FastTree is based on an approximately-maximum-likelihood approach and is used in this work because of its ability to handle millions of sequences using a reasonable amount of time and memory. PhyML in comparison can only handle up to 4,000 sequences (with default settings) and this would be insufficient for the scope of this tool. Beside the fact that FastTree can handle more sequences than PhyML it is also faster and more accurate than PhyML and RAxML 7, namely 100 to 1000 times faster. For comparison FastTree only needs 15 minutes for generating a maximum likelihood tree from 8,362 sequences, whereas PhyML and RAxML would need over 1,200 hours.

The FastTree algorithm is divided into four main parts: The first part contains of a heuristic neighbor-joining to get a rough topology. For a better speed FastTree combines three heuristics in this step - fast neighbor-joining, relaxed neighbor-joining and top hits heuristic. Unlike other tree generating tools FastTree does not generate distance matrices but profiles of internal nodes, reducing the required memory.

The second step reduces the length of the tree using nearest-neighbor interchanges (NNI) and subtree-prune-regraft (SPR) moves, called balanced minimum evolution. This step is much faster in FastTree because, as already mentioned, it is not based on distances but profiles. If the distances are not too noisy the NNI and SPR moves lead to an optimal tree [17].

The third step is used to maximize the likelihood of the tree. This step includes the improvement of both the topology and the lengths of the branches using a maximum-likelihood rearrangement. FastTree can use four different evolution models - Jukes-Cantor, generalized time-reversible models of nucleotide, Jones-Taylor-Thorton or Whelan Goldman models of amino acid evolution. This step includes the setting of the most likely category, out of 20, to each site, based on variable rates of evolution, using a Bayesian approach with gamma prior. This also prevents over-fitting small alignments.

The last step is the bootstrapping step, which uses the Shimodaira-Hasegawa test on three alternate topologies around the split. Despite using CAT approximation and not full optimizing branch lengths the resulting support values are virtually identical compared

to PhyML, which uses a SH-like local support approach. FastTree uses 1,000 resamples and does not reoptimize the branch lengths.
After performing all these mentioned steps, the generated phylogenetic tree can be assumed as optimal solution for the input data. Nevertheless it has to be noted, that the worst case still has exponential runtime. Based on this tree it is possible to infer the sequences of the mutual ancestors for a given set of sequence, up to the ancestor all sequences have in common.
If the obtained tree is **not** binary the next step is always to convert it into a binary tree, because the following steps only work for trees of this form. This means, that for exactly two leaf nodes there is only one ancestor.

2.3.2 Ancestor State Reconstruction

As it will be described in detail in Section 2.5.1, it is necessary to know the relation between all input sequences and their ancestor states. Therefore it is needed to perform an ancestor state reconstruction on the generated tree, containing the sequences of the MSA as leaf nodes. Considering the binary tree and the leaf nodes with their sequences, an ancestor state is the reconstructed sequence of the internal node. Only with knowing the ancestor states and their corresponding sequences it is possible to infer the ratio of non-synonymous and synonymous mutations (dn/ds ratios) from the input data.
Ancestor reconstruction is based on a reconstructed phylogenetic tree. Assuming a binary tree, each two sequences, represented as taxons in the tree, are linked by one node, representing the ancestor state of this two nodes. Because of this assumption the further processing gets more accurate and is faster. In this work we on the one hand use the Fitch algorithm for this step, which is based on the known sequences, given by the MSA for the leaf nodes, the rooted, binary tree and on the other hand we use an adaption of the Fitch algorithm.
Furthermore there are two different possibilities in the step where the synonymous and non-synonymous mutations are counted. Both parts will be explained in detail in the following subsections.

2.3.3 The Fitch Algorithm

This algorithm [23] is working for binary trees and minimizes the unit costs (costs of a change from one state,where a state is a amino acid in the sequence - to another) on the tree. It is an example for a parsimony algorithm and works for known trees. Because we construct the most likely phylogenetic tree with the given sequences, and therefore know it, the Fitch algorithm is the used one in this work.
The Fitch algorithm assigns possible states of characters for each inner node with a minimal number of state changes. Further we have to assume a rooted tree or introduce a root without altering the results. In dynamic programming fashion the algorithm performs two main steps for each character. The first step is the bottom-up phase, followed

by the top-down refinement. Figure 2.3 shows an example for one character, including the bottom-up and top-down phase.

Bottom-up phase

Starting from the leaves we walk along the edges to the root, such that visiting a node means the child nodes have already been visited. For each node we now collect the possible states and store them in a candidate set. This candidate set contains either both possibilities gained from the considered child nodes or when they share a common candidate it contains only this shared one. Walking along the whole tree from the leaf nodes to the root will result in candidate sets assigned to all internal nodes.

Top-down refinement

In this step the algorithm walks down the tree from the root and assigns a character from the candidate set to the internal node. If the root candidate set only comprises one element, it gets assigned, otherwise the character is randomly assigned. The assignments for internal nodes while walking down the tree are based on the parental node for the considered node. If the candidate set contains the character assigned to the parental node, this one gets assigned, otherwise the state gets picked randomly again.
Although this algorithm, containing a backtracking phase, yields in an optimal solution for the generated tree, it does not provide all co-optima.

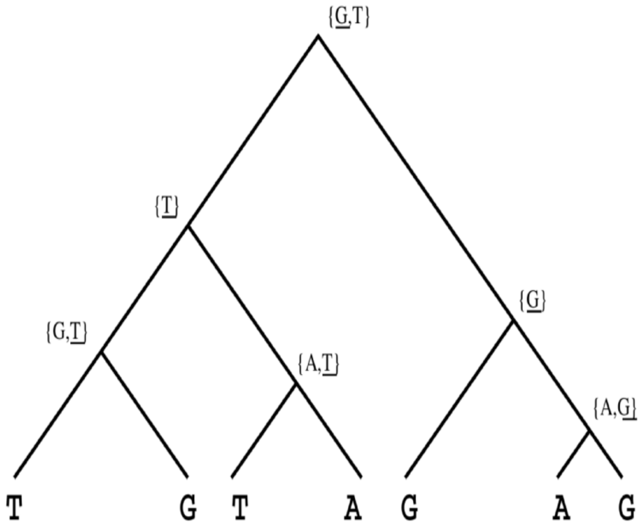

Figure 2.3: Figure showing an one character example for the Fitch algorithm. The leave nodes are labeled according to the data. Candidate sets are represented for each internal node in curly brackets. The bar indicates a possible or best solution.

As mentioned above there is another used method to determine the assigned character states for the internal nodes. This approach does not, in contrast to the usual Fitch al-

gorithm, use random assignments, but defined ones. In order to get more reproducible results this second approach always assigns the first candidate for the internal node instead of a random one out of a set. This method will therefore always lead to the same results for a dataset unless the tree gets altered in a previous step, because of the possibility of more than one optimal solutions.
Because of the properties of the these two different counting schemes, they will be named accordingly hereinafter. The first approach uses the Nei and Gobjobory algorithm with the minimal way approach and also the randomness of Fitch's algorithm and will be called NG+RF (Nei Gojoborj + random Fitch) from here on. The second approach also uses the Nei Gojoborj approach but does not use a the randomness of Fitch's algorithm but always chooses the first possibility if there is more than one. This approach will be calles NG+NRF (Nei Gojoborj + **non**random Fitch) from here on.

2.4 ACCTRAN and DELTRAN

ACCTRAN and DELTRAN, which are short forms for ACCelerates the evolutionary TRANsformation of a character and DELays the TRANsformation of a character on a tree, are two possibilities to alter results in the top down refinement phase and are therefore two possibilities to alter the results.
Using acctran will push the evolutionary transformation down the tree as far as possible. In this method reversals are favored over parallelism as long as the choice is equally parsimonious.
Deltran on the other hand will push characters up the tree as far as possible. Because it is the opposite of acctran it favors parallelism over reversals as long as the choice is equally parsimonious.
When there is no ambiguity, both methods will yield the same results.
In this thesis we used the deltran method because we want to favor parallelism over reversals. This is due to the fact that we want to find positions and/or patches under selection.

2.5 Determining Sites under Selection

As already mentioned in Chapter 1.3.2 there are three mainly used methods for determining sites under selection. We use the widely used method of counting synonymous and non-synonymous mutations, which will be introduced in the following.

2.5.1 Non-synonymous and Synonymous Mutation (dn/ds) Ratios

The calculation of dn/ds ratios is based on an inferred phylogenetic tree with know ancestor states (see Chapter 2.3.2). With this reconstructed tree it is possible to determine c_S and c_N, which are the total numbers of synonymous and non-synonymous mutations for each site, and the average numbers s_S and s_N respectively. d_S and d_N, the total number of synonymous and non-synonymous mutations, are then determined with following equations [9, 75].

$$(2.1)\quad d_S = \frac{c_S}{s_S}$$

d_S	:	ratio of synonymous mutations
c_S	:	total number of synonymous mutations
s_S	:	average number of synonymous mutations

$$(2.2)\quad d_N = \frac{c_N}{s_N}$$

d_N	:	ratio of non-synonymous mutations
c_N	:	total number of non-synonymous mutations
s_N	:	average number of non-synonymous mutations

The corresponding ω-value is then defined by

$$(2.3)\quad \omega = \frac{d_N}{d_S}$$

d_S	:	ratio of synonymous mutations
d_N	:	ratio of non-synonymous mutations
ω	:	ratio of non-synonymous to synonymous mutations

and describes the pressure of selection onto a single site.

2.6 Statistical Tests to Identify Sites under Selection

To evaluate whether a site in the considered protein mutates significantly more often than others there is a need for a statistical test.

2.6.1 Fisher's exact Test

One of the used tests is Fisher's exact test. It it used to test if the observed count of synonymous and non-synonymous mutations at a specific position are significantly higher compared to the mean value of mutations in the whole protein. This test is the most commonly used test for the determination of sites under selection.

Fisher's exact test works on a contingency table and tests for independence. Compared to the chi-square-test there is no minimal sample size needed and the results are also reliable for small test samples. Originally the test was designed for $2x2$ contingency tables, but can be used for greater ones as well. Based on a general $2x2$ contingency table, see Table 2.1,

Table 2.1: Table showing the general $2x2$ contingency table with labels used for Fisher's exact test.

	Count at Position	Mean Count whole Protein	Row Total
synonymous mutations	a	b	a+b
non-synonymous mutations	c	d	c+d
Column total	a+c	b+d	n=a+b+c+d

Fisher's exact test generally calculates the probabilities as follows [22]:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{2.4}$$

2.6.2 Z-Test

Additionally to Fisher's test the introduced workflow uses another statistical test, which is not based on the pure counts of synonymous and non-synonymous mutations of each codon but on the corresponding ω-ratio or ω-value, based on the generated MSA and phylogenetic reconstructed tree, as described in [9, 75].

In contrast to pure counts of synonymous and non-synonymous mutation, or their corresponding ω-ratio, the ω-value includes further information, e.g. codon-frequencies. This means, that the result is influenced by the frequency a codon appears and how likely it mutates into the one observed.

Considering ω-values or ω-ratios rather than counts the statistical test that is used has to be differed because Fisher's exact test is not working for discrete values. A proper way generate p-values based on the new considered omega values is the Z-test from Gauss [38].

The Gaussian hypothesis test or Z-test is a statistical test we can apply on the ω-values and ω-ratios.

Generally spoken this test exhibits a strong relationship to the t-test. The difference between both tests is the input-data. In case of the Gaussian approach it is the known standard deviation of the population. It is strongly recommended to use Gauss test and not the t-test if the standard deviation is known, which is the case in our investigations. Because of the central limit theorem (law of large numbers) we can neglect the usually needed Gaussian distribution assumption, for sample sizes approximately greater than 30.

One Sample Z-Test

The one sample Gauss test uses the arithmetic mean of a spot check, whether the expectation of the population is greater or lower compared to a given value.

Every sample $x_1, x_2, ..., x_n$ shall have characteristics of independent random variable, which is normally distributed, with unknown expectation μ and known standard deviation σ. There are three possibilities of testing.
Two sided test:

(2.5) $\mathrm{H}_0 : \mu = \mu_0$ against $\mathrm{H}_1 : \mu \neq \mu_0$

Right sided test:

(2.6) $\mathrm{H}_0 : \mu \leq \mu_0$ against $\mathrm{H}_1 : \mu > \mu_0$

Left sided test:

(2.7) $\mathrm{H}_0 : \mu \geq \mu_0$ against $\mathrm{H}_1 : \mu < \mu_0$

In all cases the value for μ_0 is predefined by the user.
Defining the arithmetic mean of the population as

(2.8) $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

the test statistic is calculated as followed:

(2.9) $z = \sqrt{n} * \frac{\bar{x} - \mu_0}{\sigma}$

Decision on the Hypotheses

For all three test the criteria for hypothesis testing and decision on whether to accept the hypothesis or not are used. Because of the fact that Z is normally distributed under the null-hypotheses, the following rules are obtained. Rejection of H_1 at significance level α if:
Two sided test:

(2.10) $|z| > u(1 - \frac{\alpha}{2})$

Right sided test:

(2.11) $z > u(1 - \alpha)$

Left sided test:

(2.12) $z < u(\alpha)$

The decision also can be based on the distribution shown in Figure 2.4, in which the significance-level is shown in correlation to the critical value or z-score.

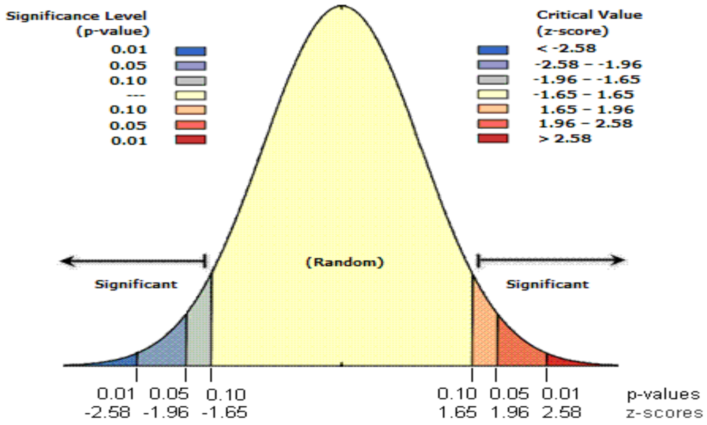

Figure 2.4: Figure showing the Gaussian distribution with additional information on the significance-level and the corresponding critical values (z-score). The colours indicate the different significance-level, on which base a decision on the hypothesis can be made. Figure taken from http://help.arcgis.com/de/arcgisdesktop/10.0/help/index.html#//005p00000006000000 accesed 18.02.2015 at 08:30.

2.7 Graph-Cut Algorithm

The graph-cut algorithm usually is employed for image smoothing, the stereo correspondence problem, but also for energy minimizing problems [8].

The graph cut in general is the application of a minimum cut on a graph. This minimum cut separates the graph in two disjoint subsets. Each of these two subsets consist of at least one edge and therefore two nodes. With this approach it is not possible to cut out single nodes. The graph cut algorithm, which was first used to smooth noisy or corrupted binary images uses this minimum cut with the difference, that the graph gets extended by two nodes, namely the source and sink node. In following source and sink node get connected with each node of the existing graph, e.g. pixels of an image. Each of the introduced edges, connecting the nodes of the initial graph with the sink and source node, are then weighted. Applying the minimum cut on the newly obtained graph will now find two disjoint graphs, of which one is minimal in a beforehand given case, e.g. energy or distance.

Adapting this algorithm for purposes of predicting sites under selection some things are different from the usual approach. The first step is to normalize the structural data retrieved by the 3D-structure of the protein and therefore the distance between individual residues, comparable to pixels in an image. Each residue then gets represented by a node. Edges get introduced for every pair of nodes, where the Euclidean distance is below a specific threshold. These edges now get weighted with the spatial distance. Therefore, close residue edges have higher weighting than distant ones. Then, comparable to the sink and source node from the originally algorithm, a negative and a positive selection node are introduced. Both are connected with each node introduced before, weighted with corresponding p-values, obtained by the dn/ds ratios and the used statistical test, for negative selection node and $1 - P(n)$ for the positive one. The graph cut then will divide the so obtained graph in two halves, one with the positive selection node and one with the negative selection node, namely minimizing the sum of weights connecting the two halves [79] (Figure 2.5).

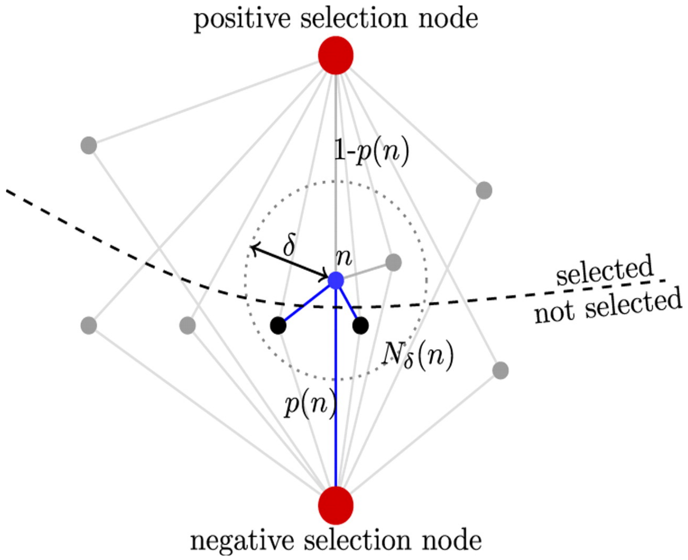

Figure 2.5: Figure showing the adapted graph-cut algorithm with the positive and negative selection node. Furthermore it shows the assigned p-values and the assigned spatial distance. The dotted line symbolizes the graph cut, dividing the graph in two halves. Picture taken from [79].

2.8 Tertiary Structure Data

Since we want to visually assign the computed patches of sites to the tertiary structure of the corresponding protein, we have chosen default 3D-structures for every protein for this purpose. The structures are chosen subtype specific to yield the best results. Although the focus in this work is on the HA protein, other proteins are already implemented for which the analysis could be done with subtype specific structures. For proteins for which no subtype specific structure is known, the background data uses a tertiary structure of another subtype. At the moment there are eleven of the fourteen proteins of influenza A set as default structures and sequences, depending on the requested job. For the proteins Pb1-N40, PA-X and M42 there are no available tertiary structures at this moment. There can be no analyses on these proteins at the moment with default background data. To run analyses for these proteins please check the possibilities in Chapter 3. Table 2.2 shows all used tertiary structures taken from the Protein Data Bank (PDB) for structure analysis.

Table 2.2: Table showing the in different subtypes used in this work with proteins that have subtype specific structures available with corresponding pdb identifier, which are implemented in the introduced workflow and are used for structure analysis and the graph cut algorithm.

