Large-scale analysis of Influenza A virus sequences reveals potential drug-target sites of NS proteins

Vivek Darapaneni, Varun K. Prabhaker, Andreas Kukol

School of Life Sciences, University of Hertfordshire, UK

Running title: Drug-target sites of influenza NS proteins

Number of words: Summary 220, main text 4524, 3 figures and 2 tables

Correspondence:
Andreas Kukol, School of Life Sciences, University of Hertfordshire, College Lane, Hatfield AL10 9AB, United Kingdom
Tel.: +44 1707 284543, Fax: +44 1707 284870 E-Mail: a.kukol@herts.ac.uk

Summary

The non-structural protein 1 (NS1) of the influenza A virus and the protein NS2, which is also known as nuclear export protein, play important roles in the infectious life cycle of the virus. The objective of this study was to find the degree of conservation in the NS proteins and to identify conserved sites of functional or structural importance, which may be utilised as potential drug target sites. The analysis was based on 2620 amino acid sequences for the NS1 protein and 1195 sequences for the NS2 protein. The degree of conservation and potential binding sites were mapped onto the protein structures obtained from a combination of experimentally available structure fragments with predicted threading models. In addition to high conservation in protein regions of known function, novel highly conserved sites have been identified, namely Glu159, Thr171, Val192, Arg200, Glu208 and Gln218 on the NS1 protein and Ser24, Leu28, Arg66, Arg84, Ser93, Ile97, and Leu103 on the NS2 protein. Using the Q-SiteFinder binding site prediction algorithm, several highly conserved binding sites were found including two spatially close sites on the NS1 protein, which could be targeted with a bivalent ligand that would interfere with double-stranded RNA binding. Altogether, this work reveals novel universally conserved residues that are candidates for protein-protein interactions and provide the basis for designing universal anti-influenza drugs.

1. Introduction

The Influenza A virus causes a respiratory disease resulting in an average death toll of 36000 people each year in the United states alone (Molinari *et al.*, 2007). Apart from annually recurring epidemics, Influenza A viruses, which infect avian and mammalian species have been responsible for devastating pandemics killing at least 40 million people in 1918/1919 (Spanish Flu, H1N1) (Johnson & Mueller, 2002) and less serious pandemics in 1957 (Asian Influenza, H2N2), 1968 (Hong Kong Influenza, H3N2) and 1977 (Russian Influenza, H1N1) (Cox & Subbarao, 2000). Influenza pandemics seem to occur when a pathogenic avian type virus acquires the capability of efficient human to human transmission (Horimoto & Kawaoka, 2005) which may occur due to mutations or reassortment of human and avian RNA segments (Lin et al, 2000). A current threat is an avian H5N1 virus, which emerged in May 1997 (Claas *et al.*, 1998; Subbarao *et al.*, 1998) and has caused almost 250 human deaths up to 2009 as reported by the World Health Organisation (2004). Due to the high mutation rate and emerging resistance against neuraminidase inhibitors (Ferraris & Lina, 2008) and amantadine/rimantadine (Rahman *et al.*, 2008) it is of utmost importance to investigate viral proteins as potential drug targets.

The Influenza A virus belongs to the family *Orthomyxoviridae*. It is a lipid enveloped virus with a negative strand RNA genome organised into eight separate segments, which code for eleven proteins (reviewed in Kobasa & Kawaoka, 2005; Obenauer et al, 2006). The segment eight encodes the non-structural (NS) proteins, namely NS1 and NS2 (NEP). The major role of NS1 is to antagonise the antiviral response of the host by preventing the activation of NF-kB and induction of Alpha/Beta interferon (Wang *et al.*, 2000). However, NS1 is a multifunctional protein that is additionally involved in (i) inhibiting the pre-mRNA 3'-end processing (Fortes *et al.*, 1994) by binding to two 3' end processing factors, namely

cleavage and polyadenylation specificity factor and poly (A)-binding protein II (Chen *et al.*, 1999; Nemeroff *et al.*, 1998); (ii) blocking the post-transcriptional processing and nuclear export of cellular mRNA (Fortes *et al.*, 1994); (iii) stimulating the translation of matrix (M1) proteins (Enami *et al.*, 1994; Marion et *al.*, 1997); (iv) inhibiting the activation of a protein kinase that phosphorylates the elf-2 translation initiation factor by binding to double stranded RNA (Lu *et al.*, 1995; Aragón *et al.*, 2000); (v) induction of the phosphatidylinositol-3-kinase (PI3K)/Akt signalling pathway in order to support virus replication (Ehrhardt et *al.*, 2006; 2007). The NS1 protein is 230-237 amino acid residues long dependent on the strain and has a molecular mass of approximately 26 kDa (reviewed in Hale *et al.*, 2008b). Residues 1-73 form an N-terminal RNA binding domain (RBD) and residues 74-230 form a C-terminal effector domain (ED). The full-length protein exists most likely as a homodimer (Nemeroff *et al.*, 1995). Specific regions of functional importance that have been identified are (Hale *et al.*, 2008b): RNA binding region (residues 1-73); cleavage and polyadenylation specificity factor 4 (CPSF4) binding region (residues 175-210); poly (A)-binding protein (PABPN1) binding region (residues 218-225); nuclear localization signal 1 (residues 34-38); nuclear localization signal 2 (residues 211-216); nuclear export signal (residues 132-141); regions that interact with the regulatory p58β-subunit of PI3K (residues 89-93, 137-142, 164-167) (reviewed in Ehrhardt & Ludwig, 2009).

The non-structural protein 2 (NS2) was previously considered to be present in infected cells only until it was shown to exist in the viral particle (Richardson & Akkiina, 1991). Therefore NS2 is not strictly a non-structural protein and also referred to as nuclear export protein (NEP) according to its role in mediating the export of viral ribonucleoproteins from the nucleus to the cytoplasm through nuclear export signals and independent interaction with human chromosome region maintenance protein Crm1 (O'Neill *et al.*, 1998;

Neumann *et al.*, 2000). The NS2 protein is potentially involved in viral assembly through its interaction with the M1 protein that plays a key role in virus assembly (reviewed in Schmitt and Lamb, 2005). It consists of 121 amino acid residues and has a molecular mass of approximately 15 kDa (Greenspan *et al.*, 1985). Regions of functional importance in the NS2 protein that have been reported (Boutet *et al.*, 2007; O'Neill *et al.*, 1998) are the Influenza M1 protein binding region (residues 59-116) (Ward et al., 1995; Akarsu *et al.*, 2003) and a nuclear export signal (residues 11 to 23).

