Evolutionary conservation of influenza A PB2 sequences reveals potential target sites for small molecule inhibitors.

Patel, H. and Kukol, A

https://dx.doi.org/10.1016/j.virol.2017.06.009

Virology – Regular manuscript

Title: Evolutionary conservation of influenza A PB2 sequences reveals potential target sites for small molecule inhibitors.

Hershna Patel[a] & Andreas Kukol#[a]

[a] School of Life and Medical Sciences, University of Hertfordshire, Hatfield, AL10 9AB, United Kingdom

#Corresponding author: Dr Andreas Kukol (a.kukol@herts.ac.uk, 01707 284 543)

First author: Hershna Patel (h.patel28@herts.ac.uk)

Abstract

The influenza A basic polymerase protein 2 (PB2) functions as part of a heterotrimer to replicate the viral RNA genome. To investigate novel PB2 antiviral target sites, this work identified evolutionary conserved regions across the PB2 protein sequence amongst all sub-types and hosts, as well as ligand binding hot spots which overlap with highly conserved areas. Fifteen binding sites were predicted in different PB2 domains; some of which reside in areas of unknown function. Virtual screening of ~50,000 drug-like compounds showed binding affinities of up to -10.3 kcal/mol. The highest affinity molecules were found to interact with conserved residues including Gln138, Gly222, Ile529, Asn540 and Thr530. A library containing 1738 FDA approved drugs were screened additionally and revealed Paliperidone as a top hit with a binding affinity of -10 kcal/mol. Predicted ligands are ideal leads for new antivirals as they were targeted to evolutionary conserved binding sites.

Keywords: Influenza A, PB2, sequence evolution, binding site, virtual screening, drug discovery, conservation, Paliperidone

Introduction

The influenza A virus causes one of the most prevalent and significant respiratory viral infections worldwide with previous pandemics resulting in remarkably high fatality rates (Taubenberger and Kash, 2010). This is largely due to continuous genome evolution and the zoonotic nature of the virus which enables rapid transmission of new re-assortant strains (Reperant et al., 2012). The main form of prevention against influenza is annual vaccination; however this may not always guarantee extensive protection or control of the virus (Chambers et al., 2015). Therefore treatment with antiviral drugs such as the neuraminidase inhibitors is heavily relied on, while there is widespread resistance against matrix protein2 (M2) inhibitors. Since the adoption of these drugs, influenza A subtypes in circulation have shown varying levels of sensitivity due to amino acid mutations in the drug target site (Hayden and De Jong, 2011; Samson et al., 2013). For this reason the M2 inhibitors are no longer recommended for clinical use (Harper et al., 2009). Consequently, the discovery and search for novel antivirals which are unlikely to be affected by resistance mutations is a priority, with several candidate inhibitors having emerged from recent studies (reviewed in Naesens et al., (2016); Patel and Kukol, (2016)).

The infectious life cycle of the virus requires several functional proteins encoded by eight RNA segments, which are released into the host cell, allowing the virus to replicate its genome and suppress the immune response (Bouvier and Palese, 2008). The influenza A polymerase basic protein 2 (PB2) is encoded by RNA segment one. It is one of the largest influenza proteins consisting of 759 amino acids and is a constituent subunit of the trimeric viral polymerase complex in addition to polymerase basic protein 1 (PB1) and the acidic polymerase (PA). Transcription and replication of the viral genome occurs in the host cell nucleus, and involves a series of stages before translation of viral mRNA in the cytoplasm (Fodor, 2013). During transcription, the PB2 protein is mainly responsible for generating the cap structure for viral mRNA from the 5' end of 7-methyl guanosine triphosphate (mGTP) capped host mRNA. The PB2 'cap snatching' mechanism involves residues between positions 318-482, which recognise methylated guanosine in order to bind the host cell RNA strand. The endonuclease subunit of the PA then cleaves the RNA leaving a 10-13 nucleotide primer to initiate transcription by PB1 (Fodor, 2013). In complex, the N-terminal 249 residues of the PB2 subunit are associated with the C-terminal subunit of PB1, which is a critical interaction to trigger PA endonuclease activity (Sugiyama et al., 2009). A structural study of the PB2 protein from a H5N1 avian virus had found that following translation, the C-terminal domain (residues 536-759) undergoes large conformational re-organisation between open and closed states. This flexibility enables the nuclear localization signal (NLS) peptide in the 686-759 region to bind with host importin-α, enabling PB2 entry into the target cell nucleus to catalyze further RNA transcription (Das et al., 2010; Delaforge et al., 2015). Other PB2 conformational changes occurring in connection with the cap-snatching mechanism and the kind of RNA bound have also been described in the context of the full polymerase complex (Reich et al., 2014; Thierry et al., 2016).

In addition to mutations in the haemagglutinin (HA) and neuraminidase (NA) proteins, changes in the sequences of polymerase proteins are also considered as major determinants of host range and adaptation (Mehle and Doudna, 2009; Neumann and Kawaoka, 2015). The characteristic PB2 host determining residue at position 627 (with lysine being prevalent in human strains, glutamic acid present in avian strains and serine in bat strains) is situated in a loop region, which along with the cap-binding domain does not make extensive contact with the PB1 and PA subunits (Kuzuhara et al., 2009; Pflug et al., 2014). The 535-684 domain has also been shown to have RNA binding activity which is affected by the E627K mutation and is unrelated to the cap-snatching function (Kuzuhara et al., 2009).

As the PB2 protein plays multiple essential roles in the virus life cycle, it is a valid target for antiviral drugs. Several crystal structures are available in the protein data bank (PDB) for specific PB2 subunit domains in holo and apo forms which can aid with structure based drug discovery studies. Despite some of the PB2 surface area being inaccessible due to trimer assembly, inhibitors of the 'cap snatching' function to prevent capped host mRNA binding have been identified showing potent effects against several influenza strains *in vitro* (Boyd et al., 2015; Clark et al., 2014; Pautus et al., 2013). The aim of this work was to identify highly conserved regions of PB2 using an algorithm superior to counting conservation from multiple sequence alignments and to identify overlap with potential ligand binding sites. Structure-based virtual screening was used to predict small drug-like molecules that may bind to those sites. These findings could help in studies aimed at obtaining novel insight into PB2 functions as well as provide a starting point for further *in vitro* investigations of replication inhibitors that are effective in a variety of hosts and lack the potential of inducing influenza antiviral resistance.

Results and discussion

PB2 sequence conservation

12,459 PB2 sequences were obtained from the NCBI influenza virus resource database. This included 31% from human, 16% from swine and 50% from avian hosts. 702 sequences remained after removing redundant sequences at 98.5% identity, indicating that a large proportion of PB2 sequences deposited are highly similar which would reflect bias upon conservation scoring (Valdar, 2002). The conservation scores calculated from the multiple sequence alignment of the non-redundant sequences shows that there is a high level of amino acid conservation throughout the entire protein sequence. The scores ranged from 0.789 (lowest) to 1.0 (highest) and the majority of amino acids had a score between 0.95 and 1.0 as shown in Fig. 1.

For display purposes the conservation scores were re-scaled and mapped on to the PB2 structure (Fig. 2). Overall, the key functional regions of PB2 were found to be highly conserved and consisted of several residues scoring 0.95 or above. This includes the N-terminal residues 1-37 which form three short α-helices comprising the PB1 binding interface required for effective polymerase activity (Sugiyama et al., 2009). The mGTP cap binding

domain (318-482) is also well conserved, albeit with moderately conserved residues at position 339, 340, 453 and 456. Substitution of Lys339 to Thr339 of certain subtypes has been found to prevent binding of the phosphate group of mGTP capped mRNA, reducing RNA synthesis, and thereby regulating PB2 activity (Liu et al., 2013). Val414, Arg415 and Gly416 are highly conserved and are required for PB2-acetyl-CoA interaction to maintain transcription activity (Hatakeyama et al., 2014). The 424-loop region is suggested to have an allosteric role in regulating PB1 activity, whilst other conserved residues are expected to contribute to the domains structurally distinct fold which allows formation of intermolecular contacts specific for mGTP cap binding activity (Guilligay et al., 2008). The 1-269 and 580-683 segments which are reported to be capable of binding the nucleoprotein (NP) (Poole et al., 2004), also consist of long stretches of conserved residues such as Ser592-Thr612. A total of 42 amino acids were found to be 100% conserved and could therefore be the most resistant to change due to evolutionary adaption of the virus. This includes Leu744 located on a surface exposed loop region and Gly693, which we suggest to be key residues in the NLS region due to their high conservation, enabling PB2 nuclear entry from the cytoplasm via binding importin-α. Other highly conserved regions with unassigned functions identified in this work may be of interest with regards to antiviral drug discovery.

