Large-scale analysis of Influenza A virus nucleoprotein sequence conservation reveals potential drug-target sites

Andreas Kukol[#] & David John Hughes[1]
School of Life and Medical Sciences, University of Hertfordshire, Hatfield, UK

#corresponding author: a.kukol@herts.ac.uk
Tel.: +44-(0)1707 284 543
FAX: +44-(0)1707 285 046

[1] Present address: Rothamsted Research, Harpenden, UK

Running title: Drug target sites of the influenza nucleoprotein

Abstract

The nucleoprotein (NP) of the influenza A virus encapsidates the viral RNA and participates in the infectious life cycle of the virus. The aims of this study were to find the degree of conservation of NP among all virus subtypes and hosts and to identify conserved binding sites, which may be utilised as potential drug target sites. The analysis of conservation based on 4430 amino acid sequences identified high conservation in known functional regions as well as novel highly conserved sites. Highly variable clusters identified on the surface of NP may be associated with adaptation to different hosts and avoidance of the host immune defence. Ligand binding potential overlapping with high conservation was found in the tail-loop binding site and near the putative RNA binding region. The results provide the basis for developing antivirals that may be universally effective and have a reduced potential to induce resistance through mutations.

Key words: evolution, protein structure, drug target, binding site, sequence conservation

Introduction

The influenza A virus causes a respiratory disease resulting in annually recurring epidemics affecting an estimated 3-5 million people and about 250,000 to 500,000 deaths worldwide (WHO, 2009). At the same time the influenza virus causes substantial losses among domesticated birds (Leibler et al., 2009). Apart from epidemics, influenza A viruses have been responsible for devastating pandemics killing at least 40 million people in 1918/1919 (Spanish Flu, H1N1) (Johnson and Mueller, 2002) and less serious pandemics in 1957 (Asian Influenza, H2N2), 1968 (Hong Kong Influenza, H3N2), 1977 (Russian Influenza, H1N1) (Cox and Subbarao, 2000) and the latest pandemic (Swine Flu, H1N1) killing an estimated 284,000 people (Dawood et al., 2012). Influenza pandemics seem to occur when a pathogenic animal-type virus acquires the capability of efficient human to human transmission (Horimoto and Kawaoka, 2005), which may occur due to mutations or reassortment of human and animal RNA segments (Lin et al., 2000). A current threat, besides the avian H5N1 virus, is an avian H7N9 virus emerging in China that has caused 104 confirmed infections including 21 deaths (WHO, 2013). Due to the high mutation rate and emerging resistance against neuraminidase inhibitors (Ferraris and Lina, 2008) and amantadine/rimantadine (Rahman et al., 2008), it is of utmost importance to investigate other viral proteins as

potential drug targets for example the non-structural proteins (Darapaneni, Prabhaker, and Kukol, 2009).

The influenza A virus belongs to the family Orthomyxoviridae. It is a lipid-enveloped virus with a negative-strand RNA genome organized into eight separate segments, which code for eleven or more proteins dependent on the particular virus strain (reviewed in e.g. in Das et al. (2010)). Based on the antigenic properties of the proteins haemagglutinin and neuraminidase the virus is classified in subtypes HxNy. Segment five encodes the nucleoprotein (NP). The primary function of NP is to encapsidate the segmented RNA and bind with the three polymerase subunits, PA, PB1 and PB2, to form ribonucleoprotein particles (RNPs) for RNA transcription, replication and packaging (Du, Cross, and Zhou, 2012). The molecular structure of NP from an H1N1 virus (Ye, Krug, and Tao, 2006) and an H5N1 virus (Ng et al., 2008) has been elucidated by x-ray crystallography. NP is composed of a head domain, a body domain, and a flexible tail loop (figure 1A). In each RNP, the viral RNA wraps around individual NP molecules, in an arginine-rich RNA-binding groove (Marklund et al., 2012; Ng et al., 2008) (figure 1A). NP oligomers are formed by insertion of the tail loop of one NP molecule into the tail loop binding pocket of another, via a crucial salt bridge between Glu339 and Arg416 (Coloma et al., 2009; Shen et al., 2011).

Influenza NP is the most abundantly expressed protein during the course of infection, with multiple functionalities including viral RNA synthesis and RNP trafficking, polymerase regulation and interaction with cellular polypeptides, including actin (Li et al., 2009; Portela and Digard, 2002). Various nuclear localisation signals (NLS) in the amino acid sequence of NP are critical for nuclear import of the whole polymerase complex (reviewed in (Hutchinson and Fodor, 2012)). The N-terminus contains an unconventional NLS, which becomes deactivated after Ser-3 phosphorylation (Wu and Pante, 2009). A potential bipartite NLS was identified in the RNA binding groove betwee residues 198 and 216 (Weber et al., 1998) and a third NLS between residues 320 and 400 was identified based on deletion studies (Wang, Palese, and ONeill, 1997). Overall it appears that the N-terminal unconventional NLS is the major determinant of nuclear import (Cros, Garcia-Sastre, and Palese, 2005), while a nuclear accumulation signal between residues 327 and 345 potentially acts to retain NP in the nucleus (Davey, Dimmock, and Colman, 1985). NP has been identified as a target for drug discovery, and some potential inhibitors acting against the NP protein were identified, namely nucleozin that triggered aggregation of NP (Kao et al., 2010), NP oligomerisation

inhibitors (Shen et al., 2011) and the RNA-binding inhibitor naproxen (Lejal et al., 2013). Analyses of NP sequences at the amino acid or nucleotide level have been conducted previously. A large scale study by Xu et al. (2011) of based on 5094 NP nucleotide sequences split into different evolutionary linages identified six highly variable sites and a few hundred conserved sites, the number of which depend on the evolutionary linage and methodology used to calculate conservation. A review article reported the fraction of conserved amino acid residues in secondary structure elements (Ng, Wang, and Shaw, 2009), but no clear definition of conservation was given. Most references to sequence conservation of NP in the literature (e.g. Mena et al., 1999; Ng, Wang, and Shaw, 2009) go back to a study conducted by Shu et al. (Shu, Bean, and Webster, 1993) based on multiple alignment of 49 sequences.

The aims of the current study were to identify the degree of conservation of NP among all influenza A virus subtypes from all hosts in order to identify sites of universal conservation. The mapping of the conservation scores onto the three-dimensional structure together with an analysis of the small-molecule binding potential suggests potential binding sites for antivirals that may be universally applicable without leading to resistance. Furthermore, our results suggest highly conserved sites with unknown function which could be further investigated experimentally.

Methods

Sequence analysis

The protein sequences of the influenza A virus NP were obtained from the National Center for Biotechnology information (NCBI) influenza virus resource (Bao et al., 2008). Sequences from all virus subtypes and all hosts were chosen. Sequence redundancy was removed by identifying clusters of sequences at 99% identity and replacing a cluster with a representative sequence using the CD-HIT suite (Huang et al., 2010). After the removal of sequences containing undefined amino acid residues the remaining sequences were subjected to multiple alignment with MUSCLE software (Edgar, 2004a; Edgar, 2004b) using an iterative refinement until convergence of the sum-of-pairs score was achieved, which yields the most accurate alignment. After conversion of the alignment to the Phylip format (Futami et al., 2008), a phylogenetic tree was calculated with PhyML 3.0 based on the maximum-likelihood method (Guindon et al., 2010) using the influenza-specific amino acid

substitution model FLU (Cuong et al., 2010). Multiple sequence alignments were visualised with Jalview (Waterhouse et al., 2009).

