Extreme Evolutionary Conservation of Functionally Important Regions in H1N1 Influenza Proteome

Samantha Warren[1], Xiu-Feng Wan[2], Gavin Conant[3,4], Dmitry Korkin[1,4,5*]

1 Department of Computer Science, University of Missouri, Columbia, Missouri, United States of America, **2** Department of Basic Sciences, Mississippi State University, Mississippi State, Mississippi, United States of America, **3** Division of Animal Sciences, University of Missouri, Columbia, Missouri, United States of America, **4** Informatics Institute, University of Missouri, Columbia, Missouri, United States of America, **5** Bond Life Science Center, University of Missouri, Columbia, Missouri, United States of America

Abstract

The H1N1 subtype of influenza A virus has caused two of the four documented pandemics and is responsible for seasonal epidemic outbreaks, presenting a continuous threat to public health. Co-circulating antigenically divergent influenza strains significantly complicates vaccine development and use. Here, by combining evolutionary, structural, functional, and population information about the H1N1 proteome, we seek to answer two questions: (1) do residues on the protein surfaces evolve faster than the protein core residues consistently across all proteins that constitute the influenza proteome? and (2) in spite of the rapid evolution of surface residues in influenza proteins, are there any protein regions on the protein surface that do not evolve? To answer these questions, we first built phylogenetically-aware models of the patterns of surface and interior substitutions. Employing these models, we found a single coherent pattern of faster evolution on the protein surfaces that characterizes all influenza proteins. The pattern is consistent with the events of inter-species reassortment, the worldwide introduction of the flu vaccine in the early 80's, as well as the differences caused by the geographic origins of the virus. Next, we developed an automated computational pipeline to comprehensively detect regions of the protein surface residues that were 100% conserved over multiple years and in multiple host species. We identified conserved regions on the surface of 10 influenza proteins spread across all avian, swine, and human strains; with the exception of a small group of isolated strains that affected the conservation of three proteins. Surprisingly, these regions were also unaffected by genetic variation in the pandemic 2009 H1N1 viral population data obtained from deep sequencing experiments. Finally, the conserved regions were intrinsically related to the intra-viral macromolecular interaction interfaces. Our study may provide further insights towards the identification of novel protein targets for influenza antivirals.

Citation: Warren S, Wan X-F, Conant G, Korkin D (2013) Extreme Evolutionary Conservation of Functionally Important Regions in H1N1 Influenza Proteome. PLoS ONE 8(11): e81027. doi:10.1371/journal.pone.0081027

Editor: Dhanasekaran Vijaykrishna, Duke-NUS Gradute Medical School, Singapore

Received June 8, 2013; **Accepted** October 8, 2013; **Published** November 25, 2013

Funding: We acknowledge funding from National Science Foundation (DBI-0845196 to DK) and the Department of Education GAANN Fellowship (P200A100053 to SW). G.C.C. is supported by the Reproductive Biology Group of the Food for the 21st Century program at the University of Missouri. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

* Email: korkin@korkinlab.org

Introduction

Influenza type A, a member of the *Orthomyxoviridae* family, is responsible for a large majority of human flu-related illnesses [17], including seasonal epidemics and four documented pandemic outbreaks. The genome is comprised of eight negative-strand RNA segments, encoding 11-12 protein products [18,19]. As a segmented RNA virus, influenza A has two major evolutionary events that define its genomic diversity: replication errors and reassortment [17,20]; these facilitate the emergence of highly pathogenic strains [21–25]. Reassortment, or the exchange of one or more discrete RNA segments into multipartite viruses, occurs frequently between influenza A viruses [26–30] and is critical for the generation of epidemic and pandemic influenza strains. The 2009 H1N1 pandemic virus is a reassortment of genomic segments from distinct swine influenza virus lineages and from human and avian influenza viruses [32].

In addition to rapid mutations and frequent reassortments, co-circulating antigenically divergent H1N1 influenza strains significantly complicate vaccine development and use. All these H1N1 viruses have been found to be genetically linked to the 1918 H1N1 pandemic virus [19,31,33,34]. While the hemagglutinin (HA) proteins of the H1N1 virus strains circulating in human populations have evolved considerably since the 1918 pandemic, those in swine have mutated much more slowly [32]. This disparity is evidenced by the structural

and antigenic similarities between HA proteins from the 2009 and 1918 outbreaks [19,31,33].

The dynamic patterns of the evolutionary changes in the influenza genome are not uniform [17]. One of the best-studied influenza proteins is HA, which has strikingly different patterns of substitution across its sequence, with a handful of residues having high substitution rates [35]. Conservation patterns and their origins in the other influenza proteins are less-studied. Here, we suggest that one source of this variation in substitution rates is the different protein structure context of the residues.

The link between conservation in the influenza proteome and the function of the respective proteins has been studied both computationally and experimentally for several decades [36–40]. Several early studies determined a high sequence similarity between the HA proteins of Influenza A and B, including similarity in the known structural features such as the hydrophobic regions of HA [36]. Another study describing the strong similarity of the epitopes located in the head region of the HA proteins across all subtypes of Influenza A and Influenza B viruses inspired the design of the new broadly neutralizing antibodies [40]. Experimental studies of HA proteins from the H1N1 strains in swine obtained between 1986 and 1991 found the proteins to be highly conserved, both antigenically and genetically [38]. Investigation of the evolution of the influenza A nucleoprotein (NP) gene across five host species using a classical, sequence-similarity based, approach found the evidence of adaptive changes in function among host-specific NPs [37]. Recently, a large-scale study evaluating sequences of 11 viral proteins across Influenza A isolates from avian and human hosts over the last 30 years, isolated 55 conserved sequence fragments with conservation ranging from 80% to 100% and linked many of them to HLA class I or class II epitopes [39].

Most of the approaches considered above are sequence-based and do not consider the structural information about the proteins. Several recent approaches, however, have demonstrated that introducing the information about the three-dimensional structure of an influenza protein may provide additional insights into their evolution and specifically the conservation of protein's structural fragments, which may be sequentially noncontiguous. For instance, two studies reported that structural conservation of human influenza A HA epitope was responsible for interaction with sialoglycans; similarly, conserved influenza B HA epitopes were successfully targeted by the monoclonal antibodies [40,41]. However a structure-based evolutionary analysis of the entire influenza proteome is yet to be done.

Despite their rapid evolution, influenza proteins participate extensively in intra- and inter-species interactions. Human antibodies, for example, recognize at least four antigenic sites on the viral HA protein of H1N1 influenza A virus [42]. Other interactions involve viral macromolecules exclusively, such as PB1, which interacts simultaneously with PA and PB2 [16,43]. More generally, these intra-viral interactions fall into three categories: heteromers, homomers, and protein-RNA interactions. Cataloging the complete human-influenza interactome was a significant challenge, and has been completed only recently [44]. The next step is to better understand the biology of these numerous, newly identified interactions, linking them with the evolutionary mechanisms in influenza.

In this work, we sought to understand the role of protein three-dimensional structure in the evolution of H1N1 genomes. Specifically, we aimed to answer two questions: (1) whether the surface residues of the majority of H1N1 proteins are diverging faster than are the core residues and (2) whether there are regions of surface residues that are completely conserved, in spite of the anticipated rapid divergence of the protein surface. We began by using a phylogenetic analysis to model the dependence of patterns of amino acid substitutions in the proteins on their exposure to the solvent using the 3D models of the H1N1 proteins. Next, we developed a computational pipeline integrating sequence and structure data in order to identify conserved regions of the proteins' three-dimensional structures. Each region is a structurally connected "patch" of residues that may not necessarily be sequentially consecutive. The pipeline determines surface residues that are 100% conserved in the sequence alignments, clusters them with respect to their structural positioning, and calculates the probability of observing such a region under a random distribution of conserved residues. Finally, we associated the identified regions with known functional sites, and mapped the mutation sites collected from the viral population data of the pandemic 2009 H1N1 influenza, obtained from deep sequencing experiments.

Results

Data collection

The initial set of H1N1 strains included 1,100 unique genomes, each containing ten sequences (*Methods*). We employed a redundancy filter with a whole-genome sequence identity threshold of 95% which yielded a final set of 75 strains (see *Methods*; Table S1), including 10 avian, 34 human, and 31 swine strains, with all strains dating between 1933 and 2009. The 1918 'Spanish flu' strain was not included in the final set due to several of its proteins having 100% sequence identity with the corresponding proteins in other strains, but was included as a case study. Nevertheless, the conservation of the surface regions between the 1918 and 2009 H1N1 pandemic strains was analyzed in detail (see below). The average sequence identity between the individual proteins in our dataset varied from 88.7% to 96.4%. As expected, there were pairs of strains sharing identical or near-identical proteins, even when other proteins in these strains were less than 95% identical. No strains shared the same proteins with less than 60% protein sequence identity (Table 1).

Phylogenetic Analysis of Structure-Based Evolution

For each of the ten proteins, we computed the maximum likelihood phylogeny using PhyML [45]. We then fitted several models of evolution to these alignments (*Methods*). The most basic, $M_{uniform}$, requires all nucleotides in the sequence to evolve at the same rate. We compared that model to M_{scaled}, where the evolution rate of positions corresponding to surface

Table 1. Strain conservation across the ten proteins.

	HA	M1	M2	NA	NP	NS1	NS2	PA	PB1	PB2
Average	89	97	89	89	95	87	93	96	96	96
Minimum	77	92	71	77	86	60	76	88	86	90
Maximum	100	100	100	100	100	100	100	100	100	100

The proteins vary in their conservation. When removing redundancy we first calculated the pairwise conservation percentage for each individual protein. From this we calculated the average pairwise conservation. We also determined the minimum and maximum conservation.

doi: 10.1371/journal.pone.0081027.t001

Table 2. Protein evolutionary rates and patch information.