An intermediate level of conservation for the host specific residue at position 627 was reflected in the alignment with a conservation score of 0.885. Due to the majority of sequences being from avian hosts, glutamic acid was the prevalent residue based on the consensus sequence. Whilst a range of amino acid residues can be tolerated at the 627 position shown by mutagenesis (Chin et al., 2014), the E627K mutation is well known for determining virulence by increasing polymerase activity and replication in mammals. This prime example of host adaptation is thought to be due to glutamic acid being able to bind the avian version of the host cell factor ANP32A; whereas substitution to lysine allows the polymerase to bind to the mammalian version of this host factor (Long et al., 2016; Moncorgé et al., 2010). . However, some avian viruses carrying the E627 variant can efficiently replicate in mammalian cells due to compensatory mutations found in the PB1 protein of H5N1 strains (Xu et al., 2012). A mutation study of the 627 domain has also identified specific conserved residues to be essential for general PB2 activity (Arg597, Pro620, Phe621, Arg646 and Arg650), as well as non-essential residues such as Pro625, Pro626 and Gln628 which are also highly conserved (Kirui et al., 2014). Furthermore, the positive charge of the highly conserved Arg630 (in the presence of NP R150), or Lys627

promotes PB2-NP interaction, which is essential for the ribonucleoprotein complex to provide structural maintenance and regulate viral transcription (Labadie et al., 2007; Ng et al., 2012).

Low (scores below 0.85) or moderate conservation was identified mainly at single amino acid positions such as 64, 107, 147, 271, 292, 453, 483, 559, 588, 590, 591, 613, 661 and 676. The lowest conservation score was 0.789 at position 147. The residues at these positions are all located on the exterior surface of the protein (Fig. 2(b)), which is consistent with the finding that surface residues evolve faster than those in the protein core (Warren et al., 2013). Adaptive mutations to Ala271, Arg591, and Ser590 have been found to enhance polymerase activity and virus replication in mammals (Bussey et al., 2010; Mehle and Doudna, 2009; Yamada et al., 2010). The remaining non-conserved positions with uncharacterised mutation effects may also be associated with determining host range, virulence, PB2 cellular localization, or with no particular function. Additionally, a conservation study on the influenza PA protein has shown that residues classed as non-conserved may indeed be biologically important, and that functional residues are not always conserved (Wu et al., 2015), which could also to be the case for PB2. The residues neighbouring less conserved positions were generally found to be highly conserved, as well as 16% of residues located in the interior of the protein. Their restricted variability is presumably essential for maintaining the protein structure, in particular the subdomains.

The protein sequence dataset analysed contains two sequences isolated from bats (including the H17N10 strain for which a crystal structure has been resolved, PDB ID: 4WSB), which are noticeably different to the consensus sequence. Influenza protein sequences isolated from bats have shown less similarity overall to sequences from other hosts (Tong et al., 2013). Despite these differences , the H17N10 sequence for PB2 remains evolutionary close to human and avian strains (Pflug et al., 2014) and is therefore unlikely to result in major structural differences.

Protein structure modeling

The PB2 sequence of a human H5N1 isolate, for which an N-terminal structural fragment (PDB ID: 3L56) exists, was used to construct a full length structural model using the I-TASSER modelling server (Yang and Zhang, 2015; Zhang, 2008). This sequence was modelled as it is from a virus isolated from a human host, while the existing full length PB2 structure is from an H17N10 bat virus. The model had a template modelling (TM) score of

0.92, and was largely built on this H17N10 template (PDB ID: 4WSB, chain C). Model co-ordinates were replaced with the 3L56 N-terminal fragment and then refined by energy minimisation. The 3L56 fragment is structurally very similar to that of H17N10 with a backbone RMSD of 1.05 Å, justifying the approach of using the bat H17N10 structure as a template for modelling the structure of full-length H5N1 PB2 sequence. The alignment of the two sequences covering parts of the polypeptide chain in the target site for virtual screening is shown in Fig. 3. The overall percentage identity between the two full length sequences is 68% calculated using the Clustal2.1 percent identity matrix, implying that the H17N10 sequence is an appropriate template..

Predicted ligand binding hot spots

Fifteen ligand binding hot spots which represent favourable binding regions with small organic molecules were predicted using the FTMap web server in different domains of the protein; many of which were located in highly conserved areas. The locations of the top ten binding hot spots are shown in Fig. 4. The most surface accessible binding regions are spots three, four, five, nine, six, seven, eight, fourteen and fifteen. Hot spots two, eleven, twelve and ten appear to be partially buried when viewing the structure in a spacefill representation, with spots one and thirteen being the least exposed to the outer surface on the protein. Although, based on previous structural information reported and the flexibility of viral polymerase subunits in general (Reich et al., 2014; Thierry et al., 2016), the surface accessibility of some of these binding hot spots may change upon trimer formation or subdomain rotation. Structural alignment of our H5N1 PB2 structure with 4WSB (chain C) suggests that the accessibility of these sites would be unaffected as they are not directly blocked by PA/PB1 in complex. Whereas alternative configurations of the heterotrimer (reviewed by Pflug, Lukarska, Resa-Infante, Reich, & Cusack, (2017)) have shown that depending on the RNA promoter bound, the PB2 cap-binding, 627 and NLS domains may exist in several states in influenza B and C polymerases. This suggests that accessibility of all hot spots (except for six, nine, ten and thirteen which are not located within these domains) could change.

Spots seven, fourteen and fifteen are clustered closely together forming the conserved mGTP cap binding site; and considering the functional importance of this region, the low ranking assigned is probably due to the affinity for the highly charged RNA molecules, which are not well represented by the library of organic solvents used in the docking with the FTmap

algorithm. However, these spots are near the binding site for the inhibitors identified by Clark *et al.*, (2014) (methylguanine derivatives) and Pautus *et al.*, (2013) and consist of residues involved in hydrogen bonding. The highest number of different probes were found to bind at hot spot one, which indicates that this area has good binding potential with a variety of functional groups. Amino acids surrounding spots one, seven, fourteen and fifteen are not involved in heterotrimer formation and are located at positions set apart from the PA/PB1 subunit interactions. Spot six is the only one located within the N-terminal third and could be implicated in trimer association. Spot three is close (within ~6.0Å) to the intermediately conserved residue Val613. The region encompassing hot spot twowas selected as the target site for docking, as the residues closely surrounding this hot spot display high conservation so are less likely to mutate. Also this hot spot is second ranked as several different probe types were predicted to bind there, suggesting it is an important site of the protein (Brenke et al., 2009), and this site has not previously been targeted by virtual screening experiments Furthermore, there are currently no identified inhibitors which target this hot spot. In relation to PB2 structure and function, the residues surrounding this spot may be associated with rotation of the C-terminal domain or contribute towards interactions with PB1.

Virtual screening - benchmarking

A virtual screening benchmark was performed against the mGTP capped RNA binding site. The ability of the docking software to identify five known inhibitors among the top 10 predictions out of 180 compounds with similar molecular weight was tested. Results showed that AutoDock Vina alone and a combination with AutoDock4 were the best method to retrieve active PB2 inhibitor compounds at the top positions of the rank list as both methods were able to identify one inhibitor (Table 1). For simplicity, the single AutoDock Vina software was used for the PB2 target site screening. The binding affinities of test compounds using Autodock Vina ranged from -7.4 kcal/mol to -4.5 kcal/mol and from -7.4 kcal/mol to -2.9 kcal/mol with AutoDock4.