Protein Structure

The crystal structure of NP from the influenza A virus subtype H5N1 (Ng et al., 2008) was obtained from the RCSB Protein Data Bank (Berman et al., 2000) with the identifier 2QO6. This structure was chosen, because the number of missing residues (from 79-86) was less than in than in a similar NP structure from an H1N1 virus with the identifier 2IQH (Ye, Krug, and Tao, 2006). The coordinate file was manually split into a monomer, and a continuous model from Ala22 to Tyr496 was obtained from the i-TASSER server 2.0 (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) (Zhang, 2008) explicitly specifying the structure of 2QO6 as a template. None of the templates identified covered the missing region 79-86, thus according to the i-Tasser protocol, *ab-initio* modelling was applied with a knowledge-based force field (Wu, Skolnick, and Zhang, 2007). The final protein structure used for further analysis was obtained by copying the structural information of residues Tyr78 to Lys87 obtained from I-TASSER into the monomer crystal structure.

Conservation analysis

The conservation of amino acid residues was obtained and projected onto the protein structure using the ConSurf 2010 server (Ashkenazy et al., 2010; Celniker et al., 2013; Landau et al., 2005). The ConSurf algorithm takes into account the evolutionary relationships between protein sequences giving a greater weight to evolutionary more distant sequences. This is essential for producing meaningful conservation scores, since the protein sequences show high similarity due to originating from the same species. The protein structure pdb-file, the multiple sequence alignment and the phylogenetic tree calculated with the FLU substitution model (Cuong et al., 2010) were uploaded manually and the Bayesian statistic method was chosen. The resulting conservation scores are standard scores with an average of zero and a standard deviation of one. A score below zero denotes higher than average conservation. Confidence intervals of conservation scores were calculated with Bayesian statistics (Mayrose et al., 2004). For assigning conservation grades (1-9), the scores below and above zero are divided into 4½ intervals each, which form the nine conservation grades.

The interval for the lowest conservation grade 1 is expanded to include residues with the highest variability (Ashkenazy et al., 2010). A table of detailed results including statistical confidence intervals is available as supporting information.

Binding site prediction

Potential binding sites were predicted by two different methods, namely computational solvent mapping with FTMap that involves docking a library of small organic solvent-like molecules to the protein structure (Brenke et al., 2009) and QsiteFinder that scans the protein surface with a methyl van Waals probe for favourable interactions (Fuller, Burgoyne, and Jackson, 2009). Clusters of docked solvent molecules in case of FTMap or clusters of methyl probes in case of QSiteFinder can identify potential hot spots of binding. The FTMap binding sites are ranked based on the number of overlapping clusters of solvent molecules, while the QsiteFinder binding sites are ranked according to the total interaction energy. Both rankings implicitly take into account the volume of the binding site. The predicted binding sites and conservation scores were combined by manually transferring the coordinates of binding sites to the coordinate file (pdb-file) obtained from the ConSurf server. Protein structures with conservation scores and predicted ligand binding sites are available as supporting information. Figures of protein structures were prepared with Rasmol (Sayle and Milnerwhite, 1995).

Results

Influenza virus sequences

Initially 4430 sequences of influenza A virus NP of all subtypes and hosts were obtained. After clustering of sequences at 99% identity threshold, 815 sequences remained that showed at least 1% sequence difference. Among the sequences there were 33% from human, 26% from swine, 15% from chicken and 12% from duck hosts; the remaining 14% were from a variety of species including turkey, quail, mallard, horse, dogs as well as a variety of birds.

Conserved amino acid residues

The evolutionary conservation of amino acid residues in the NP protein were identified using the ConSurf server (Ashkenazy et al., 2010; Landau et al., 2005). Conservation scores were

obtained between -0.78 (highest conservation) and 4.8 (highest variability) and assigned to grades between 9 (highest conservation) and 1 (highest variablity) by the ConSurf server. Highly conserved residues (grades 8–9) and variable residues (grades 1–3) are shown in table 1. Conservation grades were mapped to the protein backbone (figure 1A, B). Overall the protein is highly conserved with 59% of residues in the highest conservation grades 7-9, while 21% of residues are highly variable with grades 1-3. A significant number of 38 residues highlighted in table 1 show no variation at all among the 4430 sequences analysed.

Figure 1C and figure 2 show the conservation mapped onto the protein structure. The first N-terminal helix shows a pattern of conservation that is mainly confined to one face of the helix starting with Ile25 (grade 8) followed by Ser28 (grade 9), Met32, Ile36, Tyr40, Gln42-Glu46. Another helix inside the body domain that packs against the single antiparallel beta-sheet present in NP has high conservation at residues Gln58, Ser60 (grade 8), Thr62, Met 66 and Ser69 (all grade 9). The following loop until Glu81 contains several highly conserved residues, such as Glu73, Arg74, Arg75, Asn76, and Glu80. Note that a part of this sequence from Leu79-Gly86 was not resolved in the original structure, but is based on computational modelling. This high conservation is mirrored by an opposite loop from Gly169 (grade 9) to Ser176 (grade 8). Both loops overlook a highly conserved α-helix from Gly132 to Thr147. The end of this α-helix marks the transition into the head domain. Two short α-helices Thr151-Thr157 and Asp160-Ser165 with highly conserved features precede the second loop just described. This loop is followed by a four-helix bundle from Gly185-Leu264 after which the chain transits into the body domain. The four helix-bundle contains some variable as well as highly conserved residues, notably Gly185, Arg208, Gly212, Arg221, Cys223, Glu252, Glu254, Arg261, Ser262 and Ala263 are all conserved at grade nine. Inside the body domain from Lys273 until Arg391 some occurrences of highly variable residues are observed. The amino acid chain then forms the tail loop from approximately Arg400 to Lys430. Notably, some residues in the tail loop show high conservation at grades eight or nine, such as Ser407, Gln409, Pro410, Phe412, Ser413, Val414, Gln415 and Pro419. The tail-loop binding pocket is formed by residues not adjacent in sequence and its conservation is shown in table 3 and discussed further in the next section. The tail loop is then followed by another α-helix with highly conserved Ile445, Glu449 and Ser450. The chain traverses again to the body domain, where it reveals a well-resolved structure composed of largely non-repetitive secondary structure elements. Very high conservation is found in this region at Ser457, Gly460, Val463, Glu465, Ala471, Val476, Pro477 and Ala493.

On the other hand we found clusters of highly variable regions. Ser50, Gln52 and Phe313 form a small highly variable (grade 1) cluster on the surface of the body domain. A continuous patch of variable sites is formed by Arg98, Arg100, Asp101, Gly102, Lys103, Val105, Glu372, Ala373, Met374, Asp375 and Pro318. Another area of variable sites is formed by Thr350, Arg351, Ile353 and Tyr496 (figure 2).