Protein	r_e/r_i	Exterior / Interior	N of patches	Template coverage	Intra-viral interactions	
					Literature	Structure
HA	1.2	0.53	3	88%	[1]	
M1	2.2	0.63	1	63%	[2,3]	
M2	2.3	6.60	1	38%	[4,5,6,7]	
NA	1.5	0.33	1	82%		3B7E
NP	1.2	0.56	2	94%	[8,9,10]	
NS1	1.5	0.97	1	83%	[11]	
NS2	1.1	1.45	(1)	40%	[12,13,14]	
PA	1.9	0.65	5	91%	[15,16]	2ZNL
PB1	1.2	1.15	(3)	7%		3A1G, 2ZNL
PB2	1.2	0.70	2 (2)	47%	[31]	3A1G

The ratio of protein evolutionary rates for the exterior and interior residues (r_e/r_i) was determined using HyPhy. Shown are the ratios for entire proteins. The significant regions are shown in the following column with regions that are biologically significant, but must be explained structurally rather than statistically in parentheses. For some viral proteins the homology models of do not cover the entire sequence due to the limited coverage of their templates. Shown is the percentage of the protein sequence coverage for each structural model. The last column summarizes the evidence for the intra-viral interactions in recent literature and from DOMMINO.

doi: 10.1371/journal.pone.0081027.t002

residues was allowed to be more dissimilar than for interior residues. As expected, all ten proteins showed higher rates of substitution for surface positions (the rate ratio, r_e/r_i was greater than 1.0; Table 2, likelihood ratio test, see *Methods*). We then investigated whether this pattern was the result of differing surface to interior constraints on the various branches of Figure 1, but found no such pattern. Similarly, r_e/r_i varied only slightly for seven of the proteins: 1.1 for NS2; 1.2 for HA, NP, PB1, and PB2; 1.5 for NA and NS1. The ratio was considerable higher for the other three proteins: 1.9 for PA, 2.2 for M1, and 2.3 for M2.

The obtained phylogenetic trees were clearly separated into the host-specific lineages with occurrences of a few strains from other species (Figures 1, S1, S2). The human lineage in both HA and NA trees exhibits a strong 'trunk-like' temporal pattern that has been previously observed in the phylogenetic trees generated from whole-sequence alignments [46,47] (Figure 1, S1). In the case of PA, this pattern is less evident (Figure S2). A few human strains were found as a part of the swine clade, and a recent swine strain was found as a part of the human clade across all three analyzed proteins, indicating the bi-dimensional transmission of influenza A viruses between the animal and human interface. Interestingly, we found that after 1984, the surface-to-core ratio of human HA and NA proteins, but not PA proteins becomes significantly higher. This indicates the increasing selective pressure on the surface residues of the former two proteins.

Unlike the human lineage, the swine and avian lineages of HA and NA trees did not exhibit the trunk-like pattern. Instead, the swine lineage was divided into two clades, one comprised primarily of North American strains and another comprised of Eurasian strains. Moreover, while Eurasian swine strains had a surface-to-core ratio that was generally higher than in North American strains, we did not observe the same sudden increase in the ratio values as a function of time, as we did in the human lineages. Finally, several human strains, namely Mexico/2009, Iowa/2005, and New Jersey/1976 were included in swine lineages of HA and NA proteins (Figures 1, S1), representing spillover cases of H1N1 virus from swine to human. This was not necessarily the case for other influenza proteins, which may have originated in different hosts (Figures S2-S9, S11-S13).

Homology modeling of the individual influenza proteins

The structural analysis of H1N1 protein surfaces using homology modeling is challenged by the limited structural template coverage of some influenza proteins. Three-dimensional structures of several influenza A proteins have been modeled before and used for functional and evolutionary studies [48–52]. Unfortunately, for some influenza proteins (M2, NS2, PB1, PB2) the templates cover only a small portion of the target sequence, while for other influenza proteins the entire sequence is covered by a single template or a number of templates with a little or no structural overlap (HA, M1, NA, NP, NS1, PA). Therefore, we used a single template as the basis for our models for seven proteins and a multiple-template strategy for the remaining three (Table 1). As a result, we obtained models covering almost entire sequences of 6 H1N1 proteins, with the exception of small N-terminal and C-terminal regions. Sequences of 3 proteins were partially covered by two or more fragments (PA, PB1, and PB2). Only one protein (M2) did not have a significant portion of its sequence (residues 23-60) covered by any structural template (Table 1); these regions were not modeled structurally. The average target-template sequence identity was 91% (minimal sequence identity was 45%). This high sequence identity, thus, allowed for an accurate determination of surface and core residues of H1N1 proteins based on the homology models.

Conserved regions on H1N1 proteins surface are associated exclusively with intra-viral interactions

Each H1N1 protein was found to have at least one evolutionary conserved region that was also statistically significant (Figures 2A, 2C, Table S2). The literature search and a search of DOMMINO database of macromolecular

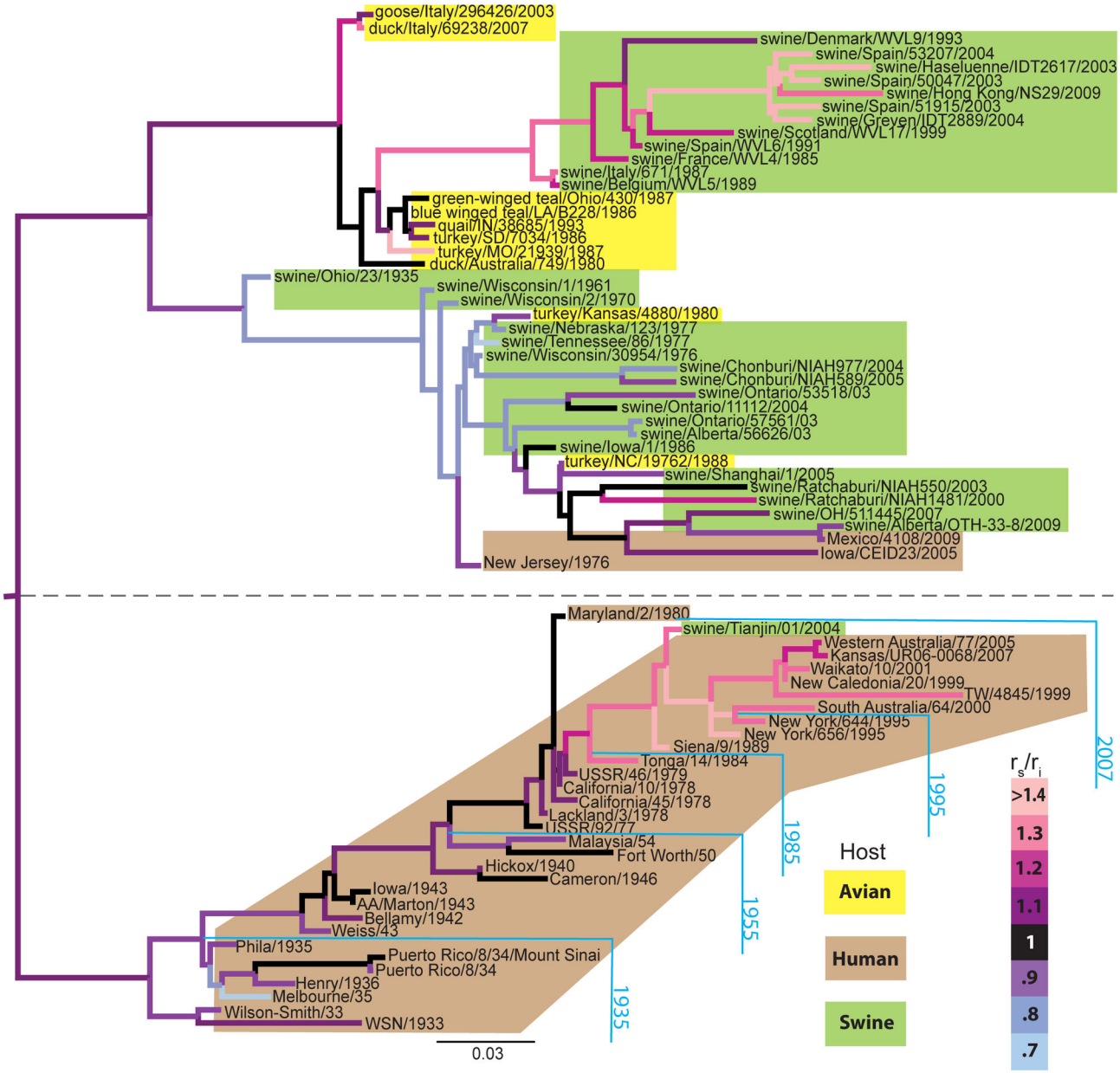


Figure 1. Phylogenetic relationships, species derivation and relative evolutionary rates for 75 accessions of H1N1 influenza. Shown is the topology inferred for the HA protein (see subsection *Inference of patterns of molecular evolution for surface* and *interior residues* in *Methods*); other proteins show somewhat differing relationships (Supporting Information). We also show the ratio of surface-to-interior amino acid substitutions (r_e/r_i), calculated as the difference between the branch lengths estimated from the exterior and interior residues. Variation in r_e/r_i is illustrated from low to high with colors from blue to pink. Each colored box represents the organism of origin: Avian (yellow), Human (beige), and Swine (green). We note that the lower clade (separated by a dashed line) is composed almost entirely of human-derived strains, with the exception of one swine accession (Tianjin/01/2004). This clade also shows a fairly clear timeline (cyan). The upper clade, however, does not give such clear indications of timing.