Table 1. Results of benchmarking three docking software for virtual screening.

Docking software	Number of true ligands found	binding affinity of ligand (kcal/mol)
AutoDock Vina	1	-7.4
AutoDock 4	0	
AutoDock Vina + AutoDock 4	1	-7.4 + -6.0

Virtual screening – PB2 target site

The binding affinities of 46,926 compounds from the NCI library screened against the target site surrounding hot spot two (Fig. 4) ranged from -10.3 kcal/mol to +13.7 kcal/mol. A large proportion of compounds were predicted to bind between the range of -5.0 kcal/mol and -7.0 kcal/mol. Pan Assay Interference compounds (PAINS), which appear as frequent hitters in bioassays and are considered as problematic screening compounds (Baell and Holloway, 2010), were removed from the rank list to increase the reliability of the predictions. The top 75 hits (supplementary material, Table S1) had binding affinities below -9.0 kcal/mol, and were all aromatic compounds. Some of the key amino acid residues found to interact with the top ten compounds via hydrogen bonding and hydrophobic interactions include: Gln138, Gly222, Ile529, Ile539, Asn540, Gly541, Tyr531 and Thr530; all of which are highly conserved. The predicted binding conformations show that some of these compounds bind partially inside a deep pocket formed by these residues (Fig. 5). Compound 1 (ZINC01617371) forms hydrophobic contacts with eighteen residues and a single hydrogen bond with Ile529 at a distance of 3.12Å. The compound bends around the 531-541 loop region causing the phenyl ring and nitrile group to be entirely buried within the protein; the methyl group at the other end is surface exposed. Also three aromatic groups of compound 4 (ZINC03954617) form hydrogen bonds with Gly222, Gln241 and Ile529 and are surrounded by eleven residues forming hydrophobic contacts. The top ten compounds share the common scaffold of an aromatic group at one or both ends (Fig. 6), occupying the binding pocket in a similar orientation as compound 2 (Fig. 5) supported by van der Waals and electrostatic interactions between atoms. Binding of compounds may have biological or functional significance with regard to host protein interactions with PB2, or interfere with trimer assembly by restricting

or inducing conformation changes, and synthesis of RNA (Thierry et al., 2016); the ligands predicted in the current study could serve as tools to investigate such functions.

The Approved DrugBank library, containing 1738 FDA-approved small molecule drugs, was also screened against the same target site in order to find any approved drugs that may also target the PB2 protein. The binding affinities ranged from -10.0 kcal/mol to +53.8 kcal/mol with the largest proportion of compounds predicted to bind between the range of -5.0 kcal/mol and -7.0 kcal/mol. None of the approved drugs had a significantly stronger predicted binding affinity than the top ranked compound of the NCI library; however, the highest ranked drug paliperidone (ZINC04214700) had the same binding affinity as three compounds of the NCI library ranked within the top ten positions. The chemical structure and docking models of a top hit compound from both libraries are shown in Fig. 5. The nitrogen atom of the central pyridine ring of paliperdone is able to form a hydrogen bond with the oxygen atom of Glu241, and the drug-protein complex is maintained via hydrophobic contacts with sixteen surrounding residues of the target site. Paliperidone binds to the dopamine and serotonin receptors, although the exact mechanism of action is not known (reviewed in Corena-McLeod, (2015)). Paliperidone is approved by the FDA for the treatment of schizophrenia and related disorders. The results of this study may be useful for repurposing this drug or derivatives as a treatment for influenza infection. The chemical properties of compounds identified as top hits from the screening of both libraries are listed in Table 2 where numbers in brackets refer to the chemical structures shown in Fig. 6. These compounds may have a tendency to bind and block other viral proteins that display similar structure and properties to the PB2 target site.

Table 2. Chemical properties and binding affinity (ΔG) of predicted top hit compounds identified from virtual screening of the NCI and DrugBank library obtained from the ZINC database. Properties include molecular mass (Mol M), predicted partition coefficient (xLogP), no. of hydrogen bond donors and acceptors, hydrophobic sites and total polar surface area (tPSA) at pH7.

Compound (ZINC ID)	ΔG (Kcal/mol)	Mol M (g/mol)	xLogP	H-bond donors	H-bond acceptors	Hydrophobic sites	tPSA ($Å^2$)
NCI library							
ZINC01617371 (1)	-10.3	390.42	2.69	1	7	4	115
ZINC05543024 (2)	-10.2	291.31	2.46	2	3	3	74
ZINC01612458 (3)	-10.1	328.76	4.26	2	6	5	80
ZINC03954617 (4)	-10.1	288.31	1.31	2	6	4	83
ZINC01040450 (5)	-10.0	354.32	3.46	2	9	2	103
ZINC08651894 (6)	-10.0	446.93	3.18	2	9	6	131
ZINC13212434 (7)	-10.0	359.84	2.51	4	6	5	82
ZINC01624487 (8)	-9.9	369.83	3.80	1	5	5	82
ZINC01612446 (9)	-9.8	404.73	3.66	2	12	4	171
ZINC01614027 (10)	-9.8	318.40	4.27	1	3	4	41
DrugBank library							
ZINC04214700 (11)	-10.0	427.50	1.97	2	7	6	86
ZINC01481956 (12)	-9.9	427.50	1.97	2	7	6	85
ZINC00538312 (13)	-9.4	411.50	2.96	1	6	6	65
ZINC13540266 (14)	-9.4	397.39	4.08	2	9	3	147
ZINC18456289 (15)	-9.4	439.39	-2.37	5	13	3	219
ZINC05844788 (16)	-9.0	406.45	3.10	4	5	4	75
ZINC01851132 (17)	-8.9	425.40	-1.53	5	11	3	197
ZINC01548097 (18)	-8.8	427.50	3.95	1	6	7	66
ZINC03964126 (19)	-8.8	435.89	2.53	1	8	3	88
ZINC02568036 (20)	-8.7	314.26	1.75	1	9	2	121

CONCLUSION

This work has identified potential binding sites of high conservation that could be further investigated to identify novel interactions between PB2 and other proteins and/or cellular metabolites. In addition drug-like compounds were predicted to bind with strong affinity to a region of the PB2 protein consisting of highly conserved residues that were not targeted in other studies. The predicted compounds could serve as laboratory tools to investigate PB2 functions and/or be developed into antivirals. Due to the low probability of the targeted region undergoing genetic changes among different virus subtypes and hosts, such antiviral compounds may remain viable long-term as universal influenza inhibitors.

METHODS

PB2 sequence analysis and calculation of conservation

PB2 sequences were downloaded from the National Centre for Biotechnology Information (NCBI) Influenza Virus Resource database (Bao et al., 2008). Full length non-identical sequences were chosen from all hosts, regions and subtypes until January 2016. The CD-HIT web server (Huang et al., 2010) was used to reduce redundancy and cluster sequences meeting a similarity threshold of 98.5% as this gave an acceptable number of sequences for further analysis. For each cluster a representative sequence was retained. After removal of sequences with undefined residues, multiple sequence alignment was performed with Clustal Omega version 1.2.1 (Sievers and Higgins, 2014). The default settings for all parameters remained unchanged whereby the number of guide tree iterations and Hidden Markov Model iterations were coupled to produce the most accurate alignment. The sequence alignment editor Jalview Version 2 (Waterhouse et al., 2009) was used to analyse and edit the alignments. The Valdar scoring method was applied for the calculation of sequence conservation scores as it incorporates sequence redundancy (Valdar, 2002), which we consider critical for conservation scoring of influenza virus sequences. For the purpose of mapping the Valdar scores onto the protein structure as beta-factors the scores were re-scaled between zero and 100 (zero being low conservation and 100 being high conservation) using the following formula, where *vs* is the original Valdar score, *min* is the minimum conservation score in the dataset and 1 is the maximum conservation score:

$$\text{Re-scaled conservation score} = (vs\text{-}min)*(100 / (1\text{-}min))$$

Protein modelling

The structure of a full length amino acid sequence of the PB2 polymerase isolated from a human host (A/Viet Nam/1203/2004 (H5N1)) was predicted with the I-TASSER server (Yang and Zhang, 2015; Zhang, 2008). Residues 483-490 and 742-759, which were not covered by any template, were modelled ab-initio. The I-TASSER model was aligned with the crystal structure of A/VietNam/1203/2004 (H5N1) (PDB ID: 3L56) and the experimentally known co-ordinates of residues 542-673 and 690-738 were copied into the model. Missing atoms and side chains were fixed with Swiss PDB Viewer version 4.1.0. To remove atomic clashes from the model energy minimization was performed with 1000 steps of the steepest descent algorithm using Gromacs version 4.6.5. The AMBER99SB – ILDN force field was selected along with the TIP3P water model. The protein was solvated in water

in a cubic box with periodic boundary conditions, and the charge of the chemical system was neutralised with 21 sodium ions and additional NaCl at 100mM concentration. The particle mesh ewald algorithm for calculating electrostatic interactions was used with a real space cut-off distance of 1.0 nm and the cut-off for van der Waals interactions was 1.0 nm. Residues with an exposed surface area above 2.5 $Å^2$ were considered as exterior residues.