There are currently no licensed drugs available targeting the NS proteins of the influenza A virus but compounds acting against the NS1 protein were identified using a yeast based assay (Basu *et al.*, 2009) and high-throughput screening was used for targeting the interaction between the NS1 protein and viral RNA (Maroto *et al.*, 2008). The aims of the current study were to identify the degree of conservation among all subtypes of influenza A viruses so as to suggest potential drug target sites. Furthermore, models of the full-length protein structures were obtained and the degree of conservation was mapped onto the protein structures. Together with a binding pocket analysis, we suggest potential binding sites, which may define in-vivo interaction sites as well as provide starting points for the design of novel anti-influenza drugs.

2. Methods

2.1 Sequence analysis and protein structure prediction

The protein sequences of NS proteins were obtained from the National Centre for Biotechnology Information (NCBI) influenza virus resource (Bao *et al.*, 2008). Full-length sequences were chosen from all subtypes, all hosts and all lineages until year 2008 (inclusive), while identical sequences were removed. Both alleles A and B of the NS1

protein (Baez et *al.*, 1981) were analysed together. Multiple sequence alignment was carried out with MUSCLE 3.6 (Edgar, 2004) using the default parameters for most accurate alignment that involved multiple refinements of the alignment until no further improvement was achieved typically resulting in 20-24 iterations. The overview of the multiple sequence alignment was made with Jalview (Clamp *et al.*, 2004). Sequence conservation plots were made with the plotcon program of the EMBOSS package (Rice *et al.*, 2000) using a window size of 10 and the default comparison matrix EBLOSUM62. A full length protein structure was predicted for the NS1 sequence of the isolate influenza A/Indonesia/CDC1032N/2007(H5N1) with the I-TASSER server (http://zhang.bioinformatics.ku.edu/I-TASSER/) (Zhang, 2008). The isolate was selected as a recent example of an H5N1 virus that infected a human. I-TASSER uses multiple threading alignments and iterative simulations and was the top-ranked prediction method in the recent Critical Assessment of Protein Structure Prediction (CASP8) exercise (Zhang, 2008). The reliability of structure prediction results are reported with the TM-score that is a number between [0, 1], with a TM score smaller than 0.17 indicating a random model, while a TM-score > 0.5 corresponds to two structures of similar topology (Zhang & Skolnic, 2004). The final model of the NS1 protein structure was generated by replacing 82% of the residues of the predicted structure (residues 5-74 and 80-197) with the experimentally determined structure PDB-ID 3F5T (Bornholdt & Prasad, 2008). A model of the NS2 protein was obtained from the NS2 sequence of the same isolate with I-TASSER. The final model was generated by the replacing 48% of the predicted structure (residues 59-116) with the experimental structure PDB-ID 1PD3 (63-116) (Akarsu *et al.*, 2003).

2.2 Identification of conserved regions

Conserved regions were identified and mapped onto the protein structures using the web based ConSurf server (http://consurf.tau.ac.il/) (Glaser *et al.*, 2003; Landau *et al.*, 2005) providing as input the multiple sequence alignment and the protein structure file. The degree of conservation is subdivided into nine grades, with grade 1 being lowest and grade 9 being highest conserved. Ligand binding sites (LBS) were identified with the Q-SiteFinder (Laurie & Jackson, 2005) (http://www.modelling.leeds.ac.uk/qsitefinder/) that evaluates the interaction energy between a $-CH_3$ van der Waals probe and the protein (Laurie & Jackson, 2005). The resulting binding sites were formed from clusters of probes and ranked according to the sum of total binding energies for each cluster The ConSurf and Q-SiteFinder results were combined so as to show conservation and LBS simultaneously. The ConSurf-Q-SiteFinder results were obtained by projecting the ConSurf result on the structures containing the LBS obtained from Q-SiteFinder server (referred to as Q-SiteFinder-ConSurf method).

Protein atomic structures with conservation scores and predicted ligand binding sites will be made available upon request.

3. Results

3.1 Protein sequence multiple alignment

For the NS1 protein 2620 sequences were obtained from the NCBI Influenza virus resource (Bao *et al.*, 2008), which were mainly from avian (64%) and human (24%) viruses. For the NS2 protein 1195 sequences were obtained, which were mainly composed of avian (66%) and human (24%) sequences. The overview of the multiple sequence alignment shown in figure 1 reveals that sequences had a high degree of similarity despite being from different lineages of influenza A virus. Overall, the sequence conservation of both proteins

was of the same level, but the NS1 protein showed larger variations of the conservation pattern than the NS2 protein. In particular, at residue 79 of the NS1 sequence there was a low conservation as indicated by the low similarity value as well as light colours in figure 1A. This was caused by a deletion of the amino acid sequence [T,A][I,M]ASV at residue 79, which is the case in 20% of the analysed sequences including the sequence of the experimental NS1 protein structure (PDB-ID 3F5T). When discussing individual residues we include the count of the 5-residue deletion according to current conventions, i.e. NS1 residue 105 in our structure is counted as residue 110. Both proteins showed a low conservation at the C-terminus.

3.2 Protein structures

The predicted protein structure for the NS1 protein was obtained with a TM-score of 0.57 ± 0.15 from the protein prediction server (Zhang, 2008). In order to improve accuracy of the model most parts of the predicted model were replaced with the experimentally defined structure (Bornholdt & Prasad, 2008) leaving the N-terminal residues 1-5 and C-terminal residues 198-225 from the modelled structure. For that reason, the NS1 structure was virtually identical with the previous x-ray crystallography structure that is described here briefly in order to aid further analysis. The two-domain structure is evident from the three-dimensional structure shown in figure 2A. The N-terminal domain RNA-binding domain (RBD) shown on the right hand side is composed of two long antiparallel α-helices joined by a turn followed by a third shorter helix at an angle of 90° to the first two-helix bundle. The domain linker extends from approx. Glu70 to Tyr84. The C-terminal effector domain (ED) is formed from a 5-strand antiparallel beta-sheet, while a shorter sixth strand is oriented in a parallel fashion. An α-helix from Thr170 to Asn188 packs against this surface formed by the β-sheets. Additionally there are shorter α-helical and β-sheet

elements in this structure. The C-terminal part from Trp203 to Val230 that was based on prediction forms a short anti-parallel β-sheet followed by an α-helix from Gln218 to Glu229.