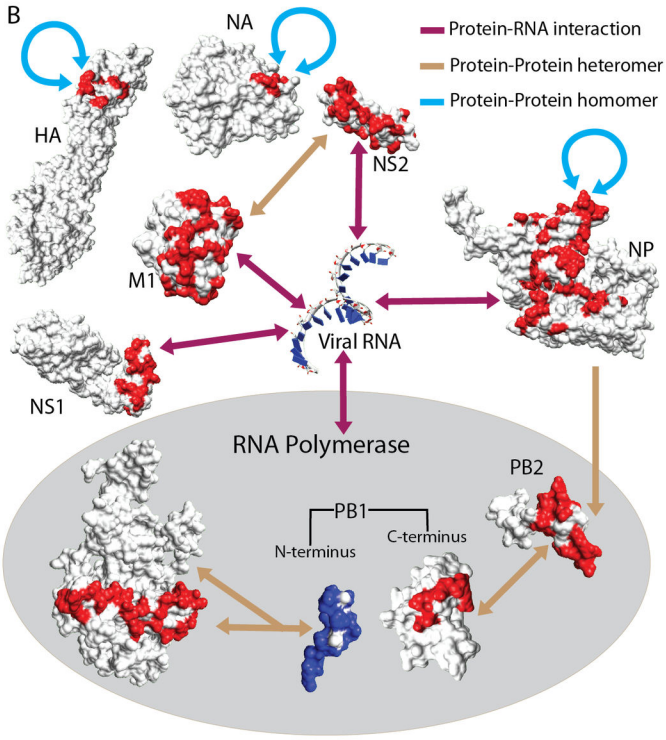
doi: 10.1371/journal.pone.0081027.g001

interactions [53] resulted in 8 proteins with regions that had been previously functionally described in the literature (17 papers in total) and 4 proteins that contained regions characterized by structural data (5 PDB structures in total) (Table 2). Even though each protein contained a significant region, some proteins had regions that required structural

A

Protein	Interaction type	Patch Size	P-value	Connectivity
HA	Homotetramer	9	0.005	2.4
M1	Hetrodimer and RNA binding	32	0.05	2.8
NA	Heterotetramer	4	0.05	1.5
NP	Homodimer, Heterodimer, and RNA binding	50	0.00005	3.0
NS1	RNA binding	14	0.00005	3.7
NS2	Heterodimer and RNA binding	20	0.06	3.8
PA	Heterotrimer/ RNA Polymerase formation	31	0.001	3.5
PB1	Heterotrimer/ RNA Polymerase formation	7	0.08	3.1
		11	0.3	4.4
PB2	Heterotrimer/ RNA Polymerase formation	13	0.5	5.1

B



C

Protein patches	HA	M1	M2	NA	NP	NS1	NS2	PA	PB1	PB2
Significant	3	1	1	1	2	1	1	5	3	3
Insignificant	3	2	0	2	5	1	0	7	1	8

Figure 2. Conserved regions are exclusively associated with known intra-viral interaction positions. A) Eight of the ten viral proteins have regions that are involved in known intra-viral interactions. For each interaction, we list the type of interaction, the size of the patch, $E_n(x)$, and the patch connectivity. We determine $E_n(x)$ as the expected number of randomly generated regions of a given size. We calculate the connectivity of the regions as the average number of neighbors each residue has in the patch. The color of the three right-most columns match to the color of the regions in panel B. B) Each of the eight proteins forms a unique interaction with (i) a copy of itself (indicated by a blue arrow), (ii) viral RNA (purple arrow), or (iii) another viral protein (tan arrow). Some conserved regions participate in more than one interaction. A uni-directional arrow indicates an interaction occurring between two proteins, but is not necessarily characterized by conserved regions on both proteins. The three proteins of RNA polymerase, PA, PB1, and PB2, are grouped by a grey oval. Shown is the interaction between the polymerase complex and the viral RNAs. C) The distribution of significant ($E_n(x) \leq 0.05$), marginal ($0.05 < E_n(x) \leq 0.1$), and insignificant ($E_n(x) > 0.1$) regions across all ten proteins.
doi: 10.1371/journal.pone.0081027.g002

explanation, such as NS2, PB1, and PB2. The distributions of random patch sizes obtained for these proteins did not fit well using an exponential distribution. Specifically, the distribution of random patch sizes for NS2 closely resembled a linear stepwise function, and for structurally modeled fragments of PB1 and PB2 the underlying distributions favored the regions of maximum size. This can be explained by the large percentage of surface residues that are classified as conserved. Indeed, since a large number of surface residues are conserved, it is difficult to create several isolated regions of small size; thus, the typical regions are large. For M1, we also obtained a random patch size distribution, which appeared almost exponential with the exception of an additional peak. Finally, the M2 protein, with a similarly high substitution rate, had a significant region on its surface. However, the small size of the M2 structural model covers only part of its sequence, possibly giving rise to a spurious patch. The location of the modeled structure in the transmembrane region increases the likelihood of existence of such a patch.

Intriguingly, the functional annotations of the significant regions reveal that all the regions are exclusively associated with the intra-viral protein-protein and protein-RNA interactions (Figure 2B), with the exception of a single residue from a region on NP (Table S2). The protein-protein interactions include both homomers (self-interactions of proteins M1 [3], M2 [5,6], NP [10], and NS1 [11]) and heteromers (interactions mediated by proteins M1 [2], NP [8], NS2 [12,13], PA [15,16], PB1 (PDB: 2ZNL, 3A1G), and PB2 (PDB: 3A1G)). Several of these proteins, including M1 [2], NP [9], NS1 [11], NS2 [14], PB2[31], also had significant surface regions associated with

protein-RNA interactions. The conserved regions share several interesting properties. First, we found that all interactions involved in the assembly of the RNA-polymerase complex included at least one region of extreme conservation. Second, while regions usually occurred on only one binding site of the interaction interface, we also found protein-protein interactions with the regions included in both binding sites (interactions between proteins PB1 and PA [16] (PDB: 2ZNL), PB2 and PB1 (PDB: 3A1G), and M1 and NS2 [2,12,13]). Finally, we found that NS2 had a conserved region annotated with multiple functions: the region from residues 65-72 is involved in both viral RNA and M1 binding, while residues 74-79 are involved exclusively in M1 binding (Table 2).

The inferred regions were only slightly affected when we additional sequences were introduced to the original non-redundant set of 75. Specifically, it took between 900 and 1000 sequences to introduce even a single non-synonymous polymorphism into a single influenza protein. Most strains experienced exactly one such mutation across their entire proteome. A particular set of outlying strains caused at most 5 polymorphisms and affected at most 3 different proteins (Table S3). This set contains 14 proteomes that can be grouped by geographic location and close years, and within these groups, the sequence identity ranges from 97% to 100%. This indicates that these grouped sequences are in the same redundancy cluster, during the redundancy removal procedure and thus could have only a minimal effect, if any, on the analysis.

Regions of extreme conservation in 1918 and 2009 pandemics

Following the findings by Xu et al [33], which identified nearly identical functional sites shared between HA proteins of the 1918 and 2009 H1N1 pandemics, we compared our identified regions of extreme conservation across strains from both pandemics. Notably, all identified regions across all proteins were identical between the 1918 and 2009 strains. This finding is in agreement with the fact that the 2009 swine origin pandemic influenza A virus is thought to originate from a recent inter-species reassortment from swine to human, and another observation that same extreme regions were found not only between human H1N1 strains but also across swine and avian strains.

We finally sought to understand the relationship between the identified regions of extreme conservation and the evolutionary dynamics of the virus when treated with antiviral drugs. Specifically, we used recently reported viral population data obtained from an immunosuppressed patient infected with 3 variants of H1N1/2009 influenza and treated with neuraminidase inhibitors [54]. The data included a set of ten mutation sites from four proteins obtained using a deep sequencing approach: HA (Val_6, Asn_{55}, Val_{125}, Thr_{220}), NA (Ile_{106}, Asp_{199}, Asp_{248}, His_{275}), NP (Ile_{100}), and NS1 (Ile_{123}). These sites were mapped onto the homology models of the proteins and compared to locations of conserved regions (Figure 3). We found that none of the ten mutation sites belonged to any of the conserved regions. Interestingly, NA's mutation site Ile_{106} was in close proximity to residues 107 and 108, which belonged to a conserved region. However, the mutation reported at this position (I106V) [54] is unlikely to cause any changes in the function associated with the conserved region due to similar properties of the residues.

Discussion

Overview of the addressed problem and result highlights

The conservation of functionally important residues on protein surfaces has been well documented [55,56]. In particular, several studies, both general and targetting specific protein families, determined the sequence and structure conservation of residues in the protein binding sites mediating intra-species protein-protein interactions [55,57,58]. However, the impact of the purifying selection on the protein binding sites in viral proteins is not clear, due to the intrinsic relationship between the intra-viral and viral-host protein-protein interactions. Unexpectedly, we gained new insight into the evolution of viral binding sites while addressing more general questions related to influenza protein evolution. The first question is whether the surface residues of the proteins evolve faster than the core residues, and whether this is seen equally across all influenza proteins. The second question is whether, in spite of the rapid evolution of surface residues in influenza proteins, there are any "extreme" protein regions that are fully conserved. To answer these questions, our approach integrated the data from evolutionary genomics, structural bioinformatics, and deep sequencing. The developed automatic pipeline (Figure 4) has allowed for the first time to detect statistically significantly conserved regions in the entire influenza proteome that are structurally connected but may not necessarily be sequentially contiguous. The pipeline is readily available to study proteomes of other viral families.