Prediction of binding hot spots

Binding hot spots were identified with the FTMap web server (Brenke et al., 2009) (http://ftmap.bu.edu/) which docks 16 different small organic molecular probes onto a protein surface to locate favourable binding regions or 'druggable' sites. These regions are ranked based on average free energy; low energy sites where several probe clusters overlapped (consensus) are considered as potential binding hot spots.

Virtual screening - benchmarking

Five previously identified compounds which were shown to inhibit influenza replication *in vitro* and are reported to bind the PB2 polymerase (Clark et al., 2014; Pautus et al., 2013) were used to benchmark virtual screening methods to find the best method of identifying true positives. These compounds were downloaded from the PDB and converted to the pdbqt-format with the AutoDock screening preparation tool Raccoon. The three methods tested were: AutoDock Vina version 1.1.1 (Trott and Olson, 2010), AutoDock 4 (Morris and Huey, 2009) and a consensus method using both (Kukol, 2011). A number of 180 decoy molecules with similar molecular weight from the National Cancer Insitute (NCI) Plated 2007 library were selected for the benchmarking. The grid parameters for the benchmarking were set around the mGTP binding pocket and remained the same as those reported in the publication by Pautus *et al,* (2013). The top ten positions of all 185 compounds ranked according to their binding affinity were considered for identifying the known inhibitors.

Virtual screening – target site

For the identification of new inhibitors the 3D chemical structures of all molecules at pH 6-8 were downloaded from the NCI Plated 2007 compound library of the ZINC database (http://zinc.docking.org/). The compounds from this library are available without charge for the scientific community. A chemically diverse subset of clustered molecules based on structural similarity within a Tanimoto cut-off selected at 80% was extracted from this library to give a total of 52,172 molecules. The compound library was filtered using the software

Openbabel version 2.3.1 (O'Boyle et al., 2011) to eliminate compounds with molecular weight over 500g/mol and partition coefficient (logP) over five. The remaining 46,926 chemical compounds were split into individual ligand files and saved in .pdbqt format using the AutoDock screening preparation tool Raccoon. The grid box dimensions for docking against the target site are shown in the supplementary material (S1). Compounds considered as 'frequent hitters' were removed by passing the compound library through the Pan Assay Interference Compounds (PAINS) filter with the online FAF-Drugs3 (Free ADME-Tox Filtering Tool) program (Baell and Holloway, 2010; Lagorce et al., 2015). A total of 42,348 compounds remained. The DrugBank-approved compound library (Law et al., 2014) was downloaded from the ZINC database and also screened against the PB2 target site. Molecular interactions were analysed with Ligplot+ version 1.4.5 (Laskowski and Swindells, 2011).

Acknowledgments

This work has made use of the University of Hertfordshire high-performance computing facility. We thank Jamie Stone for technical assistance.

Funding Information

This work was funded by the School of Life and Medical Sciences, University of Hertfordshire.

Conflicts of interest

The authors declare no conflict of interest.

Abbreviations

polymerase basic protein 2, PB2; polymerase basic protein 1, PB1; polymerase acidic, PA; pan assay interference compounds, PAINS; National Cancer Institute, NCI; National Centre for Biotechnology Information, NCBI; nucleoprotein, NP; methylguanosine triphosphate, mGTP, NLS; nuclear localisation signalling

References

Baell, J.B., Holloway, G.A., 2010. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. J. Med. Chem. 53, 2719–2740. doi:10.1021/jm901137j

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J.,

Lipman, D., 2008. The influenza virus resource at the National Center for Biotechnology Information. J. Virol. 82, 596–601. doi:10.1128/JVI.02005-07

Bouvier, N.M., Palese, P., 2008. The biology of influenza viruses. Vaccine 26 Suppl 4, D49–53.

Boyd, M.J., Bandarage, U.K., Bennett, H., Byrn, R.R., Davies, I., Gu, W., Jacobs, M., Ledeboer, M.W., Ledford, B., Leeman, J.R., Perola, E., Wang, T., Bennani, Y., Clark, M.P., Charifson, P.S., 2015. Isosteric replacements of the carboxylic acid of drug candidate VX-787: Effect of charge on antiviral potency and kinase activity of azaindole-based influenza PB2 inhibitors. Bioorg. Med. Chem. Lett. 25, 1990–4. doi:10.1016/j.bmcl.2015.03.013

Brenke, R., Kozakov, D., Chuang, G.Y., Beglov, D., Hall, D., Landon, M.R., Mattos, C., Vajda, S., 2009. Fragment-based identification of druggable "hot spots" of proteins using Fourier domain correlation techniques. Bioinformatics 25, 621–627. doi:10.1093/bioinformatics/btp036

Bussey, K.A., Bousse, T.L., Desmet, E.A., Kim, B., Takimoto, T., 2010. PB2 residue 271 plays a key role in enhanced polymerase activity of influenza A viruses in mammalian host cells. J. Virol. 84, 4395–406. doi:10.1128/JVI.02642-09

Chambers, B.S., Parkhouse, K., Ross, T.M., Alby, K., Hensley, S.E., 2015. Identification of Hemagglutinin Residues Responsible for H3N2 Antigenic Drift during the 2014-2015 Influenza Season. Cell Rep. 12, 1–6. doi:10.1016/j.celrep.2015.06.005

Chin, A.W.H., Li, O.T.W., Mok, C.K.P., Ng, M.K.W., Peiris, M., Poon, L.L.M., 2014. Influenza A viruses with different amino acid residues at PB2-627 display distinct replication properties in vitro and in vivo: revealing the sequence plasticity of PB2-627 position. Virology 468, 545–555. doi:10.1016/j.virol.2014.09.008

Clark, M.P., Ledeboer, M.W., Davies, I., Byrn, R.A., Jones, S.M., Perola, E., Tsai, A., Jacobs, M., Nti-Addae, K., Bandarage, U.K., Boyd, M.J., Bethiel, R.S., Court, J.J., Deng, H., Duffy, J.P., Dorsch, W.A., Farmer, L.J., Gao, H., Gu, W., Jackson, K., Jacobs, D.H., Kennedy, J.M., Ledford, B., Liang, J., Maltais, F., Murcko, M., Wang, T., Wannamaker, M.W., Bennett, H.B., Leeman, J.R., McNeil, C., Taylor, W.P., Memmott, C., Jiang, M., Rijnbrand, R., Bral, C., Germann, U., Nezami, A., Zhang, Y., Salituro, F.G., Bennani, Y.L., Charifson, P.S., 2014. Discovery of a novel, first-in-class, orally bioavailable azaindole inhibitor (VX-787) of influenza PB2. J. Med. Chem. 57, 6668–6678. doi:10.1021/jm5007275

Corena-McLeod, M., 2015. Comparative pharmacology of risperidone and paliperidone. Drugs R D 15, 163–174. doi:10.1007/s40268-015-0092-x

Das, K., Aramini, J.M., Ma, L.-C., Krug, R.M., Arnold, E., 2010. Structures of influenza A proteins and insights into antiviral drug targets. Nat. Struct. Mol. Biol. 17, 530–538.