A consensus sequence calculated from the multiple alignment was compared to the human influenza A pandemic NP sequences in figure 3. At residue 100 there is a striking difference between the consensus sequence and pandemic sequences, namely the consensus residue Arg100, is replaced by hydrophobic residues Ile and Val in human pandemic sequences. At all other variable positions the same or a similar residue to the consensus is found in at least one of the pandemic sequences.

Small-molecule ligand binding potential

The binding potential of NP for small molecule drugs was predicted based on clusters of energetically favourable interaction sites with a methyl-probe (Q-SiteFinder (Laurie and Jackson, 2005)) and docking of a library of small organic solvent molecules (FtMap (Brenke et al., 2009)). The spatial arrangement of binding sites is shown in Figure 4 for both methods. The binding sites are mainly located in the body domain with some sites at the interface between head and body domain. FTMap identified a number of sites within the binding pocket of the tail loop, namely F1, F2, F4, F4, F5, F9 ('F' denotes binding sites identified by FTMap, while the number indicates the rank assigned by the algorithm). Further binding sites F3 and F8 were found in the RNA binding region, while F7 is located in a different region of the protein. Q-SiteFinder identified binding sites in similar as well as different locations. Similar binding sites are Q3 and Q7, which are located in the tail loop binding pocket. The highest-ranked binding Q1 is located in the RNA binding region. Q4 and F7 are in similar regions, while there is not much agreement between the other binding sites.

An important aim of the present work was to identify binding sites that are spatially close to conserved amino acid residues. Table 2 identifies residues, which are in close proximity of the binding sites. The highest ranked sites have a large volume, thus spatially distant residues are included. The number of residues included in table 2 does not relate to the ranking of binding sites, but to the number of residues which are in close proximity based on the arbitrary cut-off of 0.3 nm. As shown in table 2, most predicted binding sites are close to

one or more highly conserved residues, except F4, Q3, Q8 and Q9. Notably the three highest ranked sites F1, F2 and F3 are in close proximity to conserved residues, while two highest ranked sites of the ConSurf method Q1 and Q2 are close to conserved residues.

Discussion

Sequence conservation

The objective of this study was to determine the degree of conservation of NP among all sequenced influenza A viruses isolates and to identify small-molecule binding sites in conserved regions, which may form potential target sites for future antiviral drug discovery. The protein structure of NP was based on the recently published X-ray structure (Ng et al., 2008) of an H5N1 isolate with the missing amino acids residues 79-86 build by ab-initio modelling with the i-Tasser server (Zhang, 2008). Sequence conservation may arise due to functionally important sites, such as interaction sites or targeting signals, sites important in maintaining the protein structure (Schueler-Furman and Baker, 2003) or on the RNA level as packaging signals (Gog et al., 2007). Variable sites on the other hand may arise due to adaptation to different hosts or due to evolutionary pressure to escape the host immune system. Significant amounts of NP were detected on the surface of infected cells, which enables the detection by the host immune system (Yewdell, Frank, and Gerhard, 1981).

In established functionally important regions we find high conservation, for example in the bipartite nuclear localisation signal (residues 198 to 216) (Wang, Palese, and ONeill, 1997), Trp207, Arg208, Gly209 and Gly212 are conserved at grade 9. In addition it was shown that Arg204, Trp207 and Arg208 bind to the viral polymerase (Marklund et al., 2012). In the nuclear accumulation signal (residues 327 to 345) (Davey, Dimmock, and Colman, 1985) Ala332, Ala337 and Glu339 are conserved at grade 9. Significant conservation was found near the putative RNA binding grove in two flaps overlooking a highly conserved α-helix (figure 1). Previously it was found that mutation of Asp72, Arg74, Lys113, Arg156, Arg174, Arg175, Arg195, Arg199, Lys325 and Arg361 to alanine impaired the incorporation of viral RNA (Li et al., 2009). All of those residues except Arg174 and Arg195 are conserved at grade 8 or 9. Another region of conservation can be found in the tail loop (residues 402-428), which facilitates NP oligomerisation by binding to a corresponding pocket formed between the NP body and head domain. Gln409, Pro410, Phe412, Gln415 and Pro419 in the tail loop

are conserved at grade 9, while many other residues including Arg416 show conservation at grade 8. The mutation Phe412Ala was shown to decrease the transcription of viral RNA (Li et al., 2009). Figure 5 shows the binding of the tail loop into a cavity between head and body domain.

Table 3 shows the conservation grades of residues of the tail loop binding pocket, which are in close contact with the tail loop. Notable the mutation Arg267Ala decreased the assembly of the viral NP by 50%, while the Glu339Ala completely abolished NP formation (Chan et al., 2010). Also mutation of the highly conserved Glu449 located near to the tail loop in the helix of the head domain decreases assembly of NP (Chan et al., 2010). It was recently shown that isolated NP exists in equilibrium between monomers and trimers and mutations Arg416Ala, Glu339Ala stabilise the monomeric form, while Tyr148Ala shifts the equilibrium to the trimeric form (Chenavas et al., 2013). Arg416 is located in the tail loop and shows a limited residue variety of Arg, Lys and Gly with a conservation grade of 8. While the wild-type monomer cannot be crystallised, a crystal structure of the monomeric mutant Arg416Ala showed that the tail loop is folded onto its own body domain covering the tail loop binding site (Chenavas et al., 2013). Furthermore, Chenavas et al (Chenavas et al., 2013) postulated that the wild-type monomer is stabilised by phosphorylation of Ser165 and showed using biophysical measurements that a phosphomimetic mutation Ser165Asp leads to a monomeric NP. This explains the high conservation of Ser165 (table 3) that only allows for a substitution by a Thr residue in some virus isolates.

Additionally we identified areas on the surface of NP that showed significant variation (figure 2). Some of the variation can be explained by adaptation to different hosts. It was shown that mutations of the residue 313 confers resistance to the human interferon-induced Mx1 protein (Manz et al., 2013). Mx1 has a known antiviral activity against RNA viruses by binding to their ribonucleocapsid (Haller, Staeheli, and Kochs, 2007). In addition residues 53, 100, 283 and 289 were shown to cause Mx1 resistance (Manz et al., 2013). Of those only residues 100 and 283 are highly variable, while Gly53 is conserved at grade 7 and Tyr289 is conserved at grade 6, albeit there was insufficient data to reliably assign the conservation of Tyr289. Notably, pandemic influenza strains have hydrophobic Ile or Val at position 100, while the consensus is the positively charged Arg residue (figure 3). The significance of other highly variable clusters remains to be established. Notably, NP was detected on the surface of infected cells, which enables the detection by the host immune system (Yewdell, Frank, and

Gerhard, 1981). Evolutionary pressure to avoid the host immune system may explain some of this variation.