Subtype	Encoded Protein	PDB Identifier
seasonal H1N1	HA M1 NP PA C-terminal PB1 middle domain PB2	2WRG [43] 4PUS [62] 2IQH [92] 2ZNL [57] 3A1G [73] 4J2R [78]
H3N2	HA M1 NP NS1 PB2 NA2 NEP M2	3HMG [82] 1EA3 [3] 2WFS [16] 3O9T [30] 2VY6 [76] 4GZO [96] 1PD3 [1] 2L0J [65]
H5N1	HA M1 PA endonuclease NS1 PB2	2IBX [87] 2Z16 Saijo et al. unpublished! 3HW3 [95] 3F5T [6] 3KC6 [88]
pandemic H1N1	HA M1 Pa endonuclease NS1 PB2	3AL4 [94] 3MD2 Liu et al. unpublished! 4AVQ [37] 3M5R Fremont et al. unpublished! 3KHW [88]
H7N7	HA	4DJ6 [89]
H7N9	HA PA C-terminal	4N5J [86] 4P9A [50]

2.9 Alignment of all different Subtypes

This alignment is necessary because of the different numbering schemes and the different insertions/deletions the different subtypes have. Introducing this "master" alignment makes it possible to have one numbering scheme for all subtypes, which makes the results more comparable.

```
H3N2        ------QDLPGNDNSTATLCLGHHAVPNGTLVKTITDDQIEVTNATELVQSSSTGKICNN
H5N1        MEKIVLLFAIVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQDILEKTHNGKLCDL
sH1N1       ----------------DTICIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDSHNGKLCKL
pH1N1       ----------ADLGSRDTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDKHNGKLCKL
                              :*:*:**  .   *.*: :.:: **.: ::::.. .**:*.

H3N2        PH-RILDGIDCTLIDALLGDPHCDVFQN-ETWDLFVERSK-AFSNCYPYDVPDYASLRSL
H5N1        DGVKPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVEKANPVNDLCYPGDFNDYEELKHL
sH1N1       KGIAPLQLGKCNIAGWLLGNPECDLLLTASSWSYIVETSNSENGTCYPGDFIDYEELREQ
pH1N1       RGVAPLHLGKCNIAGWILGNPECESLSTASSWSYIVETPSSDNGTCYPGDFIDYEELREQ
                 *   .*.:   :**:* *: : .   *. :**  .     *** *. ** .*:

H3N2        VASSGTL---EFITEGFTWTGVTQ-NGGSNACKRGPGSGFFSRLNWLTKSGSTYPVLNVT
H5N1        LSRINHFEKIQIIPK-SSWSSHEASLGVSSACPYQGKSSFFRNVVWLIKKNSTYPTIKRS
sH1N1       LSSVSSFEKFEIFPKTSSWPNHETTKGVTAACSYAGASSFYRNLLWLTKKGSSYPKLSKS
pH1N1       LSSVSSFERFEIFPKTSSWPNHDSNKGVTAACPHAGAKSFYKNLIWLVKKGNSYPKLSKS
            ::    :   ::: :  :*       * : **     ..*: .: ** *. .:** :. :

H3N2        MPNNDNFDKLYIWGIHHPSTNQEQTSLYVQASGRVTVSTRRSQQTIIPNIGSRPWVRGQS
H5N1        YNNTNQEDLLVLWGIHHPNDAAEQTKLYQNPTTYISVGTSTLNQRLVPRIATRSKVNGQS
sH1N1       YVNNKGKEVLVLWGVHHPPTGTDQQSLYQNADAYVSVGSSKYNRRFTPEIAARPKVRDQA
pH1N1       YINDKGKEVLVLWGIHHPSTSADQQSLYQNADTYVFVGSSRYSKKFKPEIAIRPKVRDQE
              * .  : * :**:***    :* .** :    : *.:   .: : *.*. *  *. *

H3N2        SRISIYWTIVKPGDVLVINSNGNLIAPRGYFKM-RTGKSSIMRSDAPIDTCISECITPNG
H5N1        GRMEFFWTILKPNDAINFESNGNFIAPEYAYKIVKKGDSTIMKSELEYGNCNTKCQTPMG
sH1N1       GRMNYYWTLLEPGDTITFEATGNLIAPWYAFALNRGSGSGIITSDAPVHDCNTKCQTPHG
pH1N1       GRMNYYWTLVEPGDKITFEATGNLVVPRYAFAMERNAGSGIIISDTPVHDCNTTCQTPKG
            .*:. :**:::* * : :::.**::.*   : : : . * *: *:     * : * ** *

H3N2        SIPNDKPFQNVNKITYGACPKYVKQNTLKLATGMRNVPEKQT-
H5N1        AINSSMPFHNIHPLTIGECPKYVKSNRLVLATGLRNSPQRE--
sH1N1       AINSSLPFQNIHPVTIGECPKYVRSTKLRMATGLRNIPSRQS-
pH1N1       AINTSLPFQNIHPITIGKCPKYVKSTKLRLATGLRNIPSIQSR
            :* .. **:*:. :* * *****:.. * :***:** *. :
```

Figure 2.6: Alignment used to get a consistent numbering between all different investigated influenza A subtypes. The sequences used are the same as introduced in the methods. So it is sequence related to the pdb identifier 2WRG for the seasonal H1N1, 3AL4 fpr the pandemic H1N1, 3HMG for the H3N2 subtype and 2IBX for subtype H5N1. Alignment was made using MUSCLE [19].

2.10 Different Approaches

After introducing the different counting schemes as well as the different statistical tests, we now introduce the different approaches IPoSuS uses.
First of all, the AdaPatch approach, which uses the graph-cut algorithm based on p-values obtained by Fisher's exact test. For this approach there are both different counting schemes possible, resulting in the approaches AdaPatch (AP) with NG+RF and AdaPatch with NG+NRF. This approach uses the pure count of synonymous and non-synonymous mutations to calculate p-values on base of Fishers exact test.
The second approach uses ω-ratios, and is therefore named OmegaRatio (OR). This approach again can be combined with both NG+RF and NG+NRF. This approach uses the p-values obtained by conversion of the z-values received by the Gaussian Z-test. This approach is the first one that uses the ratio of the counts of synonymous and non-synonymous mutations.
The third and last approach uses ω-values, therefore is named OmegaValue (OV) and is only be combined with the NG+NRF counting scheme. This approach uses the same p-value calculation as the OR approaches. This approach also uses the ratio of synonymous to non-synonymous mutations, but also adds additional information, such that the content of information rises.

2.11 Implementation

The workflow is written in Python (www.python.org) and consists of several scripts, modularizing the application. There are many free available packages, which were used to generate this application. All of those used packages that have not been mentioned in separate chapters are listed in Table 2.3. The only part not implemented as a Python script is the graph-cut algorithm, which is implemented in C++.

Table 2.3: Table showing the dependencies for correct python implementation.

Name	Version	Source	Used Package
Python	2.7.8	https://www.python.org/	os, time, argparse, ftplib, subprocess, sys, shutil
Scipy	0.15.1	http://www.scipy.org/	scipy.stats.*
Numpy	1.8.2	http://www.numpy.org/	numpy.std, numpy.array, numpy.sum

3 Workflow

This chapter serves the understanding of the workflow and will introduce every single step, one after another, that is needed for the whole workflow to work and detect patches of sites under selection. Figure 3.1 additionally shows the whole process in one diagram.

At first you have to choose, via input parameters, what kind of analysis you want to be run, because there are several different possibilities. The first is to have the predefined analysis with default settings by just using filters and a specific time span for the analysis. While choosing the individual filtering parameter it is advisable to specify a subtype, e.g. H3N2. For the timespan it could be useful to use the known seasons of influenza A viruses. For this purpose either the timespan from April to September of a year for the southern hemisphere or from October to March of the following year for the northern hemisphere are useful. But every other imaginable timespan can be used and investigated. Furthermore, in this step the user manually alters input parameters, which are used to define the filters to generate the desired output, only containing the requested coding sequences. This parameters can consist of strings for special cities, subtypes of influenza A viruses or other specifications that the user wants to make.

After the input of the individual parameter the first step of the default working process is the automated data-download via FTP from the Influenza Virus Resource database (IVR), hosted by the NCBI (see Section 2.1).

The second step uses MUSCLE to calculate the MSA (see Section 2.2 and 2.2.1) based on the filtered coding sequences from the previous step with followed editing by TrimAl (see section 2.2.2). TrimAl in this default setting is used to filter for spurious sequences which should not be retrieved by the database. But due to the filtering parameters and the automated nature of the workflow it can happen that sequences of other proteins get downloaded as well as the desired ones. This, for example, can happen when the identifier of the sequence contains one of the filtering parameters, e.g. NS or PB. TrimAl recognizes such sequences and deletes them from the MSA. The output then is given in fasta format.

The subsequent third step contains the creating of a phyolgenetic tree reconstruction, which is processed by FastTree (see Section 2.3.2), using the MSA file from the previous step. Furthermore the leafs of the tree are labeled according to the sequences in the MSA and their phylogenetic relation to each other. Based on this output, given as Newick-String, the ancestor state reconstruction is done by an implemented Fitch algorithm (see Section 2.3.3). But before this step can be processed in the correct way, the tree has to be binary. This means that two adjacent nodes always have exactly one ancestor. If the generated tree does not fulfill this criteria, it automatically gets converted into a binary tree.

Afterwards the synonymous and non-synonymous mutations are counted (see Section 2.5.1) on which basis the ω-values and the ω-ratios are calculated. For this step the two already in Section 2.5.1 introduced counting schemes are available to chose from. Based on the counts of synonymous and non-synonymous mutations per codon Fisher's exact test calculates the statistical significance according to the mean value of corresponding counts in the whole dataset. This statistical significance is represented by a p-value and gets assigned to every codon for synonymous and non-synonymous mutations. Furthermore there is a similar calculation based on the transformation of Z-values into p-values for the ω-values and the ω-ratio approaches.

The second last step is the graph-cut algorithm for detection of patches of sites under positive selection [79] (see Section 2.7), based on the p-values derived from the previous step combined with a calculated spatial distance between each amino acid in the 3D-structure of the considered protein. This step merges single amino acids to patches of sites under selection if they fulfill the requirements.

The final step visualizes all detected patches on the protein structure using PyMol [64]. If there is need for a more individual analysis the workflow provides several parameters which can be set to individualize the results. Thus another possibility is to choose whether you want to have both introduced counting methods (see Chapter 2.5.1) or only one of them. Next you can choose whether you only want human, only animal strains or a mix of both. If choosing analysis for the subtype H1N1 there also is the possibility to choose, whether the analysis shall be done for seasonal, pandemic or a mix of both strains.

Usually the workflow uses predefined default background data for the analysis to be done, but IPoSuS also offers the possibility to choose own background data just by defining the paths to the location of the files. For this possibility to work four different parameters and therefore four different files have to be introduced. These four files contain a reference sequence, the surface accessibility using netsurf [59], the fasta sequence on which bases the accessibility has been calculated and the 3D-structure of the protein.

Furthermore you can choose if you really want the automated data download and therefore the use of an external database or if you want analysis for your own data to be done. If analyzing own data, it can be chosen if the analysis should be run with default background data or some specified own data again consisting of the previous mentioned four needed files. If the option for analysis of own data is chosen, the previous mentioned second step is the first one that is performed, therefore IPoSuS starts with generating the MSA. Combining the two options to use own background data and own sequences, IPoSuS also provides another alternative, the one in which the analysis are completely based on your own data, giving the chance to also analyze other proteins, e.g. of other viruses. Here, again, the second step will be the first one that gets executed.

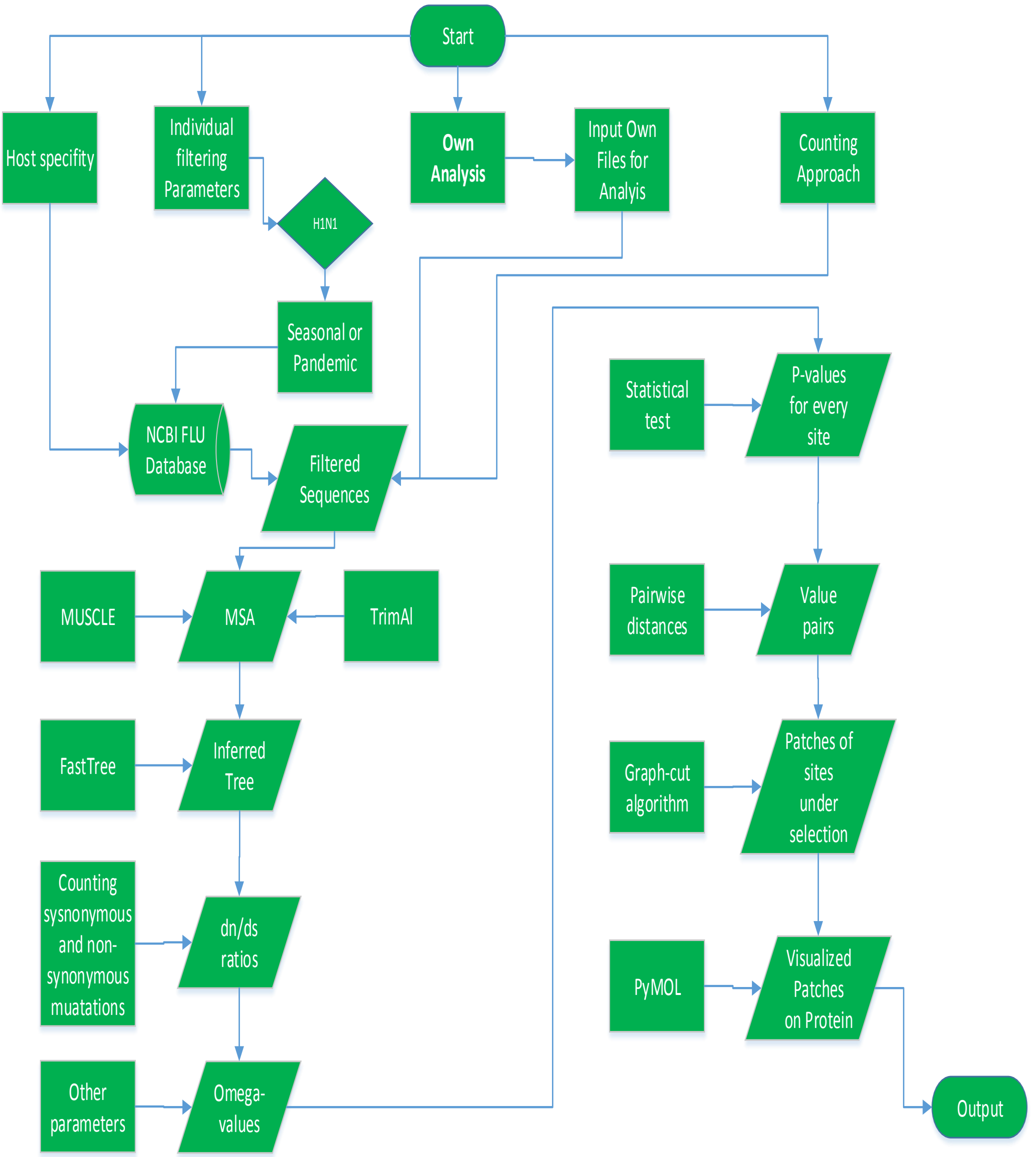

Figure 3.1: Diagram showing the procedure of the used methods taken together as one algorithm. Rectangles represent the used tools or processes. Parallelograms show the results/outputs of each processing step. The blue arrows indicate possible steps and therefore the different possible parameters.

4 Results

In this chapter all obtained results using the in Chapter 3 introduced workflow, will be gathered and shown. This includes the depiction of the seasonal results for the different counting schemes, introduced in 2.5.1, and different subtypes as raw results and therefore represent the identified patches. Furthermore the results will be shown in a depatched way, which makes it possible to count the occurrence of single positions. After all that there is also an evaluation of the obtained results.

4.1 Seasonal Patches for Human Host

In this section the results using the developed framework will be presented, divided by the subtype. All results will be shown in tables showing all different approaches.

4.1.1 Subtype H1N1 - seasonal

AdaPatch Approaches

Table 4.1: Results for the seasonal H1N1 subtype comparing the two AdaPatch approaches, differed by the counting method. For Seasons the introduced algorithm did not find patches of sites under selection the entries are left empty and consecutive seasons without found patches are taken together into one entry. The investigated seasons still remain the same.

	Approach	
Season	AP:NG+RF	AP:NG+NRF
2000/04-2000/09	248+163+199+198+196 187+189+188+193 482+476+475+474	197+198+196+248 276+274+53 142+141+144 131+132+156
2000/10-2001/09		
2001/10-2002/03	248+163+199+198+196 187+189+188+193 482+476+475+474	276+275+274+53 197+198+196+248 142+141+144
2002/04-2002/09		

2002/10-2003/03	248+163+199+198+196 197+192+158+156 187+189+188+193 279+276+275+274	279+276+275+274 193+196+199+198 219+189+188+187 158+156+132 142+141+77
2003/04-2003/09	248+163+199+198+196 248+163+199+198+196 279+276+275+274	279+276+275+274 193+196+199+198 219+189+188+187 131+156+132 271+91+93+65
2003/10-2004/03	248+163+199+198+196 187+189+188+193 279+276+275+274	279+276+275+274 193+196+199+198 219+189+188+187 271+91+93+65 142+141+77
2004/04-2004/09		
2004/10-2005/03	248+163+199+198+196 187+189+188+193 279+276+275+274	271+91+269+93+65 219+188+187+189 279+276+275+274 193+196+199+198 244+212+210
2005/04-2005/09	248+163+199+198+196+192 187+189+188+219 279+276+275+274	248+163+199+198+196 279+276+275+274 219+188+187 158+132+131
2005/10-2006/03	198+196+189+188 279+53+276	188+189+198+196 279+53+276 101+104+103 275+274+273 80+78+77
2006/04-2006/09	248+163+199+198+196+197 187+189+188+192+193	197+196+192+199+198+163+248 142+141+144 193+189+190 275+274+273 219+188+187
2006/10-2007/03	248+199+198+197+196 285+274+273+50	196+193+158+156+248+199+198 279+53+275+276 285+274+273+50 187+192+190 271+91+93+65 219+189+188 142+141+77

2007/04-2007/09	193+192+198+187+190 156+160+159+158 132+133+131	193+192+188+198+197+196 132+133+131 278+279+56+54 141+145+144
2007/10-2008/03	192+193+189+190+187 165+163+129+156 274+56+53+276 244+210+207+208	193+192+187+189+190 275+274+276+53+56 188+199+198+196+197 248+163+165 131+156+129 141+144+145 476+475+474 210+207+208
2008/04-2008/09	193+192+197+196+190+189	187+189+190+193+192+196 278+279+275+274+273 188+199+198
2008/10-2009/03	189+190+187 192+193+188+197 158+196+156 312+45+46 276+275+274	248+163+158+156+196 193+188+189+190+187 404+400+401
2009/04-2009/09	187+190+188+189 196+131+156 144+141+145	219+187+188+189+190 263+264+262+261 144+145+141 193+196+156 133+132+131 445+443+439
2009/10-2010/03	199+198+196+192+189+187+188 145+144+141+142 156+132+131	248+163+199+198+196+192 142+141+145+144 219+188+189+187 156+158+132 279+275+274
2010/04-2010/09		
2010/10-2011/03		276+57+53+54 291+289+390
2011/04-2011/09		
2011/10-2012/03		285+269+273
2012/04-2012/09		144+141+145 290+46+45
2012/10-2015/03		

1900/01-2015/12	188+187+189+190+192+193+196	197+199+198+188+189+190+187+196+192+193
	53+56+275+276+273	274+273+276+275
	142+141+144+145	142+141+144+145
	241+169+240+173	169+241+240+173
	158+156+129	91+269+271+92
	244+165+163	244+165+163
	46+45+312	158+156+129
	262+261+264	262+261+264
	214+199+198	46+45+312
		226+227+225
		401+404+400

OmegaRatio Approaches

Table 4.2: Results for the seasonal H1N1 subtype comparing the two OmegaRatio approaches, differed by the counting method. For Seasons the introduced algorithm did not find patches of sites under selection the entries are left empty and consecutive seasons without found patches are taken together into one entry. The investigated seasons still remain the same.

Season	OR:NG+RF	OR:NG+NRF
2000/04-2000/09	240+172+173+171	240+172+173
2000/10-2001/09		
2001/10-2002/03	172+173+171	240+172+173
2002/04-2002/09		
2002/10-2003/03	240+172+173	240+172+173
2003/04-2003/09	240+172+173+171	240+172+173+171
2003/10-2004/03	240+172+173+171	240+172+173+171
2004/04-2004/09		
2004/10-2005/03	240+172+173+171	240+172+173+171
2005/04-2005/09	240+172+173+171 143+144+136	240+172+173+171 143+144+136
2005/10-2006/03	143+141+73+135+136 172+173+171+240+169	172+173+171+169+240
2006/04-2006/09	240+239+172+173+171+169 143+142+141+73+136	240+239+171+169+172+173 143+142+141
2006/10-2007/03	172+173+171+240+169	172+173+240+171+169
2007/04-2007/09	240+169+172+173+171 136+141+144 225+227+186	172+171+173+169 225+186+227 136+141+144
2007/10-2008/03	143+141+144	143+141+144

2008/04-2008/09	240+172+173+171+169	240+169+172+173+171
2008/10-2009/03	238+172+171+240+169 143+141+144	239+240+238+171+169 143+144+145+137
2009/04-2009/09	240+238+173+172+171+169 200+248+199 224+225+100	200+199+248 100+227+186+225 172+171+169 240+238+173
2009/10-2010/03	240+172+173+171	240+172+173
2010/04-2015/03		
1900/01-2015/12	171+169+238+240 212+211+210	157+158+132 171+169+238 136+141+145

OmegaValue Approach

Table 4.3: Results for the seasonal H1N1 subtype using the OmegaValue approach. For Seasons the introduced algorithm did not find patches of sites under selection the entries are left empty and consecutive seasons without found patches are taken together into one entry. The investigated seasons still remain the same.