Delaforge, E., Milles, S., Bouvignies, G., Bouvier, D., Boivin, S., Salvi, N., Maurin, D., Martel, A., Round, A., Lemke, E.A., Ringkjøbing Jensen, M., Hart, D.J., Blackledge, M., 2015. Large-Scale Conformational Dynamics Control H5N1 Influenza Polymerase PB2 Binding to Importin α. J. Am. Chem. Soc. 137, 15122–15134. doi:10.1021/jacs.5b07765

Fodor, E., 2013. The RNA polymerase of influenza A virus: mechanisms of viral transcription and replication. Acta Virol. 57, 113–122. doi:10.4149/av

Guilligay, D., Tarendeau, F., Resa-Infante, P., Coloma, R., Crepin, T., Sehr, P., Lewis, J., Ruigrok, R.W., Ortin, J., Hart, D.J., Cusack, S., 2008. The structural basis for cap binding by influenza virus polymerase subunit PB2. Nat. Struct. Mol. Biol. 15, 500–506. doi:10.1038/nsmb.1421

Harper, S.A., Bradley, J.S., Englund, J.A., File, T.M., Gravenstein, S., Hayden, F.G., McGeer, A.J., Neuzil, K.M., Pavia, A.T., Tapper, M.L., Uyeki, T.M., Zimmerman, R.K., 2009. Seasonal Influenza in Adults and Children—Diagnosis, Treatment, Chemoprophylaxis, and Institutional Outbreak Management: Clinical Practice Guidelines of the Infectious Diseases Society of America. Clin. Infect. Dis. 48, 1003–1032. doi:10.1086/598513

Hatakeyama, D., Shoji, M., Yamayoshi, S., Hirota, T., Nagae, M., Yanagisawa, S., Nakano, M., Ohmi, N., Noda, T., Kawaoka, Y., Kuzuhara, T., 2014. A novel functional site in the PB2 subunit of influenza A virus essential for acetyl-CoA interaction, RNA polymerase activity, and viral replication. J. Biol. Chem. 289, 24980–94. doi:10.1074/jbc.M114.559708

Hayden, F.G., De Jong, M.D., 2011. Emerging influenza antiviral resistance threats. J. Infect. Dis. 203, 6–10. doi:10.1093/infdis/jiq012

Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W., 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26, 680–2. doi:10.1093/bioinformatics/btq003

Kirui, J., Bucci, M.D., Poole, D.S., Mehle, A., 2014. Conserved features of the PB2 627 domain impact influenza virus polymerase function and replication. J. Virol. 88, 5977–86. doi:10.1128/JVI.00508-14

Kukol, A., 2011. Consensus virtual screening approaches to predict protein ligands. Eur. J. Med. Chem. 46, 4661–4664. doi:10.1016/j.ejmech.2011.05.026

Kuzuhara, T., Kise, D., Yoshida, H., Horika, T., Murazaki, Y., Nishimura, A., Echigo, N., Utsunomiya, H., Tsuge, H., 2009. Structural basis of the influenza A virus RNA polymerase PB2 RNA-binding domain containing the pathogenicity-determinant lysine 627 residue. J. Biol. Chem. 284, 6855–6860. doi:10.1074/jbc.C800224200

Labadie, K., Dos Santos Afonso, E., Rameix-Welti, M.-A., van der Werf, S., Naffakh, N., 2007. Host-range determinants on the PB2 protein of influenza A viruses control the interaction between the viral polymerase and nucleoprotein in human cells. Virology 362, 271–282. doi:10.1016/j.virol.2006.12.027

Lagorce, D., Sperandio, O., Baell, J.B., Miteva, M.A., Villoutreix, B.O., 2015. FAF-Drugs3: A web server for compound property calculation and chemical library design. Nucleic Acids Res. 43. doi:10.1093/nar/gkv353

Laskowski, R.A., Swindells, M.B., 2011. LigPlot+: Multiple Ligand–Protein Interaction Diagrams for Drug Discovery. J. Chem. Inf. Model. 51, 2778–2786. doi:10.1021/ci200227u

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z.T., Han, B., Zhou, Y., Wishart, D.S., 2014. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 42, D1091–7. doi:10.1093/nar/gkt1068

Liu, Y., Qin, K., Meng, G., Zhang, J., Zhou, J., Zhao, G., Luo, M., Zheng, X., 2013. Structural and functional characterization of K339T substitution identified in the PB2 subunit cap-binding pocket of influenza A virus. J. Biol. Chem. 288, 11013–11023. doi:10.1074/jbc.M112.392878

Long, J.S., Giotis, E.S., Moncorgé, O., Frise, R., Mistry, B., James, J., Morisson, M., Iqbal, M., Vignal, A., Skinner, M.A., Barclay, W.S., 2016. Species difference in ANP32A underlies influenza A virus polymerase host restriction. Nature 529, 101–104. doi:10.1038/nature16474

Mehle, A., Doudna, J.A., 2009. Adaptive strategies of the influenza virus polymerase for replication in humans. Proc. Natl. Acad. Sci. U. S. A. 106, 21312–6. doi:10.1073/pnas.0911915106

Moncorgé, O., Mura, M., Barclay, W.S., 2010. Evidence for avian and human host cell factors that affect the activity of influenza virus polymerase. J. Virol. 84, 9978–86. doi:10.1128/JVI.01134-10

Morris, G., Huey, R., 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J. Comput. Chem. 30, 2785–2791. doi:10.1002/jcc.21256.AutoDock4

Naesens, L., Stevaert, A., Vanderlinden, E., 2016. Antiviral therapies on the horizon for influenza. Curr. Opin. Pharmacol. 30, 106–115. doi:10.1016/j.coph.2016.08.003

Neumann, G., Kawaoka, Y., 2015. Transmission of influenza A viruses. Virology 479-480, 234–46. doi:10.1016/j.virol.2015.03.009

Ng, A.K.-L., Chan, W.-H., Choi, S.-T., Lam, M.K.-H., Lau, K.-F., Chan, P.K.-S., Au, S.W.-N., Fodor, E., Shaw, P.-C., 2012. Influenza polymerase activity correlates with the strength of interaction between nucleoprotein and PB2 through the host-specific residue K/E627. PLoS One 7, e36415. doi:10.1371/journal.pone.0036415

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., Hutchison, G.R., 2011. Open Babel: An open chemical toolbox. J. Cheminform. 3, 33. doi:10.1186/1758-2946-3-33

Patel, H., Kukol, A., 2016. Recent discoveries of influenza A drug target sites to combat virus replication. Biochem. Soc. Trans. 44, 932–936. doi:10.1042/BST20160002

Pautus, S., Sehr, P., Lewis, J., Fortune, A., Wolkerstorfer, A., Szolar, O., Guilligay, D., Lunardi, T., De, J., Cusack, S., 2013. New 7 □ Methylguanine Derivatives Targeting the In fl uenza Polymerase PB2 Cap-Binding Domain. J. Med. Chem. 56, 8915–8930.