Evolutionary dynamics of H1N1 and our hypothesis

It was recently shown that reassortment with swine strains resulted in nearly identical regions of conserved antigenic residues in HA protein of the 1918 and 2009 H1N1 strains [33,59]. However, that conservation is in striking contrast to the 50% sequence divergence between strains from 2007 and the 1940's [33] and appears the result of the replacement of H1N1 genes from the human strains with those from swine strains, which are much slower evolving in the protein sequence [32]. This combination of rapid evolution and reassortment is the principal reason for the lack of conserved regions around the HA antigenic sites, when considering H1N1 strains of different years. The result points to a more general conclusion: the evolutionary conserved surface regions, should any exist, are unlikely to occur in the regions mediating the viral-host interactions, for which the host proteins may be subject to selection against viral replication. Indeed, host-viral interactions may give rise to Red-Queen/arms-race type dynamics [60].

Insights to obtained exterior-to-interior evolutionary rates across different proteins

In addition to confirming a higher rate of evolution on the surface of viral proteins when compared to the interior, our

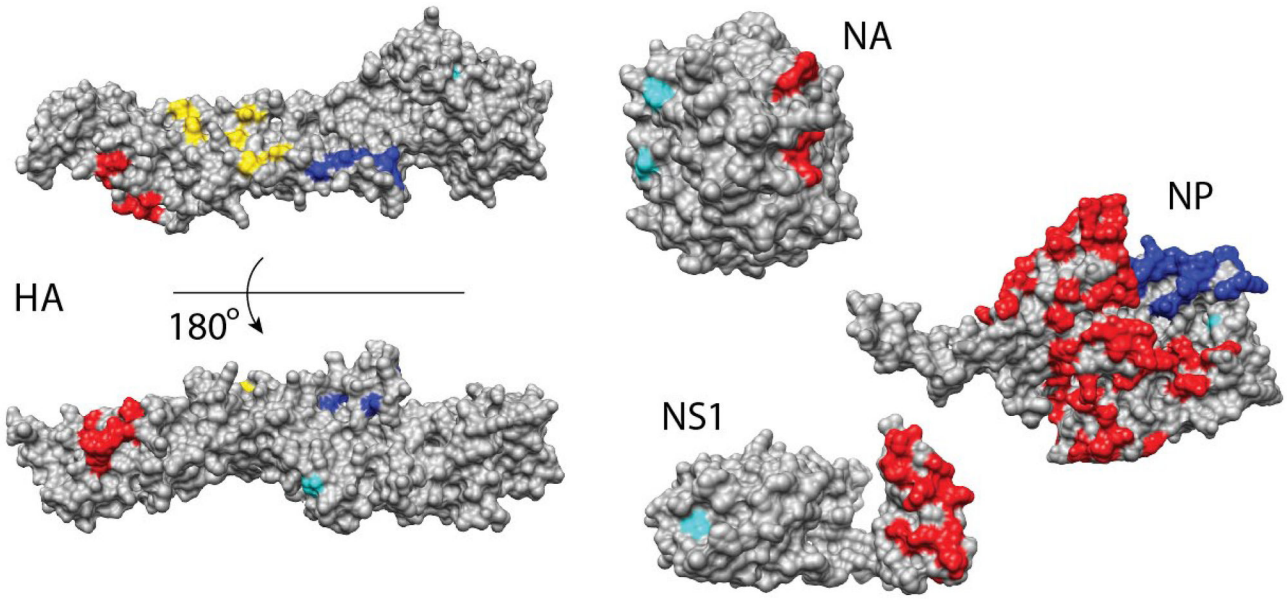


Figure 3. Genetic variation of the viral population data obtained from a patient does not affect regions of extreme conservation. Shown are ten mutation sites (cyan) from four proteins, HA, NA, NS1, and NP, obtained using a deep sequencing approach. The mutation sites were mapped onto the structural models and their locations were compared to the conserved regions. Individual regions of extreme conservation were coloured red, blue and yellow.
doi: 10.1371/journal.pone.0081027.g003

phylogenetic study revealed signals of viral reassortment in influenza strains from other hosts {Brown, 2000 #37;Li, 2005 #1580;. As a result, each protein has a unique gene-tree topology (although we did not assess the phylogenetic uncertainty inherent in these trees, since the tree inference was not a primary goal of our study). The source of the variation in exterior-to-interior residue rate ratios (r_e/r_i) is less straightforward to explain. While most values were between 1.1 and 1.5, PA (1.9), M1 (2.2), and M2 (2.3) were significantly higher. One possible reason is that PA and M2 were both incomplete structures, thus residues that are buried in the full structure could be assigned as "exterior" residues. Thus, structural data for M2 was limited to the helix-linker-helix structural fragment of the transmembrane region, resulting in 33 “exterior” residues and only 5 “interior” residues, even though all of these residues would be buried in a membrane *in vivo*.

Structure-based phylogenetic analysis provides insights into the multi-species evolution of H1N1 virus

Using structure-driven phylogenetic analysis, we found that the human lineage of HA and NA phylogenetic trees of the H1N1 virus had a trunk-like structure while swine and avian lineages did not, indicating that the topological diversities of phylogenetic trees for H1N1 viral proteins can reflect the difference of selective pressures in human and animals. Indeed, due to a longer life span and fewer limitations on geographical barriers, the human influenza virus can be further exposed to herd immunity. As a result, one strain can be easily circulated globally. On the other hand, multiple sublineages of influenza viruses can be co-circulating in different and geographically separated animal populations. In contrast to the surface proteins, the human lineage of internal H1N1 proteins, e.g. PA, do not have trunk-like structures. This is likely due to the frequent reassortments {Nelson, 2008 #1539;Zinder, 2013 #1540}, and these proteins can have different animal origins and evolutionary histories.

The fact that there are viruses from multiple hosts located at the same lineage indicates frequent bi-dimensional transmission of influenza A viruses at human-animal, and animal-animal interfaces. For example, Mexico/2009, Iowa/2005, and New Jersey/1976 are three well-documented swine-origin influenza A viruses [61–63]. Nevertheless, the comparative analysis of the structural patterns in the phylogenetic trees of individual proteins suggests that these reassortments were different in their nature: for HA, all three strains are clustered together within North American swine lineage; for NA, Iowa/2005 and New Jersey/1976 strains are clustered with North American, while Mexico/2009 is clustered with European clade; finally for PA, Iowa/2005 and Mexico/2009 are clustered with a larger clade that includes avian and European swine lineages, while New Jersey/1976 is clustered together with other human strains.

An interesting feature of the human lineage is that the surface-to-core ratios of HA and NA proteins have increased significantly since 1984 (Figs.1, S1). Such increase could be due to H1N1-specific herd immunity from accumulating infections of H1N1 since 1977 as well as vaccine-derived immunity, as the first nation-wide vaccination was introduced in the U.S. at the end of 1976 [64,65]. This observation was only among the surface proteins HA and NA, but not internal

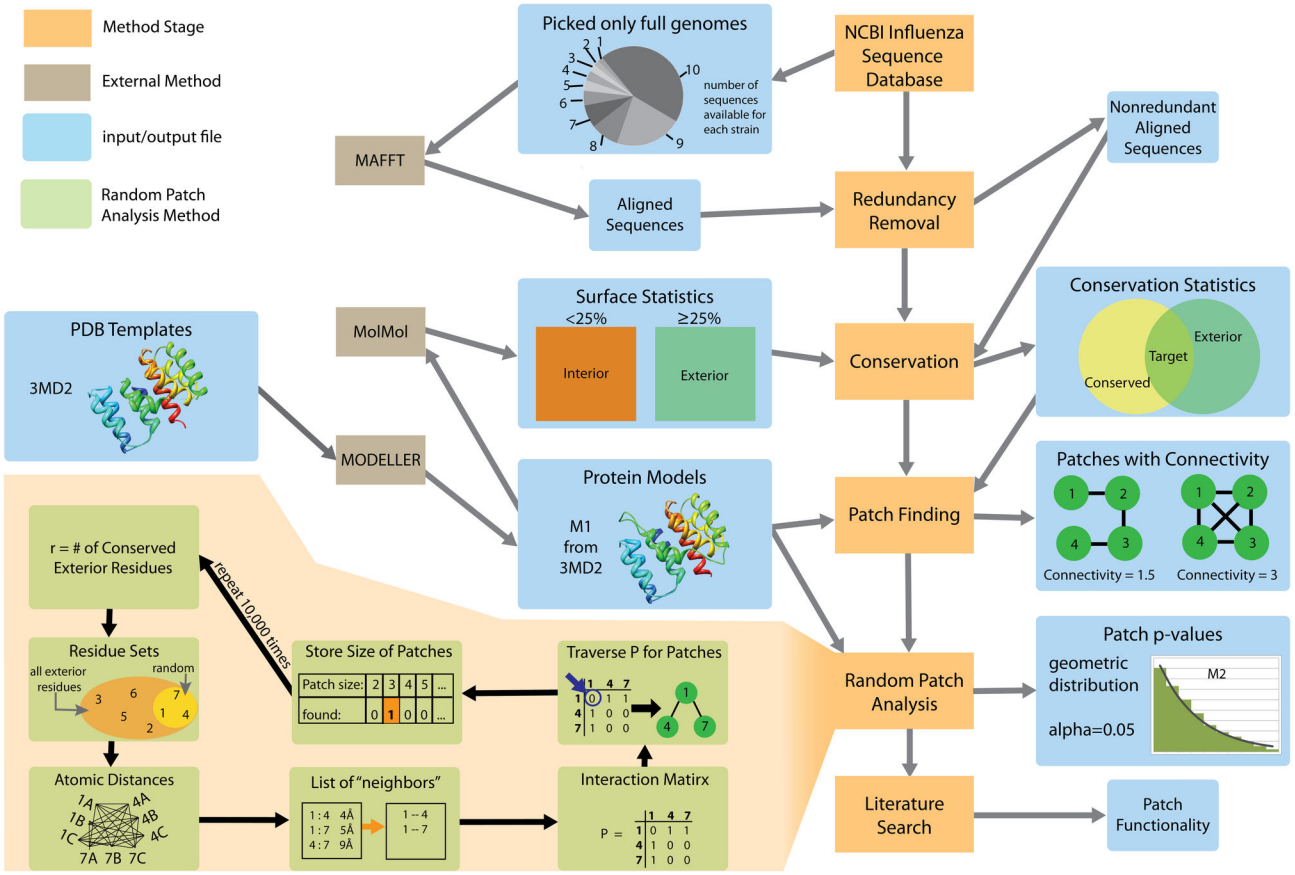


Figure 4. Method pipeline. Our conserved patch analysis method consists of six stages (orange boxes): data collection, redundancy removal, conservation detection, patch finding, random patch analysis, and functional annotation of the conserved regions. The method integrates data from multiple sources (blue) and employs four previously developed software packages (grey): MAFFT, MolMol, and MODELLER. The random patch analysis stage is described in more detail (peach).

doi: 10.1371/journal.pone.0081027.g004

proteins because HA and NA are the primary target of the immunological system.