Season	OmegaValues
2000/04-2000/09	240+172+173
2000/10-2001/09	
2001/10-2002/03	240+172+173
2002/04-2002/09	
2002/10-2003/03	240+172+173
2003/04-2003/09	240+171+173
2003/10-2004/03	240+171+173
2004/04-2004/09	
2004/10-2005/03	240+171+173
2005/04-2005/09	240+171+173 143+144+136
2005/10-2006/03	171+173+240+169
2006/04-2006/09	240+239+171+173+169
2006/10-2007/03	171+173+240+169
2007/04-2007/09	
2007/10-2008/03	143+141+144
2008/04-2008/09	240+169+171+173
2008/10-2009/03	239+240+171+169 186+227+189
2009/04-2009/09	240+171+169+173 200+199+248

2009/10-2010/03	240+172+173
2010/04-2015/03	
1900/01-2015/12	167+171+169 239+238+240

4.1.2 Subtype H1N1 - pandemic

AdaPatch Approaches

Table 4.4: Results for the pandemic H1N1 subtype comparing the two AdaPatch approaches, differed by the counting method. The tables start with the first occurrence of the pandemic H1N1 subtype in 2009. For Seasons the introduced algorithm did not find patches of sites under selection the entries are left empty and consecutive seasons without found patches are taken together into one entry. The investigated seasons still remain the same.

Season	AP:NG+RF	AP:NG+NRF
2009/04-2009/09	54+49+52 60+62+90	54+49+52 90+60+62 288+273+274
2009/10-2010/03		133+135+163+160
2010/04-2010/09	200+203+202	200+203+202 160+159+133
2010/10-2011/03	163+133+160 244+241+245+172	200+203+202+251 163+160+133
2011/04-2011/09		202+203+200 163+133+160
2011/10-2012/03		
2012/04-2012/09		200+203+202
2012/10-2013/03		143+144+138 288+273+274 135+133+136
2013/04-2013/09	291+286+43+42	291+286+43+42 202+203+251
2013/10-2014/03	200+203+202+251 273+274+88	200+203+202 190+222+188
2014/04-2015/03		
1900/01-2015/12	264+178+176+177 228+227+230+225 42+41+40 268+267+266	178+176+264+177 42+41+40 268+267+266 228+227+230

OmegaRatio Approaches

Table 4.5: Results for the pandemic H1N1 subtype comparing the two OmegaRatios approaches, differed by the counting method. The tables start with the first occurrence of the pandemic H1N1 subtype in 2009. For Seasons the introduced algorithm did not find patches of sites under selection the entries are left empty and consecutive seasons without found patches are taken together into one entry. The investigated seasons still remain the same.

2009/04-2009/09		
Season	OR:NG+RF	OR:NG+NRF
2009/10-2010/03	159+163+132+165	274+273+275 49+50+284+283
2010/04-2010/09		
2010/10-2011/03		
2011/04-2011/09	282+283+54	282+283+54
2011/10-2012/03	199+200+202 265+266+113	199+200+202 265+266+113
2012/04-2012/09	199+200+159	199+200+159
2012/10-2013/03	99+232+219	99+100+232
2013/04-2013/09		
2013/10-2014/03	134+136+159	134+136+159 89+90+88
2014/04-2014/09		
2014/10-2015/03		
1900/01-2015/12		

OmegaValue Approach

Table 4.6: Results for the pandemic H1N1 OmegaValue approach. The table starts with the first occurrence of the pandemic H1N1 subtype in 2009. For Seasons the introduced algorithm did not find patches of sites under selection the entries are left empty and consecutive seasons without found patches are taken together into one entry. The investigated seasons still remain the same.

Season	OmegaValues
2009/04-2009/09	
2009/10-2010/03	274+273+275 49+50+284+283
2010/04-2011/09	
2011/10-2012/03	199+200+202
2012/04-2012/09	199+200+159
2012/10-2013/03	99+100+232
2013/04-2013/09	
2013/10-2014/03	134+136+159
2014/04-2015/03	
1900/01-2015/12	

4.1.3 Subtype H3N2

AdaPatch Approaches

Table 4.7: Results for the H3N2 subtype comparing the two AdaPatch approaches, differed by the counting method. For Seasons the introduced algorithm did not find patches of sites under selection the entries are left empty and consecutive seasons without found patches are taken together into one entry. The investigated seasons still remain the same.

Season	AP:NG+RF	AP:NG+NRF
2000/04-2003/09		
2003/10-2004/03	325+21+37+23	
2004/04-2005/09		
2005/10-2006/03		239+174+173
2006/04-2009/03		
2009/04-2009/09	328+22+21	94+91+95
2009/10-2010/03		
2010/04-2010/09	160+158+156	21+328+22 41+23+24
2010/10-2011/03		160+156+158 280+289+47
2011/04-2011/09		239+175+173+171 21+22+328
2011/10-2012/03	156+158+160 325+37+23+22	
2012/04-2012/09	158+156+160	158+156+160 242+241+239
2012/10-2013/03	156+158+160 21+22+328 262+173+175	156+158+160 21+22+328 276+275+278 173+175+239
2013/04-2013/09		
2013/10-2014/03	328+22+23	328+22+23 189+187+188
2014/04-2014/09		
2014/10-2015/03		47+290+289

1900/01-2015/12	37+23+24+22+21 239+238+175+169 295+296+307+293 189+186+185+196 198+199+200 264+263+266 62+63+92+271 322+25+35 328+326+325	326+325+328+22+21 239+238+175+169 295+296+307+293 197+200+198+199 322+37+23+24 190+185+186+227 262+264+263+266 289+280+287 188+189+196 58+82+57

OmegaRatio Approaches

Table 4.8: Results for the H3N2 subtype comparing the two OmegaRatio approaches, differed by the counting method. For Seasons the introduced algorithm did not find patches of sites under selection the entries are left empty and consecutive seasons without found patches are taken together into one entry. The investigated seasons still remain the same.

Season	OR:NG+RF	OR:NG+NRF
2000/04-2003/09		
2003/10-2004/03	244+241+242	244+241+242
2004/04-2009/03		
2009/04-2009/09	227+222+229	227+222+229
2009/10-2011/03		
2011/04-2011/09	240+242+241	240+242+241
2011/10-2012/09		
2012/10-2013/03	229+227+222	295+293+315
2013/04-2015/03		
1900/01-2015/12		

OmegaValue Approach

Table 4.9: Results for the H3N2 subtype for the OmegaValue approach. For Seasons the introduced algorithm did not find patches of sites under selection the entries are left empty and consecutive seasons without found patches are taken together into one entry. The investigated seasons still remain the same.

Season	OmegaValues
2000/04–2003/09	
2003/10-2004/03	244+241+242
2004/04-2007/09	

2007/10-2008/03	321+322+37
2008/04-2011/03	
2011/04-2011/09	240+242+241
2011/10-2012/09	
2012/10-2013/03	240+238+241
2013/04-2015/03	
1900/04-2015/12	

4.1.4 Subtype H5N1

AdaPatch Approaches

Table 4.10: Results for the H5N1 subtype comparing the two AdaPatch approaches, differed by the counting method. For Seasons the introduced algorithm did not find patches of sites under selection the entries are left empty and consecutive seasons without found patches are taken together into one entry. The investigated seasons still remain the same.

Season	AP:NG+RF	AP:NG+NRF
2000/04-2005/03		
2005/04-2005/09		145+142+144
2005/10-2006/03		
2006/04-2006/09		142+145+144
2006/10-2007/03	158+131+130+133	
2007/04-2009/03		
2009/04-2009/09	216+189+185+188	216+189+185+188 280+50+278
2009/10-2010/03	189+185+188+186	216+188+186
2010/04-2010/09	280+50+278+53	280+50+278+53 188+189+185
2010/10-2011/03		158+159+160 269+88+285 227+187+188 145+71+144 127+128+124
2011/04-2011/09		
2011/10-2012/03		159+158+131
2012/04-2013/09		
2013/10-2014/03		269+87+88+91
2014/04-2015/03		

1900/01-2015/12	226+227+189+187 128+127+255 160+159+158 269+88+285 193+196+192	188+186+189+187 128+127+124 160+159+158 193+196+192 135+133+131

OmegaRatio Approaches

Table 4.11: Results for the H5N1 subtype comparing the two OmegaRatio approaches, differed by the counting method. For Seasons left out of the table the introduced algorithm did not find patches of sites under selection and consecutive seasons without found patches are taken together into one entry. The investigated seasons still remain the same.

Season	OR:NG+RF	OR:NG+NRF
2000/04-2015/03		
1900/01-2015/12	141+144+142	59+43+285 279+278+280 141+142+144

OmegaValue Approach

Table 4.12: Results for the H5N1 subtype for the OmegaValue approach. For Seasons the introduced algorithm did not find patches of sites under selection the entries are left empty and consecutive seasons without found patches are taken together into one entry. The investigated seasons still remain the same.

Season	OmegaValues
2000/04-2015/03	
1900/01-2015/12	279+278+280

4.1.5 Subtypes H7

For the Subtype H7N9, and also H7N7, no patches under selection have been detected using IPoSuS.

4.2 Evaluation

In this section all results will be evaluated against literature known amino acids under selectional pressure. Therefore the results of each subtype are put together so that there is a relative count of appearances in patches of this amino acid position. Furthermore all results will get introduced to the H3N2 numbering. According to the resulting numbering the results are evaluated, whether they are located in a known patch, are mentioned in other literature or are newly found. Other literature in this case means that these positions are not in patches, but are also have experimental proof to be under selection. This differentiation is needed, because the positions of other literature are not used for evaluation in other publications. For this reason the reported patches will be used as well as all results based on other literature. The evaluation data for subtypes of H3 is given in Table 1.2 and table 1.3 shows the data used for subtype H1.

4.2.1 Pandemic H1N1

Combining all resulting amino acids contained in patches, we get 64 different positions which are found to be under selectional pressure as shown in Table 4.13.

Table 4.13: Results for the pandemic H1N1 subtype with dispatched patches. The first column indicates the amino acids position of the patches found by the introduced algorithm, the second column shows the according H3N2 numbering, the third column shows how often this amino acid position was found in a patch and the fourth column shows whether this specific amino acids position is located in a known patch - therefore characterized by a capital letter as denoted in Table 1.2 and Table 1.3- or was reported in any other publication as position under selection. Furthermore, "new" indicates that this position was newly found as position under selectional pressure. Patches are assigned using the H3 numbering. Patches from the evaluation data are labeled by Sa, Sb, Ca1, Ca2, Cb for H1 and A, B, C, D and e for H3 subtype.

Amino Acid Position	H3 Numbering	Count	Reported
40	39	2	Li2011 [42] + Arunachalam2013 [2]
41	40	2	Arunachalam2013 [2]
42	41	4	Arunachalam2013 [2]
43	42	2	Arunachalam2013 [2]
49	48	3	Arunachalam2013 [2] + Ding2010 [18]
50	49	1	Arunachalam2013 [2]
52	51	2	Arunachalam2013 [2]
54	53	4	Arunachalam2013 [2]
60	58	2	Arunachalam2013 [2]
62	60	2	Arunachalam2013 [2]
88	85	2	new
89	86	1	Li2011 [42]
90	87	3	new
99	95	2	new
100	96	1	new
113	109	2	new

132	125	1	new
133	126	6	new
134	127	3	new
135	128	2	new
136	129	4	new
138	131	1	new
143	135	1	new
144	136	1	new
159	151	8	new
160	152	5	new
163	155	5	new
165	157	1	new
172	164	1	new
176	168	2	new
177	169	2	Ca1
178	170	2	new
188	180	1	Lee2015 [41]
190	182	1	new
199	191	6	new
200	192	13	new
202	194	11	new
203	195	8	new
219	211	1	new
222	214	1	Zehender2012 [93] + Ding2010 [18]
225	217	1	new
227	219	2	new
228	220	2	Lee2015 [41]
230	222	2	Li2011 [42]
232	224	2	Li2011 [42]
241	233	1	new
243	235	1	new
244	236	2	new
245	237	2	Li2011 [42]
251	243	3	new
264	256	2	new
265	257	2	new
266	258	4	new
267	259	2	new
268	260	2	new
273	264	5	new
274	265	5	new

275	266	2	new
282	273	2	new
283	274	3	new
284	275	1	new
286	277	2	new
288	279	2	new
291	282	2	Li2011 [42]

For the pandemic H1N1 subtype no patch data is available and therefore the results have to be compared to only literature data.

Of the 64 found amino acid positions under selection 18 are reported in the publications [2, 18, 41, 42, 93]. According to the results 28.13% of the found 64 amino acid positions are already known to be under selection. On the other hand there are 46 newly identified candidate positions found which equals 71.87% as shown in Figure 4.1.

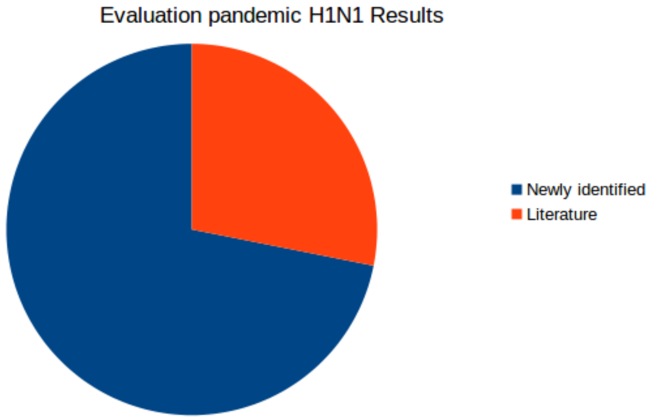

Figure 4.1: Figure showing the results using subtype specific evaluation data. The blue area corresponds to the amount of positions that are newly identified and the orange area represents positions that already have been reported in literature.

Figure 4.2.1 shows all findings mapped onto the 3D-structure as well as all positions which are exposed on the outer surface of the protein. Both combined show the difference between the found and exposed positions of the protein, giving a better understanding of the used algorithm.

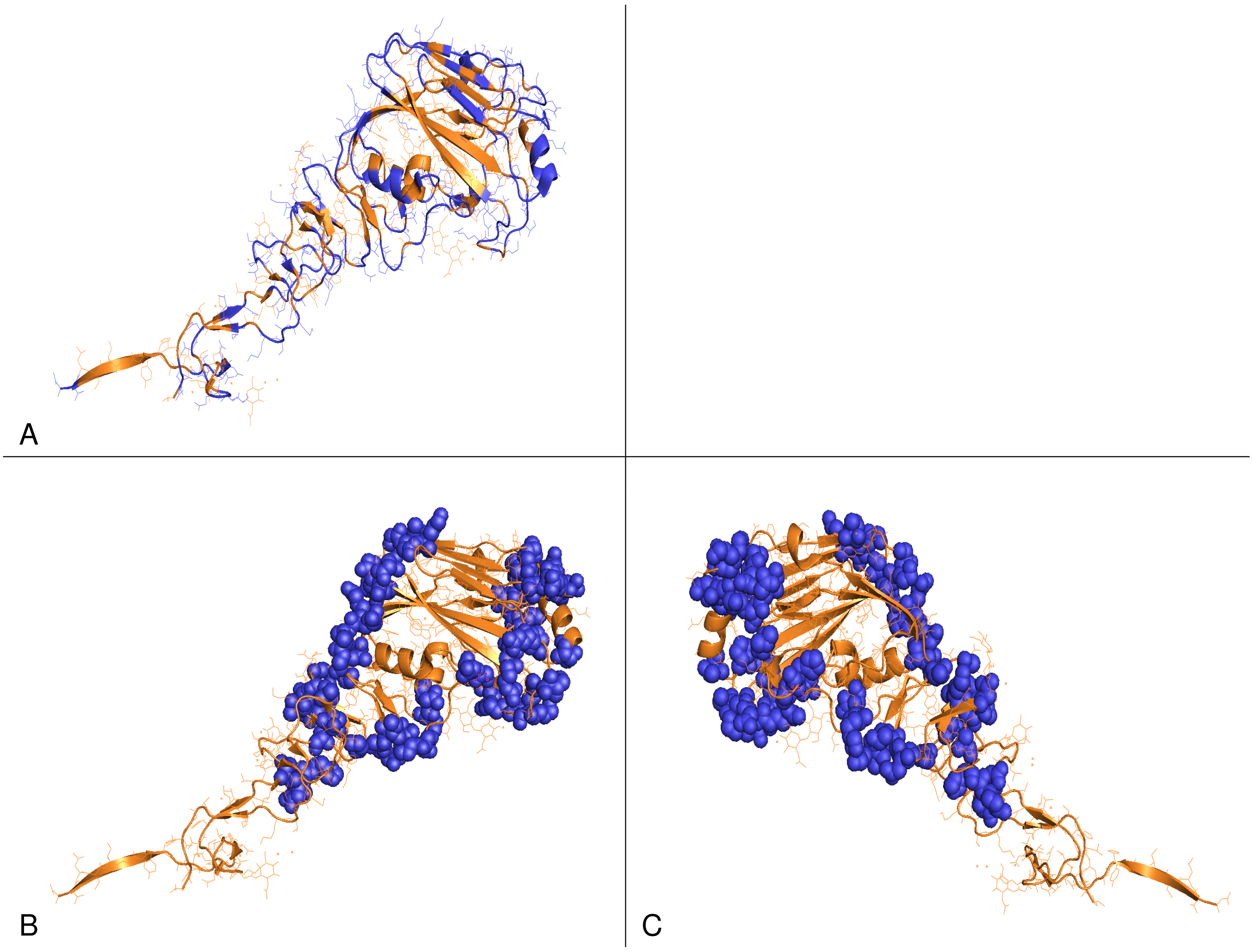

Figure 4.2: Figure showing different amino acid positions on the 3D-structure of the HA protein for the pandemic H1N1 subtype in orange. Picture on the top left (A) shows all positions that are exposed on the outer surface in blue. Pictures on the bottom, left (B) and right (C), show the found positions using IPoSuS as spheres in blue.

4.2.2 Seasonal H1N1

Taking all results together there are 92 different positions which are found to be under selectional pressure as shown in Table 4.14.

Table 4.14: Results for the seasonal H1N1 subtype with dispatched patches. The first column indicates the amino acids position of the patches found by the introduced algorithm, the second column shows the according H3N2 numbering, the third column shows how often this amino acid position was found in a patch and the fourth column shows whether this specific amino acids position is located in a known patch - therefore characterized by a capital letter as denoted in Table 1.2 and Table 1.3 - or was reported in any other publication as position under selection. Furthermore a "new" indicates that this position was newly found as position under selectional pressure. Patches are assigned using the H3 numbering. Patches from the evaluation data are labeled by Sa, Sb, Ca1, Ca2, Cb for H1 and A, B, C, D and e for H3 subtype.

Amino Acid Position	H3 Numbering	Count	Reported
45	51	4	new
46	52	4	new
50	56	1	Gianfrani2000 [24]
53	58	4	new
54	59	1	new
56	61	2	Gianfrani2000 [24]
57	62	1	new
65	70	1	new
73	78	2	Cb
78	82	1	Cb
80	84	1	Cb
91	94	1	new
92	95	1	new
93	96	1	new
100	103	2	new
101	104	1	new
103	106	1	new
104	107	1	new
129	129	3	Sa
131	131	5	new
132	132	4	new
135	134	1	new
136	135	8	new
137	136	1	Gianfrani2000 [24]
141	140	17	Ca2
142	141	6	new

143	142	11	new
144	143	17	Ca2
145	144	9	new
156	155	10	new
157	156	1	Sb
158	157	7	new
159	158	1	Sa
160	159	1	Sb
163	162	11	Sa
165	164	3	new
167	166	1	Sa
169	168	25	new
171	170	37	new
172	171	33	new
173	172	43	new
186	185	4	new
187	186	16	new
188	187	15	new
189	188	18	new
190	189	8	new
192	191	11	new
193	192	13	Sb
196	195	16	new
197	196	6	Sb
198	197	14	new
199	198	15	Sb
200	199	3	new
207	206	1	new
208	207	1	Ca1
210	209	2	new
211	210	1	Ping2011 [60]
212	211	1	new
214	213	1	new
219	218	3	new
224	223	1	new
225	224	5	Ca2
226	225	1	Ca2
227	226	5	new
238	237	7	new
239	238	5	new
240	239	46	new

241	240	2	Ca1
244	243	3	new
248	247	12	new
261	260	3	new
263	261	1	new
264	262	3	new
269	267	2	Jones1994 [29],Stern1994 [71]
271	269	1	Jones1994 [29],Stern1994 [71]
273	271	4	new
274	272	9	new
275	273	8	Ca1
276	274	10	new
278	276	10	new
279	277	6	new
285	283	2	new
289	287	1	new
290	288	1	new
291	289	1	new
312	310	3	new
390	390	1	new
400	400	1	new
401	401	1	new
404	404	1	new
439	439	1	Gianfrani2000 [24]
443	443	1	Gianfrani2000 [24]
445	445	1	Gianfrani2000 [24]
474	474	2	new
475	475	2	new
476	476	2	new
482	482	2	new
133	-	2	new
262	-	3	new
77	-	1	new

Of this 100 amino acids 19 are located patches of H1 subtype evaluation data, while nine are reported in other publications. This means that a total of 28 amino acids known. On the other hand this means that 72 positions are newly identified candidate positions.