The NS2 model was obtained from I-TASSER with a TM-score 0.61 ± 0.14, which indicates a reliable model with an estimated root mean square deviation of the atomic coordinates to the true structure of 6.1 ± 3.8 Å. The protein structure after replacing residues 63-116 with the experimental determined fragment (Akaruso et *al*., 2003) is shown in figure 2B. The main structural feature is a four-helix bundle composed of two pairs of antiparallel α-helices. The predicted structure of the N-terminal residues 1-62 is composed of two antiparallel α-helices (helix one and two) that pack against the other two antiparallel helices (three and four). Between helix two and three there is a stretch of residues from Glu47 to Gln67 forming a short bend α-helix followed by a loop.

3.3 Conserved residues

The conserved residues in the NS1 protein and NS2 protein were identified using ConSurf server (Glaser *et al.*, 2003; Landau *et al.*, 2005) and are shown in the table 1. Highly conserved residues of the grades 7-9 and variable residues of the grades 1-3 were grouped together in table 1. Conservation scores were mapped to the protein backbone in figure 2 and on a spacefill model in figure 3. In the N-terminal RBD of the NS1 protein highly conserved residues (grade 9) are located on the inside part of the two-helix bundle facing the C-terminal ED. In particular, helix two from Asp29 to Thr49 shows a helical periodicity in the number of highest conserved residues, namely Pro31, Arg35, Asp39, Ser42, Arg46 and Thr49. A similar but less distinguished pattern of conservation is present on the first N-terminal helix. The linker region shows a substantially lower degree of conservation with only Leu69 and Glu72 in the grade 7 category. The C-terminal ED shows conserved

residues and variable residues at various positions. A highly conserved (grade 9) residue is Met93, which is with three exceptions present in all sequences analysed. This is followed by a short α-helix that contains the conserved Glu97-Arg100 (table 1). After a succession of highly conserved (grade 9) residues Trp102, Met104 and Pro107 that are highly surface exposed, we found significant conservation throughout the third, fourth and fifth anti-parallel β-sheets, namely Leu130-Val136, Leu146-Thr151 and Val157-Pro162. An α-helix packs against the surface formed by the antiparallel β-sheets. Conserved residues in this helix tend to be located at the packing interface between the helix and the surface formed by β-sheets, in particular Val174, Ala177, Leu181 and Gly184 . These residues most likely stabilise the structure by hydrophobic interactions. At the end of this helix we found a conserved feature of Glu186-Asn188 that is not involved in intra-protein interactions with Trp187 in the grade 9 category. Other highly conserved residues are Asn190, Gln199, Arg200 and Gln218, which is the only conserved residue at the beginning of the C-terminal α-helix.

In the NS2 protein the highest conserved part is the C-terminal helix from Ser93 to Arg114. The other parts of the predicted structure contain highly conserved residues at various positions including the turns and loops joining the α-helices, namely Ser24 and Trp65-Arg66. Some highly conserved residues are accessible from the outside, such as Ser17, Leu21, Trp65, Arg66, Glu75, Trp78, while others are buried inside.

3.4 Predicted ligand binding sites

Among the ten best ligand binding sites (LBS) reported by Q-SiteFinder (Laurie & Jackson, 2005), five LBS were chosen on the basis of the number of conserved residues around the LBS, which are shown in table 2. Note that the sites are not numbered sequentially, but the numbers represent the original ranking reported by Q-SiteFinder, whereby the highest-

ranking binding site is denoted as site number one. In the NS1 protein, site 1 located on the effector domain (ED) contains five highly conserved (grade 9) residues, namely Tpr102, Met104, Asp119, Arg148 and Glu159. Site 2 is located in the N-terminal RBD and contains eight conserved residues out of which Asp12, Arg19, Pro31, Arg35 and Asp39 are highly conserved. They are located on inside of the first two helices of the RBD facing the ED. Site 5 is located on the ED forming a cleft between two of the anti-parallel β-sheets and the long α-helix. Met104 and Asp120 surrounding this site are in the highest conservation category, while Gln109, Lys110, Ile117, Gln121 and Gly184 are also conserved. Site 6 is another binding site identified on the N-terminal RBD domain that contains ten conserved residues out of which Ser8, Arg38, Asp39, Ser42 and Arg46 are the highly conserved. Similar to site 2, site 6 is located on the side of the two long N-terminal helices facing the ED. Site 9 is formed by residues from the RBD from the long α-helix and short α-helix that is oriented towards the domain linker region. It contains six conserved residues out of which Leu69 is the highly conserved residue.

In the NS2 protein, site 1 forms a deep cleft inside the four-helix bundle and is accessible from the apex of the bundle. Site 1 is surrounded by nineteen conserved residues out of which Met1, Leu38, Tyr41, Glu75, Thr98, Ala102 and Leu103 are highly conserved. Site 2 is buried inside the helix bundle but accessible from the side. It contains nine conserved residues out of which Leu38 and Tyr41 are highly conserved residues. Site 7 is located at the protein surface in an apical position and contains four conserved residues out of which Arg84 is highly conserved. Site 8 is located at the apex opposite to site 7. It contains two conserved residues out of which Trp65 and Arg66 are the highly conserved residues. Notably, site 7 and site 4 have a substantial positive charge. Site 10 is formed from residues of the two N-terminal helices and is open to the outside. It contains five

conserved residues out of which Gly61 is the highly conserved residue. The combined results of sequence conservation and predicted binding sites, here referred to as the Q-SiteFinder-ConSurf method, are shown in figure 3.

4. Discussion

4.1 Protein structures

The objective of this study was to find the degree of conservation in the non-structural proteins among the influenza A viruses and to identify previously unknown sites of functional or structural importance, which may also form potential drug target sites. The protein structure of the NS1 protein was based largely on a recently published x-ray structure (Bornholdt & Prasad, 2008). For the NS2 protein only the structure of the C-terminal fragment residues 63-116 forming two antiparallel α-helices was know previously (Akarsu *et al.*, 2003), while the N-terminal fragment withstood all crystallisation efforts. In the x-ray structure the C-terminal fragment was obtained as a dimer, while the full-length NS2 protein is monomeric (Lommer & Luo, 2002). This was convincingly explained by the interaction between clustered hydrophobic residues on one side of the helical hairpin. Our predicted full-length model explains the existence of monomers, since the hydrophobic surface of the C-terminal helical hairpin is matched by a hydrophobic surface of the N-terminal hairpin. Our NS2 model of the N-terminal fragment is most likely one of several conformations, as it was shown to exist in a very flexible conformation (Lommer & Luo, 2002).