Small-molecule ligand binding potential

A significant binding potential for small molecule ligands was found in the tail loop binding pocket. Some of the highest ranked sites, such as F1 and F2, were found here in proximity of highly conserved residues (table 2). This pocket has already been exploited for the development of influenza virus inhibitors with the strongest inhibitory activity (IC) of IC_{50} = 2.7 µM shown by an organic molecule (Shen et al., 2011). Based on our results one can predict, that compounds targeting this pocket are unlikely to be become ineffective due to virus resistance, since residues in this pocket are highly conserved among virus subtypes and different host organisms. A study by Kao et al. (2010) identified nucleozin and analogues as a replication inhibitor with an IC_{50} between 0.07 and 0.33 µM for different virus isolates. A mutation Tyr289His was identified as the only mutation contributing to resistance of an escape mutant after five passages of selection indicating that the binding site for nucleozin is located close to Tyr289. This corresponds to the binding site Q9 identified in the present study (table 9), which is not located close to highly conserved region explaining the emergence of escape mutants in the previous study.

Another potential binding site for small molecule ligands in the RNA binding region was detected as sites F3, F8 and Q1, Q6. These binding sites are close to a highly conserved helix, thus resistance against inhibitors is unlikely to develop. The residue Tyr78 close to Q1 is located on a flexible loop, which can enable the access of small molecule ligands to Q1. It was shown that flexibility of the two loops overlooking Q1 is required for NP activity (Tarus et al., 2012). Thus any ligand binding to Q1 that interacts with loop residues such as Tyr78 could reduce loop flexibility and inhibit RNA binding. That binding site has not yet been explored for development of influenza virus inhibitors. However, recently the anti-inflammatory drug naproxen has been shown to inhibit RNA binding to NP and reduced viral titres with EC_{50} = 11 µM (Lejal et al., 2013). Based on molecular docking and site-specific mutagenesis a binding site in proximity to Q2 was detected. The mutations Tyr148Ala and Arg361Ala abolished the ability of naproxen to inhibit RNA binding. Furthermore, no escape mutants were detected after six passages of selection, as it would be expected based on our conservation analysis of the Q2 binding site shown in table 2.

Other binding sites are also in contact with conserved amino acid residues, such as Q10 and F9. While these binding sites received low ranks from the detection algorithm, the nucleozin example mentioned above (Kao et al., 2010) shows that even low ranking binding site have a potential to bind small organic molecules, possibly due to a conformational change of the protein, which the binding site detection algorithms do not take into account.

Some lower ranking binding sites may be involved in NP-NP interactions, since NP exists as oligomers in the ribonucleprotein complex.

Conclusion

A study of 4430 nucleoprotein sequences reveals a high sequence conservation that overlaps with potential binding sites rendering the nucleoprotein an excellent drug-target for small-molecule antiviral drug discovery. Targeting the tail loop binding pocket or potential binding sites near the RNA binding region are the most promising strategies for antivirals that are unlikely to lead to resistance in the future, since the binding sites are conserved among different subtypes and hosts. Furthermore, antivirals targeting these sites are likely to be universally effective against viruses of avian, swine or human origin.

Acknowledgments

This work was supported by the School of Life and Medical Sciences at the University of Hertfordshire, UK, and it has made use of the Science and Technology Research Institute high-performance computing facility.

References

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* **38**(Web Server issue)**,** W529-33.

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., and Lipman, D. (2008). The influenza virus resource at the National Center for Biotechnology Information. *J Virol* **82**(2)**,** 596-601.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**(1)**,** 235-42.

Brenke, R., Kozakov, D., Chuang, G. Y., Beglov, D., Hall, D., Landon, M. R., Mattos, C., and Vajda, S. (2009). Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* **25**(5)**,** 621-627.

Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., Pupko, T., and Ben-Tal, N. (2013). ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. *Israel Journal of Chemistry* **53**(3-4)**,** 199-206.

Chan, W. H., Ng, A. K. L., Robb, N. C., Lam, M. K. H., Chan, P. K. S., Au, S. W. N., Wang, J. H., Fodor, E., and Shaw, P. C. (2010). Functional Analysis of the Influenza Virus H5N1 Nucleoprotein Tail Loop Reveals Amino Acids That Are Crucial for Oligomerization and Ribonucleoprotein Activities. *Journal of Virology* **84**(14)**,** 7337-7345.

Chenavas, S., Estrozi, L. F., Slama-Schwok, A., Delmas, B., Di Primo, C., Baudin, F., Li, X., Crepin, T., and Ruigrok, R. W. (2013). Monomeric nucleoprotein of influenza A virus. *PLoS Pathog* **9**(3)**,** e1003275.

Coloma, R., Valpuesta, J. M., Arranz, R., Carrascosa, J. L., Ortin, J., and Martin-Benito, J. (2009). The Structure of a Biologically Active Influenza Virus Ribonucleoprotein Complex. *Plos Pathogens* **5**(6).

Cox, N. J., and Subbarao, K. (2000). Global epidemiology of influenza: Past and present. *Annual Review of Medicine* **51,** 407-421.

Cros, J. F., Garcia-Sastre, A., and Palese, P. (2005). An unconventional NLS is critical for the nuclear import of the influenza A virus nucleoprotein and ribonucleoprotein. *Traffic* **6**(3)**,** 205-13.

Cuong, C. D., Le, Q. S., Gascuel, O., and Vinh, S. L. (2010). FLU, an amino acid substitution model for influenza proteins. *Bmc Evolutionary Biology* **10,** 99.

Darapaneni, V., Prabhaker, V. K., and Kukol, A. (2009). Large-scale analysis of influenza A virus sequences reveals potential drug target sites of non-structural proteins. *Journal of General Virology* **90,** 2124-2133.

Das, K., Aramini, J. M., Ma, L. C., Krug, R. M., and Arnold, E. (2010). Structures of influenza A proteins and insights into antiviral drug targets. *Nature Structural & Molecular Biology* **17**(5)**,** 530-538.

Davey, J., Dimmock, N. J., and Colman, A. (1985). Identification of the sequence responsible for the nuclear accumulation of the influenza virus nucleoprotein in Xenopus oocytes. *Cell* **40**(3)**,** 667-75.

Dawood, F. S., Iuliano, A. D., Reed, C., Meltzer, M. I., Shay, D. K., Cheng, P. Y., Bandaranayake, D., Breiman, R. F., Brooks, W. A., Buchy, P., Feikin, D. R., Fowler, K. B., Gordon, A., Hien, N. T., Horby, P., Huang, Q. S., Katz, M. A., Krishnan, A., Lal, R., Montgomery, J. M., Molbak, K., Pebody, R., Presanis, A. M., Razuri, H., Steens, A., Tinoco, Y. O., Wallinga, J., Yu, H. J., Vong, S., Bresee, J., and Widdowson, M. A. (2012). Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. *Lancet Infectious Diseases* **12**(9)**,** 687-695.

Du, J., Cross, T. A., and Zhou, H. X. (2012). Recent progress in structure-based anti-influenza drug design. *Drug discovery today* **17**(19-20)**,** 1111-1120.

Edgar, R. C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5,** 1-19.

Edgar, R. C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**(5)**,** 1792-1797.