According to the results 19.00% of the found 100 amino acid positions are located in known patches of H1 evaluation data and 9.00% are are known by other literature, which makes a total of 28.00% amino acid positions known. That makes a total of 72.00% newly identified candidate positions found as shown in Figure 4.3.

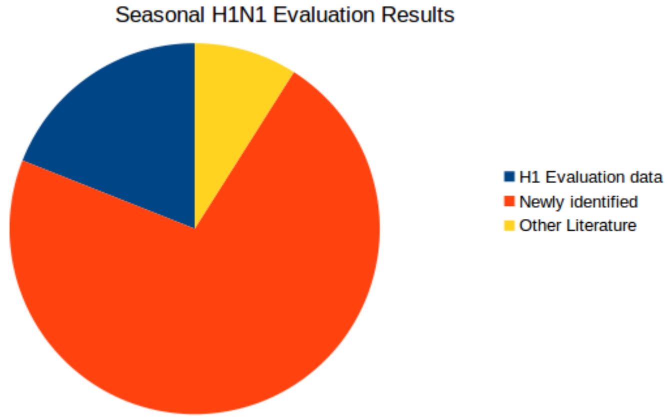

Figure 4.3: Figure showing the results using subtype specific evaluation data. The blue area corresponds to the amount of positions in known patches, the orange part represents the amount of newly identified positions and the yellow area represents positions that already have been reported in other literature.

Figure 4.2.2 shows all findings mapped onto the 3D-structure of the used protein and also shows all amino acid positions that are exposed on the outer surface of the protein. Both combined show the difference between the found and exposed positions of the protein, giving a better understanding of the used algorithm.

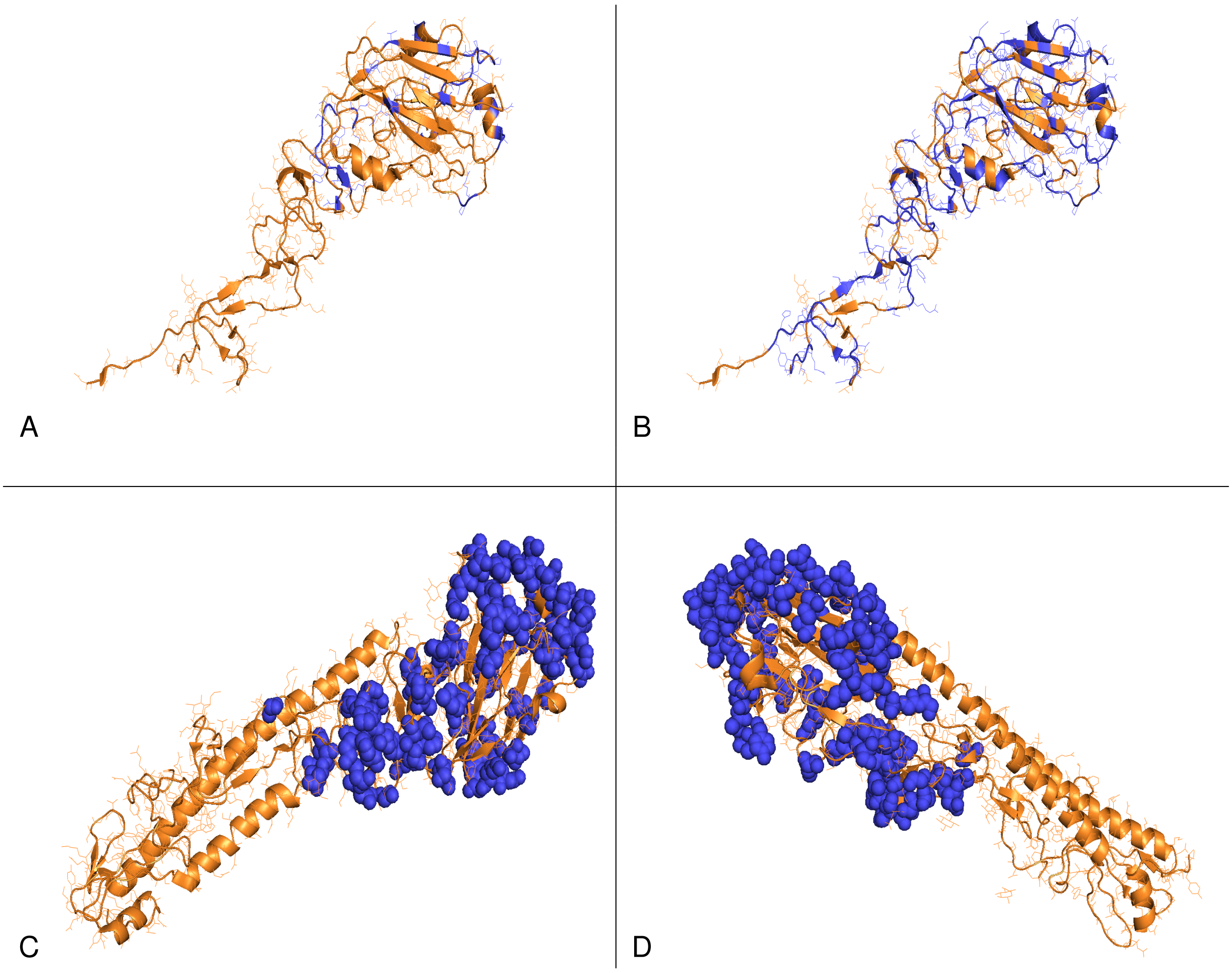

Figure 4.4: Figure showing different amino acid positions on the 3D-structure of the HA protein for the seasonal H1N1 subtype in orange. Picture on the top left (A) shows the positions that have to be found using H1 evaluation data in blue. Picture on the top right (B) shows all positions that are exposed on the outer surface in blue. Pictures on the bottom left (C) and right (D) show the found positions using IPoSuS as spheres in blue.

4.2.3 H3N2

Depatching all resulting patches IPoSuS has identified 69 different positions which are under selectional pressure, as shown in Table 4.15.

Table 4.15: Results for the H3N2 subtype with dispatched patches. The first column indicates the amino acids position in the patches found by the introduced algorithm, the second column shows how often this amino acid position was found in a patch and the third column shows whether this specific amino acids position is located in a known patch - therefore characterized by a capital letter as denoted in Table 1.2 and Table 1.3 - or was reported in any other publication as position under selection. Furthermore a "new" indicates that this position was newly found as position under selectional pressure. Patches are assigned using the H3 numbering. Patches from the evaluation data are labeled by Sa, Sb, Ca1, Ca2, Cb for H1 and A, B, C, D and E for H3 subtype.

Amino Acid Position	Count	Reported
21	8	new
22	10	new
23	7	new
24	3	new
25	1	new
35	1	new
36	3	new
37	2	new
41	1	new
47	2	C
57	1	Gianfrani 2000 [24]
58	1	Gianfrani 2000 [24]
62	1	E
63	1	E
82	1	E
91	1	E
92	1	E
94	1	E
95	1	new
156	6	B
158	7	B
160	7	B
169	3	new
171	1	D
173	4	D
174	1	D

175	5	D
185	2	new
186	2	B
187	1	B
188	2	B
189	2	B
190	1	B
196	1	B
197	1	B
198	2	B
199	2	new
200	2	new
222	3	new
227	4	D
229	3	D
238	3	D
239	6	new
240	4	D
241	8	new
242	7	D
244	3	D
262	2	E
263	2	Jones 1994 [29], Stern 1994 [71]
264	2	E
266	2	Jones 1994 [29], Stern 1994 [71]
271	1	new
275	1	C
276	1	C
278	1	C
280	2	C
287	1	new
289	3	new
290	1	new
293	3	new
295	3	new
296	2	new
307	2	C
315	1	Carmichael 1997 [11]
321	1	Carmichael 1997 [11]
322	3	Carmichael 1997 [11]
325	3	new

326	2	new
328	9	new

Of this 69 amino acids 35 are located in known patches of H3 subtype evaluation data, whereas 7 are reported in other publications. This means that a total of 27 amino acids are newly found. These are 21, 22, 23, 24, 25, 35, 36, 37, 41, 95, 185, 199, 200, 222, 239, 241, 271, 287, 289, 290, 293, 295, 296, 325, 326 and 328

According to the results 50.72% of the found 69 amino acid positions are located in known patches. Additionally 10.15% are known by other literature, which makes a total of 60.87% amino acid positions known. Furthermore a total of 39.13% are newly found, as shown in Figure 4.5.

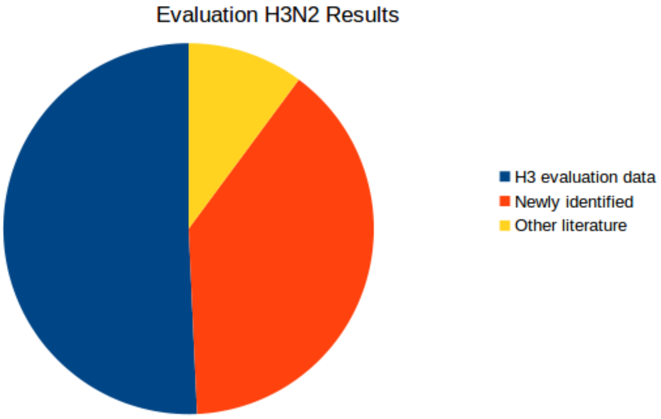

Figure 4.5: Figure showing the results using subtype specific evaluation data. The blue area corresponds to the amount of positions in known patches, the orange part represents the amount of newly identified positions and the yellow area represents positions that already have been reported in other literature.

Figure 4.2.3 shows all amino acid positions that are exposed on the outer surface of the protein, all amino acid positions that are in the H3 evaluation data and all findings using IPoSuS. The figure combined shows the difference between the found and exposed positions of the protein, giving a better understanding of the used algorithm.

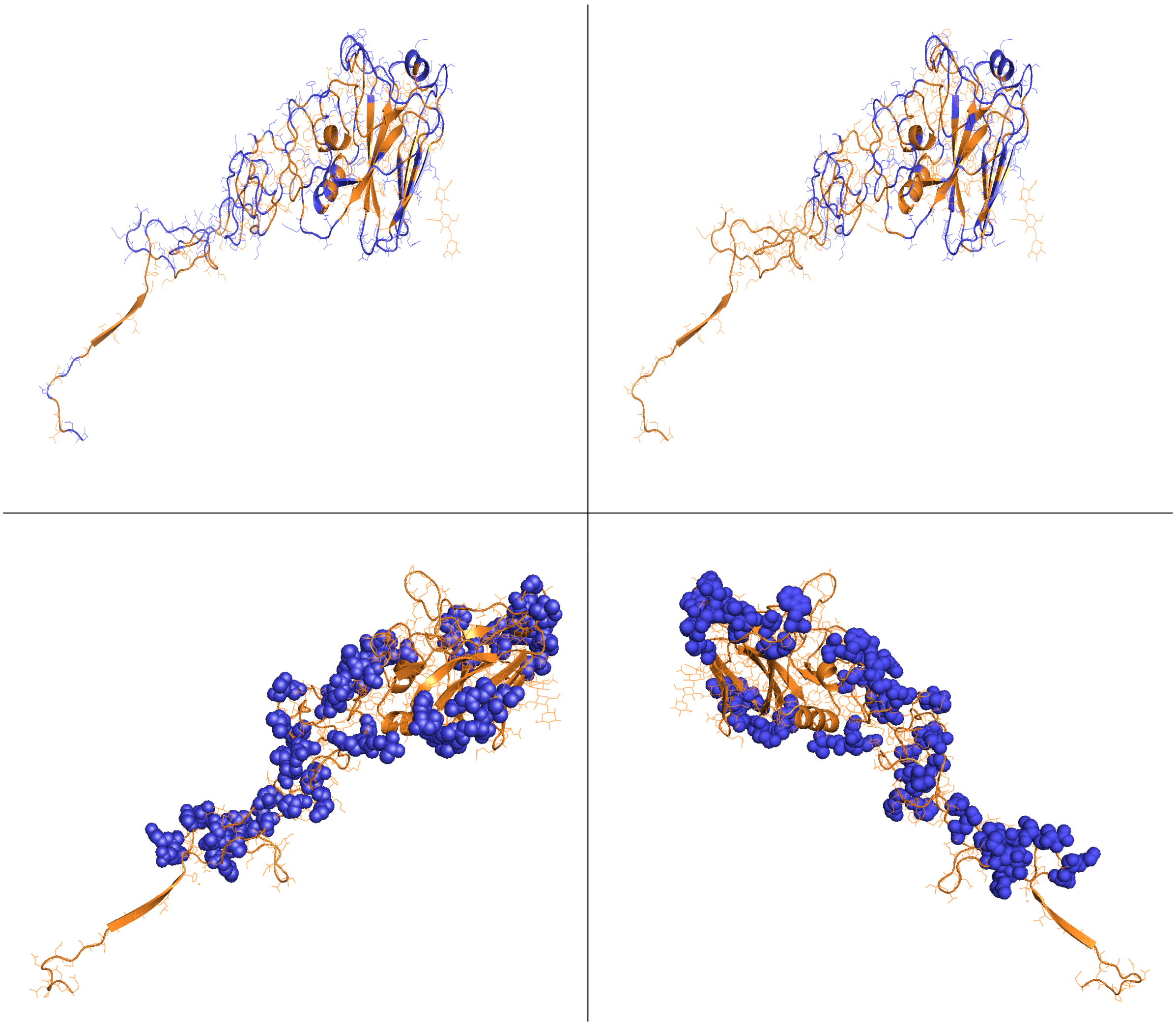

Figure 4.6: Figure showing different amino acid positions on the 3D-structure of the HA protein for subtype H3N2 in orange. Picture on the top left shows the positions that are exposed on the outer surface in blue. Picture on the top right shows the positions that have to be found in using H3 evaluation data in blue. Pictures on the bottom, left and right, show the found positions using IPoSuS as spheres in blue.

4.2.4 H5N1

Considering all patches, we get 40 different positions which are found to be under selectional pressure for this subtype, as shown in Table 4.16.

Table 4.16: Results for the H5N1 subtype with dispatched patches. The first column indicates the amino acids position of the patches found by the introduced algorithm, the second column shows the according H3N2 numbering, the third column shows how often this amino acid position was found in a patch and the fourth column shows whether this specific amino acids position is located in a known patch - therefore characterized by a capital letter as denoted in Table 1.2 and Table 1.3 - or was reported in any other publication as position under selection. Furthermore a "new" indicates that this position was newly found as position under selectional pressure. Patches are assigned using the H3 numbering. Patches from the evaluation data are labeled by Sa, Sb, Ca1, Ca2, Cb for H1 and A, B, C, D and e for H3 subtype.

Amino Acid Position	H3N2 numbering	Count	Reported
43	49	1	new
50	55	1	new
52	57	2	new
53	58	1	new
55	60	3	new
59	64	1	new
71	76	1	new
87	91	1	new
88	92	1	new
91	94	1	new
92	95	3	new
124	125	2	new
127	128	3	new
128	129	2	Kongchanagul2008 [35]
130	131	1	new
131	132	3	new
133	133	2	Kongchanagul2008 [35]
135	135	1	new
141	141	2	new
142	142	4	new
144	144	5	new
145	145	3	new
158	158	5	Kongchanagul2008 [35]
159	159	4	new
160	160	3	new
169	169	2	new

185	185	4	new
186	186	3	Kongchanagul2008 [35]
187	187	3	new
188	188	7	new
189	189	6	new
192	192	2	new
193	193	2	new
196	196	2	new
216	216	3	new
226	226	1	new
227	227	2	Kongchanagul2008 [35]
255	255	1	new
269	268	1	new
278	277	4	new
279	278	2	new
280	279	5	new
285	284	3	new

For subtype H5N1 no individual patch data is available, therefore we evaluate the results without the H1 and H3 evaluation data. Without using them, there are only five positions of the 43 found, that are already reported to be under selection by Kongchanagul [35]. This means that the remaining 38 positions found are newly identified candidates. This equals 11.63% known positions and 89.37% newly found ones, as shown in Figure **??**.

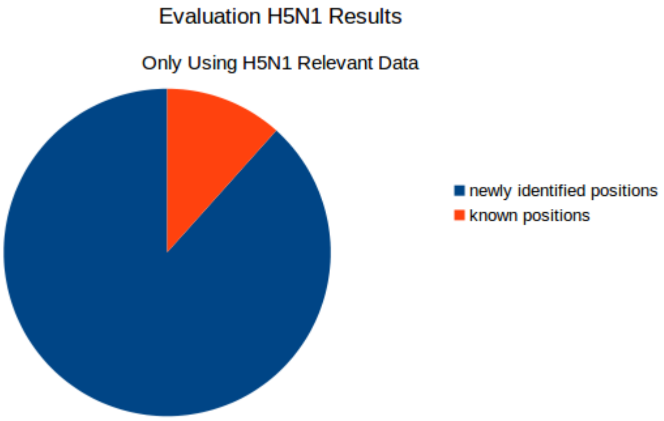

Figure 4.7: Graphic showing the results of the evaluation of the results of subtype H5N1 only using H5N1 relevant data. The blue area corresponds to the amount of newly identified positions, while the orange part represents the amount of found positions that already have been reported.

Figure 4.2.4 shows all findings mapped onto the 3D-structure of the used protein as well as all exposed amino acids on the outer surface. Both combined show the difference between the found and exposed positions of the protein, giving a better understanding of the used algorithm.

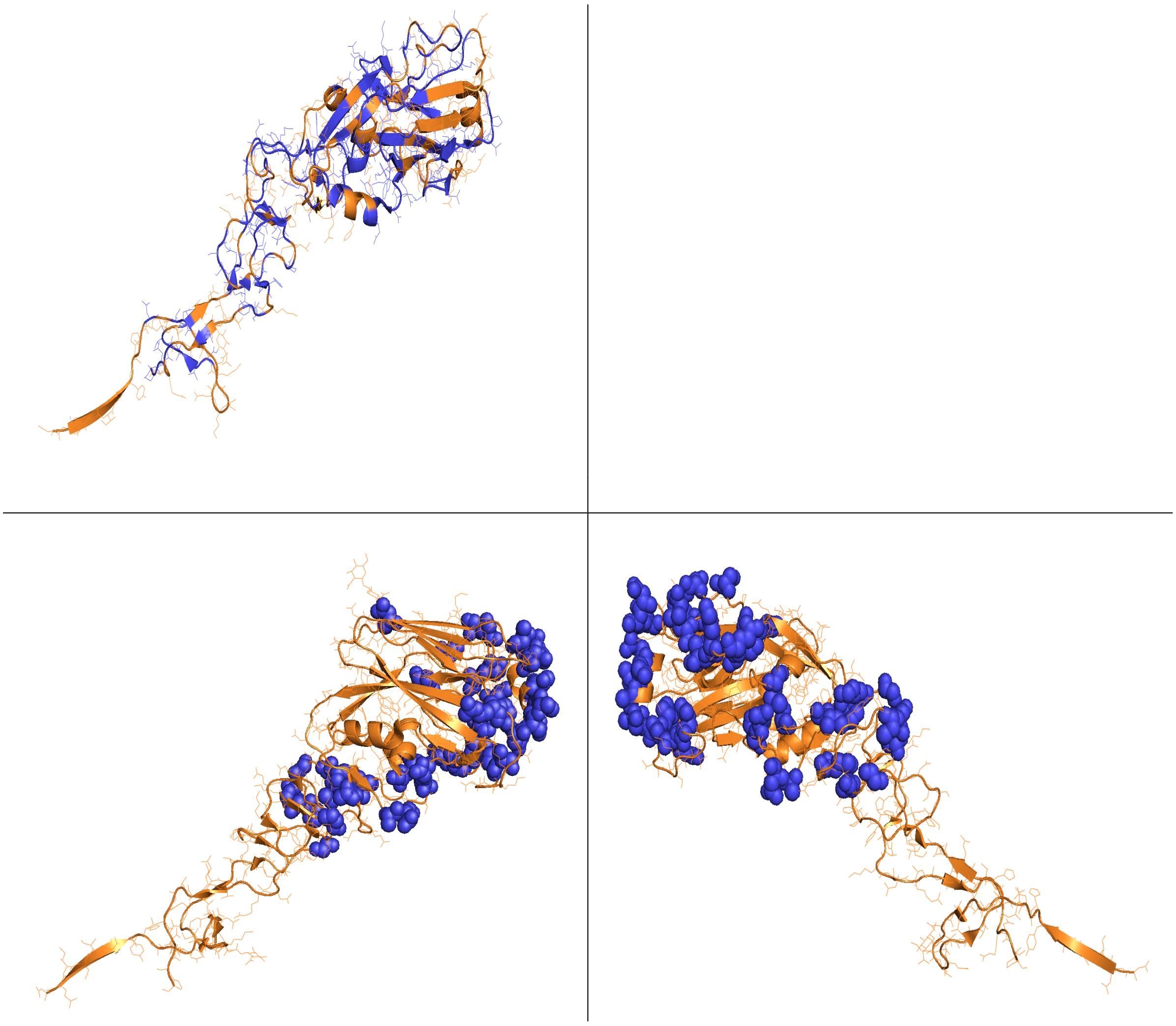

Figure 4.8: Figure showing different amino acid positions on the 3D-structure of the HA protein for subtype H5N1 in orange. Picture on the top left shows the positions that are exposed on the outer surface of the protein in blue. Pictures on the bottom, left and right, show the found positions using IPoSuS as spheres in blue.