4.2 Sequence conservation and binding sites of the NS1 protein

Both alleles A and B of the NS1 genes of all hosts were analysed together in order to identify universally conserved features of potentially pandemic viruses that might arise

from re-assortment of human and avian viruses. The conserved residues detected on the NS proteins may have a functional significance in interaction with other proteins and RNA, may be important in stabilising the protein structure (Schueler-Furman & Baker, 2003) or may be conserved on the level of nucleotides as packaging signals (Gog et *al.*, 2007). For the NS1 protein RBD residues Arg38 and Lys41 are implicated in dsRNA binding (Wang *et al.*, 1999) and together with Arg35 they form a nuclear localisation sequence (Greenspan *et al.*, 1988). Arg35 and Arg38 are highly conserved (grade 9), while Lys41 is with grade 5 not among the highest conserved residues. Arg38 and Lys41 are also implicated in the inhibition of β-interferon mRNA production by direct interaction with an RNA helicase, the cytoplasmic pathogen sensor RIG-I (Opitz *et al.*, 2007; Pichlmair *et al.*, 2006) . Furthermore it has been shown that mutation of the highly conserved Ser42 can reduce virus virulence (Donelan *et al.*, 2003) possibly by antagonising the host's interferon response pathways. The clusters of highly conserved basic and hydrophilic residues on the inside of the RBD facing the ED, are involved in dsRNA binding (Wang *et al.*, 1999; Yin *et al.*, 2007). The effector domain (ED) shows a similar degree of conservation. Some of the highly conserved residues of the ED, which are buried inside, most likely fulfil a role in stabilising the protein structure such as Ala132, Leu144, Ala149, Ile160, Ala177 and Leu181. Residues 81-113 were reported to interact with a translation initiation factor eIF4F (Aragón *et al.*, 2000). Based on the present analysis we postulate that the conserved residues Tyr89, Ser99, Met104, Leu105, Pro107, Gln109 and Lys110 are involved in interaction with eIF4F, while other highly conserved residues in this region such as Met93 and Trp102 are stabilising the structure, since they are not as exposed to the surface as the previous residues. Together with Tyr89 a Met93 residue has been proposed to bind to the C-terminal SH2-domain p85β isoform of the lipid kinase PI3K (Hale *et al.*, 2006). Indeed Tyr89 and Met93 are

conserved, while Met93 is partially buried inside the protein but Tyr84 is highly surface exposed. Other residues involved in p85β interaction include 159 and 162 (Shin et *al*., 2007). Both Pro159 and Pro162 are surface exposed yet not highly conserved, but they are surrounding a highly surface exposed and conserved Ser160, which may be the key site involved in this interaction.

Additionally the avian influenza virus was shown to hyperactivate PI3K by binding to the signalling proteins Crk and/or CrkL (Heikkinen *et al.*, 2008). This involves residues 207-212, which are surface exposed yet highly variable. The high variability of these residues clearly demonstrates that this interaction is not universal for all influenza A viruses, but restricted to avian influenza viruses as reported previously. The NS1 protein has been implicated in the inhibition of mRNA maturation via interaction with the poly(A)-binding protein II (PABPII) (Chen *et al.*, 1999) and the cleavage and polyadenylation specificity factor (CPSF30) (Twu *et al.*, 2006). Among the PABPII interacting residues reported (218-232), there is no significant conservation. Some of the residues reported to interact with the CPSF30 subunit of the polyadenylation complex are highly conserved, namely Leu144, Gly184, Glu186, Asn188 and the highly surface exposed Tpr187. Residues 123-127 were shown to interact with a dsRNA-dependent serine/threonine protein kinase R (PKR) (Min *et al.*, 2007). Those residues are surface exposed and Lys126 is conserved (grade 8), while the other residues are variable. In the recent crystal structure of the NS2 protein (Bornholdt & Prasad, 2008) the formation of long chain oligomers was observed, based on electrostatic interactions and hydrogen bonding between residues Lys131/Glu97, Thr91/Arg193, Glu196/Arg200, Glu152/Leu95, Glu96/Glu152. Among those residues we found only Glu97 and Arg200 highly conserved (grade 9), while there was moderate conservation (grade 7) of residues Lys131 and Glu196. Residues Glu152 and Leu95 were highly variable. It is

therefore less likely that the formation of oligomers is a common feature of NS1 proteins *in vivo*.

In addition to identifying residues involved in known interactions significant insight into new protein-protein interactions and new potential drug target sites can be gained from focussing on highly conserved residues, which are not known from previous studies to be involved in any interactions. Examples of those are Glu159 located on the ED domain in a β-sheet, Thr191/Val192 located in an exposed loop, Arg200 in a short α-helix, Glu208 located in an exposed turn and Gln218, the only highly conserved residue in the C-terminal α-helix (figure 3A).

In addition to surface exposed residues on the NS1 protein that may fit into binding sites of other proteins, the NS1 protein contains binding pockets surrounded by conserved residues (figure 3A). The highest ranking binding site one reported by Q-SiteFinder is surrounded by the highly conserved residues Trp102, Met104, Asp120, Arg148, Gly158 and Glu159, some of which are in the region that interact with translation initiation factor eIF4F (see above). This site could form a potential target for small molecule inhibitors of protein-protein interactions (Twu *et al.*, 2006). Binding sites two and six are located in-between highly conserved residues on the RBD such as Arg19, Arg35, and Arg46. These binding sites are spatially close with a distance of about 1 nm. This would make it possible to design bifunctional inhibitors that target both sites simultaneously and potentially interfere with dsRNA binding. Binding site five close to the conserved residues Gln109, Lys110, Gln121 and Gly184, some of which may interact with the translation initiation factor eIF4F. Overall, the sequence conservation around binding site five is not that high as for the previous binding sites. Binding site nine is located close to the moderately

conserved Phe14 but other residues surrounding this site do not show a high degree of conservation.