Finally, when comparing the surface-to-core rates between Eurasian and North American swine lineages, two differences were noticed. The first difference, the fact that Eurasian swine lineage is clustered together with the avian lineage, while North American swine lineage is not, can be explained by the well-documented multiple transmission events of the avian H1N1 virus to pigs in continental Europe and later in Asia [66–68]. The second difference, the consistently higher surface-to-core ratios in Eurasian swine lineage, compared to the North American lineage, has not been previously reported. One explanation may be that unlike the classical swine flu in North American lineage, the swine influenza virus from the European lineage, once transmitted from the avian host, required fast adaption to the swine host. In addition, the rate difference may be associated with the suggested difference in epizootiology between the U.S. and European swine influenza, since in Europe herds may harbor the virus while showing no clinical symptoms [69,70]. A further analysis with a more detailed reassortment history between the avian and swine lineages may be required to confirm this hypothesis.

We note that sampling bias of the strains could also be an influencing factor in our analysis, since it is one of the most common problems in influenza sequence analysis in general. For instance, the most diverse of large clusters of similar strains defined during the redundancy removal is likely to have more random mutations than those of small clusters. Thus, the higher r_e/r_i ratio of the Eurasian swine lineage compared with the North American lineage could be a byproduct of sampling bias. Unfortunately, sampling bias is difficult to avoid, so one should be cautious not to overinterpret the changes of r_e/r_i ratio over time in such cases. To handle sampling bias, several approaches could be explored in the future. For instance, one could look at the correlation of the r_e/r_i ratios of strains with the number of redundant strains they represent or at the average values of r_e/r_i ratio per year versus number of samples in the dataset before and after redundancy removal for the same year.

Effects of positive and negative selection on the protein surface in H1N1 proteins

While all of the virus proteins are subject to evolutionary change, the extent to which each protein allows certain changes depends on several factors such as location of the protein in the virion, the protein's function, and the fact that some genes are encoded on the same genomic segment. For instance, HA is expressed on the surface of the virion, is involved in host binding, and is located on its own gene segment [71]. Thus, HA is subject to a stronger selective pressure compared to the internal proteins, such as M1, which serves a structural purpose as well as RNA binding, and shares a coding region with M2 [71]. Because of this shared coding region, each mutation risks causing a detrimental change in the other gene. There is also variation within a given protein: HA's antigenic sites are subject to positive selection due to host immune pressure, yet the stem region is subject to purifying selection due to its role in trimer formation. This mixed selection is seen in essentially all of the proteins: there exist regions that are subject to positive selection due to their role in viral-host interactions and there exist regions that are subject to negative selection due to their role in intra-viral interactions.

High conservation of H1N1 functional regions have been previously reported

There have been several studies that have found high but not necessarily 100% conserved regions on the surfaces of the influenza proteins. For instance, it has been found that the dsRNA binding track of NS1 consists of conserved binding residues [72]. Additionally, the conservation of the surface regions has been determined near the stem region of HA protein, [73]. Since HA evolves considerably faster than NS1, it is of note that both of these structures are known to have conserved binding regions. The regions found in the present study overlapped with regions identified, experimentally, to be conserved but did not overlap with them entirely.

Analysis of extremely conserved regions

In concert with the above findings, we found that all of the detected conserved regions were associated with the intra-viral macromolecular complexes, including protein homomers, heteromers, or protein-viral RNA interactions. Interestingly, each region covered a part of, but never the entire, binding site. This type of co-localization suggests that though most of an intra-viral binding site is conserved, variable residues exist perhaps under weaker selective pressure than their conserved neighbors. In the case of M1, NP, and NS2, the conserved regions are co-localized with multiple binding sites. Note that each of these interactions buries the exposed residues of conserved regions in the interaction interface, effectively making them the interior residues. However, while some interactions are more long-term than others, none are bound for the entire viral life cycle. In contrast to the situation with host-viral interactions, natural selection is expected to stabilize intra-viral interactions [74], which accounts for their conservation. Alternatively, there could be co-evolution between the interacting residues, such as found in some host-viral interactions [75,76]. While each significant region has been associated with at least one known functional region, there are portions of each region that do not overlap with any functional sites. Those regions may be involved in undiscovered intra-viral interactions. This hypothesis is plausible, given that very few known interactions have been comprehensively characterized on the residue level. The geographic scale and time scale, together with the degree of observed extreme conservation in the influenza proteins allows one to suspect that these conserved regions would also occur across viral strains in any given year. Consistent with this hypothesis, our mapping of genetic variation obtained from an individual carrying three genetic variants from two distinct phylogenetic clades did not find a single mutation in any of the conserved regions. However, further studies involving multiple subjects and larger viral populations are necessary to provide a stronger linkage between the temporal and population-wise conservation of the functional regions in influenza proteins.

Our findings may provide insights into new influenza drug targets

Attaining total protection against Influenza A virus through the development of universal antivirals and vaccines has been a challenging task due to the increasing resistance to the treatments of new viral strains as well as the enormous diversity of the viral population. Recently, a number of promising approaches have been identified, including human monoclonal antibodies and antivirals inhibiting the activity of influenza proteins. Both vaccines and antiviral are capable of neutralizing a wide range of influenza A and often B strains [77–83], but they have been focused thus far on only a few protein targets: the vaccines for HA and antivirals for M2 and NA. Moving beyond these targets, the design of new protein inhibitors of influenza polymerase has been recently suggested as a potential direction in the development of new antivirals [84]. Our study may provide further insight towards identifying new protein targets for influenza antivirals or antibodies, pinpointing the key binding regions that are conserved across a wide range of current and past influenza strains and thus likely to be preserved in future strains. One example from our data is the PB1 to PB2 interaction, which, if disrupted, could result in the loss of viral RNA replication function [85]. One of the main challenges in targeting the regions of extreme conservation, however, comes from their intrinsic property: the regions become inaccessible upon intra-viral macromolecular interactions. Understanding the dynamics of such interactions may provide further insight into this challenge as well as the evolutionary mechanisms behind the extreme conservation.

Methods

Data selection and alignment

Our data selection protocol was carried out in three stages. First, a set of 1,100 complete genomes of H1N1 influenza was selected from the NIH Influenza Virus Resource (Table S1). All 100% identical sequences were filtered. Because most genomes had only PB1-F2 sequence fragments, we chose to use only the other ten proteins. During the second stage, the redundant strains were identified: we defined two strains as

redundant if the sequence identity for each of the ten pairs of proteins was greater than 95%. Sequence identity was calculated based on the sequence alignment obtained using MAFFT [86]. Finally, the strains were clustered into redundancy clusters, relative to their redundancy with each other, and a representative was selected for each cluster, resulting in 75 non-redundant strains (Table S1). Using the remaining 1,025 sequences, we analyzed how the addition of sequences to the non-redundant set of 75 affected site-specific conservation. This analysis was also done using multiple sequence alignment software MAFFT [86].

Protein Structure Prediction and Surface Analysis

The accurate identification of the surface residues for each influenza protein is a critical step in our approach. The ideal method for inferring each surface residue is to compute a homology model for each protein sequence and using the model structure to define the accessible surface residues. However, making such inferences for each sequence is computationally expensive. Therefore, in our protocol, a single target sequence was randomly chosen from the selected strains of each of the ten proteins, and a corresponding protein structure was predicted using the comparative modeling software MODELLER (Table 3) [87]. Next, for each modeled protein, we identified exterior residues using the CalcSurface subroutine. This routine calculates the solvent accessible surface area (SASA) using the MolMol software package [88]. Residues with a SASA greater than 25% were defined as exterior. This threshold has been previously used to identify a protein's surface residues [89]. Finally, the surface residues of the remaining 74 strains were mapped from the modeled strain using sequence alignment.

Table 3. Coverage and sequential similarity of protein templates.