4.3 Approach Analysis

In this section the obtained results for each subtype and approach will be shown.

4.3.1 Seasonal H1N1

For the two different counting methods using the AdaPatch approach the differences in the results for both approaches are only determined by the approach using the counting NG+NRF. Therefore every single amino acid occurring in the AdaPatch approach using NG+RF counting is present in the one using NG+NRF counting. There are 30 different amino acid positions found, as shown in Table 4.17.

Table 4.17: Differences and similarities between the two different counting schemes used, considering the AdaPatch (AP) approach, which means that the test statistic used is the exact Fisher test, for the subtype sH1N1. The used numbering is according to the subtype, because a direct comparison is possible and needs no shifting to match other subtypes.

	AP:NG+RF Specific	AP:NG+NRF Specific	AP Similarities
Position	-	54, 57, 65, 77, 78, 80, 91, 92, 93,101, 103, 104, 212, 225, 226, 227, 263, 269, 271, 278, 289, 290, 291, 390, 400, 401, 404, 439, 443, 445	45, 46, 50, 53, 56, 129, 131, 132, 133, 141, 142, 144, 145, 56, 158, 159, 160, 163, 165, 169, 173, 187, 188, 189, 190, 192, 193, 196, 197, 198, 199, 207, 208, 210, 214, 219, 240, 241, 244, 248, 261, 262, 264, 273, 274,275, 276, 279,285, 312, 474, 475, 476, 482

For the two different counting methods using the OmegaRatio approach there are differences in both approaches. Unique for the approach using NG+RF counting are six positions, while the approach using NG+NRF counting has five unique positions. Similar are 19 positions, as shown in Table 4.18.

Table 4.18: Differences and similarities between the two different counting schemes used, considering the OmegaRatio (OR) approach, which means that the test statistic used is the Gaussian Z-Test, for the subtype sH1N1. The used numbering is according to the subtype, because a direct comparison is possible and needs no shifting to match other subtypes.

	OR:NG+RF specific	OR:NG+NRF Specific	OR Similarities
Position	73, 135, 210, 211, 212, 224	132, 137, 145, 157, 158	100, 136, 141, 142, 143, 144,169, 171, 172, 173, 186, 199, 200, 225, 227, 238, 239, 240, 248

Considering these four approaches, together with the OmegaValue approach, all these have an overlap in the amino acids 141, 144, 169, 173, 199, 240 and 248.

4.3.2 Pandemic H1N1

Using the AdaPatch approach with its two different counting schemes, we get differences in both approaches. For the approach using NG+RF counting there are six positions that are unique, while using NG+NRF counting results in ten unique positions. Both approaches are similar in 32 positions, as shown in Table 4.19.

Table 4.19: Differences and similarities between the two different counting schemes used, considering the AdaPatch (AP) approach, which means that the test statistic used is the exact Fisher test, for the subtype pH1N1. The used numbering is according to the subtype, because a direct comparison is possible and needs no shifting to match other subtypes.

	AP:NG+RF Specific	AP:NG+NRF Specific	AP Similarities
Position	25, 88, 172, 241, 244, 245,	135, 136, 138, 143, 144, 159, 188, 190, 222, 288	40, 41, 42, 43, 49, 52, 54, 59, 60,62, 90, 133, 160, 163, 176, 177, 178, 200, 202, 203, 227, 228, 230, 251, 264, 266, 267, 268, 273, 274, 286, 291

For the two different counting methods using the OmegaRatio approach there are differences in both approaches. Unique for the approach using NG+RF counting there are six unique positions, while the NG+NRF counting results in nine unique positions. Both approaches are similar in 12 positions, as shown in Table 4.20.

Table 4.20: Differences and similarities between the two different counting schemes used, considering the OmegaRatio (OR) approach, which means that the test statistic used is the Gaussian Z-Test, for the subtype pH1N1. The used numbering is according to the subtype, because a direct comparison is possible and needs no shifting to match other subtypes.

	OR:NG+RF Specific	OR:NG+NRF Specific	OR Similarities
Position	99, 132, 163, 165, 219, 232	49, 50, 88, 89, 90, 273, 274, 275, 284	54, 113, 134, 136, 159, 199, 200, 202, 265,266, 282, 283

Considering these four approaches, together with the OmegaValue approach, all these have an overlap in the amino acids 200 and 202.

4.3.3 H3N2

For the two different counting methods using the AdaPatch approach there are differences in both approaches. Unique for the approach using NG+RF counting are six, while the approach using NG+NRF counting has 24 unique positions. Similar are 32.

Table 4.21: Differences and similarities between the two different counting schemes used, considering the AdaPatch (AP) approach, which means that the test statistic used is the exact Fisher test, for the subtype H3N2. The used numbering is according to the subtype, because a direct comparison is possible and needs no shifting to match other subtypes.

	AP:NG+RF Specific	AP:NG+NRF Specific	AP Similarities
Position	25, 35, 62, 63, 92, 271	41, 47, 57, 58, 82, 91, 94, 95, 171,174, 187, 188, 190, 197, 227, 241, 242, 275, 276, 278, 280,287, 289, 290	21, 22, 23, 24, 37, 156, 158, 160, 169, 173, 175, 185, 186, 189, 196, 198, 199, 200, 238, 239, 262, 263, 264, 266, 293, 295, 296, 307, 322, 325, 326, 328

For the two different counting methods using the OmegaRatio approach the differences are only determined by the approach using the counting NG+NRF. Therefore every single amino acid occurring in the AdaPatch approach using NG+RF counting is present in the approach using NG+NRF counting. The different Amino acids are: 293, 295 and 315.

Table 4.22: Differences and similarities between the two different counting schemes used, considering the OmegaRatio (OR) approach, which means that the test statistic used is the Gaussian Z-Test, for the subtype H3N2. The used numbering is according to the subtype, because a direct comparison is possible and needs no shifting to match other subtypes.

	OR:NG+RF Specific	OR:NG+NRF Specific	OR Similarities
Position	-	293, 295, 315	222, 227, 229, 240, 241, 242, 244

Considering all five different approaches there is no overlap.

4.3.4 H5N1

For the two different counting methods using the AdaPatch approach there are differences in both approaches. Unique for the approach using NG+RF counting are the positions 130, 226 and 255, while the positions 71, 87, 91, 124, 135, 142, 144 and 145 are unique for the approach using NG+NRF counting. Similar are the positions are 50, 53, 88, 127, 128, 131, 133, 158, 159, 160, 185, 186, 187, 188, 189, 192, 193, 196, 216, 227, 269, 278, 280 and 285.

Table 4.23: Differences and similarities between the two different counting schemes used, considering the AdaPatch (AP) approach, which means that the test statistic used is the exact Fisher test, for the subtype H5N1. The used numbering is according to the subtype, because a direct comparison is possible and needs no shifting to match other subtypes.

	AP:NG+RF Specific	AP:NG+NRF Specific	AP Similarities
Position	130, 226, 255	71, 87, 91, 124, 135, 142, 144, 145	50, 53, 88, 127, 128, 131, 133, 158, 159, 160, 185, 186, 187, 188, 189, 192, 193, 196, 216, 227, 269, 278, 280, 285

For the two different counting methods using the OmegaRatio approach the differences are only determined by the approach using the counting NG+NRF. Therefore every single amino acid occurring in the AdaPatch approach using NG+RF counting is present in the approach using NG+NRF counting. The different Amino acids are: 43, 59, 278, 279, 280 and 285 using NG+NRF counting scheme.

Table 4.24: Differences and similarities between the two different counting schemes used, considering the OmegaRatio (OR) approach, which means that the test statistic used is the Gaussian Z-Test, for the subtype H5N1. The used numbering is according to the subtype, because a direct comparison is possible and needs no shifting to match other subtypes.

	OR:NG+RF Specific	OR:NG+NRF Specific	OR Similarities
Position	-	43, 59, 278, 279, 280, 285	141, 142, 144

Considering all five different approaches there is no overlap.

4.3.5 Overlap pH1N1 and sH1N1

Considering the two subtypes of H1N1 - pandemic and seasonal - we get 24 amino acid positions that are similar over all approaches using the H3 numbering. These positions are 51, 58, 95, 96, 129, 131, 135, 136, 155, 157, 164, 168, 170, 191, 192, 195, 211, 224, 237, 243, 260, 273, 274 and 277.

4.3.6 Overlap H3N2 and H5N1

Comparing the obtained results of the subtypes H3N2 and H5N1 there is an overlap in 17 positions. These positions are 57, 58, 91, 92, 94, 95, 158, 160, 169, 185, 186, 187, 188, 189, 196, 227 and 278.

4.3.7 Overlap H5N1 and seasonal H1N1

Comparing the obtained results of the subtypes H3N2 and H5N1 there is an overlap in 21 positions. These positions are 58, 94, 95, 129, 131, 132, 135, 141, 142, 144, 158, 159, 185, 186, 187, 188, 189, 192, 196, 226 and 277.

4.3.8 Overlap H5N1 and pandemic H1N1

Comparing the obtained results of the subtypes H3N2 and H5N1 there is an overlap in 13 positions. These positions are 49, 58, 60, 95, 125, 128, 129, 131, 135, 169, 192, 277 and 279.

4.3.9 Overlap H3N2 and seasonal H1N1

Comparing the obtained results of the subtypes H3N2 and H5N1 there is an overlap in 25 positions. These positions are 58, 62, 82, 94, 95, 156, 158, 171, 185, 186, 187, 188, 189, 196, 197, 198, 199, 238, 239, 240, 262, 271, 276, 287 and 289.

4.3.10 Overlap H3N2 and pandemic H1N1

Comparing the obtained results of the subtypes H3N2 and H5N1 there is an overlap in eight positions. These positions are 41, 58, 95, 169, 222, 264, 266 and 275.

4.3.11 Overlap pH1N1 and sH1N1 compared to other Subtypes

Considering only the overlap of the seasonal and the pandemic H1N1 subtype and compare them to the results obtained for the other subtypes there is also an overlap. Compared to the H3N2 subtype the overlap is two amino acid positions great. These two positions are 58 and 95. Comparing the overlap to the H5N1 subtype there also is an overlap, which is seven amino acids great. These amino acids positions are, all numbered according to the introduced H3 numbering, 58, 95, 129, 131, 135, 192 and 277.

4.3.12 All approaches of all proteins

According to the HA masteralignment of all different subtypes we have some shifts in the results. Using the new results - H3N2 numbering - we get the following singularities and similarities.

Similarities

Pooling the results of the different approaches for the different subtypes - five approaches for each subtype - lead to an overlap in two amino acid position, namely positions **58** and **95**. This result is based on the here used H3 numbering.

Differences

After pooling the results, every subtype has its own unique results when comparing them to the other subtypes, based on the used H3 numbering. The pandemic H1N1 has 29, the seasonal H1N1 has 44, subtype H3N2 has 32 and subtype H5N1 has ten unique results, as shown in table 4.25.

Table 4.25: Table showing the subtype specificities after pooling the results of all 5 different approaches for each subtype.

-	pH1N1 Specific	sH1N1 specific	H3N2 specific	H5N1 specific
Position	39, 40, 42, 48, 53, 85, 86, 87, 109, 126, 127, 151, 152, 180, 182, 194, 214, 217, 219, 220, 233, 235, 236, 256, 257, 258, 259, 265, 282	52, 56, 59, 61, 70, 78, 84, 103, 104, 106, 107, 134, 140, 143, 162, 166, 172, 206, 207, 209, 210, 213, 218, 223, 225, 247, 261, 267, 269, 272, 390, 400, 401, 404, 439, 443, 445, 474, 475, 476, 482	21, 22, 23, 24, 25, 35, 36, 37, 47, 63, 173, 174, 175, 190, 200, 229, 241, 242, 244, 263, 280, 290, 293, 295, 296, 307, 315, 321, 322, 325, 326, 328	55, 64, 76, 133, 145, 193, 216, 255, 268, 284

4.3.13 Evaluation

The first of the amino acid positions overlap, which is present in every subtype, is not located in a patch, but lies between position 94, which is part of patch E in the H3 evaluation data and position 96 which is part of patch D, also in H3 evaluation data. Using the H1 evaluation data there is no result for this position, such that it can be classified as new finding.

For position 58 the situation is equivalent, it is not located in a reported patch, but lies between two positions - 57 and 59 - of patch E, using H3 evaluation data. Taking H1 evaluation data into account there is again no result, such that this position can also be seen as new finding. Not using known patch data but also using literature data position 58 is a known position, because Gianfrani has already reported this position [24].

5 Discussion

In this section all results are discussed. This includes the evaluation, as well as analysis for the different approaches. Furthermore all new findings will be shown and the used evaluation data will be discussed. In the end there also is a recommendation for the usage of IPoSuS, such that it yields best results.

5.1 Evaluation

Because [79] already applied the graph-cut algorithm to data of the HA protein of influenza A, we expect results of at least the same quality with comparable resulting amino acid positions. Furthermore [79] did not consider other publications, but only known patch data. Considering the results they had 17 out of 35 positions in patch data for the seasonal H1N1 and 33 of 35 for subtype H3N2. The results for the pandemic H1N1 have never been evaluated because no patch data is available. This means that [79] obtained 48.57% known positions for the seasonal H1N1 and 94.29% for subtype H3N2. In this thesis we obtained 28.00% of known positions for the seasonal H1N1, combining H1 evaluation data and other literature. For the subtype H3N2 we obtain a value of 60.87% known positions, again combining literature and patch data. Comparing these results, we obtained results of lower quality than [79] using the same graph-cut algorithm. The main factor, leading to such a high difference between the quality of the results is probably the amount of considered data and approaches which lead to negative, or better, new results. This is due to the algorithm, which finds also neighbors and, because a patch consists of at least three amino acid positions, it is likely to have a high rate of new positions found. Furthermore [79] only uses one single, specific, manually curated dataset from another database, which could also lead to some shifting in the results.

Getting more into detail, IPoSuS works as intended. This can be seen by comparing the obtained results with the ones published by [79]. The results obtained by using IPoSuS are overlapping in 21 positions with the published results using the graph cut algorithm. Ten of these positions are classified as new findings and are found in different datasets, suggesting them as new findings and not as wrong findings. For subtype H3N2 the situation is similar, because 18 positions are overlapping in both results. Of these positions 17 are located in patches and one position is newly found. Considering the results of the pandemic H1N1 subtype the overlap is six positions great.

Beside the overlap with already reported positions using the graph-cut algorithm the main question is, how meaningful all newly identified positions are.

First of all, all identified positions are located on the outer surface of the protein, such that every new identified position could be a possible target of antibodies. Second, taking the homologue nature of the different subtypes into account, it could be a possible approximation for the meaningfulness of newly identified positions, when they are reported to be under selection in other subtypes. Taking this into account, the results for the seasonal H1N1 subtype also contain 51 positions that are located in patches for the H3N2 subtype and 37 of them are not overlapping with H1 evaluation data. For the pandemic H1N1 subtype eight positions are located in H1 patches and 31 in H3 patches. Also for subtype H3N2 one position is reported in H1 evaluation data. For subtype H5N1 nine positions are in H1 evaluation data and 29 in H3. These mentioned positions are, in another subtype, relevant for the virus evasion of the immune system of the host and could possibly be of interest, when considering other subtypes, as well. Regarding all these findings, most of the results are reasonable and could have been under selection in the datasets they have been found for. Further this means that IPoSuS and the used graph-cut algorithm are working as intended with good results.

But there are some things that should additionally be considered. At first it has to be reconsidered how up to date all the used evaluation data is. This is due to the fact that the H3 evaluation data is from 1981 and the H1 evaluation data from 1983. In this study we used this data from 1983 for analysis of a pandemic H1N1 subtype, because it is the best evaluation data we got. Without knowing of the generality of the obtained results in 1983 there is no way to state the newly obtained results as wrong or right, because it can be that there have been changes that did not occur before. Furthermore the influenza A virus is a fast evolving virus, which is able to constantly evade the hosts immune system. Because of this, it could be possible, that the data obtained in 1981 or 1983 does not have as much impact on todays evolution, because the hosts immune system, and the virus as well, has adapted. It would be an interesting thing to see how the obtained quality measurements change when we could use up to date data.

As a second point it should be mentioned that the applied tool does not look to find single amino acids, but patches of amino acids. This means that an amino acid with strong selective pressure can compensate for amino acids which are under only weak selective pressure, probably resulting in a patch with an amino acid reported in patches and additional amino acids not reported, probably leading to a higher false positive rate.

A third thing that should be considered is the amount of evaluation data for either H1 or H3. As shown in the results section (for example see Section 4.2.1) there are 31 reported amino acid positions in H1 patches and 131 reported amino acids in H3 patches, which makes 137 unique amino acid positions in patches in total.

The fourth thing to rethink is that this study investigates the different seasons and not a whole dataset at all, such that it was likely to get new results compared to other publications using entire datasets without the differentiation of seasons.

Another issue is the used numbering, which is obtained by a MSA and the resulting mapping of all subtypes to one specific numbering. This is done to get the possibility to compare all results with one another. As shown in Figure 5.2 the HA proteins of influenza A are very similar and structural they are nearly the same, but nevertheless saying one amino acid position is the same as in another subtype, only because of a MSA, is again more of an approximation than a real fact. This also could lead to small shifts and therefore to false positives because the resulting amino acid position is slightly not reported.

Another question is, how much sense it makes to use the standard evaluation process for an approach which, first of all is likely to find surrounding amino acid positions and not only the reported one and second wants to find new amino acid positions with this help. keeping this in mind there probably should be something new to determine the quality of a tool. This also makes sense in respect of the amount of the evaluation data and will be discussed further in Section 5.3.

5.2 Approach Analysis

Now taking all results into account it should be possible to state which of the above introduced approaches is the best one or if they should be run all together to achieve the best results.

As shown in Chapter 4 all five approaches differ from one another. But there are also some similarities to be mentioned. First of all the AdaPatch approaches with differing counting statistics are as similar to one another as the OmegaRatio approaches are. This is due to the same test statistic used, but also shows that the difference in the counting statistics is not of such a high influence. Second, the OmegaValue approach, based on only one counting statistic, is most similar to the OmegaRatio approaches, which could be due to the same statistical test used. An example could be either the season 2002/10-2003/03 for the OmegaRatios **and** OmegaValue approach for the seasonal H1N1 subtype, all three resulting in a patch of the amino acid positions 172+173+240, or the season 2005/10-2006/03 for the AdaPatch approaches, resulting in the same patches of 198+196+189+188 and 279+53+276, but differing in the other patches found.

This last mentioned season also leads to another finding, such that the AdaPatch approach using NG+NRF counting statistics is more likely to lead to patches than NG+RF counting is. Therefore there are way more found patches using NG+NRFs counting than using NG+RF. It even leads to results where the other counting statistic is left without any. As an example serves the season 2010/10-2011/03 of the seasonal H1N1 subtype. Considering the OmegaValue approaches the difference in the amount of resulting patches is not that high, but with a slight advance to NG+NRF counting.

Comparing the amount of results of all three different approaches, neglecting the counting method, there are way more resulting patches for the AdaPatch approaches than there are for the OmegaRatio approaches. The OmegaValue approach only leads to very few results. This difference can be explained with the test statistic that gets used. While the AdaPatch approaches use the pure counts, and therefore has a low amount of real zeros in the table, the OmegaRatio and OmegaValue approaches use fractions of these counts, leading to zeros and not significant results in many cases. The OmegaValue approach even uses two fractions and therefore the amount of zeros is even higher than in the OmegaValue approach, leading to way fewer results.