4.3 Sequence conservation and binding sites of the NS2 protein

The NS2 protein is known to interact with the cellular nuclear export protein Crm1 through the exposed Trp78 residue surrounded by several glutamate residues (Akarsu *et al.*, 2003). Indeed, Tpr78 is highly conserved as well as two of the four glutamate residues, namely Glu74 and Glu75. A further region of functional importance is a nuclear export signal formed by residues 11 to 23 (O'Neill *et al.*, 1998), of which Ile12, Arg15, Met16, Ser17 and Leu21 are conserved (grades 7-9). Among those residues Ser17 and Leu21 are highest conserved and surface exposed, which we propose to be the key residues of the nuclear export signal. Other highly conserved surface exposed residues that have not been assigned to any function are found throughout the protein, namely Ser24, Leu28, Arg66, Arg84, Ser93, Ile97, and Leu103 (figure 3B). Some other highly conserved residues are found in buried positions that may have a role in stabilising the structure or may become exposed upon conformational change, since the N-terminal domain has been reported as highly flexible.

A variety of binding sites was predicted between the interface of N-terminal and C-terminal helix dimer forming the four-helix bundle. Since the N-terminal part of our structure was based on a computational prediction and furthermore it was shown previously to be very flexible, the predicted binding sites may be an artefact of incomplete packing between the two helix hairpins. Further molecular dynamics simulations of the predicted structure are required to assess the structural stability of the predicted four-helix bundle and the significance of the predicted binding sites. We will therefore briefly discuss binding sites found on the outside of the structure. Binding site seven is located

closely to the conserved residues Arg84, Glu91 and Glu96. This could define a previously unknown interaction site, which may be investigated as a potential drug target site. Binding site eight is exposed at the apex of the helix bundle and close to the highly conserved residues Trp65 and Arg66. Binding site 10 is surrounded by less conserved residues and shows a very low site volume (table 2). It would not form an ideal drug target site. Contrary to the NS1 protein, the NS2 protein does not reveal binding sites on its surface that are close enough in space so they could be targeted with a bivalent compound. In conclusion it was found that both non-structural proteins showed a pattern of variable and conserved residues among all influenza A virus subtypes and hosts. Binding sites near conserved residues revealed in this work are potential targets for developing universal anti-influenza drugs, which at the same time are less likely to become ineffective due to a mutation of the virus into a drug-resistant form. Further work emerging from this study needs to characterise the function of the so-far unknown highly conserved residues as well as develop lead compounds that target some of the predicted binding sites in highly conserved regions. The significance of some binding sites on the NS2 protein should be analysed with molecular dynamics simulations taking into account the high flexibility of the full-length structure, which has so far withstood all experimental high-resolution methods.

Acknowledgments

This work was supported by the School of Life Sciences, University of Hertfordshire, UK.

References

Akarsu, H., Burmeister, W. P., Petosa, C., Petit, I., Muller, C. W., Ruigrok, R. W. H. & Baudin, F. (2003). Crystal structure of the M1 protein-binding domain of the influenza A virus nuclear export protein (NEP/NS2). *EMBO J* **22**, 4646-4655.

Aragon, T., de la Luna, S., Novoa, I., Carrasco, L., Ortin, J. & Nieto, A. (2000). Eukaryotic translation initiation factor 4GI is a cellular target for NS1 protein, a translational activator of influenza virus. *Mol Cel Biol* **20**, 6259-6268.

Baez, M., Zazra, J.J., Elliott, R.M., Young, J.F. & Palese, P. (1981). Nucleotide sequence of the influenza A/duck/Alberta/60/76 virus NS RNA: conservation of the NS1/NS2 overlapping gene structure in a divergent influenza virus RNA segment. *Virology* **113**, 397-402.

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

Basu, D., Walkiewicz, M. P., Frieman, M., Baric, R. S., Auble, D. T. & Engel, D. A. (2009). Novel influenza virus NS1 antagonists block replication and restore innate immune function. *J Virol* **83**, 1881-1891.

Bernstein, H. J. (2000). Recent changes to RasMol, recombining the variants. *TTIBS* **25**, 453-455.

Bornholdt, Z. A. & Prasad, B. V. V. (2008). X-ray structure of NS1 from a highly pathogenic H5N1 influenza virus. *Nature* **456**, 985-U985.

Chen, Z. Y., Li, Y. Z. & Krug, R. M. (1999). Influenza A virus NS1 protein targets poly(A)-binding protein II of the cellular 3 '-end processing machinery. *EMBO J* **18**, 2273-2283.

Claas, E. C. J., Osterhaus, A., van Beek, R., De Jong, J. C., Rimmelzwaan, G. F., Senne, D. A., Krauss, S., Shortridge, K. F. & Webster, R. G. (1998). Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus. *Lancet* **351**, 472-477.

Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. (2004). The Jalview Java alignment editor. *Bioinformatics* **20**, 426-427.

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

Donelan, N. R., Basler, C. F. & Garcia-Sastre, A. (2003). A recombinant influenza A virus expressing an RNA-binding-defective NS1 protein induces high levels of beta interferon and is attenuated in mice. *J Virol* **77**, 13257-13266.

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

Ehrhardt, C. & Ludwig, S. (2009). *Cell Microbiol*, available online, doi: 10.1111/j.1462-5822.2009.01309.x

Ehrhardt, C., Marjuki, H., Wolff, T., Nurnberg, B., Planz, O., Pleschka, S. & Ludwig, S. (2006). Bivalent role of the phosphatidylinositol-3-kinase (PI3K) during influenza virus infection and host cell defence. *Cell Microbiol* **8**, 1336-1348.

Ehrhardt, C., Wolff, T., Pleschka, S., Planz, O., Beermann, W., Bode, J.G., Schmolke, M., Ludwig, S. (2007). Influenza A virus NS1 protein activates the PI3K/Akt pathway to mediate antiapoptotic signaling responses. *J Virol* **81**, 3058-3067.

Enami, K., Sato, T. A., Nakada, S. & Enami, M. (1994). Influenza-virus NS1 protein stimulates translation of the M1 protein. *J Virol* **68**, 1432-1437.

Ferraris, O. & Lina, B. (2008). Mutations of neuraminidase implicated in neuraminidase inhibitors resistance. *J Clin Virol* **41**, 13-19.

Fortes, P., Beloso, A. & Ortin, J. (1994). Influenza virus NS1 protein inhibits pre-messenger RNA splicing and blocks messenger RNA nucleocytoplasmic transport. *EMBO J* **13**, 704-712.

Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E. & Ben-Tal, N. (2003). ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. *Bioinformatics* **19**, 163-164.

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

Greenspan, D., Krystal, M., Nakada, S., Arnheiter, H., Lyles, D. S. & Palese, P. (1985). Expression of influenz virus NS2 nonstructural proteins in bacterial and localization of NS2 in infected eukaryotic cells. *J Virol* **54**, 833-843.