Ferraris, O., and Lina, B. (2008). Mutations of neuraminidase implicated in neuraminidase inhibitors resistance. *Journal of Clinical Virology* **41**(1)**,** 13-19.

Fuller, J. C., Burgoyne, N. J., and Jackson, R. M. (2009). Predicting druggable binding sites at the protein-protein interface. *Drug Discovery Today* **14**(3-4)**,** 155-161.

Futami, R., Llorens, C., Vicente-Ripolles, M., and Moya, A. (2008). The alignment format converter server 1.0. *Biotechvana Bioinformatics* **Collections 2008,** Scripts.

Gog, J. R., Afonso, E. D., Dalton, R. M., Leclercq, I., Tiley, L., Elton, D., von Kirchbach, J. C., Naffakh, N., Escriou, N., and Digard, P. (2007). Codon conservation in the influenza A virus genome defines RNA packaging signals. *Nucleic Acids Res* **35**(6)**,** 1897-1907.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**(3)**,** 307-21.

Haller, O., Staeheli, P., and Kochs, G. (2007). Interferon-induced Mx proteins in antiviral host defense. *Biochimie* **89**(6-7)**,** 812-818.

Henikoff, S., and Henikoff, J. G. (1992). Amino-Acid Substitution Matrices from Protein Blocks. *Proc Natl Acad Sci U S A* **89**(22)**,** 10915-10919.

Horimoto, T., and Kawaoka, Y. (2005). Influenza: Lessons from past pandemics, warnings from current incidents. *Nature Reviews Microbiology* **3**(8)**,** 591-600.

Huang, Y., Niu, B. F., Gao, Y., Fu, L. M., and Li, W. Z. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**(5)**,** 680-682.

Hutchinson, E. C., and Fodor, E. (2012). Nuclear import of the influenza A virus transcriptional machinery. *Vaccine* **30**(51)**,** 7353-7358.

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

Kao, R. Y., Yang, D., Lau, L. S., Tsui, W. H. W., Hu, L. H., Dai, J., Chan, M. P., Chan, C. M., Wang, P., Zheng, B. J., Sun, J. A., Huang, J. D., Madar, J., Chen, G. H., Chen, H. L., Guan, Y., and Yuen, K. Y. (2010). Identification of influenza A nucleoprotein as an antiviral target. *Nature Biotechnology* **28**(6)**,** 600-U88.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., and Ben-Tal, N. (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* **33**(Web Server issue)**,** W299-302.

Laurie, A. T. R., and Jackson, R. M. (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **21**(9)**,** 1908-1916.

Leibler, J. H., Otte, J., Roland-Holst, D., Pfeiffer, D. U., Magalhaes, R. S., Rushton, J., Graham, J. P., and Silbergeld, E. K. (2009). Industrial Food Animal Production and Global Health Risks: Exploring the Ecosystems and Economics of Avian Influenza. *Ecohealth* **6**(1)**,** 58-70.

Lejal, N., Tarus, B., Bouguyon, E., Chenavas, S., Bertho, N., Delmas, B., Ruigrok, R. W. H., Di Primo, C., and Slama-Schwok, A. (2013). Structure-Based Discovery of the Novel Antiviral Properties of Naproxen against the Nucleoprotein of Influenza A Virus. *Antimicrobial Agents and Chemotherapy* **57**(5)**,** 2231-2242.

Li, Z., Watanabe, T., Hatta, M., Watanabe, S., Nanbo, A., Ozawa, M., Kakugawa, S., Shimojima, M., Yamada, S., Neumann, G., and Kawaoka, Y. (2009). Mutational Analysis of Conserved Amino Acids in the Influenza A Virus Nucleoprotein. *Journal of Virology* **83**(9)**,** 4153-4162.

Lin, Y. P., Shaw, M., Gregory, V., Cameron, K., Lim, W., Klimov, A., Subbarao, K., Guan, Y., Krauss, S., Shortridge, K., Webster, R., Cox, N., and Hay, A. (2000). Avian-to-human transmission of H9N2 subtype influenza A viruses: Relationship between H9N2 and H5N1 human isolates. *Proc Natl Acad Sci U S A* **97**(17)**,** 9654-9658.

Manz, B., Dornfeld, D., Gotz, V., Zell, R., Zimmermann, P., Haller, O., Kochs, G., and Schwemmle, M. (2013). Pandemic influenza A viruses escape from restriction by human MxA through adaptive mutations in the nucleoprotein. *PLoS Pathog* **9**(3)**,** e1003279.

Marklund, J. K., Ye, Q. Z., Dong, J. H., Tao, Y. J., and Krug, R. M. (2012). Sequence in the Influenza A Virus Nucleoprotein Required for Viral Polymerase Binding and RNA Synthesis. *Journal of Virology* **86**(13)**,** 7292-7297.

Mayrose, I., Graur, D., Ben-Tal, N., and Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. *Molecular Biology and Evolution* **21**(9)**,** 1781-1791.

Mena, I., Jambrina, E., Albo, C., Perales, B., Ortin, J., Arrese, M., Vallejo, D., and Portela, A. (1999). Mutational analysis of influenza A virus nucleoprotein: Identification of mutations that affect RNA replication. *Journal of Virology* **73**(2)**,** 1186-1194.

Ng, A. K., Zhang, H., Tan, K., Li, Z., Liu, J. H., Chan, P. K., Li, S. M., Chan, W. Y., Au, S. W., Joachimiak, A., Walz, T., Wang, J. H., and Shaw, P. C. (2008). Structure of the influenza virus A H5N1 nucleoprotein: implications for RNA binding, oligomerization, and vaccine design. *FASEB J* **22**(10)**,** 3638-47.

Ng, A. K. L., Wang, J. H., and Shaw, P. C. (2009). Structure and sequence analysis of influenza A virus nucleoprotein. *Science in China Series C-Life Sciences* **52**(5)**,** 439-449.

Portela, A., and Digard, P. (2002). The influenza virus nucleoprotein: a multifunctional RNA-binding protein pivotal to virus replication. *Journal of General Virology* **83,** 723-734.

Rahman, M., Bright, R. A., Kieke, B. A., Donahue, J. G., Greenlee, R. T., Vandermause, M., Balish, A., Foust, A., Cox, N. J., Klimov, A. I., Shay, D. K., and Belongia, E. A. (2008). Adamantane-resistant influenza infection during the 2004-05 season. *Emerging Infectious Diseases* **14**(1)**,** 173-176.

Sayle, R. A., and Milnerwhite, E. J. (1995). Rasmol - Biomolecular Graphics for All. *Trends in Biochemical Sciences* **20**(9)**,** 374-376.

Schueler-Furman, O., and Baker, D. (2003). Conserved residue clustering and protein structure prediction. *Proteins-Structure Function and Genetics* **52**(2)**,** 225-235.

Shen, Y. F., Chen, Y. H., Chu, S. Y., Lin, M. I., Hsu, H. T., Wu, P. Y., Wu, C. J., Liu, H. W., Lin, F. Y., Lin, G., Hsu, P. H., Yang, A. S., Cheng, Y. S. E., Wu, Y. T., Wong, C. H., and Tsai, M. D. (2011). E339 ... R416 salt bridge of nucleoprotein as a feasible target for influenza virus inhibitors. *Proc Natl Acad Sci U S A* **108**(40)**,** 16515-16520.