5.2.1 OmegaValue Approach

The OmegaValue approach is, also it results in only a few findings, with rather good quality depending on the investigated subtype, because many found amino acid positions are located in patches, regarding to the evaluation data. Also IPoSuS does not tend to find single specific amino acid under selective pressure, but patches, making it likely to have also structural neighbors, determined by the graph cut algorithm, which gets introduced in Section 2.7, we only evaluate the real findings.
For the pandemic H1N1 subtype the OmegaValue approach leads to 16 different amino acid positions of which 3 are reported, which makes 18.75%. Using the OmegaValue approach for the seasonal H1N1 subtype we obtain 18 different positions of which 4 are reported, that makes 22.22%. For subtype H3N2 the OmegaValue approach leads to eight different positions of which six are reported, which makes 75.00%. Applying the OmegaValue on the datasets of subtype H5N1, it results in three different amino acid positions of which none is reported. This makes 0.00%. In average the OmegaValue approach results in 28.99% reported positions.

5.2.2 OmegaRatio Approaches

For the OmegaRatio approach the two different counting schemes are available, which additionally are compared.
The usage of the NG+RF counting scheme and the OmegaRatio approach for the pandemic H1N1 subtype leads to 18 different amino acid positions of which two are reported, which makes 11.11%. Using the NG+NRF counting scheme instead the datasets result in 24 different amino acid positions of which five are reported, which makes 20.83%.
Using the OmegaRatio approach for the seasonal H1N1 subtype, using NG+RF counting scheme, we obtain 25 different positions of which six are reported, that makes 24.00%. Using the NG+NRF counting scheme instead the datasets result in 24 different amino acid positions of which six are reported, which makes 25.00%.
For subtype H3N2 the OmegaRatio approach leads to seven different positions of which five are reported, which makes 71.43%. Using the NG+NRF counting scheme instead the datasets result in ten different amino acid positions of which six are reported, which makes 60.00%.
Applying the OmegaRatio on the datasets of subtype H5N1, it results in three different amino acid positions of which none is reported. This makes 0.00%. Using the NG+NRF counting scheme instead the datasets result in nine different amino acid positions of which none is reported, which makes 0.00%.
In average the OmegaValue approach using NG+RF counting scheme results in 26.64% reported positions, while using the NG+NRF counting scheme results in 26.46% reported positions.

5.2.3 AdaPatch Approaches

Using the AdaPatch approaches leads to all left open results and additionally to nearly all introduced ones in Chapter 4. Furthermore, the different used counting schemes are compared.

The usage of the NG+RF counting scheme and the AdaPatch approach for the pandemic H1N1 subtype leads to 37 different amino acid positions of which 14 are reported, which makes 37.84%. Using the NG+NRF counting scheme instead the datasets result in 41 different amino acid positions of which 15 are reported, which makes 36.59%.

Using the AdaPatch approach for the seasonal H1N1 subtype, using NG+RF counting scheme, we obtain 54 different positions of which 15 are reported, that makes 27.78%. Using the NG+NRF counting scheme instead the datasets result in 80 different amino acid positions of which 22 are reported, which makes 27.50%.

For subtype H3N2 the AdaPatch approach, using NG+RF counting, leads to 39 different positions of which 19 are reported, which makes 48.72%. Using the NG+NRF counting scheme instead the datasets result in 55 different amino acid positions of which 34 are reported, which makes 61.82%.

Applying the AdaPatch on the datasets of subtype H5N1, it results in 27 different amino acid positions of which five are reported. This makes 18.52%. Using the NG+NRF counting scheme instead the datasets result in 31 different amino acid positions of which four are reported, which makes 12.90%.

In average the AdaPatch approach using NG+RF counting scheme results in 33.22% reported positions, while using the NG+NRF counting scheme results in 34.70% reported positions.

Considering all this result should make it easy to decide which approach is the best one. Generally spoken the results using NG+RF counting scheme are slightly better than the one using NG+NRF. For the AdaPatch approach applied on the subtype H3N2 the NG+NRF counting scheme results in a way higher precision than the one of NG+RF. Also the OmegaRatio approach seems to have results of quite high quality, although they are a little bit worse than the mean of the OmegaValue approach. Overall the OmegaRatio approach can be considered as better, because this approaches yield more results. Nevertheless, the AdaPatch approach has the best quality regarding the results. With the mean value of the obtained results, the default usage of IPoSuS should be the AP:NG+NRF approach, because it yields the best and the most results combined over all subtypes, as shown in Table 5.1.

On the other hand there are results that are, in this study, approach dependent and do not occur in the other ones and are correct, applying the H3 evaluation data. For example does the AdaPatch approach using NG+RF counting scheme have 6 specifici-

Table 5.1: Table showing the results of the approaches for the different subtypes. The results shown are the percentage (in %) of correct identified positions under selection for the subtype and approach.

		Subtype				
		pH1N1	**sH1N1**	**H3N2**	**H5N1**	**Mean**
Approach	**AP:NG+RF**	37.84	27.78	48.72	18.52	33.22
	AP:NG+NRF	36.59	27.50	61.82	12.90	34.70
	OR:NG+RF	11.11	24.00	71.43	0.00	26.64
	OR:NG+NRF	20.83	25.00	60.00	0.00	26.46
	OmegaValue	18.75	22.22	75.00	0.00	28.99

ties compared to the counting scheme NG+NRF. Of these six positions are 3 located in patches and two of them - 62 and 63 - are only detected using this approach. The other three positions are newly found.

Beside these specificities in and the amount of the resulting positions the seasonal specificity has to be taken into account. This is a point regarding to the used dataset, because for each season there has been the same dataset for all five approaches for each subtype investigated. For example does the OmegaValue approach finds a patch for the season 2007/10-2008/03 of the subtype H3N2, whereas the other 4 approaches do not find anything in this season. On the other hand only the AdaPatch approach using NG+NRF counting scheme finds patches for subtype H3N2 for the season 2011/04-2011/09. Using the counting scheme of NG+RF and the AdaPatch approach for the same subtype (and also the same dataset as just before), it finds a unique patch for the season 2011/10-2012/03. Only the OmegaRatio approach does always find patches in the same seasons for each counting scheme and each subtype. Also they do not find results if the AdaPatch approach does not find some.

In the end the recommendation is to use the AP:NG+NRF approach if new datasets are investigated. If knowledge is existent, the OmegaValue approach could probably be of more interest, because it yields yet less but perhaps more interesting and new results. Furthermore, the usage of more approaches at once lead to a better understanding and an instant comparability of the results and also a first analysis, because in the best case all five approaches totally overlap.

Additionally it has to be reconsidered whether the data foundation is good enough or not to make a statement like this. Especially for the subtype H5N1 there is very few data regarding the evaluation as well as the sequence data. But without this subtype there would be only three left of which two are H1N1, differed by seasonal and pandemic. Excluding subtype H5N1 from this analysis, because the data basis is not satisfiable, the mean values would be a little different. The OmegaValue approach would reach mean value of 38.66%, the OR:NG+RF a value of 35.51%, approach OR:NG+NRF reaches a value of 35.28%, the mean value of approach AP:NG+RF equals 38.11% and the AP:NG+NRF approach yields 41.97% of correct identified positions in the mean. Considering this case, only investigating subtypes which have adequate evaluation data available, the recommendation gets a little bit more clear. The approach AP:NG+NRF yields the best results but also contains a high number of new identified positions. The OmegaValue approach leads to only slightly worse results, but also yields in only a few results, compared to the AP:NG+NRF approach.

5.3 Evaluation Data

First of all there is a huge overlap between the reported H1 and H3 positions, second there is a huge difference in the amount of reported amino acids, because the H1N1 subtype has not been in the human for as long as the H3N2 subtype. Only considering the H3 evaluation data now would mean that any tool or analysis of the H3N2 subtype would have to find 131 amino acids of 324, which is the length of the protein. This is a total of 40.43% of the protein reported to be in a patch. If we would now exclude the first 40 amino acids, which are located in the membrane, and therefore should not have any impact on the evasion of the immune system, a tool would have to find 46.13% of the protein, which would make it nearly to throwing a coin to decide whether an amino acid position is under selective pressure or not. Despite of that it could be that some amino acid changes have been arisen before but have been of only temporary existence meaning they disappeared again but can probably arise again. This even gets more to a random process or even worse when there would be an inclusion of additional data provided by other literature.

To this point we always evaluated everything we got against the H3 evaluation data, because of the fact, that the results fit and lead to a good quality of IPoSuS. One thing that should be reconsidered in this case is, how much can the difference between two so similar proteins be and why are there 131 amino acid positions in patches for H3 and only 32 for H1. As already mentioned before, of these 32 positions, only six differ from the H1 evaluation data. This six different amino acid positions are probably subtype specific differences. The main difference is still the amount of known positions in patches. This could be either to real differences, meaning IPoSuS is only working properly for the H3N2 subtype or because there have not been enough investigations on the amino acid positions forming the patches of H1. Another possible explanation could be that the positions for the H3 evaluation data is just not nuanced enough and has way too much false positives in it, leading to such a high amount of positions. The problem with this is that a tool does have to reach this high amount of positions to get a high quality, but without having any false positives. But as already mentioned, when leaving out the membrane parts of the protein, it is nearly by chance whether a position is in a patch or not. Furthermore this leads to a point where nearly every finding is a good or a positive one, because nearly half of the protein is a correct finding. This in the end would lower the quality of every developed algorithm, because no one can be sure what is correct and what is incorrect. Again considering the positive and good quality results as well as the high overlap in H1 and H3 evaluation data, it is more likely to be the first possibility, that there has been not enough investigation about the H1 subtypes. Especially thinking about the first real appearance of the pandemic H1N1 subtype, it is reasonable that there is not such a high amount of known data for now. The situation regarding the amount of evaluation data peaks for subtype H5N1. This is due to the rare occurrence of H5N1 subtype viruses in humans and the therefore few amount of investigations about the adaptations and the selectional pressure. This in the very end leads again to the possibility to use the H3 evaluation data for every other subtype, at least as an approach, leaving the subtype specific positions of the newly investigated subtype open. The overlap between newly identified positions of the subtype H5N1 and the pandemic H1N1 and the known patch data for H1 and H3 have been shown before and highly suggest this method.

5.4 New Findings

Beside the in Section 5.1 presented concerns regarding the results, there are some interesting new findings provided by this study.

First of all, there are two amino acid overlap in all four considered subtypes. Because this study only used sequences of human hosts, this two amino acid positions can be seen as relevant for the human adaption or the continuance in the human host.

A second very interesting thing that appeared because of the seasonal view is that it seems that there are some amino acid positions which are mainly altered and therefore have a high number of appearances in patches over the seasons, accompanied by structural neighbors which seem to be exchangeable. Even so it is not surprising, that there are surroundings found by the algorithm, but the amount of counts is interesting. Taking them in to account this finding gets clearer. Figure 5.1 shows this findings as a graphical plot.

Mathematically this even makes sense regarding to the main effort the virus has - evading the immune system of the host. Now thinking of an amino acid position, which is mainly altered to evade the immune system of the host. The virus only has 20 different possibilities to change this particular position and in the worst case it would only last for 20 seasons before it extinct because the immune system would "know" all different possibilities. But if the virus randomly would change the amino acid positions in an area of round about ± 3 positions the virus would have $20^7 = 1280000000$ possibilities to alter this area. Having multiple of such areas highly increases the survivability of the virus, regarding to the evasion of the hosts immune system, because the immune system can not have such an high amount of different antibodies against every single possibility of such a highly alterable area.

Another new finding is, that there is not such of a difference (see Section 5.2) in the highly mutable sites as expected and that the evaluation data of H1 and H3 could and

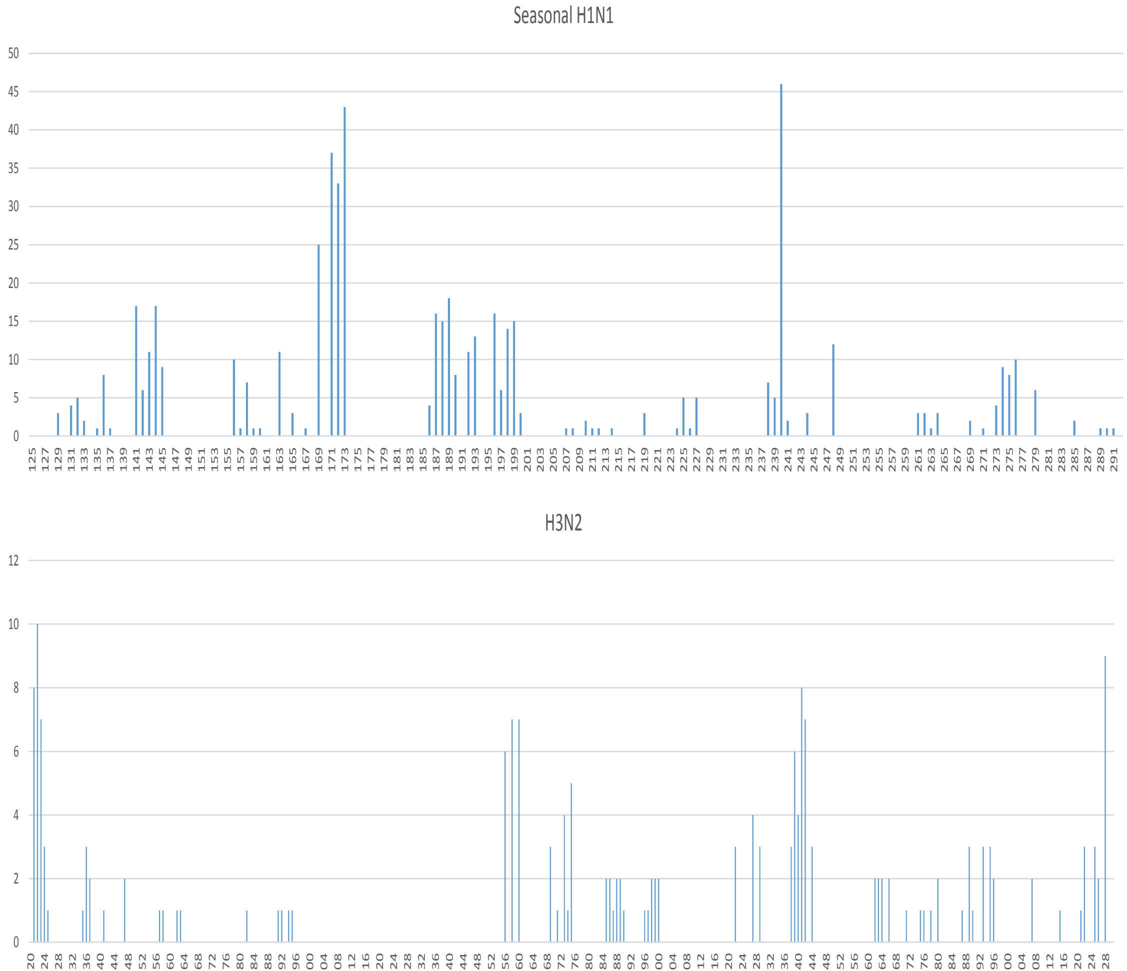

Figure 5.1: Figure exemplarily showing the results for the seasonal H1N1 and the H3N2 subtype. On the x-axis are the amino acid positions while on the y-axis is the count, how often this position was found in a patch, showing that there are clearly favored amino acid positions which get surrounded by less favored ones that are exchangeable.

should be combined. Furthermore the results show that the combination of H1 and H3 subtype evaluation data can be applied to evaluate the results of subtype H5. This finding should make it also possible to analyze other subtypes.

Reconsidering the results and the fact that many of the resulting findings are one position beside a known patch-position shows that it is of major importance, if a MSA for position comparison is used and if so, which sequences it contains. Using another MSA or a MSA with different sequences could lead to another position shifting, matching them to the subtype H3N2, and make therefore some mismatches to matching ones. On the other hand this also could lead some findings to shift and be no findings at all. Another point is the already mentioned H3 evaluation data, which seems to cover way too man positions and makes it therefore very hard to make precise statements. But for this particular problem the amount is only a small piece of the puzzle. The other main part is, that most of the evaluation data is either consecutive, like 140 to 145, which makes shifted results likely to be right, no matter the shifting, or they are like 100, 102, 104 etc.

which makes a shifting either right or wrong. For comparability we therefore created our own MSA.

Another new finding is the overlap between the seasonal and the pandemic H1N1 subtype, after withdrawing the positions this overlap has with the findings of the other two subtypes. After doing so only the positions 51, 96, 136, 155, 157, 164, 168, 170, 191, 195, 211, 224, 237, 243, 260, 273 and 274 are left over. All of these positions can be considered as subtype specific for the subtype H1N1. Evaluating them against the evaluation data two are located in a patch using H1 evaluation data and additional one position is reported in another publication for the seasonal H1N1 and two positions are reported for the pandemic H1N1 subtype, which makes it in combination 12 newly found amino acid positions - 51, 96, 155, 157, 164, 168, 170, 211, 237 and 260. Taking also the amount of occurrences in patches into account the positions with high occurrences should be analyzed further. The ones that should be considered further are occurring, combining both subtypes, 15, 17, 24, 6 and 13 times for position 155, 191, 195, 243 and 274 respectively. The positions 168 and 170 on the other hand seem to be specific for the seasonal H1N1 because they are found 25 and 37 times in a patch, whereas these positions only been found to be in a patch 2 times each for the pandemic H1N1 subtype.

But there are not only new findings in pooling of all obtained results, but there are also some matchless results for each subtype, shown in Table 4.3.12. Evaluating these specificities to the H3 evaluation data, could yield some subtype specific adaptations. For the pandemic H1N1, there are eleven positions not reported or close to reported positions, and are therefore new findings - 40, 42, 99, 113, 232, 251, 267, 282, 283, 286, 288. For the seasonal H1N1 subtype there are 16 new findings - 65 73, 224, 269, 285, 390, 400, 401, 404, 439, 443, 445, 474, 475, 476, 482. For the subtype H3N2 there are 15 new findings - 21 22, 23, 24, 25, 35, 36, 37, 266, 287, 321, 322, 325, 326, 328. For the subtype H5N1 there is only one finding, position 255. Thinking of these all as subtype specific positions there are huge differences in the host adaption or the alterations to stay in the host between the different subtypes investigated in this study. Furthermore it is also of interest, that for the subtype H3N2 there are findings in the membrane part of the protein HA, which is usually thought of not being under selective pressure, because it not located on the outer surface of the protein and therefore has no direct contact with the immune system of the host. Because of the overall results, which seem to be very good and of a quite good quality, this should probably be checked in wet-lab experiments, if these positions located in the membrane do have an impact on the virus, either in virulence or replication or even something else. Second, regarding to the results of the seasonal H1N1 subtype, there should be some further analysis over the second chain of the HA protein, because it seems that there quite some interesting parts under selective pressure. Also there should be some analysis about the positions 282, 283, 285 and 287, because they could form a little patch closing the distance between position 280 and 294 of patch C in the H3 evaluation data. Additionally position 282 was already found by some other case studies by Li et. al in 2011 [42].

5.4.1 Positions and Counts

Only until the last section we only looked at the fact that a position has been found by the algorithm, neglecting the amount of findings, which already has been noted down in Chapter 4. We now want to consider only amino acid positions that have been significantly found more often over all approaches for one subtype. Therefore we calculate the mean value for each subtype and the standard deviation and only choose positions with counts higher than the sum of both because we are only interested in these.

Pandemic H1N1

For the pandemic H1N1 subtype we have 176 counts in total, with a mean value of 2.75 and a standard deviation of 2.34, which makes it 5.09 as a border. So we only consider positions with a count value of 5 or higher. For this subtype there are only ten positions fulfilling this criteria: 133, 159, 160, 163, 199, 200, 202, 203, 273 and 274 after their own numbering or 126, 151, 152, 155, 191, 192, 194, 195, 264 and 265 after H3 numbering. Of these ten positions, considering H1 numbering, all positions are new and are also not reported elsewhere. All of these position have been found significantly often and have not been reported yet to be under selective pressure, which would make it reasonable to further analyze them.

Seasonal H1N1

For the seasonal H1N1 subtype we have 607 counts in total, with a mean value of 6.07 and a standard deviation of 8.60, which makes it 14.67 as a border. So we only consider positions with a count value of 15 or higher. For this subtype there are only 12 positions fulfilling this criteria: 141, 144, 169, 171, 172, 173, 187, 188, 189, 196, 199 and 240 after their own numbering or 140, 143, 168, 170, 171, 172, 186, 187, 188, 195, 198 and 239 after H3 numbering. Of these 12 positions, considering H3 numbering, positions - 168, 170, 171, 172, 186, 187, 188, 195 and 239 - are new, three positions are located in a patch considering H1 evaluation data- 140, 143 and 198 - and no position has been reported elsewhere. Because three of these position are fairly known to be under selective pressure, the most interesting findings are the positions 168, 170, 171, 172, 186, 187, 188, 195 and 239 (H3 numbering). This positions have been found significantly often and have not been reported yet to be under selective pressure, which would make it reasonable to further analyze it. But furthermore there should be analysis of position 239 (H3 numbering) because it has been found 46 times and has the highest value of all and is not stated as a position in a patch.