Greenspan, D., Palese, P. & Krystal, M. (1988). Two nuclear localisation signals in the influenza virus NS1 nonstructural protein. *J Virol* **62**, 3020-3026.

Hale, B.G., Jackson, D., Chen, Y.H., Lamb, R.A. & Randall, R.E. (2006). Influenza A virus NS1 protein binds p85beta and activates phosphatidylinositol-3-kinase signaling. *Proc Natl Acad Sci* **103**, 14194-14199.

Hale, B. G., Batty, I. H., Downes, C. P. & Randall, R. E. (2008a). Binding of influenza A virus NS1 protein to the inter-SH2 domain of p85 beta suggests a novel mechanism for phosphoinositide 3-kinase activation. *J Biol Chem* **283**, 1372-1380.

Hale, B. G., Randall, R. E., Ortin, J. & Jackson, D. (2008b). The multifunctional NS1 protein of influenza A viruses. *J Gen Virol* **89**, 2359-2376.

Heikkinen, L. S., Kazlauskas, A., Melen, K., Wagner, R., Ziegler, T., Julkunen, I. & Saksela, K. (2008). Avian and 1918 Spanish influenza a virus NS1 proteins bind to Crk/CrkL src homology 3 domains to activate host cell signaling. *J Biol Chem* **283**, 5719-5727.

Horimoto, T. & Kawaoka, Y. (2005). Influenza: Lessons from past pandemics, warnings from current incidents. *Nature Rev Microbiol* **3**, 591-600.

Johnson, N. & Mueller, J. (2002). Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic. *Bull Hist Med* **76**, 105-115.

Kobasa, D. & Kawaoka, Y. (2005). Emerging influenza viruses: Past and present. *Curr Mol Med* **5**, 791-803.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T. & Ben-Tal, N. (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucl Acids Res* **33**, W299-W302.

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

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., & Hay, A. (2000). Avian-to-human transmission of H9N2 subtype influenza A viruses: Relationship between H9N2 and H5N1 human isolates. *Proc Natl Acad Sci* **97**, 9654–9658.

Lommer, B. S. & Luo, M. (2002). Structural plasticity in influenza virus protein NS2 (NEP). *J Biol Chem* **277**, 7108-7117.

Lu, Y., Wambach, M., Katze, M. G. & Krug, R. M. (1995). Binding of the influenza-virus NS1 protein to double-stranded RNA inhibits the activation of the protein kinase that phosphorylates the ELF-2 translation initiation factor. *Virology* **214**, 222-228.

Molinari, N.M., Ortega-Sanchez, I.R., Messonnier, M.L., Thompson, W.W., Wortley, P.M., Weintraub, E., et al. (2007). The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine* **25**, 5086-5096.

Marión, R.M., Aragón, T., Beloso A., Nieto, A., Ortín, J. (1997). The N-terminal half of the influenza virus NS1 protein is sufficient for nuclear retention of mRNA and enhancement of viral mRNA translation. *Nucleic Acids Res* **25**, 4271-4277.

Maroto, M., Fernandez, Y., Ortin, J., Pelaez, F. & Cabello, M. A. (2008). Development of an HTS assay for the search of anti-influenza agents targeting the interaction of viral RNA with the NS1 protein. *J Biomol Screen* **13,** 581–590.

Min, J. Y., Li, S. D., Sen, G. C. & Krug, R. M. (2007). A site on the influenza A virus NS1 protein mediates both inhibition of PKR activation and temporal regulation of viral RNA synthesis. *Virology* **363**, 236-243.

Nemeroff, M. E., Barabino, S. M. L., Li, Y. Z., Keller, W. & Krug, R. M. (1998). Influenza virus NS1 protein interacts with the cellular 30 kDa subunit of CPSF and inhibits 3 ' end formation of cellular pre-mRNAs. *Molecular Cell* **1**, 991-1000.

Nemeroff, M. E., Qian, X. Y. & Krug, R. M. (1995). The influenza virus NS1 protein forms multimers in-vitro and in-vivo. *Virology* **212**, 422-428.

Neumann, G., Hughes, M. T. & Kawaoka, Y. (2000). Influenza A virus NS2 protein mediates vRNP nuclear export through NES-independent interaction with hCRM1. *EMBO J* **19**, 6751-6758.

Obenauer, J.C., Denson, J., Mehta, P.K., Su, X., Mukatira, S., Finkelstein, D.B., Xu, X., Wang, J., Ma, J., Fan, Y., Rakestraw, K.M., Webster, R.G., Hoffmann, E., Krauss, S., Zheng, J., Zhang, Z. & Naeve, C.W. (2006). Large-scale sequence analysis of avian influenza isolates, *Science* **311**, 1576–1580.

O'Neill, R. E., Talon, J. & Palese, P. (1998). The influenza virus NEP (NS2 protein) mediates the nuclear export of viral ribonucleoproteins. *EMBO J* **17**, 288-296.

Opitz, B., Rejaibi, A., Dauber, B., Eckhard, J., Vinzing, M., Schmeck, B., Hippenstiel, S., Suttorp, N. & Wolff, T. (2007). IFN beta induction by influenza A virus is mediated by RIG-I which is regulated by the viral NS1 protein. *Cell Microbiol* **9**, 930-938.

Pichlmair, A., Schulz, O., Tan, C. P., Naslund, T. I., Liljestrom, P., Weber, F. & Sousa, C. R. E. (2006). RIG-I-mediated antiviral responses to single-stranded RNA bearing 5 '-phosphates. *Science* **314**, 997-1001.

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. & Belongia, E. A. (2008). Adamantane-resistant influenza infection during the 2004-05 season. *Emerg Infect Dis* **14**, 173-176.

Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: The European molecular biology open software suite. *Trends in Genet* **16**, 276-277.

Richardson, J.C., Akkina, R.K. (1991). NS2 protein of influenza virus is found in purified virus and phosphorylated in infected cells. *Arch Virol* **116**, 69-80.

Sayle, R. A. & Milnerwhite, E. J. (1995). RASMOL – biomolecular graphics for all. *Trends Biochem Sci* **20**, 374-376.

Schmitt, A.P., Lamb, R.A. (2005). Influenza Virus Assembly and Budding at the Viral Budozone. *Adv Virus Res* **64**, 383-416.