Pflug, A., Guilligay, D., Reich, S., Cusack, S., 2014. Structure of influenza A polymerase bound to the viral RNA promoter. Nature 516, 355–60. doi:10.1038/nature14008

Pflug, A., Lukarska, M., Resa-Infante, P., Reich, S., Cusack, S., 2017. Structural insights into RNA synthesis by the influenza virus transcription-replication machine. Virus Res. doi:10.1016/j.virusres.2017.01.013

Poole, E., Elton, D., Medcalf, L., Digard, P., 2004. Functional domains of the influenza A virus PB2 protein: identification of NP- and PB1-binding sites. Virology 321, 120–33. doi:10.1016/j.virol.2003.12.022

Reich, S., Guilligay, D., Pflug, A., Malet, H., Berger, I., Crépin, T., Hart, D., Lunardi, T.,

Nanao, M., Ruigrok, R.W.H., Cusack, S., 2014. Structural insight into cap-snatching and RNA synthesis by influenza polymerase. Nature 516, 361–6. doi:10.1038/nature14009

Reperant, L.A., Kuiken, T., Osterhaus, A.D.M.E., 2012. Adaptive pathways of zoonotic influenza viruses: from exposure to establishment in humans. Vaccine 30, 4419–34. doi:10.1016/j.vaccine.2012.04.049

Samson, M., Pizzorno, A., Abed, Y., Boivin, G., 2013. Influenza virus resistance to neuraminidase inhibitors. Antiviral Res. 98, 174–85. doi:10.1016/j.antiviral.2013.03.014

Sievers, F., Higgins, D.G., 2014. Clustal omega. Curr. Protoc. Bioinformatics 48, 3.13.1–3.13.16. doi:10.1002/0471250953.bi0313s48

Sugiyama, K., Obayashi, E., Kawaguchi, A., Suzuki, Y., Tame, J.R.H., Nagata, K., Park, S.-Y., 2009. Structural insight into the essential PB1-PB2 subunit contact of the influenza virus RNA polymerase. EMBO J. 28, 1803–11. doi:10.1038/emboj.2009.138

Taubenberger, J.K., Kash, J.C., 2010. Influenza virus evolution, host adaptation, and pandemic formation. Cell Host Microbe 7, 440–451. doi:10.1016/j.chom.2010.05.009

Thierry, E., Guilligay, D., Kosinski, J., Bock, T., Gaudon, S., Round, A., Pflug, A., Hengrung, N., El Omari, K., Baudin, F., Hart, D.J., Beck, M., Cusack, S., 2016. Influenza Polymerase Can Adopt an Alternative Configuration Involving a Radical Repacking of PB2 Domains. Mol. Cell 61, 125–137. doi:10.1016/j.molcel.2015.11.016

Tong, S., Zhu, X., Li, Y., Shi, M., Zhang, J., Bourgeois, M., Yang, H., Chen, X., Recuenco, S., Gomez, J., Chen, L.-M., Johnson, A., Tao, Y., Dreyfus, C., Yu, W., McBride, R., Carney, P.J., Gilbert, A.T., Chang, J., Guo, Z., Davis, C.T., Paulson, J.C., Stevens, J., Rupprecht, C.E., Holmes, E.C., Wilson, I.A., Donis, R.O., 2013. New world bats harbor diverse influenza A viruses. PLoS Pathog. 9, e1003657. doi:10.1371/journal.ppat.1003657

Trott, O., Olson, A., 2010. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J. Comput. Chem. 31, 455–461. doi:10.1002/jcc.21334.AutoDock

Valdar, W.S.J., 2002. Scoring residue conservation. Proteins Struct. Funct. Genet. 48, 227–241. doi:10.1002/prot.10146

Warren, S., Wan, X.F., Conant, G., Korkin, D., 2013. Extreme evolutionary conservation of functionally important regions in H1N1 influenza proteome. PLoS One 8, 1–14. doi:10.1371/journal.pone.0081027

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J., 2009. Jalview AVersion 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189–91. doi:10.1093/bioinformatics/btp033

Wu, N.C., Olson, C.A., Du, Y., Le, S., Tran, K., Remenyi, R., Gong, D., Al-Mawsawi, L.Q., Qi, H., Wu, T.T., Sun, R., 2015. Functional Constraint Profiling of a Viral Protein Reveals Discordance of Evolutionary Conservation and Functionality. PLoS Genet. 11, 1–27. doi:10.1371/journal.pgen.1005310

Xu, C., Hu, W.-B., Xu, K., He, Y.-X., Wang, T.-Y., Chen, Z., Li, T.-X., Liu, J.-H., Buchy, P., Sun, B., 2012. Amino acids 473V and 598P of PB1 from an avian-origin influenza A virus contribute to polymerase activity, especially in mammalian cells. J. Gen. Virol. 93, 531–40. doi:10.1099/vir.0.036434-0

Yamada, S., Hatta, M., Staker, B.L., Watanabe, S., Imai, M., Shinya, K., Sakai-Tagawa, Y., Ito, M., Ozawa, M., Watanabe, T., Sakabe, S., Li, C., Kim, J.H., Myler, P.J., Phan, I., Raymond, A., Smith, E., Stacy, R., Nidom, C.A., Lank, S.M., Wiseman, R.W., Bimber, B.N., O'Connor, D.H., Neumann, G., Stewart, L.J., Kawaoka, Y., 2010. Biological and structural characterization of a host-adapting amino acid in influenza virus. PLoS Pathog. 6, e1001034. doi:10.1371/journal.ppat.1001034

Yang, J., Zhang, Y., 2015. Protein Structure and Function Prediction Using I-TASSER. Curr. Protoc. Bioinforma. 52, 5.8.1–15. doi:10.1002/0471250953.bi0508s52

Zhang, Y., 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9, 40. doi:10.1186/1471-2105-9-40

Figures

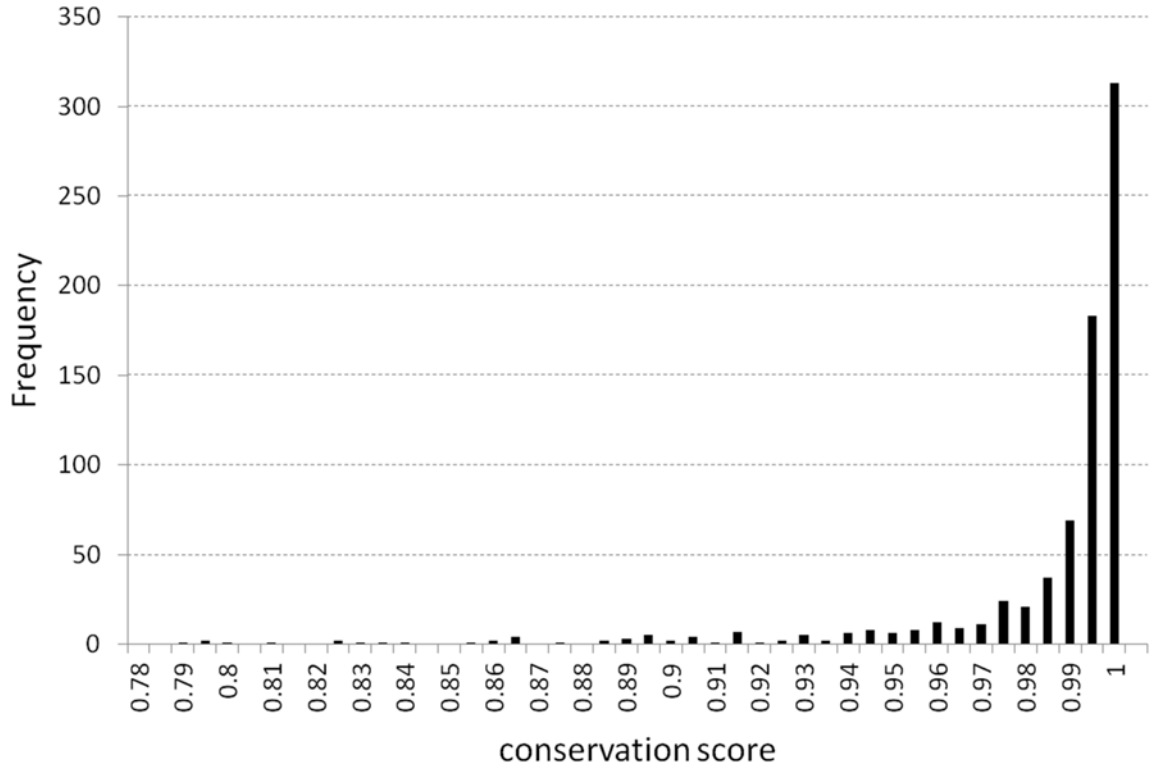

Fig.1. Frequency distribution of PB2 amino acid conservation scores obtained after alignment of 702 non-redundant influenza A sequences from mainly human, avian and swine hosts using the Valdar scoring formula.