H3N2

For the subtype H3N2 we have 188 counts in total, with a mean value of 2.72 and a standard deviation of 2.23, which makes it 4.95 as a border. So we only consider positions with a count value of 5 or higher. For this subtype there are only 11 positions fulfilling this criteria: 21, 22, 23, 156, 158, 160, 175, 239, 241, 242 and 328 after their own numbering. Of these 11 positions, considering H3 numbering, five positions - 21, 22, 23, 239, 241 and 328 - are new, five positions are located in a patch of the H3 evaluation data - 156, 158, 160, 175 and 242 - and no position has been reported elsewhere. Because five of these position are fairly known to be under selective pressure, the most interesting findings are positions 21, 22, 23, 239, 241 and 328 (H3 numbering). These positions have been found significantly often and have not been reported yet to be under selective pressure, which would make it reasonable to further analyze it. Again, there should be further analysis of position 239 (H3 numbering). Interesting, considering the new findings for this subtype, are that four of the five positions are located in the membrane part and only positions 239 and 241 are not. This clearly shows that there have to be more analysis about the membrane part of the HA protein because it seems to have, at least for this subtype, more impact or at least is under more selective pressure than thought. The reasons for that should be investigated.

H5N1

For the pandemic H1N1 subtype we have 176 counts in total, with a mean value of 2.53 and a standard deviation of 1.50, which makes it 4.03 as a border. So we only consider positions with a count value of 4 or higher. For this subtype there are only nine positions fulfilling this criteria: 142, 144, 158, 159, 185, 188, 189 and 278 after their own numbering or 142, 144, 158, 159, 185, 188, 189 and 277 after H3 numbering. Of these nine positions, considering H3 numbering, only position 158 is reported to be under selection. All other positions have not been reported to be so yet, but have been found significantly often, which would make it reasonable to further analyze them.

5.5 Applicability for other Influenza A Proteins

As described in the work flow, see Figure 3.1, IPoSuS already has the possibility to be used for other proteins. To do the analysis for all different proteins of the influenza A virus it has to be mind, that there is not a solved protein for every subtype and therefore the analysis cant be done properly. There is the possibility to use the homologue solved proteins of other subtypes. Figure 5.2 shows an example of the differences which the four subtypes H1N1, H3N2, H5N1 and H7N7 have for the HA protein.

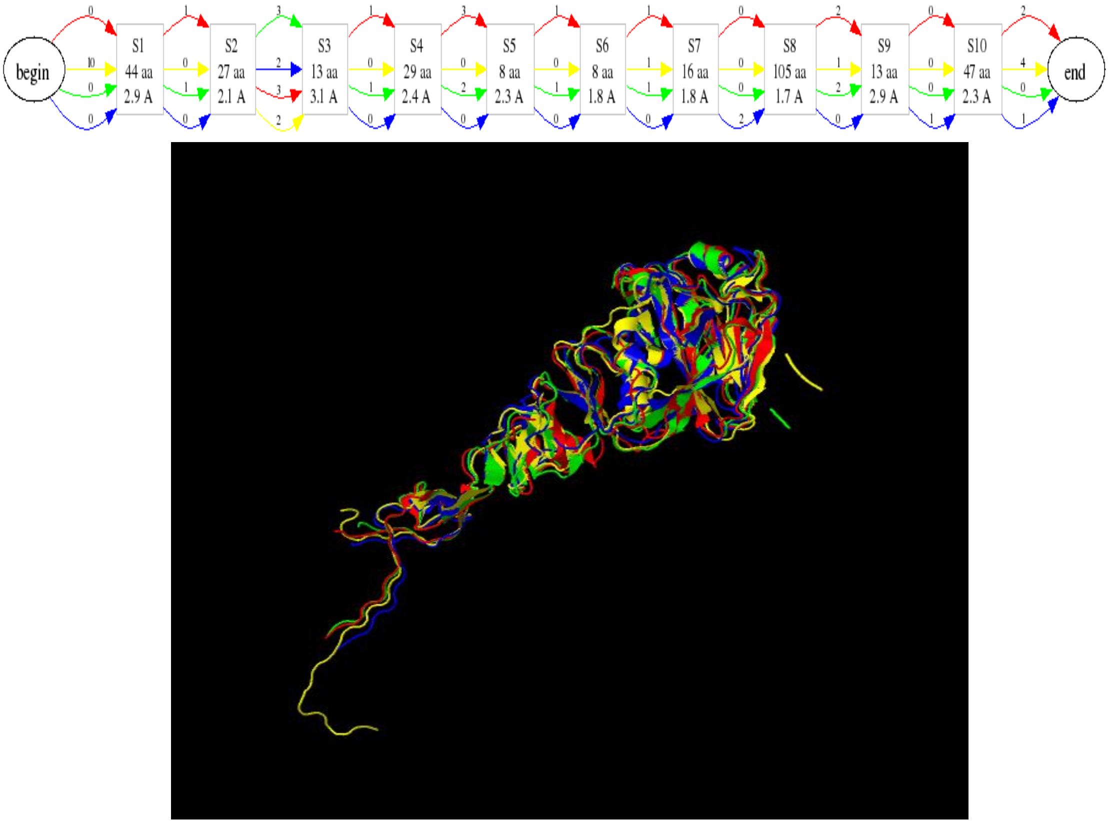

Figure 5.2: Figure showing the tertiary structure similarities and differences of the four subtypes H1N1, H3N2, H5N1 and H7N7 with the corresponding pdb id's 2WRG, 2HMG 2IBX and 4DJ6 and the corresponding colours red, yellow, green and blue, on the bottom of the figure. The top of the figure shows the run of the sequences of the different structures. Areas with high similarities are shown in rectangles in addition with the length of this segment and the RMSD value for it. Arrows combining these segments have a higher difference or are insertions/deletions in other structures. Colors are the same as in the bottom part of the figure.

As seen in the figure most of segments are very similar with a root-mean-square-deviation (RMSD) value, a measure of the average distance between atoms, ranging from 1.7 to a maximum of 3.1 and are only divided by the several inputs or deletions some of the subtypes feature. This indicates, that the used structures are very similar to each other, sometime even more similar than the resolutions of the proteins are. With this knowledge the analysis for every other protein of influenza A can be done but are not totally correct, but can be seen as a first approach.

This overlap in the 3D structure also supports the theory that the evaluation data should be put together and at least the protein HA should be considered as a whole, as long as analysis only investigate one specific host. To get knowledge about host dependencies there have to be more analysis.

5.6 Used Data

Regarding the quality of the data there is only one statement, that the quality is good, because of the probity of the provider of the data and their reputation on this field of work.

The second point is the amount of data which gets even more critical when using the OmegaValue and the OmegaRatio approaches, because using them assumes at least 30 sequences to have generality for the Z-value calculation. Having less than 30 sequences can also result in findings but there is no certainty about them and the quality of the obtained results. The amount of data is also the reason why there are no results represented in this work for the subtypes H7N7 and H7N9, because there was not enough different data to obtain results. Also the amount of data is the reason why there sometimes are no results at all for some seasons. The reason for that is not only the pure amount of data but the amount of different data with different mutations. This arises from the fact that these subtypes usually occur in birds and therefore do not have a possibility to infect humans and the whole data is from one single outbreak in one region in one season. Therefore all of the gathered sequences are nearly similar and did not had enough time to evolve further. Furthermore the needed adaption to infect humans would not mutate any further because the adaption already took place. For this lack of information the only possible thing to do would be to consider both, animal hosts and human hosts together, compared to only animal host, to see the differences which occurred between these two approaches.

5.7 Positions Not Under Selective Pressure

In the previous sections the discussion was all about the findings of the developed algorithm. Now, on the other hand there is need for a discussion of the positions that have not been identified to be under selection. There are at least two imaginable possibilities for these positions. The first possibility is the most easy one, these positions just do not have any impact on the evasion of the hosts immune system. This could be either due to the structural position they are at, because if they do not lie exposed they probably do not have to alter or it is because they are necessary for the structure to be formed correctly and may not alter or otherwise the protein will misfold. The second possibility is the more speculative one. Assuming that there are special positions in the protein, that enables the virus to infect other species if they are in the right combination or they did alter in a specific way, these positions should not alter anymore, because otherwise the virus could loose its ability to infect the host. With this assumption positions that are

not altered could be the primary reason for host switches or adaption in a new host. Furthermore these positions would only alter once and could therefore not been detected by this, or even any other, algorithm, that is looking for amino acids under selection, because they would not be remarkable in terms of p-values or pure counts of alterations.
Beside these speculations these positions also provide another opportunity. The reason why known vaccines do not work for much longer than a couple of seasons is because of the previous mentioned alterations in the amino acids exposed to the outer surface.
Knowing the positions that do not alter over such a high amount of different seasons in all the different subtypes could lead to a new vaccine that can be applied no matter the season or the influenza A strain. Additionally these positions should have to be laid on the outer surface of the protein, such that it can be targeted by for example a synthetic molecule. A possible reason why the human immune system was not able to develop such a molecule by its own could be the rarity of these positions or the structural location of all those, such that they are just to far apart. Thinking about the H3N2 subtype HA protein of 330 length, there are 156 positions that are not found by the introduced algorithm, over all results and subtypes. Now thinking about the membrane parts that are thought of not being altered, due to location, there are 99 positions left. Of these 99 residues are 32 located to the outer surface and are therefore exposed. These positions are: 50, 54, 65, 73, 75, 77, 80, 81, 83, 100, 101, 110, 113, 114, 116, 117, 119, 120, 122, 123, 124, 137, 165, 167, 208, 212, 246, 248, 285, 291, 292 and 299. Additionally it has to be factored in, that some of them are reported to be into patches and have just not been found. Taking all this into account, the ones that are reasonable for further analysis are: 65, 73, 77, 100, 101, 110, 113, 114, 116, 119, 120, 123, 285, 291 and 292, as shown in Figure 5.3
Many of these positions are directly located to a position in a reported patch, which could be a reason why it cant be targeted as well as wanted by the host immune system, because the surroundings are altering very heavily, which makes the binding to one specific position very difficult. All of these positions could have impact in the adaption or the continuance in the human as host. To further analyze this, IPoSuS needs to be used for other datasets, not including human sequences to see, whether it is human specific or HA specific. Both cases would be of appropriate interest.

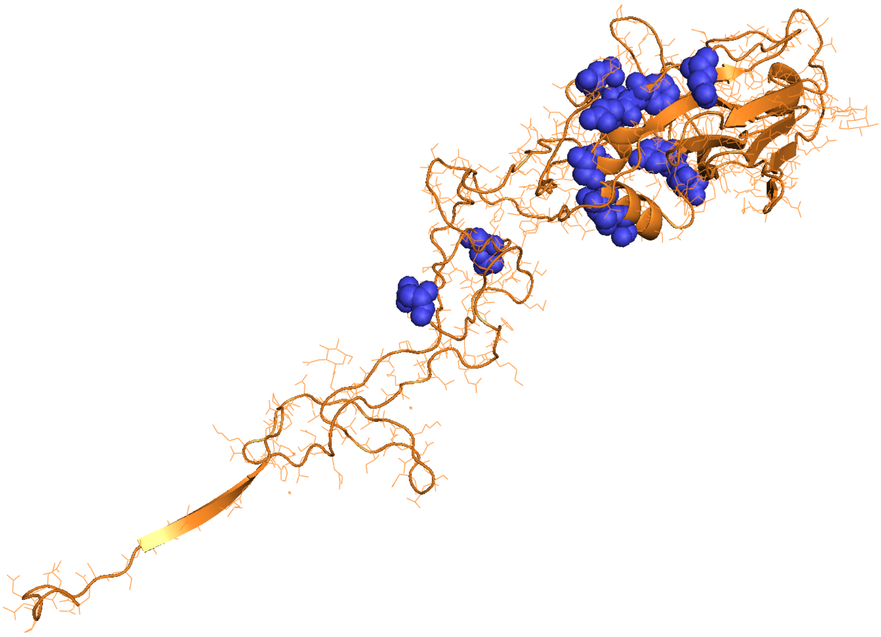

Figure 5.3: Picture showing the location of the amino acid positions that have not been found in any approach and subtype and are additionally exposed on the outer surface (H3N2 subtype structure and numbering).

6 Outlook

In this work IPoSuS was used for analysis of the protein HA of the influenza A virus, using subtype specific structures. Furthermore, it should be used to analyze this protein using structures of other subtypes, resulting in a use of a homologue protein structures of the HA protein. In the end there could be a comparison of the results to find differences and similarities in the subtypes. With this similarities and differences over all subtypes and structures there can be a statement about the quality and reliability of the usage of subtype unspecific structures. This could be used to investigate all of the other proteins of the influenza A virus. Doing this with the seasonal differentiation as already done in this work for the HA protein could provide a new insight into the dependencies of adaptive mutations and the coherence of mutations of different proteins of an organism. Further on there can be analysis about different hosts of different subtypes including or excluding the information about the geographic location, because for now the only analysis was for human infecting subtypes of the HA protein.

After using IPoSuS for a whole protein analysis of influenza A it can also be used to investigate proteins of other fast evolving viruses like the Ebola or Marburg viruses. But not only virus proteins can be investigated, but all proteins for which a tertiary structure is known, or modeled, and for which enough coding sequences are given.

All this could lead to new insights into the evolution of proteins in viruses, or other organisms, and also enlighten the field of depended mutations in different proteins. Furthermore retrospective analysis could be done on the obtained results regarding the different seasons to get even more insight into the development scheme and pattern of the proteins.

As a next point IPoSuS is very easy to expend concerning the test- and counting statistics. This means that whenever something new gets developed or found out, IPoSuS can be extended as long as these new statistics can be converted or represented as p-values. But it does not have to be new findings but can also be statistics, on which test should be done. Because of the already implemented five different approaches there is an instant feedback and comparability of the results.

Another good thing to do would be the introducing of an automated evaluation to directly have a feedback about the data, for example own input data. Furthermore it would be interesting to think about a new possibility to determine the quality of IPoSuS or a specific implemented test statistic because as already mentioned this tool is likely to find non reported amino acid positions but surroundings. Using an exponential function with different bases could perhaps be a good foundation to think about this.

As already mentioned in the discussion there is a problem with the underlying MSA, because the whole results of any analysis can shift from good to bad or from bad to good, dependent on the used MSA. All in all there should probably be a greater study about different used sequences and MSA's and the obtained results for this tool, such that it perhaps is possible to denoise the results.

Another thing, a combination of two mentioned, could be the automated evaluation and different underlying MSA's for obtaining this evaluation. In combination it should be possible to denoise the results or have the most likely results as best hits.
Also there should be further analysis taking different hosts into account, just like suggested in Section 5.6 for subtypes H7N9 and H7N7. This procedure can also been done for every other subtype as well. The differences could give new insights especially into positions which do have to alter for an initial host switch and do not alter once it happened.

7 Summary

In this work we introduced a tool, based on an already existing graph-cut approach, to detect patches of sites under selection with an automated framework. This automated framework does include the data download for specific parameters set by the user, the automated creation of a MSA, the proceedings with it and the further dn/ds counting and graph-cut for patch identification.

Furthermore we introduced a new statistical test which enables us to not only use the pure counts of synonymous and non synonymous mutations but the ratio and further calculations of this value, to obtain OmegaRatios and OmegaValues.

Considering the results, the used algorithm has found many new positions under selection which have not been reported before. Furthermore the results suggest to use a combination of the evaluation data for the subtypes H1 and H3 and also a review of the meaningfulness of this data. Beside of this, the work showed that the newly introduced OmegaRatio and OmegaValue approach are usable and could also be used as default, although the AdaPatch approach yields the best results, but plenty of them. Additionally there should be a more detailed analysis about the positions that have not been identified by IPoSuS in any subtype or approach, to specify their role for the protein and organism, as shown in Section 5.7.

Another special thing was the consideration of many different seasons and the obtained possibility to count the appearances of findings. The analyzes of these positions showed only a couple of amino acid positions that should be of further interest.

All in all we introduced a tool which works as intended, with a quite good quality, when taking into account that not reported findings are intentional. Furthermore, IPoSuS is able to detect patches under selection for every protein, as long as there is enough data and the needed 3D-structure. This includes not only proteins of influenza A viruses, but all imaginable proteins of viruses and other organisms. This makes this tool very powerful with a wide possible application range.

Bibliography

[1] Akarsu, H., Burmeister, W. P., Petosa, C., Petit, I., MÃČÂijller, C. W., Ruigrok, R. W., and Baudin, F. Crystal structure of the M1 protein-binding domain of the influenza A virus nuclear export protein (NEP/NS2). *The EMBO journal 22*, 18 (2003), 4646–4655.

[2] Arunachalam, R. Detection of site-specific positive darwinian selection on pandemic influenza a/h1n1 virus genome: integrative approaches. *Genetica 141*, 4-6 (Jun 2013), 143–155.

[3] Arzt, S., Petit, I., Burmeister, W. P., Ruigrok, R. W. H., and Baudin, F. Structure of a knockout mutant of influenza virus m1 protein that has altered activities in membrane binding, oligomerisation and binding to nep (ns2). *Virus Res 99*, 2 (Feb 2004), 115–119.

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

[5] Bokhari, S. H., and Janies, D. Reassortment Networks for Investigating the Evolution of Segmented Viruses. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 7*, 2 (Apr. 2010), 288–298.

[6] Bornholdt, Z. A., and Prasad, B. V. V. X-ray structure of ns1 from a highly pathogenic h5n1 influenza virus. *Nature 456*, 7224 (Dec 2008), 985–988.

[7] Bouvier, N. M., and Palese, P. The biology of influenza viruses. *Vaccine 26* (2008), D49–D53.

[8] Boykov, Y., Veksler, O., and Zabih, R. Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 23*, 11 (2001), 1222–1239.

[9] Bush, R. M., Fitch, W. M., Bender, C. A., and Cox, N. J. Positive selection on the h3 hemagglutinin gene of human influenza virus a. *Molecular biology and evolution 16*, 11 (1999), 1457–1465.

[10] Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics 25*, 15 (Aug. 2009), 1972–1973.

[11] Carmichael, P., Copier, J., So, A., and Lechler, R. Allele-specific variation in the degeneracy of major histocompatibility complex (mhc) restriction. *Hum Immunol 54*, 1 (Apr 1997), 21–29.

[12] Carrat, F., Vergu, E., Ferguson, N. M., Lemaitre, M., Cauchemez, S., Leach, S., and Valleron, A.-J. Time Lines of Infection and Disease in Human Influenza: A Review of Volunteer Challenge Studies. *American Journal of Epidemiology 167*, 7 (Mar. 2008), 775–785.

[13] Caton, A., Brownlee, G., Yewdell, J., and Gerhard, W. The antigenic structure of the influenza virus a/pr/8/34 hemagglutinin (h1 subtype). *Cell* (1982), 31:417–427.

[14] CDC. Key facts about influenza (flu) & flu vaccine, September 9 2014.

[15] Collier, L., Balows, A., and Sussman, M. *Topley & Wilson's Microbiology and Microbial Infections, Ninth Edition: Volume 1 - Virology*, vol. 1. Arnold, a member of the Hodder Headline Group London, Sydney, Auckland, 1998.

[16] Coloma, R., Valpuesta, J. M., Arranz, R., Carrascosa, J. L., Ortín, J., and Martín-Benito, J. The structure of a biologically active influenza virus ribonucleoprotein complex. *PLoS Pathog 5*, 6 (Jun 2009), e1000491.

[17] Desper, R. Theoretical Foundation of the Balanced Minimum Evolution Method of Phylogenetic Inference and Its Relationship to Weighted Least-Squares Tree Fitting. *Molecular Biology and Evolution 21*, 3 (Dec. 2003), 587–598.

[18] Ding, X., Jiang, L., Ke, C., Yang, Z., Lei, C., Cao, K., Xu, J., Xu, L., Yang, X., Zhang, Y., Huang, P., Huang, W., Zhu, X., He, Z., Liu, L., Li, J., Yuan, J., Wu, J., Tang, X., and Li, M. Amino acid sequence analysis and identification of mutations under positive selection in hemagglutinin of 2009 influenza a (h1n1) isolates. *Virus Genes 41*, 3 (Dec 2010), 329–340.

[19] Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research 32*, 5 (Mar. 2004), 1792–1797.