Schueler-Furman, O. & Baker, D. (2003). Conserved residue clustering and protein structure prediction. *Protein Struct Funct and Genet* **52**, 225-235.

Shin, Y.K., Liu, Q., Tikoo, S.K., Babiuk, L.A., & Zhou, Y. (2007). SH3 binding motif 1 in influenza A virus NS1 protein is essential for PI3K/Akt signalling pathway activation. *J Virol* **81**, 12730-12739.

Subbarao, K., Klimov, A., Katz, J., Regnery, H., Lim, W., Hall, H., Perdue, M., Swayne, D., Bender, C., Huang, J., Hemphill, M., Rowe, T., Shaw, M., Xu, X. Y., Fukuda, K. & Cox, N. (1998). Characterization of an avian influenza A (H5N1) virus isolated from a child with a fatal respiratory illness. *Science* **279**, 393-396.

Twu, K. Y., Noah, D. L., Rao, P., Kuo, R. L. & Krug, R. M. (2006). The CPSF30 binding site on the NS1A protein of influenza A virus is a potential antiviral target. *J Virol* **80**, 3957-3965.

Wang, W., Riedel, K., Lynch, P., Chien, C. Y., Montelione, G. T. & Krug, R. M. (1999). RNA binding by the novel helical domain of the influenza virus NS1 protein requires its dimer structure and a small number of specific basic amino acids. *RNA* **5**, 195-205.

Wang, X. Y., Li, M., Zheng, H. Y., Muster, T., Palese, P., Beg, A. A. & Garcia-Sastre, A. (2000). Influenza A virus NS1 protein prevents activation of NF-kappa B and induction of alpha/beta interferon. *J Virol* **74**, 11566-11573.

Ward, A. C., Castelli, L. A., Lucantoni, A. C., White, J. F., Azad, A. A. & Macreadie, I. G. (1995). Expression and analysis of the NS2 protein of influenza A virus. *Arch Virol* **140**, 2067-73.

World Health Organisation (2004). Avian influenza A(H5N1). *Weekly Epidemiol Rec* **79**, 65-70.

Yin, C. F., Khan, J. A., Swapna, G. V. T., Ertekin, A., Krug, R. M., Tong, L. & Montelione, G. T. (2007). Conserved surface features form the double-stranded RNA binding siteof non-structural protein 1 (NS1) from influenza A and B viruses. *J Biol Chem* **282**, 20584-20592.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinf* **9**, 40.

Zhang, Y., Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-710.

Table 1: Conserved residues and variable residues of the NS1 and NS2 proteins identified using ConSurf server (Glaser *et al.*, 2003; Landau *et al.*, 2005).

Residues	NS1 protein	NS2 protein
Conserved*	Met1, Asp2, THR5, Ser8-His17, Arg19, Lys20, Ala23, Asp24, Gly28-Phe32, Asp34, Arg35, Arg37-Asp39, Gln40, Ser42, Leu43, Gly45, Arg46, Thr49, Leu50, Ile54, Ala57, Thr58, Gly61, Ile64, Leu69, Glu72, Tyr89**, Asp92, Met93, Glu97-Arg100, Trp102, Met104, Leu105, Pro107, Gln109, Lys110, Leu115, Ile117, Asp120-Ala122, Lys126, Ile128, Leu130-Val136, Leu141, Leu144, Leu146-Thr151, Gly154, Val157-Pro162, Ser165, Gly168, His169, Glu172-Lys175, Ala177, Ile178, Leu181, Ile182, Gly184, Glu186-Asn188, Asn190-Val192, Ser195, Glu196, Gln199-Phe201, Trp203, Glu208, Gln218	Met1, Thr5, Ser8-Gln10, Ile12, Arg15-Ser17, Lys18, Gln20, Leu21, Ser24, Ser25, Leu28, Gly30, Thr33, Ser37, Leu38, Tyr41-Asp43, Leu45, Gly46, Met50, Arg51, Gly53, Asp54, Gln59, Arg61, Asn62, Trp65 Arg66, Leu69, Lys72-Glu75, Arg77-Ile80, Gly82, Arg84, Leu87, Thr90, Glu91, Ser93-Phe99, Gln101-Leu106, GLu108-Glu110, Glu112, Arg114, Ser117, Gln119, Ile121
Variable*	Ser3, Val6, Phe22, Glu26, Leu27, Lys44, Gly47, Asn48, Asp53, Glu55, Arg59, Ala60, Gln63, Arg67, Glu70, Glu71, Asp74-Leu77, Lys79-Arg83**, Leu90, Leu95, Asp101, Phe103, Val111, Ala112, Cys116, Met124, Asp125, Thr127, Ile129, Ile137-Gly139, Glu152, Glu153, Ala155, Gly171, Val180, Leu185, Thr197, Arg204-Gly206, Asp209-Asn217, Arg220, Lys221, Arg224, Ile226, Glu227, Val230	Ser3, Val6, Val14, Ala22, Ser23, Glu26, Asp27, Met31, Gln34, Gly36, Lys39, Leu40, Asp47-Val49, Met52, Phe55, Phe58, Ile60, Gly63, Lys64, Glu67, Ser70, Val83, His85, Arg86, Lys88, Ile89, Leu107, Phe116, Phe118

* Conserved residues, classified by the ConSurf server with grades 7-9 and variable residues classified by the grades 1-3 are grouped together.
** Residue numbers for the effector domain are numbered conventionally, counting the five residue deletion.

Table 2: Predicted ligand binding sites of the NS1 and NS2 proteins.

Site*	Residues in the ligand binding sites of the NS1 protein	Site volume ($Å^3$)	Site*	Residues in the ligand binding sites of the NS2 protein	Site volume ($Å^3$)
1	Trp102- Met104, Lys118-Asp120, Ile123, Arg148, Ile156-Glu159**	165	1	Met1, Ser8, Phe9, Ile12, Leu38, Leu40, Tyr41, Ser44, Leu45, Glu47, Ala48, Glu75, Ile76, Leu79, Ile80, Val83, Arg86, Glu95, Thr98, Phe99, Gln101, Ala102, Leu103, Leu105, Leu106, Val109	466
2	Asp12, Leu15, Trp16, Arg19, Pro31-Leu33, Arg35, Leu36, Asp39	229	2	Ile12, Met16, Thr33, Gln34, Ser37, Leu38, Tyr41, Lys72, Ile76, Val109, Glu112, Ile113	192
5	Met104, Lys108, Gln109, Lys109, Ile117, Lys118, Met119, Asp120, Gln121, Val180, Gly184**	188	7	Arg84, Leu87, Lys88, Glu91, Gln96, Met100	119
6	Thr5, Val6, Ser8, Phe9, Trp16, Arg38, Asp39, Ser42, Leu43, Arg46, Leu50	198	8	Lys64, Trp65, Arg66, Glu67	108
9	Phe14, Leu15, His17, Val18, Gly61, Lys62, Val65, Glu66, Leu69	146	10	Leu69-Gln71, Phe73, Glu74, Glu110, Ile113, Arg114, Ser117	92

*The sites are numbered according to the rank obtained from Q-SiteFinder. Only those sites are reported here, which are close to conserved residues identified by ConSurf.
** Residue numbers for the effector domain are numbered conventionally, counting the five residue deletion.