Protein	Strain	Template	Template subtype	Template similarity, %	Residues covered
HA	A/Fort Worth/50	1H0A (A)	-	45	19-517
M1	A/Iowa/1943	1AA7 (A)	-	97	1-158
M2	A/Iowa/1943	2KIH (A)	H5N1	89	23-60
NA	A/Iowa/1943	3B7E (A)	H1N1	91	83-467
NP	A/swine/Alberta/ OTH-33-8/2009	2Q06 (A)	H5N1	93	28-502
NS1	A/Fort Worth/50	3F5T (A)	H5N1	90	5-202
NS2	A/Iowa/1943	1PD3 (A)	H1N1	100	68-116
PA	A/Iowa/1943	3HW3 (A)	H5N1	96	1-193
		2ZNL (A)	H1N1	96	239-699
PB1	A/Iowa/1943	2ZNL (B)	H1N1	100	1-15
		3A1G (A)	H1N1	95	686-736
PB2	A/Iowa/1943	2ZTT (B)	H1N1	94	1-36
		2VQZ (A)	H3N2	95	318-457
		3R2V (A)	H3N2	93	538-720

To make the protein models, we first selected a sequence and one or more templates for each protein (many of the proteins needed multiple structures in order to cover most of the sequence). To select the templates for PA, PB1, and PB2 we chose the PDB references with the highest coverage and best resolution. For the others, we used MODWEB, which will automatically pick the best template. We picked the sequence (or strain) based on the sequence alignment. We generally selected either the sequence with the least number of gaps or the smallest number of unique gaps. The sequence similarity between the template and sequence is significantly high due to the high conservation between strains (Table 1).

doi: 10.1371/journal.pone.0081027.t003

Inference of patterns of molecular evolution for surface and interior residues

Using the structural information obtained from the comparative modeling, we explored the difference between patterns of sequence evolution of the proteins' interior and surface residues. Specifically, we fit three models of sequence evolution to these data using maximum likelihood, as implemented in the HyPhy software package [90]. The first and most restrictive model $M_{uniform}$ requires that the estimated branch lengths of the surface and interior partitions be identical. Thus, this model allows for no overall difference in the rate of evolution between the surface and interior residues. In the second model, M_{scaled}, we relaxed this assumption slightly to allow two partitions to have branch lengths that differ by a scaling constant α. Thus, each branch length for the surface partition is multiplied by α (generally <1.0) to give a corresponding length for the interior partition. In the third model, $M_{arbitrary}$, the branch lengths of the two partitions are estimated completely independently. We note that our models do not explicitly take into account rate heterogeneity. Phylogenetic analyses typically treat rate heterogeneity as a poorly understood nuisance parameter [91]. However, as we have previously discussed, a significant contributor to this variation is the variation between surface and interior residue selective constraint [92,93] that we have accounted for with our structural models.

These three models are nested with respect to each other, with model $M_{uniform}$ being a special case of both model M_{scaled} (when α=1.0) and $M_{arbitrary}$ (when the paired branch lengths for the two partitions are equal). We can thus use a likelihood ratio test [94] to ascertain whether M_{scaled} constitutes a statistically significant improvement over the null model $M_{uniform}$. The likelihood ratio compares the difference in log-likelihood between the two models to a chi-square distribution, where the number of degrees of freedom of that distribution is given by the number of excess parameters in the alternative model. For M_{scaled}, the parameter α adds one degree of freedom. Therefore, if the above test shows significant improvement for M_{scaled}, one can then explore whether the model may be further improved by allowing each branch to differ between the surface and interior residues (i.e., model $M_{arbitrary}$). We again used the likelihood ratio test: in this case there are 146 extra parameters, corresponding to the 147 extra branch lengths in $M_{arbitrary}$, minus the unnecessary α parameter.

Automated conservation analysis pipeline

We next developed an automated computational pipeline to determine structurally conserved protein regions and assess their statistical significance. This pipeline was applied to study the extremely conserved regions of the H1N1 proteome (Figure 4) and consists of four basic steps. We first determine the conserved residues shared between a set of representative protein sequences. Second, we use the homology models of the H1N1 proteins to filter out the conserved residues in the core of each protein. Third, we cluster the remaining residues that are fully conserved on the surface into regions. Finally, we determine the statistically significant regions by employing a random model that generates surface regions with similar properties. The process is further described below.

To identify the regions of extreme conservation, we aligned the set of 75 representative sequences for each of the 10 proteins and determined which of the surface residues were 100% conserved across all 75 sequences. Next, we calculated the Euclidean distance between all pairs of 100% conserved exterior residues. Pairs of conserved residues that were no farther than 6Å apart were defined as structural neighbors. The neighborhood relationship was then summarized as a binary contact matrix of a graph, and the whole set of surface residues were represented as a neighborhood graph with edges designated by the contact matrix. Finally, the surface residues were clustered into regions by defining each connected component of the neighborhood graph to be a cluster. In addition, for each region we calculated its size, contributing surface residues, and residue connectivity. The residue connectivity is defined as an average number of edges per vertex in the neighborhood graph.

To assess if the sizes of the observed regions were larger than expected by chance, we generated a sample of random patches using the corresponding MODELLER subroutine [95]. For each sample, the procedure randomly selects the same number of unique surface residues as conserved surface residues on the protein structure. We then apply the same clustering algorithm as the one discussed above to each of the randomly generated samples, obtaining a patch of neighboring residues and determining the size of the patch. We repeated this procedure 10,000 times (the number is selected as a trade-off between the sample size and computational time of the random trial procedure), yielding a distribution of patch sizes expected for randomly selected groups of exterior residues. The conserved regions obtained from the real data were compared against this distribution, identifying significant regions. Specifically, we determined the *P*-value for each region size using a geometric distribution with a weighted average:

$$P-value = 1-\left(1-\frac{1}{avg(X)}\right)^{n}, \quad avg\left(X\right)=\frac{\sum_{n=1}^{L} n*y_n}{\sum_{n=1}^{L} y_n},$$

where X is the set of all random patches and frequencies, n is conserved region size, y_n is the frequency of a patch of size n, and L is the largest possible patch. For this weighted average, we also considered patches of size 1, the residues that were isolated after clustering. The addition of these residues was necessary for understanding the underlying distribution. The distribution appeared exponential, however since the distribution was of discrete values, we decided that a geometric distribution was a better choice. We then defined a region as significant if the *P*-value was less than 0.05.

Each statistically significant region was functionally annotated. Specifically, we mapped intra- and inter-species binding sites of the H1N1 influenza proteins collected from our database of macromolecular interactions DOMMINO [53] and PubMed literature search, and then determined if each of the conserved regions overlap with any of the mapped binding sites (see Table 2).

Supporting Information

Text S1. Information on conserved regions in influenza proteins.
(DOCX)

Figure S1. Phylogenetic relationships, species derivation and relative evolutionary rates inferred for the NA protein based on 75 accessions of H1N1 influenza. Shown are the inferred topology and the ratio of surface-to-interior amino acid substitutions (r_e/r_i), calculated as the difference between the branch lengths estimated from the exterior and interior residues. The coloring scheme is the same as in Figure 1.
(TIF)

Figure S2. Phylogenetic relationships, species derivation and relative evolutionary rates inferred for the PA protein based on 75 accessions of H1N1 influenza. Shown are the inferred topology and the ratio of surface-to-interior amino acid substitutions (r_e/r_i), calculated as the difference between the branch lengths estimated from the exterior and interior residues. The coloring scheme is the same as in Figure 1.
(TIF)

Figure S3. Phylogenetic relationships and relative evolutionary rates inferred for M1. The inferred topology and the ratio r_e/r_i are calculated as in Figure 1.
(TIF)

Figure S4. Phylogenetic relationships and relative evolutionary rates inferred for M2. The inferred topology and the ratio r_e/r_i are calculated as in Figure 1.
(TIF)

Figure S5. Phylogenetic relationships and relative evolutionary rates inferred for NP. The inferred topology and the ratio r_e/r_i are calculated as in Figure 1.
(TIF)

Figure S6. Phylogenetic relationships and relative evolutionary rates inferred for NS1. The inferred topology and the ratio r_e/r_i are calculated as in Figure 1.
(TIF)

Figure S7. Phylogenetic relationships and relative evolutionary rates inferred for NS2. The inferred topology and the ratio r_e/r_i are calculated as in Figure 1.
(TIF)

Figure S8. Phylogenetic relationships and relative evolutionary rates inferred for PB1. The inferred topology and the ratio r_e/r_i are calculated as in Figure 1.
(TIF)

Figure S9. Phylogenetic relationships and relative evolutionary rates inferred for PB2. The inferred topology and the ratio r_e/r_i are calculated as in Figure 1.
(TIF)

Figure S10. Overlap of binding sites and conserved regions of HA and NA proteins. Shown are conserved regions (blue), protein binding sites (gold) and the overlap of binding sites and conserved regions (red).
(TIF)

Figure S11. Phylogenetic relationships and relative evolutionary rates inferred for HA, protein shown with bootstrap values.
(TIF)

Figure S12. Phylogenetic relationships and relative evolutionary rates inferred for NA protein shown with bootstrap values.
(TIF)

Figure S13. Phylogenetic relationships and relative evolutionary rates inferred for M1 protein shown with bootstrap values.
(TIF)

Table S1. A representative set of 75 strains used in our analysis.
(DOCX)

Table S2. Conserved regions with p-value and residues. Each protein is shown with their associated regions, p-values, and residues. Residues in bold are also found in intra-viral binding regions.
(DOCX)

Table S3. Outlying strains and the regions they affect.
(DOCX)

Acknowledgements

We thank Olga Kalinina, Elodie Ghedin, and Sounak Chakraborti for critical reading of the manuscript and useful suggestions. We would also like to thank Laura Boykin and an anonymous reviewer for their detailed and insightful comments during the review process.

Author Contributions

Conceived and designed the experiments: SW GC DK. Performed the experiments: SW. Analyzed the data: SW XFW GC DK. Wrote the manuscript: SW XFW GC DK.