[20] Farris, J. S. The logical basis of phylogenetic analysis. *Advances in Cladistics 2* (1983), 7–36.

[21] Felsenstein, J. *Inferring Phylogenies*. Sinauer Associates: Sunderland, Massachusetts., 2004.

[22] Fisher, R. A. On the interpretation of $\chi 2$ from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society* (1922), 87–94.

[23] FITCH, W. M. Toward defining the course of evolution: Minimum change for a specific tree topology. *SystZool 20*, 4 (1971), 406–416.

[24] GIANFRANI, C., OSEROFF, C., SIDNEY, J., CHESNUT, R. W., AND SETTE, A. Human memory ctl response specific for influenza a virus is broad and multispecific. *Hum Immunol 61*, 5 (May 2000), 438–452.

[25] HASTINGS, W. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57 (1970), 97–109.

[26] HERFST, S., SCHRAUWEN, E. J. A., LINSTER, M., CHUTINIMITKUL, S., DE WIT, E., MUNSTER, V. J., SORRELL, E. M., BESTEBROER, T. M., BURKE, D. F., SMITH, D. J., RIMMELZWAAN, G. F., OSTERHAUS, A. D. M. E., AND FOUCHIER, R. A. M. Airborne transmission of influenza a/h5n1 virus between ferrets. *Science 336*, 6088 (Jun 2012), 1534–1541.

[27] HOFER, U. Viral evolution: Past, present and future of influenza viruses. *Nature Reviews Microbiology 12*, 4 (Mar. 2014), 237–237.

[28] ITO, T., COUCEIRO, J. N. S., KELM, S., BAUM, L. G., KRAUSS, S., CASTRUCCI, M. R., DONATELLI, I., KIDA, H., PAULSON, J. C., WEBSTER, R. G., AND OTHERS. Molecular basis for the generation in pigs of influenza A viruses with pandemic potential. *Journal of virology 72*, 9 (1998), 7367–7373.

[29] JONES, C. M., LAKE, R. A., LAMB, J. R., AND FAITH, A. Degeneracy of t cell receptor recognition of an influenza virus hemagglutinin epitope restricted by hla-dq and -dr class ii molecules. *Eur J Immunol 24*, 5 (May 1994), 1137–1142.

[30] KERRY, P. S., AYLLON, J., TAYLOR, M. A., HASS, C., LEWIS, A., GARCÍA-SASTRE, A., RANDALL, R. E., HALE, B. G., AND RUSSELL, R. J. A transient homotypic interaction model for the influenza a virus ns1 protein effector domain. *PLoS One 6*, 3 (2011), e17946.

[31] KIMURA, M. *Population Genetics, Molecular Evolution, and the Neutral Theory: Selected Papers.* University of Chicago Press, 1994.

[32] KLENK, H.-D., NAGAI, Y., AND TASHIRO, M. Influenza viruses, cell enzymes, and pathogenicity. *American Journal of Respiratory and Critical Care Medicine* (1995), 152:4 pt 2,S16–S19.

[33] KLENK, H. D., ROTT, R., ORLICH, M., AND BLODORN, J. Activation of influenza a viruses by trypsin treatment. *Virology* (1975), 68:426–439.

[34] KOEL, B. F., BURKE, D. F., BESTEBROER, T. M., VAN DER VLIET, S., ZONDAG, G. C. M., VERVAET, G., SKEPNER, E., LEWIS, N. S., SPRONKEN, M. I. J., RUSSELL, C. A., EROPKIN, M. Y., HURT, A. C., BARR, I. G., DE JONG, J. C., RIMMELZWAAN, G. F., OSTERHAUS, A. D. M. E., FOUCHIER, R. A. M., AND SMITH, D. J. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science 342*, 6161 (Nov 2013), 976–979.

[35] KONGCHANAGUL, A., SUPTAWIWAT, O., KANRAI, P., UIPRASERTKUL, M., PUTHAVATHANA, P., AND AUEWARAKUL, P. Positive selection at the receptor-binding site of haemagglutinin h5 in viral sequences derived from human tissues. *J Gen Virol 89*, Pt 8 (Aug 2008), 1805–1810.

[36] KOSAKOVSKY POND, S. L. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution 22*, 5 (Feb. 2005), 1208–1222.

[37] KOWALINSKI, E., ZUBIETA, C., WOLKERSTORFER, A., SZOLAR, O. H. J., RUIGROK, R. W. H., AND CUSACK, S. Structural analysis of specific metal chelating inhibitor binding to the endonuclease domain of influenza ph1n1 (2009) polymerase. *PLoS Pathog 8*, 8 (2012), e1002831.

[38] KRAMER/KAMPS. *Grundlagen der Wahrscheinlichkeitsrechnung und Statistik: Ein Skript für Studierende der Informatik, der Ingenieur- und Wirtschaftswissenschaften*, vol. 2. Springer, 2008.

[39] LAMB, R. A., AND CHOPPIN, P. W. The gene structure and replication of influenza virus. *Annual review of biochemistry 52*, 1 (1983), 467–506.

[40] LARKIN, M., BLACKSHIELDS, G., BROWN, N., CHENNA, R., MCGETTIGAN, P., MCWILLIAM, H., VALENTIN, F., WALLACE, I., WILM, A., LOPEZ, R., THOMPSON, J., GIBSON, T., AND HIGGINS, D. Clustal w and clustal x version 2.0. *Bioinformatics 23*, 21 (Nov. 2007), 2947–2948.

[41] LEE, A. J., DAS, S. R., WANG, W., FITZGERALD, T., PICKETT, B. E., AEVERMANN, B. D., TOPHAM, D. J., FALSEY, A. R., AND SCHEUERMANN, R. H. Diversifying selection analysis predicts antigenic evolution of 2009 pandemic h1n1 influenza a virus in humans. *J Virol 89*, 10 (May 2015), 5427–5440.

[42] LI, W., SHI, W., QIAO, H., HO, S. Y. W., LUO, A., ZHANG, Y., AND ZHU, C. Positive selection on hemagglutinin and neuraminidase genes of h1n1 influenza viruses. *Virol J 8* (2011), 183.

[43] LIU, J., STEVENS, D. J., HAIRE, L. F., WALKER, P. A., COOMBS, P. J., RUSSELL, R. J., GAMBLIN, S. J., AND SKEHEL, J. J. Structures of receptor complexes formed

by hemagglutinins from the Asian Influenza pandemic of 1957. *Proceedings of the National Academy of Sciences 106*, 40 (2009), 17175–17180.

[44] LIU, N., WANG, G., LEE, K. C., GUAN, Y., CHEN, H., AND CAI, Z. Mutations in influenza virus replication and transcription: detection of amino acid substitutions in hemagglutinin of an avian influenza virus (H1n1). *The FASEB Journal 23*, 10 (Oct. 2009), 3377–3382.

[45] LOEWE, L. Negative selection. *nature Education* (2008), 1(1):59.

[46] LU, G., ROWLEY, T., GARTEN, R., AND DONIS, R. O. FluGenome: a web tool for genotyping influenza A virus. *Nucleic Acids Research 35*, Web Server (May 2007), W275–W279.

[47] MCHARDY, A. C., AND ADAMS, B. The role of genomics in tracking the evolution of influenza a virus. *PLoS pathogens 5*, 10 (2009), e1000566.

[48] MEDINA, R. A., AND GARCÍA-SASTRE, A. Influenza a viruses: new research developments. *Nature Reviews Microbiology 9*, 8 (July 2011), 590–603.

[49] MEMORANDUM, W. A revision of the system of nomenclature for influenza viruses: a WHO memorandum. *Bull World Health Organ 58* (1980), 585–591.

[50] MOEN, S. O., ABENDROTH, J., FAIRMAN, J. W., BAYDO, R. O., BULLEN, J., KIRKWOOD, J. L., BARNES, S. R., RAYMOND, A. C., BEGLEY, D. W., HENKEL, G., MCCORMACK, K., TAM, V. C., PHAN, I., STAKER, B. L., STACY, R., MYLER, P. J., LORIMER, D., AND EDWARDS, T. E. Structural analysis of H1n1 and H7n9 influenza A virus PA in the absence of PB1. *Scientific Reports 4* (Aug. 2014).

[51] MOLLES, M. C. *Ecology Concepts and Applications*. 2010.

[52] MOUNT, D. *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY., 2004.

[53] MURAMOTO, Y., NODA, T., KAWAKAMI, E., AKKINA, R., AND KAWAOKA, Y. Identification of novel influenza a virus proteins translated from PA mRNA. *Journal of Virology 87*, 5 (Mar. 2013), 2455–2462.

[54] NEEDLEMAN, S. B., AND WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology 48*, 3 (March 1970), 443–453.

[55] NEI, M., AND GOJOBORI, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution 3*, 5 (1986), 418–426.

[56] NELSON, M. I., AND HOLMES, E. C. The evolution of epidemic influenza. *Nature Reviews Genetics 8*, 3 (Mar. 2007), 196–205.

[57] OBAYASHI, E., YOSHIDA, H., KAWAI, F., SHIBAYAMA, N., KAWAGUCHI, A., NAGATA, K., TAME, J. R. H., AND PARK, S.-Y. The structural basis for an essential subunit interaction in influenza virus rna polymerase. *Nature 454*, 7208 (Aug 2008), 1127–1131.

[58] PENG, J., YANG, H., JIANG, H., LIN, Y.-X., LU, C. D., XU, Y.-W., AND ZENG, J. The Origin of Novel Avian Influenza A (H7n9) and Mutation Dynamics for Its Human-To-Human Transmissible Capacity. *PLoS ONE 9*, 3 (Mar. 2014), e93094.

[59] PETERSEN, B., PETERSEN, T. N., ANDERSEN, P., NIELSEN, M., AND LUNDEGAARD, C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol 9* (2009), 51.

[60] PING, J., KELETA, L., FORBES, N. E., DANKAR, S., STECHO, W., TYLER, S., ZHOU, Y., BABIUK, L., WEINGARTL, H., HALPIN, R. A., BOYNE, A., BERA, J., HOSTETLER, J., FEDOROVA, N. B., PROUDFOOT, K., KATZEL, D. A., STOCKWELL, T. B., GHEDIN, E., SPIRO, D. J., AND BROWN, E. G. Genomic and protein structural maps of adaptive evolution of human influenza a virus to increased virulence in the mouse. *PLoS One 6*, 6 (2011), e21740.

[61] PRICE, M. N., DEHAL, P. S., AND ARKIN, A. P. FastTree approximately maximum-likelihood trees for large alignments. *PloS one 5*, 3 (2010), e9490.

[62] SAFO, M. K., MUSAYEV, F. N., MOSIER, P. D., ZHOU, Q., XIE, H., AND DESAI, U. R. Crystal Structures of Influenza A Virus Matrix Protein M1: Variations on a Theme. *PLoS ONE 9*, 10 (Oct. 2014), e109510.

[63] SCHOLTISSEK, C. Molecular evolution of influenza viruses. *Virus genes 11*, 2-3 (1995), 209–215.

[64] SCHRÖDINGER, LLC. The PyMOL molecular graphics system, version 1.3r1. August 2010.

[65] SHARMA, M., YI, M., DONG, H., QIN, H., PETERSON, E., BUSATH, D. D., ZHOU, H.-X., AND CROSS, T. A. Insight into the mechanism of the influenza A proton channel from a structure in a lipid bilayer. *Science 330*, 6003 (2010), 509–512.

[66] Shtyrya, Y. A., Mochalova, L. V., and Bovin, N. V. Influenza virus neuraminidase: structure and function. *Acta naturae 1*, 2 (2009), 26.

[67] Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D., and Higgins, D. G. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology 7*, 1 (Apr. 2014), 539–539.

[68] Smith, T. F., and Waterman, M. S. Identification of common molecular subsequences. *Journal of molecular biology 147*, 1 (1981), 195–197.

[69] Sokal, R. R., and Michener, C. D. *A statistical method for evaluating systematic relationships*. 38:1409-1438. University of Kansas Science Bulletin, 1958.

[70] Steinhauer, D. A. Role of hemagglutinin cleavage for the pathogenicity of influenza virus. *Virology 258* (1999), 1–20.

[71] Stern, L. J., Brown, J. H., Jardetzky, T. S., Gorga, J. C., Urban, R. G., Strominger, J. L., and Wiley, D. C. Crystal structure of the human class ii mhc protein hla-dr1 complexed with an influenza virus peptide. *Nature 368*, 6468 (Mar 1994), 215–221.

[72] Suarez, D. L. Evolution of avian influenza viruses. *Veterinary Microbiology* (2000), 74:15–27.

[73] Sugiyama, K., Obayashi, E., Kawaguchi, A., Suzuki, Y., Tame, J. R. H., Nagata, K., and Park, S.-Y. Structural insight into the essential pb1-pb2 subunit contact of the influenza virus rna polymerase. *EMBO J 28*, 12 (Jun 2009), 1803–1811.

[74] Sun, X., Shi, Y., Lu, X., He, J., Gao, F., Yan, J., Qi, J., and Gao, G. Bat-derived influenza hemagglutinin h17 does not bind canonical avian or human receptors and most likely uses a unique entry mechanism. *Cell Reports 3*, 3 (Mar. 2013), 769–778.

[75] Suzuki, Y. Natural Selection on the Influenza Virus Genome. *Molecular Biology and Evolution 23*, 10 (July 2006), 1902–1911.

[76] Tarendeau, F., Crepin, T., Guilligay, D., Ruigrok, R. W. H., Cusack, S., and Hart, D. J. Host Determinant Residue Lysine 627 Lies on the Surface of a Discrete, Folded Domain of Influenza Virus Polymerase PB2 Subunit. *PLoS Pathogens 4*, 8 (Aug. 2008), e1000136.

[77] TONG, S., ZHU, X., LI, Y., SHI, M., ZHANG, J., BOURGEOIS, M., YANG, H., CHEN, X., RECUENCO, S., GOMEZ, J., AND OTHERS. New world bats harbor diverse influenza a viruses. *PLoS pathogens 9*, 10 (2013), e1003657.

[78] TSURUMURA, T., QIU, H., YOSHIDA, T., TSUMORI, Y., AND TSUGE, H. Crystallization and preliminary x-ray diffraction studies of a surface mutant of the middle domain of pb2 from human influenza a (h1n1) virus. *Acta Crystallogr F Struct Biol Commun 70*, Pt 1 (Jan 2014), 72–75.

[79] TUSCHE, C., STEINBRUCK, L., AND MCHARDY, A. C. Detecting patches of protein sites of influenza a viruses under positive selection. *Molecular Biology and Evolution 29*, 8 (Aug. 2012), 2063–2071.

[80] WARD, M. J., LYCETT, S. J., AVILA, D., BOLLBACK, J. P., AND BROWN, A. J. L. Evolutionary interactions between haemagglutinin and neuraminidase in avian influenza. *BMC evolutionary biology 13*, 1 (2013), 222.

[81] WEBSTER, R. G., BEAN, W. J., GORMAN, O. T., CHAMBERS, T. M., AND KAWAOKA, Y. Evolution and ecology of influenza a viruses. *Microbiol Rev 56*, 1 (Mar 1992), 152–179.

[82] WEIST, W. I., BRÜNGER, A. T., SKEHEL, J. J., AND WILEY, D. C. Refinement of the influenza virus hemagglutinin by simulated annealing. *J. Mol. Biol.*, 212 (1990), 737–761.

[83] WHO. Influenza (seasonal) fact sheet number 211, March 2014.

[84] WILEY, D., WILSON, I., AND SKEHEL, J. Structural identification of the antibody-binding sites of hong kong influenza haemagglutinin and their involvement in antigenic variation. *Nature* (1981), 289:373.

[85] WISE, H. M., HUTCHINSON, E. C., JAGGER, B. W., STUART, A. D., KANG, Z. H., ROBB, N., SCHWARTZMAN, L. M., KASH, J. C., FODOR, E., FIRTH, A. E., GOG, J. R., TAUBENBERGER, J. K., AND DIGARD, P. Identification of a novel splice variant form of the influenza a virus m2 ion channel with an antigenically distinct ectodomain. *PLoS Pathogens 8*, 11 (Nov. 2012), e1002998.

[86] XU, X., ZHU, X., DWEK, R. A., STEVENS, J., AND WILSON, I. A. Structural Characterization of the 1918 Influenza Virus H1n1 Neuraminidase. *Journal of Virology 82*, 21 (Nov. 2008), 10493–10501.

[87] YAMADA, S., HATTA, M., STAKER, B. L., WATANABE, S., IMAI, M., SHINYA, K., SAKAI-TAGAWA, Y., ITO, M., OZAWA, M., WATANABE, T., SAKABE, S., LI, C., KIM, J. H., MYLER, P. J., PHAN, I., RAYMOND, A., SMITH, E., STACY, R., NIDOM, C. A.,

Lank, S. M., Wiseman, R. W., Bimber, B. N., O'Connor, D. H., Neumann, G., Stewart, L. J., and Kawaoka, Y. Biological and Structural Characterization of a Host-Adapting Amino Acid in Influenza Virus. *PLoS Pathogens 6*, 8 (Aug. 2010), e1001034.

[88] Yamada, S., Hatta, M., Staker, B. L., Watanabe, S., Imai, M., Shinya, K., Sakai-Tagawa, Y., Ito, M., Ozawa, M., Watanabe, T., Sakabe, S., Li, C., Kim, J. H., Myler, P. J., Phan, I., Raymond, A., Smith, E., Stacy, R., Nidom, C. A., Lank, S. M., Wiseman, R. W., Bimber, B. N., O'Connor, D. H., Neumann, G., Stewart, L. J., and Kawaoka, Y. Biological and structural characterization of a host-adapting amino acid in influenza virus. *PLoS Pathogens 6*, 8 (Aug. 2010), e1001034.

[89] Yang, H., Carney, P. J., Donis, R. O., and Stevens, J. Structure and receptor complexes of the hemagglutinin from a highly pathogenic h7n7 influenza virus. *J Virol 86*, 16 (Aug 2012), 8645–8652.

[90] Yang, Z., and Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends in ecology & evolution 15*, 12 (2000), 496–503.

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

[92] Ye, Q., Krug, R. M., and Tao, Y. J. The mechanism by which influenza A virus nucleoprotein forms oligomers and binds RNA. *Nature 444*, 7122 (Dec. 2006), 1078–1082.

[93] Zehender, G., Pariani, E., Piralla, A., Lai, A., Gabanelli, E., Ranghiero, A., Ebranati, E., Amendola, A., Campanini, G., Rovida, F., Ciccozzi, M., Galli, M., Baldanti, F., and Zanetti, A. R. Reconstruction of the evolutionary dynamics of the a(h1n1)pdm09 influenza virus in italy during the pandemic and post-pandemic phases. *PLoS One 7*, 11 (2012), e47517.

[94] Zhang, W., Qi, J., Shi, Y., Li, Q., Gao, F., Sun, Y., Lu, X., Lu, Q., Vavricka, C. J., Liu, D., Yan, J., and Gao, G. F. Crystal structure of the swine-origin A (H1n1)-2009 influenza A virus hemagglutinin (HA) reveals similar antigenicity to that of the 1918 pandemic virus. *Protein & Cell 1*, 5 (May 2010), 459–467.

[95] Zhao, C., Lou, Z., Guo, Y., Ma, M., Chen, Y., Liang, S., Zhang, L., Chen, S., Li, X., Liu, Y., Bartlam, M., and Rao, Z. Nucleoside monophosphate complex structures of the endonuclease domain from the influenza virus polymerase pa subunit reveal the substrate binding site inside the catalytic center. *J Virol 83*, 18 (Sep 2009), 9024–9030.

[96] ZHU, X., MCBRIDE, R., NYCHOLAT, C. M., YU, W., PAULSON, J. C., AND WILSON, I. A. Influenza virus neuraminidases with reduced enzymatic activity that avidly bind sialic acid receptors. *J Virol 86*, 24 (Dec 2012), 13371–13383.

Declaration of Authorship

I hereby certify that this thesis has been composed by me and is based on my own work, unless stated otherwise. No other person's work has been used without due acknowledgement in this thesis. All references and verbatim extracts have been quoted, and all sources of information, including graphs and data sets, have been specifically acknowledged. Furthermore, I certify that this research thesis or any part of it has not been previously submitted for a degree or any other qualification.

Where I have used thoughts from external sources, directly or indirectly, published or unpublished, in full or parts, this is always clearly attributed.

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich meine Arbeit selbstständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die Arbeit noch nicht anderweitig für Prüfungszwecke vorgelegt habe.Darüberhinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, bereits für eine andere Prüfung angefertigt wurde.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Mittweida, 23. August 2015