Shu, L. L., Bean, W. J., and Webster, R. G. (1993). Analysis of the Evolution and Variation of the Human Influenza-a Virus Nucleoprotein Gene from 1933 to 1990. *Journal of Virology* **67**(5)**,** 2723-2729.

Tarus, B., Chevalier, C., Richard, C. A., Delmas, B., Di Primo, C., and Slama-Schwok, A. (2012). Molecular Dynamics Studies of the Nucleoprotein of Influenza A Virus: Role of the Protein Flexibility in RNA Binding. *PLoS ONE* **7**(1).

Wang, P., Palese, P., and ONeill, R. E. (1997). The NPI-1/NPI-3 (Karyopherin alpha) binding site on the influenza A virus nucleoprotein NP is a nonconventional nuclear localization signal. *Journal of Virology* **71**(3)**,** 1850-1856.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**(9)**,** 1189-1191.

Weber, F., Kochs, G., Gruber, S., and Haller, O. (1998). A classical bipartite nuclear localization signal on Thogoto and influenza A virus nucleoproteins. *Virology* **250**(1)**,** 9-18.

WHO (2009). Influenza (seasonal). *Fact sheet* **No. 211**.

WHO (2013). China-WHO Joint Mission on Human Infection with Avian Influenza A(H7N9) Virus, Vol. 2013, Bejing.

Wu, S. T., Skolnick, J., and Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. *Bmc Biology* **5**.

Wu, W. W., and Pante, N. (2009). The directionality of the nuclear transport of the influenza A genome is driven by selective exposure of nuclear localization sequences on nucleoprotein. *Virol J* **6,** 68.

Xu, J. P., Christman, M. C., Donis, R. O., and Lu, G. Q. (2011). Evolutionary dynamics of influenza A nucleoprotein (NP) lineages revealed by large-scale sequence analyses. *Infection Genetics and Evolution* **11**(8)**,** 2125-2132.

Ye, Q. Z., Krug, R. M., and Tao, Y. Z. J. (2006). The mechanism by which influenza A virus nucleoprotein forms oligomers and binds RNA. *Nature* **444**(7122)**,** 1078-1082.
Yewdell, J. W., Frank, E., and Gerhard, W. (1981). Expression of Influenza a Virus Internal Antigens on the Surface of Infected P815-Cells. *Journal of Immunology* **126**(5)**,** 1814-1819.
Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9,** 40.

Tables

Table 1: Conserved residues (conservation grades 8–9) and variable residues (conservation grades 1-3) of the NP protein identified using the ConSurf server (Ashkenazy et al., 2010). Residues in bold print do not show any variation.

Residue Classification	**NP protein**
Conserved (grades 8-9)	Ile25, Ser28, Val29, **Met32**, Ile36, Gly37, **Tyr40**, Gln42, Met43, Cys44, Thr45, Glu46, Asp51, Arg55, Gln58, Asn59, Ser60, Thr62, Met66, **Ser69**, Ala70, Phe71, Asp72, Glu73, Arg74, Arg75, **Asn76**, **Tyr78**, Glu80, Pro83, Lys87, **Asp88**, **Pro89**, Lys90, **Thr92**, Gly93, Gly94, **Tyr97**, Arg106, Leu110, **Lys113**, Trp120, Ala123, Asn124, Gly132, Thr134, His135, Met137, Ile138, His140, Ser141, Asn142, Leu143, Asn144, Asp145, **Thr147**, Arg150, Thr151, Ala153, Leu154, Val155, **Arg156**, Gly158, Asp160, **Pro161**, Met163, Cys164, Ser165, Leu166, Met167, **Gly169**, **Ser170**, **Thr171**, Leu172, **Pro173**, Arg175, Ser176, Ala178, **Ala179**, Ala181, **Gly185**, Gly187, Thr188, Met196, Arg199, Asn205, Phe206, Trp207, Arg208, **Gly209**, **Gly212**, Arg213, Ala218, Glu220, **Arg221**, Cys223, Leu226, Lys227, Gly228, Lys229, Thr232, Ala233, Gln235, Met238, Asp240, Gln241, Arg243, Pro248, Asn250, Ala251, Glu252, Glu254, **Arg261**, Ser262, **Ala263**, Arg267, Ala271, **Lys273**, **Ser274**, Val285, **Gly288**, Phe291, Ser297, Val299, **Gly300**, Asp302, Pro303, Gln308, Gln311, Ser314, Arg317, Glu320, His324, Lys325, Gln327, Leu328, Ala332, Ala336, Ala337, Glu339, Asp340, Arg342, Phe346, Arg348, **Gly349**, Pro354, **Arg355**, Arg361, **Ala366**, Thr378, Ser383, Tyr385, **Ala387**, Ile388, Ser392, Gln405, Ser407, **Gln409**, **Pro410**, **Phe412**, Ser413, Val414, **Gln415**, Arg416, **Pro419**, Thr424, Ala427, Phe429, Glu434, Arg436, Asp439, Ile445, Met448, **Glu449**, Ser457, Phe458, **Gly460**, Gly462, Val463, Glu465, Ser467, Ala471, Val476, Pro477, Phe489, Asp491, Ala493
Variable (grades 1-3)	Ala22, Arg31, Val33, Gly34, Arg38, Ile41, Ser50, Gln52, Gly54, Ile61, Ile63, Arg77, Ser84, Ala85, Arg98, Arg100, Asp101, Gly102, Lys103, Val105, Ile109, Tyr111, Glu114, Arg117, Ile119, Asn125, Glu127, Asp128, Met136, Ala146, Ile183, Val186, Met189, Val190, Ile194, Ile197, Ile201, Arg214, Ile217, Phe230, Lys236, Met239, Ile253, Ile257, Leu259, Leu283, Asp290, Arg293, Ile301, Leu306, Phe313, Pro318, Ala323, Val329, Leu341, Val343, Ser344, Thr350, Arg351, Val352, Ile353, Gln357, Leu358, Val363, Val371, Glu372, Ala373, Met374, Asp375, Ser377, Arg384, Arg391, Asn397, Gln398, Arg400, Ile406, Val408, Asn417, Leu418, Arg422, Ala423, Ile425, Lys430, Asn432, Thr433, Arg446, Ser450, Arg452, Pro453, Val456, Leu466, Thr472, Asn473, Asp480, Met481, Ser482, Asn483, Gly485, Asn492, Glu494, Glu495, Tyr496

Table 2: Amino acid residues within 0.3 nm of predicted binding sites. Conserved residues with grade 8 or 9 are shown in bold face.