References

1. Stevens J, Corper AL, Basler CF, Taubenberger JK, Palese P et al. (2004) Structure of the uncleaved human H1 hemagglutinin from the extinct 1918 influenza virus. Science 303: 1866-1870. doi:10.1126/science.1093373. PubMed: 14764887.
2. Noton SL, Medcalf E, Fisher D, Mullin AE, Elton D et al. (2007) Identification of the domains of the influenza A virus M1 matrix protein required for NP binding, oligomerization and incorporation into virions. J Gen Virol 88: 2280-2290. doi:10.1099/vir.0.82809-0. PubMed: 17622633.
3. Harris A, Forouhar F, Qiu S, Sha B, Luo M (2001) The crystal structure of the influenza matrix protein M1 at neutral pH: M1-M1 protein interfaces can rotate in the oligomeric structures of M1. Virology 289: 34-44. doi:10.1006/viro.2001.1119. PubMed: 11601915.
4. Phongphanphanee S, Rungrotmongkol T, Yoshida N, Hannongbua S, Hirata F (2010) Proton transport through the influenza A M2 channel: three-dimensional reference interaction site model study. J Am Chem Soc 132: 9782-9788. doi:10.1021/ja1027293. PubMed: 20578761.
5. Stouffer AL, Acharya R, Salom D, Levine AS, Di Costanzo L et al. (2008) Structural basis for the function and inhibition of an influenza virus proton channel. Nature 451: 596-599. doi:10.1038/nature06528. PubMed: 18235504.
6. Wang J, Qiu JX, Soto C, Degrado WF (2011) Structural and dynamic mechanisms for the function and inhibition of the M2 proton channel from influenza A virus. Curr Opin Struct Biol, 21: 68–80. PubMed: 21247754.
7. Kochendoerfer GG, Salom D, Lear JD, Wilk-Orescan R, Kent SB et al. (1999) Total chemical synthesis of the integral membrane protein influenza A virus M2: role of its C-terminal domain in tetramer assembly. Biochemistry 38: 11905-11913. doi:10.1021/bi990720m. PubMed: 10508393.
8. Biswas SK, Boutz PL, Nayak DP (1998) Influenza virus nucleoprotein interacts with influenza virus polymerase proteins. J Virol 72: 5493-5501. PubMed: 9621005.
9. Kobayashi M, Toyoda T, Adyshev DM, Azuma Y, Ishihama A (1994) Molecular dissection of influenza virus nucleoprotein: deletion mapping of the RNA binding domain. J Virol 68: 8433-8436. PubMed: 7966640.
10. Elton D, Medcalf E, Bishop K, Digard P (1999) Oligomerization of the influenza virus nucleoprotein: identification of positive and negative sequence elements. Virology 260: 190-200. doi:10.1006/viro.1999.9818. PubMed: 10405371.
11. Wang W, Riedel K, Lynch P, Chien CY, Montelione GT et al. (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. doi:10.1017/S1355838299981621. PubMed: 10024172.
12. Darapaneni V, Prabhaker VK, Kukol A (2009) Large-scale analysis of influenza A virus sequences reveals potential drug target sites of non-structural proteins. J Gen Virol 90: 2124-2133. doi:10.1099/vir.0.011270-0. PubMed: 19420157.
13. Akarsu H, Burmeister WP, Petosa C, Petit I, Müller CW et al. (2003) Crystal structure of the M1 protein-binding domain of the influenza A virus nuclear export protein (NEP/NS2). EMBO J 22: 4646-4655. doi:10.1093/emboj/cdg449. PubMed: 12970177.
14. Boulo S, Akarsu H, Ruigrok RW, Baudin F (2007) Nuclear traffic of influenza virus proteins and ribonucleoprotein complexes. Virus Res 124: 12-21. doi:10.1016/j.virusres.2006.09.013. PubMed: 17081640.
15. Hemerka JN, Wang D, Weng Y, Lu W, Kaushik RS et al. (2009) Detection and characterization of influenza A virus PA-PB2 interaction through a bimolecular fluorescence complementation assay. J Virol 83: 3944-3955. doi:10.1128/JVI.02300-08. PubMed: 19193801.
16. Zürcher T, de la Luna S, Sanz-Ezquerro JJ, Nieto A, Ortín J (1996) Mutational analysis of the influenza virus A/Victoria/3/75 PA protein:

studies of interaction with PB1 protein and identification of a dominant negative mutant. J Gen Virol 77 (8): 1745-1749. doi: 10.1099/0022-1317-77-8-1745. PubMed: 8760421.
17. Nelson MI, Holmes EC (2007) The evolution of epidemic influenza. Nat Rev Genet 8: 196-205. doi:10.1038/nrg2053. PubMed: 17262054.
18. Tsai KN, Chen GW (2011) Influenza genome diversity and evolution. Microbes Infect 13: 479-488. doi:10.1016/j.micinf.2011.01.013. PubMed: 21276870.
19. Guilligay D, Tarendeau F, Resa-Infante P, Coloma R, Crepin T et al. (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. PubMed: 18454157.
20. Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK et al. (2008) The genomic and epidemiological dynamics of human influenza A virus. Nature 453: 615-619. doi:10.1038/nature06945. PubMed: 18418375.
21. Khatchikian D, Orlich M, Rott R (1989) Increased viral pathogenicity after insertion of a 28S ribosomal RNA sequence into the haemagglutinin gene of an influenza virus. Nature 340: 156-157. doi: 10.1038/340156a0. PubMed: 2544809.
22. Orlich M, Gottwald H, Rott R (1994) Nonhomologous recombination between the hemagglutinin gene and the nucleoprotein gene of an influenza virus. Virology 204: 462-465. doi:10.1006/viro.1994.1555. PubMed: 8091680.
23. Mitnaul LJ, Matrosovich MN, Castrucci MR, Tuzikov AB, Bovin NV et al. (2000) Balanced hemagglutinin and neuraminidase activities are critical for efficient replication of influenza A virus. J Virol 74: 6015-6020. doi: 10.1128/JVI.74.13.6015-6020.2000. PubMed: 10846083.
24. Suarez DL, Senne DA, Banks J, Brown IH, Essen SC et al. (2004) Recombination resulting in virulence shift in avian influenza outbreak, Chile. Emerg Infect Dis 10: 693-699. doi:10.3201/eid1004.030396. PubMed: 15200862.
25. Pasick J, Handel K, Robinson J, Copps J, Ridd D et al. (2005) Intersegmental recombination between the haemagglutinin and matrix genes was responsible for the emergence of a highly pathogenic H7N3 avian influenza virus in British Columbia. J Gen Virol 86: 727-731. doi: 10.1099/vir.0.80478-0. PubMed: 15722533.
26. McCullers JA, Saito T, Iverson AR (2004) Multiple genotypes of influenza B virus circulated between 1979 and 2003. J Virol 78: 12817-12828. doi:10.1128/JVI.78.23.12817-12828.2004. PubMed: 15542634.
27. Li C, Yu K, Tian G, Yu D, Liu L et al. (2005) Evolution of H9N2 influenza viruses from domestic poultry in Mainland China. Virology 340: 70-83. doi:10.1016/j.virol.2005.06.025. PubMed: 16026813.
28. Brown IH, Harris PA, McCauley JW, Alexander DJ (1998) Multiple genetic reassortment of avian and human influenza A viruses in European pigs, resulting in the emergence of an H1N2 virus of novel genotype. J Gen Virol 79 (12): 2947-2955. PubMed: 9880008.
29. Matsuzaki Y, Mizuta K, Sugawara K, Tsuchiya E, Muraki Y et al. (2003) Frequent reassortment among influenza C viruses. J Virol 77: 871-881. doi:10.1128/JVI.77.2.871-881.2003. PubMed: 12502803.
30. Widjaja L, Krauss SL, Webby RJ, Xie T, Webster RG (2004) Matrix gene of influenza a viruses isolated from wild aquatic birds: ecology and emergence of influenza a viruses. J Virol 78: 8771-8779. doi: 10.1128/JVI.78.16.8771-8779.2004. PubMed: 15280485.
31. Honda A, Mizumoto K, Ishihama A (1999) Two separate sequences of PB2 subunit constitute the RNA cap-binding site of influenza virus RNA polymerase. Genes Cells 4: 475-485. doi:10.1046/j.1365-2443.1999.00275.x. PubMed: 10526235.
32. Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S et al. (2009) Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. Science 325: 197-201. doi: 10.1126/science.1176225. PubMed: 19465683.
33. Xu R, Ekiert DC, Krause JC, Hai R, Crowe JE Jr. et al. (2010) Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. Science 328: 357-360. doi:10.1126/science.1186430. PubMed: 20339031.
34. Morens DM, Taubenberger JK, Fauci AS (2009) The persistent legacy of the 1918 influenza virus. N Engl J Med 361: 225-229. doi:10.1056/NEJMp0904819. PubMed: 19564629.
35. Ferguson NM, Galvani AP, Bush RM (2003) Ecological and immunological determinants of influenza evolution. Nature 422: 428-433. doi:10.1038/nature01509. PubMed: 12660783.
36. Krystal M, Elliott RM, Benz EW Jr., Young JF, Palese P (1982) Evolution of influenza A and B viruses: conservation of structural features in the hemagglutinin genes. Proc Natl Acad Sci U S A 79: 4800-4804. doi:10.1073/pnas.79.15.4800. PubMed: 6956892.
37. Gorman OT, Bean WJ, Kawaoka Y, Webster RG (1990) Evolution of the nucleoprotein gene of influenza A virus. J Virol 64: 1487-1497. PubMed: 2319644.
38. Noble S, McGregor MS, Wentworth DE, Hinshaw VS (1993) Antigenic and genetic conservation of the haemagglutinin in H1N1 swine influenza viruses. J Gen Virol 74 (6): 1197-1200. doi: 10.1099/0022-1317-74-6-1197. PubMed: 8389804.
39. Heiny AT, Miotto O, Srinivasan KN, Khan AM, Zhang GL et al. (2007) Evolutionarily conserved protein sequences of influenza a viruses, avian and human, as vaccine targets. PLOS ONE 2: e1190. doi: 10.1371/journal.pone.0001190. PubMed: 18030326.
40. Dreyfus C, Laursen NS, Kwaks T, Zuijdgeest D, Khayat R et al. (2012) Highly conserved protective epitopes on influenza B viruses. Science 337: 1343-1348. doi:10.1126/science.1222908. PubMed: 22878502.
41. Ekiert DC, Kashyap AK, Steel J, Rubrum A, Bhabha G et al. (2012) Cross-neutralization of influenza A viruses mediated by a single antibody loop. Nature 489: 526-532. doi:10.1038/nature11414. PubMed: 22982990.
42. Caton AJ, Brownlee GG, Yewdell JW, Gerhard W (1982) The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). Cell 31: 417-427. doi:10.1016/0092-8674(82)90135-0. PubMed: 6186384.
43. Biswas SK, Nayak DP (1996) Influenza virus polymerase basic protein 1 interacts with influenza virus polymerase basic protein 2 at multiple sites. J Virol 70: 6716-6722. PubMed: 8794308.
44. Shapira SD, Gat-Viks I, Shum BO, Dricot A, de Grace MM et al. (2009) A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. Cell 139: 1255-1267. doi:10.1016/j.cell.2009.12.018. PubMed: 20064372.
45. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696-704. doi:10.1080/10635150390235520. PubMed: 14530136.
46. Nelson MI, Viboud C, Simonsen L, Bennett RT, Griesemer SB et al. (2008) Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918. PLoS Pathog 4: e1000012. PubMed: 18463694.
47. Zinder D, Bedford T, Gupta S, Pascual M (2013) The roles of competition and mutation in shaping antigenic and genetic diversity in influenza. PLoS Pathog 9: e1003104. PubMed: 23300455.
48. Greenbaum JA, Kotturi MF, Kim Y, Oseroff C, Vaughan K et al. (2009) Pre-existing immunity against swine-origin H1N1 influenza viruses in the general human population. Proc Natl Acad Sci U S A 106: 20365-20370. doi:10.1073/pnas.0911580106. PubMed: 19918065.
49. Igarashi M, Ito K, Yoshida R, Tomabechi D, Kida H et al. (2010) Predicting the antigenic structure of the pandemic (H1N1) 2009 influenza virus hemagglutinin. PLOS ONE 5: e8553. doi:10.1371/journal.pone.0008553. PubMed: 20049332.
50. Wang Y-T, Chan C-h, Su Z-Y, Chen C-L (2010) Homology modeling, docking, and molecular dynamics reveal HR1039 as a potent inhibitor of 2009 A (H1N1) influenza neuraminidase. Biophys Chem 147: 74-80. doi:10.1016/j.bpc.2009.12.002. PubMed: 20045243.
51. Abdussamad J, Aris-Brosou Sp (2011) The nonadaptive nature of the H1N1 2009 Swine Flu pandemic contrasts with the adaptive facilitation of transmission to a new host. BMC evolutionary biology 11: 6
52. Tharakaraman K, Raman R, Stebbins NW, Viswanathan K, Sasisekharan V et al. (2013) Antigenically intact hemagglutinin in circulating avian and swine influenza viruses and potential for H3N2 pandemic. Sci Rep 3: 1822-. PubMed: 23661027.
53. Kuang X, Han JG, Zhao N, Pang B, Shyu CR et al. (2012) DOMMINO: a database of macromolecular interactions. Nucleic Acids Res 40: D501-D506. doi:10.1093/nar/gkr1128. PubMed: 22135305.
54. Ghedin E, Laplante J, DePasse J, Wentworth DE, Santos RP et al. (2011) Deep sequencing reveals mixed infection with 2009 pandemic influenza A (H1N1) virus strains and the emergence of oseltamivir resistance. J Infect Dis 203: 168-174. doi:10.1093/infdis/jiq040. PubMed: 21288815.
55. Guharoy M, Chakrabarti P (2010) Conserved residue clusters at protein-protein interfaces and their use in binding site identification. BMC Bioinformatics 11: 286. doi:10.1186/1471-2105-11-286. PubMed: 20507585.
56. Panjkovich A, Daura X (2010) Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery. BMC Struct Biol 10: 9. doi:10.1186/1472-6807-10-9. PubMed: 20356358.
57. Ma B, Elkayam T, Wolfson H, Nussinov R (2003) Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci U S A 100: 5772-5777. doi:10.1073/pnas.1030237100. PubMed: 12730379.