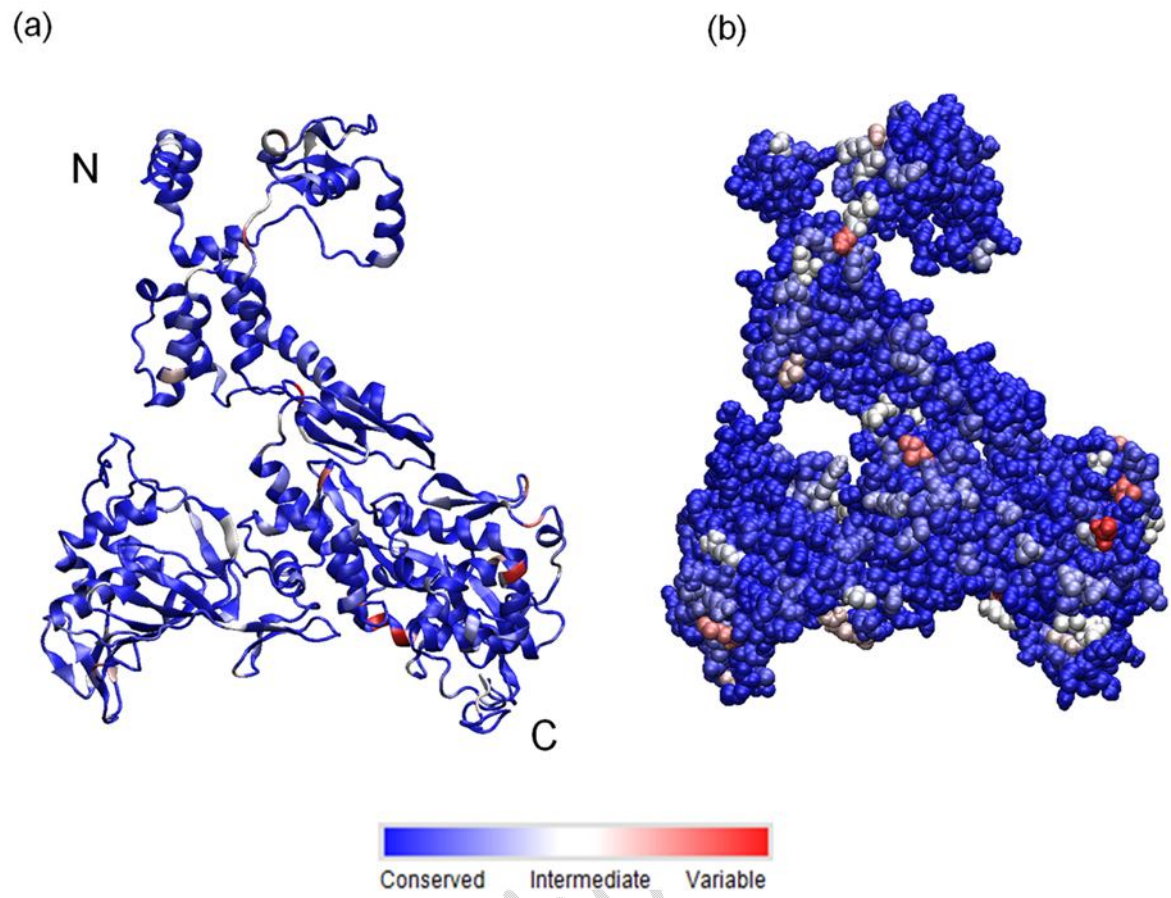

Fig. 2. Amino acid conservation mapped onto the H5N1 influenza A PB2 protein structure shown in (a) cartoon representation and (b) spacefill representation.

```
                    110         120         130         140         150
                    ....|....|....|....|....|....|....|....|....|....|
PB2_H17N10          RAGPVSDVVHYPRVYKMYFDRLERLTHGTFGPVKFYNQVKVRKRVDINPG
PB2_H5N1            RNGPATSAVHYPKVYKTYFEKVERLKHGTFGPVHFRNQVKIRRRVDINPG
                    * **.:..****:*** **:::***.*******:* ****:*:*******

                    510         520         530         540         550
                    ....|....|....|....|....|....|....|....|....|....|
PB2_H17N10          YLRVRNEKGELLISPEEVSEAQGQEKLPINYNSSLMWEVNGPESILTNTY
PB2_H5N1            FLRVRDQRGNVLLSPEEVSETQGTEKLTITYSSSMMWEINGPESVLVNTY
                    :****:::*::*:*******:** *** *.*.**:***:*****:*.***
```

Fig. 3. PB2 H5N1 (A/Viet Nam/1203/2004) and H17N10 (A/little yellow-shouldered bat/Guatemala/060/2010) sequence alignment of the region 101-150 and 501-550 covering the target site for virtual screening. An asterisk indicates identical amino acid residues, a colon indicates strong similarity, and a period indicates weak similarity.

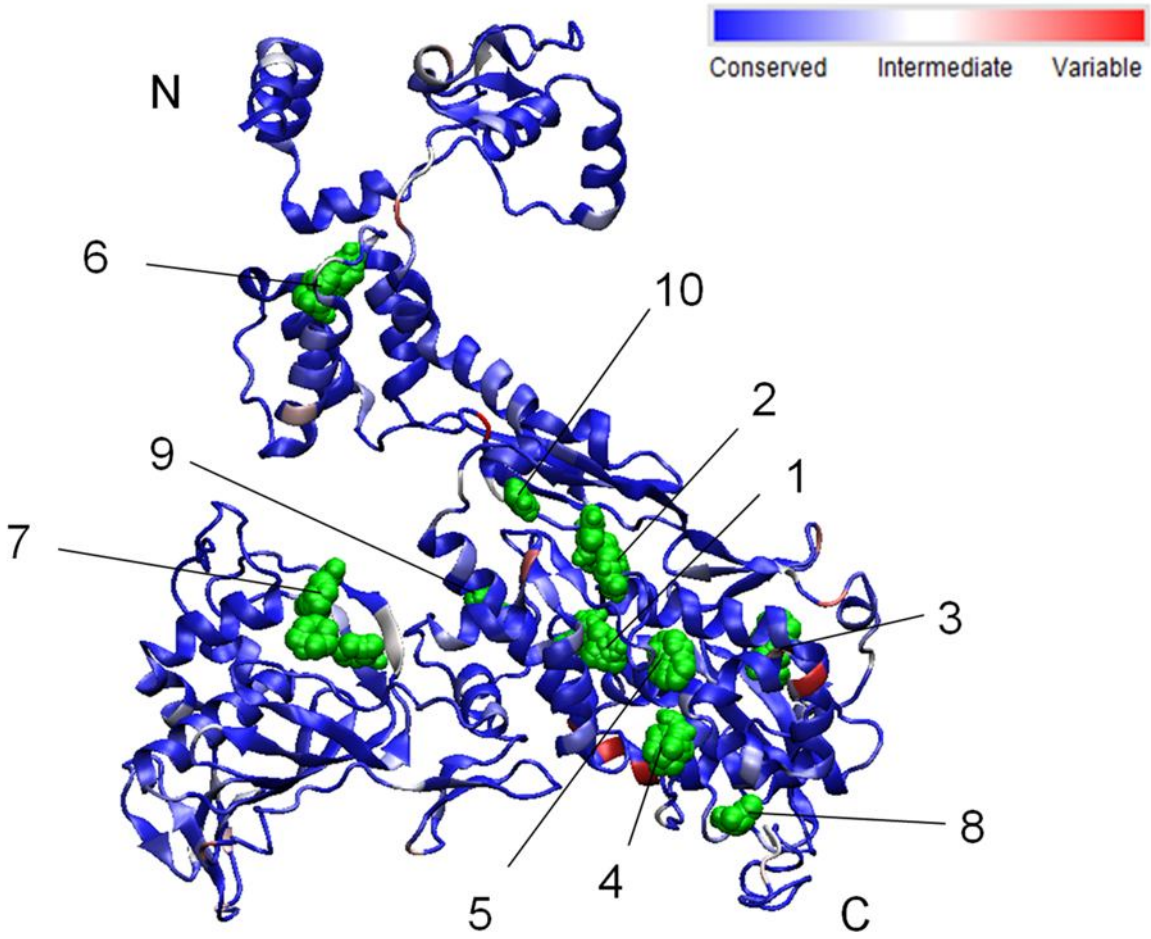

Fig. 4. Locations of the top ten ligand binding hot spots (green spheres) identified by the FTMap algorithm shown together with the degree of PB2 sequence conservation on the H5N1 PB2 structure.