FTMap		ConSurf	
Site	Amino acid residues	Site	Amino acid residues
F1	Ser344, **Ala387**, **Ile388**	Q1	**Tyr78**, **Ser141**, **Thr171**
F2	**Ser165**, Val186, **Ala263**, Gly268, Ile270, **Ser392**	Q2	**Asp145**, Tyr148, **Arg150**
F3	**Gln58, Thr62, Tyr97**, Ile365	Q3	Phe304, Trp330, His334, Ile347
F4	His272, His334, Ser335, Thr390	Q4	**Asp51**, Gly54
F5	**Ile388**, Arg461, Gly462	Q5	**Arg342**, Phe479, Asn483, Glu484
F6	**Ala337**, Phe338, **Glu339**, Ser486, **Phe489**	Q6	**Gln58**, **Thr62**, **Tyr97**
F7	**Tyr40**, Gly54	Q7	**Ser165**, Phe488, Phe489
F8	Ala131, **Thr134, His135**, **Lys273**	Q8	Leu307, Ser310, Val312, Leu279, Leu381
F9	**Glu339**, **Asp340**	Q9	Tyr289, Arg305, Asn309
		Q10	**Arg74**, Arg174, **Arg175**

Table 3: Amino acid residues in the tail loop binding pocket close (0.3 nm) to the tail loop

Residue	Conservation grade
Gln149	7
Ser165	9
Leu264	7
Arg267	9
Ile270	5
His272	9
Ser335	6
Phe338	6
Glu339	9
Arg342	8
Thr390	7
Gln459	4
Phe489	8

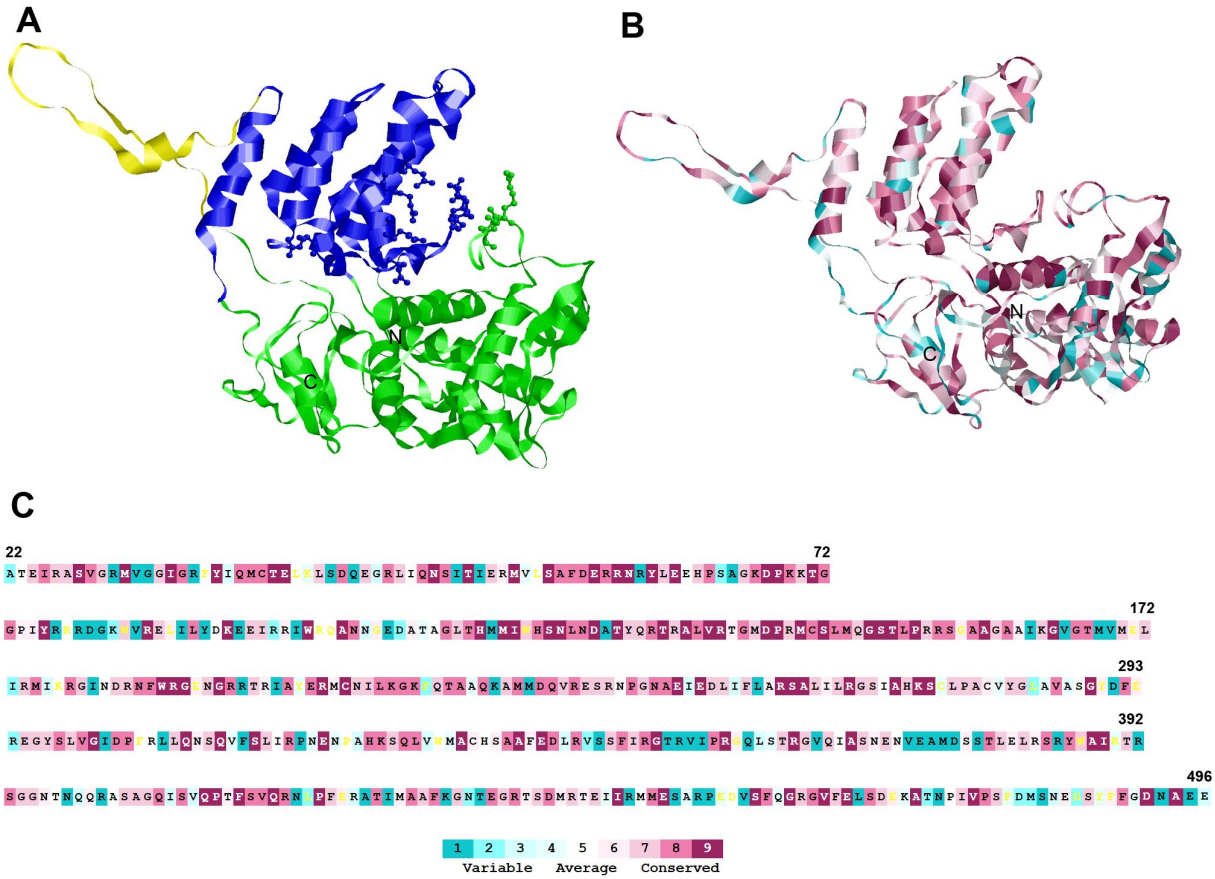

Figure 1: A) Structure of NP showing tail loop (yellow), head (blue) and body domain (green). Residues contributing to RNA binding are shown in ball-and-stick representation. B), C) Structure and amino acid sequence of NP, colour coded by conservation scores as indicated in the colour legend.

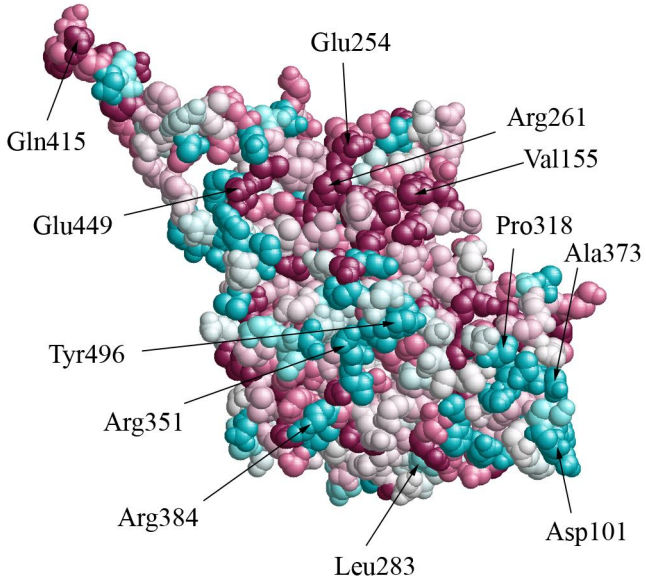

Figure 2: Spacefill representation of the NP structure (PDB-id 2Q06) with conservation grades mapped on the structure. See figure 1 for colour legend. Significant residues discussed in the text are labelled.