Figure legends

Figure 1: Overview of the multiple sequence alignment for the NS1 protein (A) and the NS2 protein (B). Darker shades in the alignment show a higher degree of sequence identity, while the line graph shows the degree of conservation within a window of ten residues. This figure was made with Jalview (Clamp *et al.*, 2004) and plotcon of the EMBOSS package (Rice *et al.*, 2000).

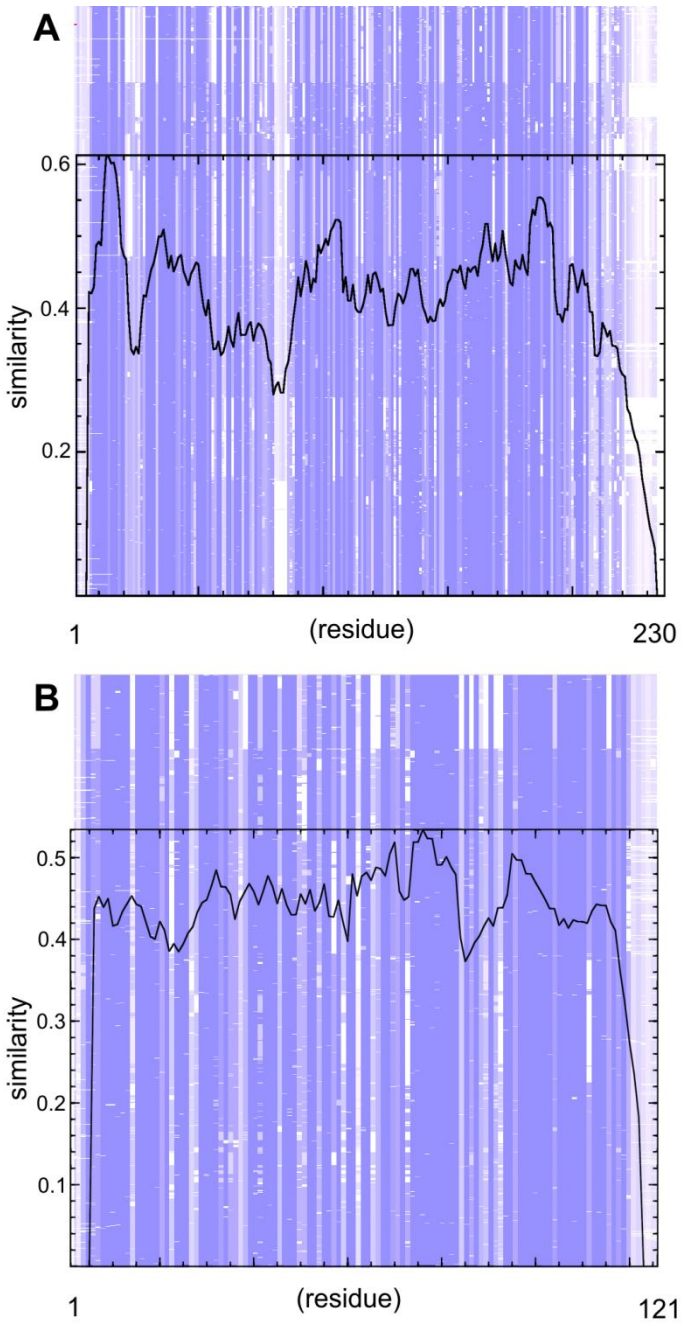

Figure 2: Backbone ribbon models of the NS1 protein (A) and NS2 protein (B). The ribbon was coloured with conservation scores obtained from ConSurf. The N- and C-termini are indicated. The colour codes are explained in figure 3. The inset shows in green the parts of the structures that are known from experiments, while the predicted parts are shown in grey. This figure was made with RasMol (Sayle & Milnerwhite, 1995; Bernstein, 2000).

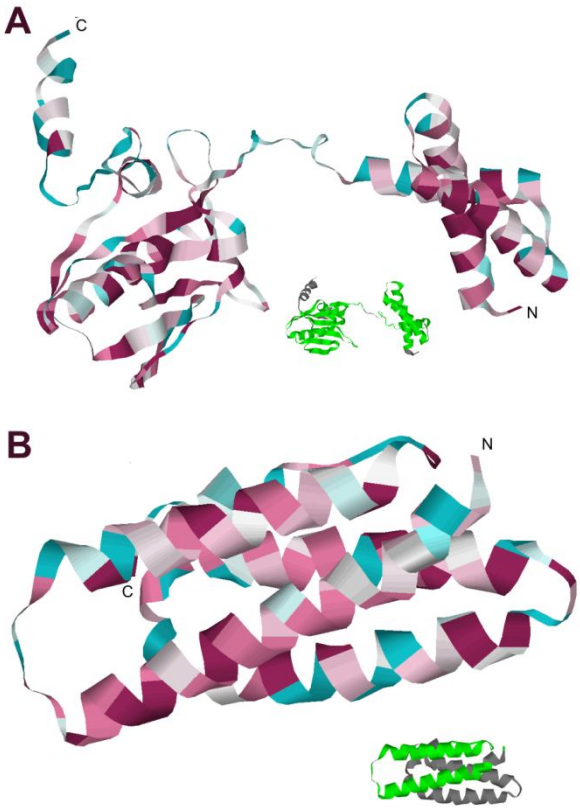

Figure 3: Conservation score and predicted binding sites shown in a spacefill representation of the NS1 protein (A) and NS2 protein (B). Each protein is shown in two orientations (180° rotated). The binding sites are shown in ball-stick format coloured green. This figure was made with RasMol [32, 33]