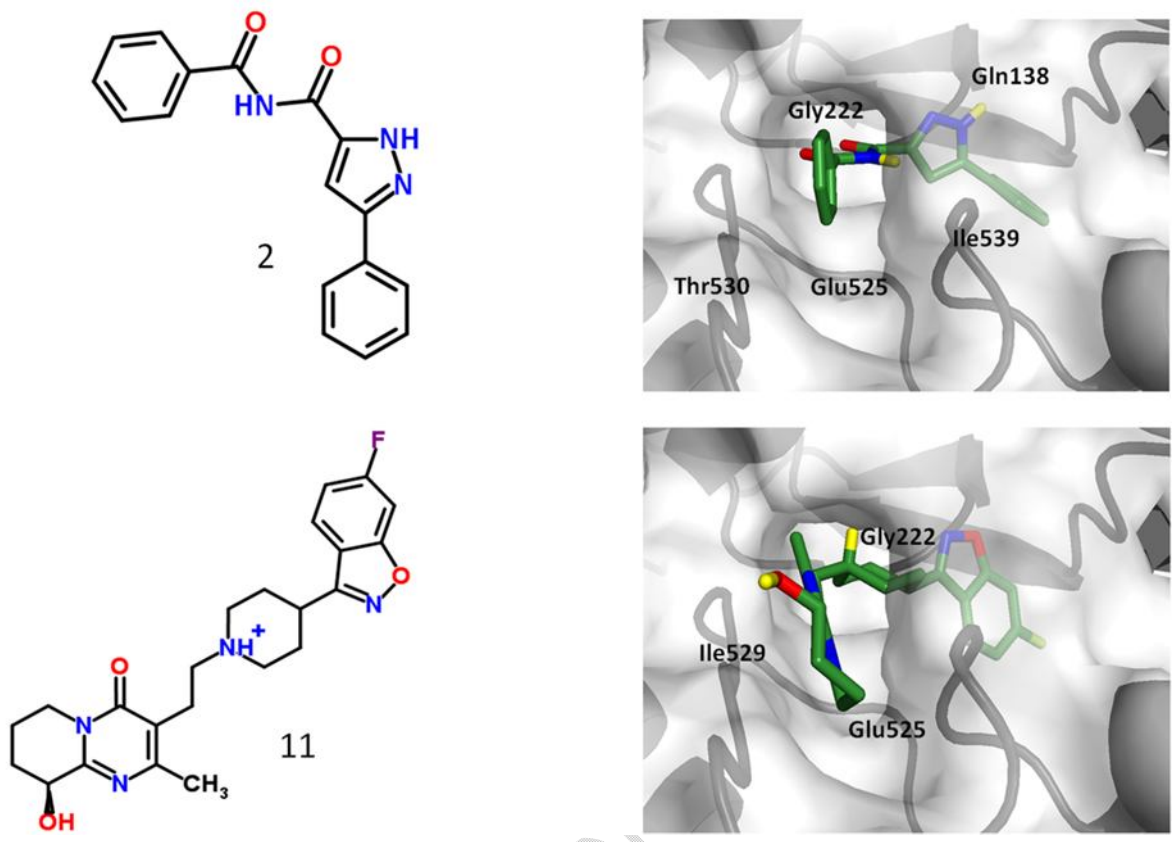

Fig. 5. Docking models of top hit compounds targeting the PB2 protein: ZINC05543024 (2) and paliperidone (11) identified by virtual screening using AutoDock Vina. Interacting PB2 residues are labelled.

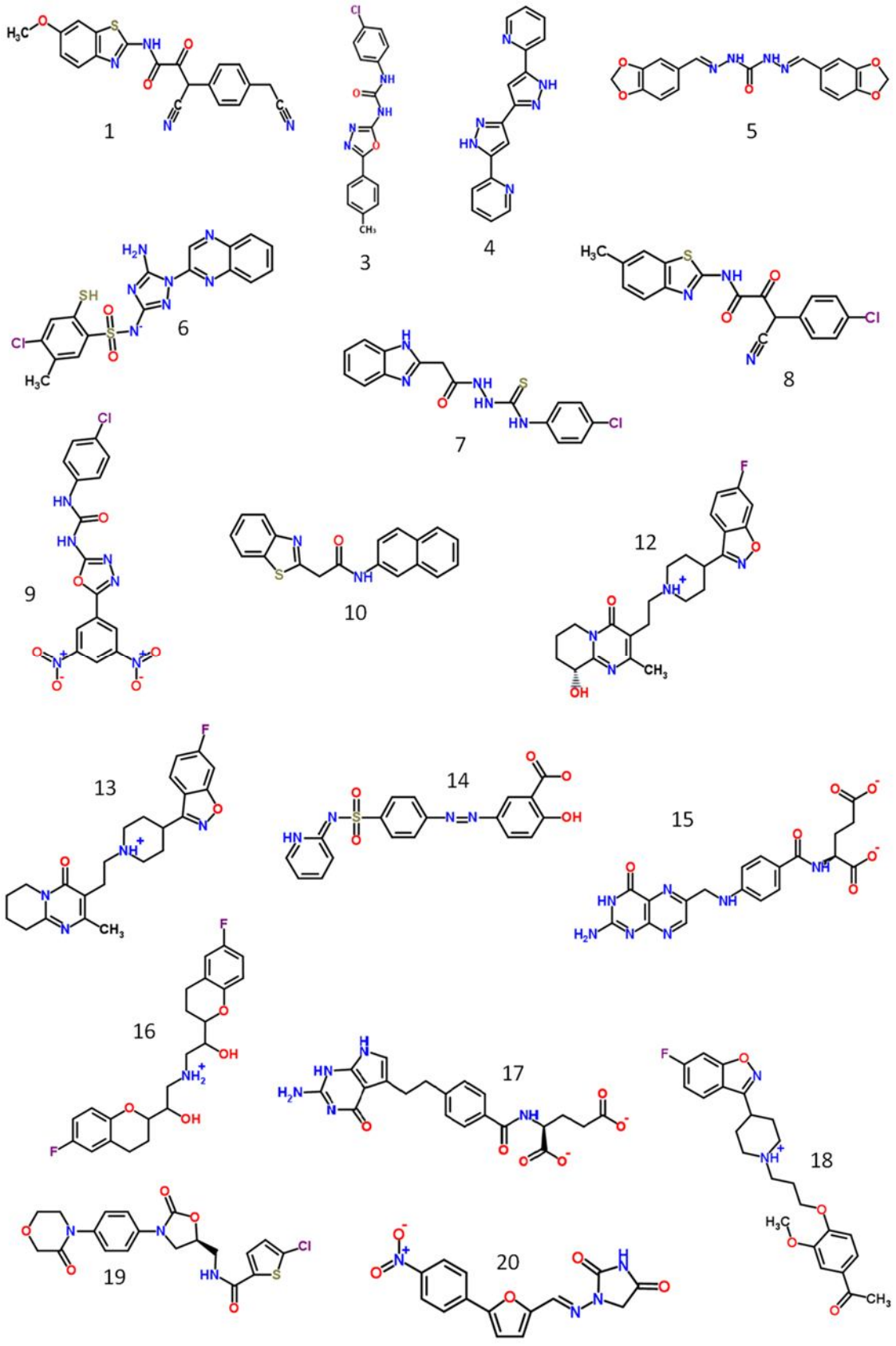

Fig. 6. Chemical structures of predicted top hit compounds from the NCI and DrugBank library. Numbers correspond to the ZINC ID shown in table 2.