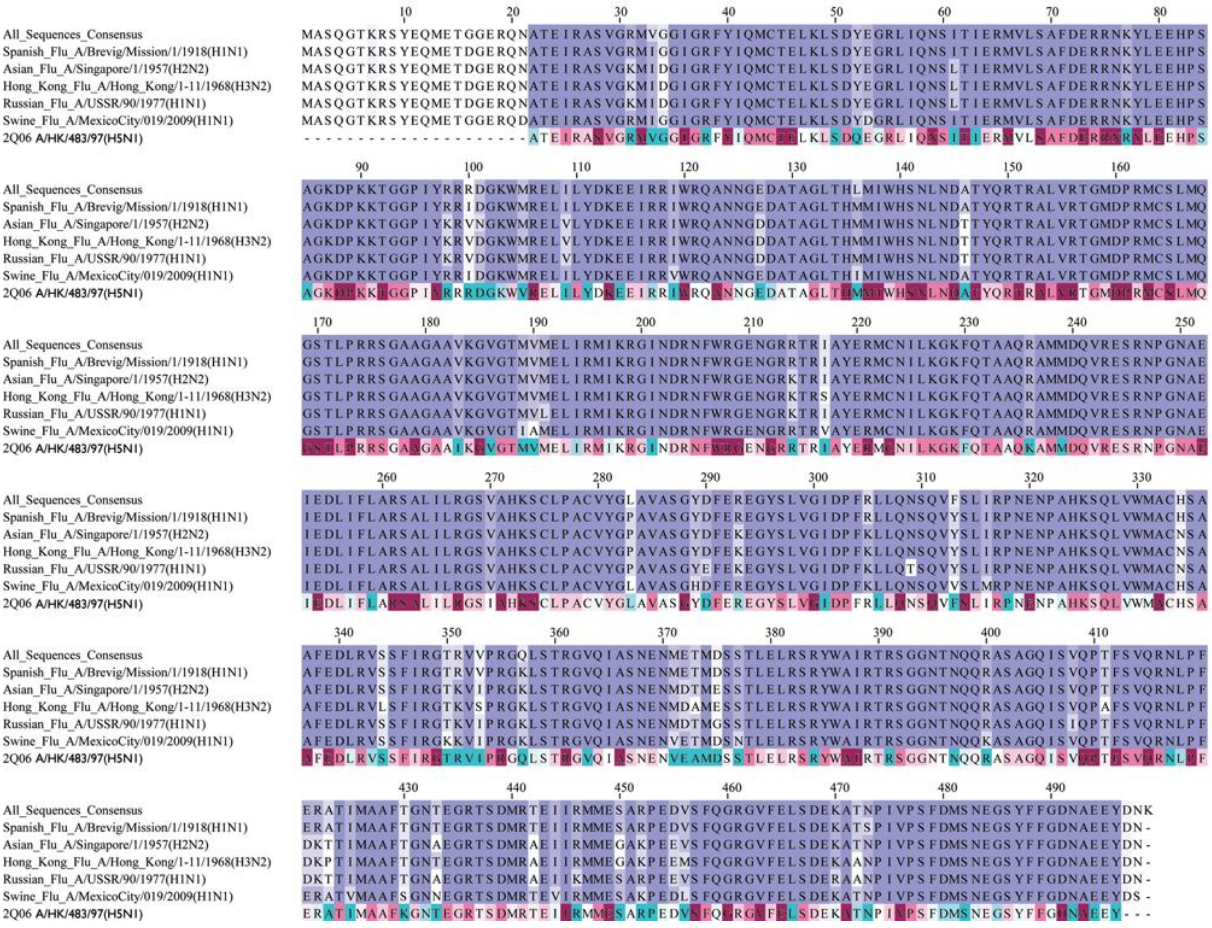

Figure 3: Multiple sequence alignment of a consensus sequence calculated from all sequences used in this study with sequences from influenza A pandemics and the sequence of the protein structure 2Q06. The sequence alignment is coloured by conservation based on the BLOSUM62 matrix (Henikoff and Henikoff, 1992). The 2Q06 sequence is coloured according to conservation as in figure 1.

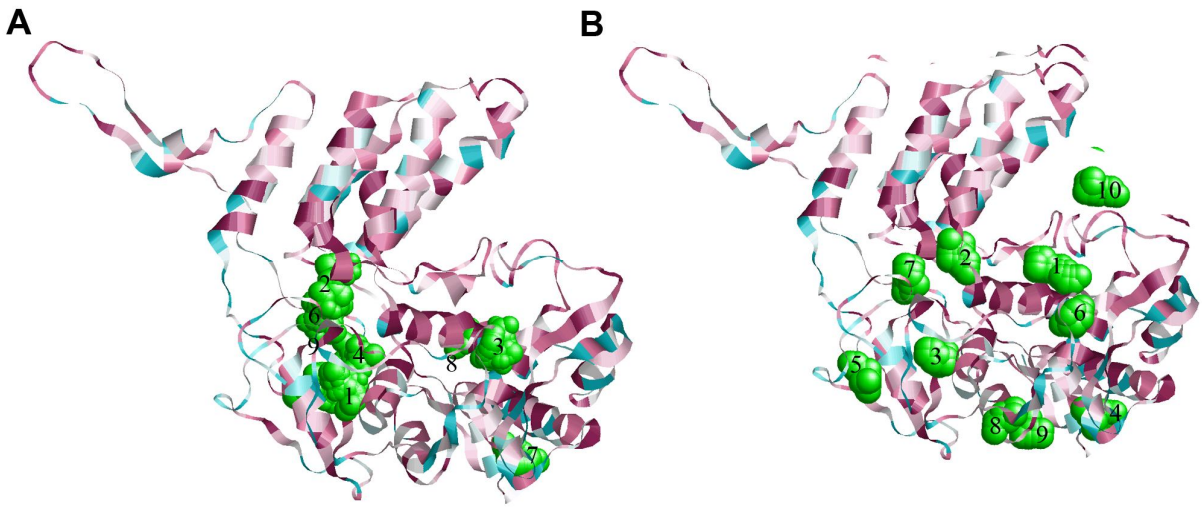

Figure 4: Ligand binding sites identified by FtMap (A) and Q-SiteFinder (B). The numbers indicate the ranking assigned by the algorithms, which '1' denoting the highest ranked binding site.

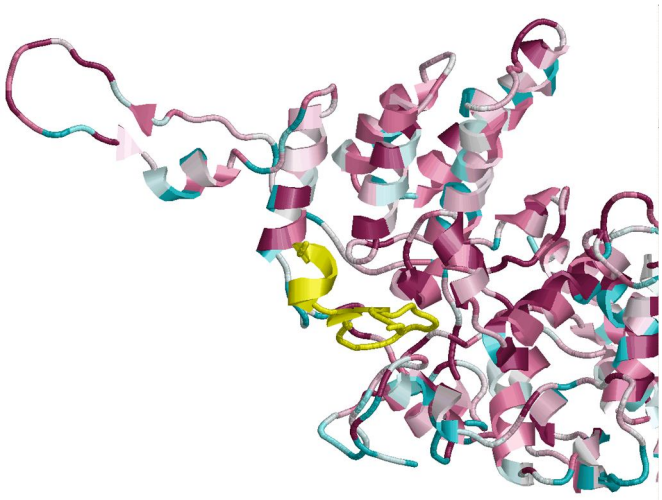

Figure 5: Illustration of the interaction between the tail loop and its binding pocket. The tail loop is based on the structure PDB-ID 2IQH (Ye, Krug, and Tao, 2006) that was fitted onto the 2QO6 structure analysed here.