58. Swapna LS, Bhaskara RM, Sharma J, Srinivasan N (2012) Roles of residues in the interface of transient protein-protein complexes before complexation. Sci Rep 2: 334. PubMed: 22451863.
59. Wei CJ, Boyington JC, Dai K, Houser KV, Pearce MB et al. (2010) Cross-neutralization of 1918 and 2009 influenza viruses: role of glycans in viral evolution and vaccine design. Sci Transl Med 2: 24ra21. PubMed: 20375007.
60. Woolhouse ME, Webster JP, Domingo E, Charlesworth B, Levin BR (2002) Biological and biomedical implications of the co-evolution of pathogens and their hosts. Nat Genet 32: 569-577. doi:10.1038/ng1202-569. PubMed: 12457190.
61. Gaydos JC, Hodder RA, Top FH, Soden VJ, Allen RG, et al. (1977) Swine Inftuenza A at Fort Dix, New Jersey (January,ÄìFebruary 1976). I. Case Finding and Clinical Study of Cases. Journal of Infectious Diseases 136: S356-S362
62. Neumann G, Noda T, Kawaoka Y (2009) Emergence and pandemic potential of swine-origin H1N1 influenza virus. Nature 459: 931-939. doi:10.1038/nature08157. PubMed: 19525932.
63. Smith GJ, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M et al. (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 459: 1122-1125. doi:10.1038/nature08182. PubMed: 19516283.
64. Fedson DS, Kessler HA (1983) A hospital-based influenza immunization program, 1977-78. Am J Public Health 73: 442-445. doi:10.2105/AJPH.73.4.442. PubMed: 6829828.
65. Sencer DJ, Millar JD (2006) Reflections on the 1976 swine flu vaccination program. Emerg Infect Dis 12: 29–33. PubMed: 16494713.
66. Pensaert M, Ottis K, Vandeputte J, Kaplan MM, Bachmann PA (1981) Evidence for the natural transmission of influenza A virus from wild ducts to swine and its potential importance for man. Bull World Health Organ 59: 75-78. PubMed: 6973418.
67. Schultz U, Fitch WM, Ludwig S, Mandler J, Scholtissek C (1991) Evolution of pig influenza viruses. Virology 183: 61-73. doi:10.1016/0042-6822(91)90118-U. PubMed: 2053297.
68. Guan Y, Shortridge KF, Krauss S, Li PH, Kawaoka Y et al. (1996) Emergence of avian H1N1 influenza viruses in pigs in China. J Virol 70: 8041-8046. PubMed: 8892928.
69. Bachmann PA, editor (1989) Swine influenza virus. Amsterdam, Netherlands: Elsevier. pp. 193-207.
70. Brown IH (2000) The epidemiology and evolution of influenza viruses in pigs. Vet Microbiol 74: 29-46. doi:10.1016/S0378-1135(00)00164-4. PubMed: 10799776.
71. Brown EG (2000) Influenza virus genetics. Biomed Pharmacother 54: 196-209. doi:10.1016/S0753-3322(00)89026-5. PubMed: 10872718.
72. Yin C, Khan JA, Swapna GV, Ertekin A, Krug RM et al. (2007) Conserved surface features form the double-stranded RNA binding site of non-structural protein 1 (NS1) from influenza A and B viruses. J Biol Chem 282: 20584-20592. doi:10.1074/jbc.M611619200. PubMed: 17475623.
73. Okuno Y, Isegawa Y, Sasao F, Ueda S (1993) A common neutralizing epitope conserved between the hemagglutinins of influenza A virus H1 and H2 strains. J Virol 67: 2552-2558. PubMed: 7682624.
74. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. Science 296: 750-752. doi:10.1126/science.1068696. PubMed: 11976460.
75. Korber BT, Farber RM, Wolpert DH, Lapedes AS (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci U S A 90: 7176-7180. doi:10.1073/pnas.90.15.7176. PubMed: 8346232.
76. Travers SA, Tully DC, McCormack GP, Fares MA (2007) A study of the coevolutionary patterns operating within the env gene of the HIV-1 group M subtypes. Mol Biol Evol 24: 2787-2801. doi:10.1093/molbev/msm213. PubMed: 17921487.
77. Corti D, Voss J, Gamblin SJ, Codoni G, Macagno A et al. (2011) A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. Science 333: 850-856. doi:10.1126/science.1205669. PubMed: 21798894.
78. Ekiert DC, Friesen RH, Bhabha G, Kwaks T, Jongeneelen M et al. (2011) A highly conserved neutralizing epitope on group 2 influenza A viruses. Science 333: 843-850. doi:10.1126/science.1204839. PubMed: 21737702.
79. Ekiert DC, Bhabha G, Elsliger MA, Friesen RH, Jongeneelen M et al. (2009) Antibody recognition of a highly conserved influenza virus epitope. Science 324: 246-251. doi:10.1126/science.1171491. PubMed: 19251591.
80. Sui J, Hwang WC, Perez S, Wei G, Aird D et al. (2009) Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. Nat Struct Mol Biol 16: 265-273. doi:10.1038/nsmb.1566. PubMed: 19234466.
81. Throsby M, van den Brink E, Jongeneelen M, Poon LL, Alard P et al. (2008) Heterosubtypic neutralizing monoclonal antibodies cross-protective against H5N1 and H1N1 recovered from human IgM+ memory B cells. PLOS ONE 3: e3942. doi:10.1371/journal.pone.0003942. PubMed: 19079604.
82. Pielak RM, Schnell JR, Chou JJ (2009) Mechanism of drug inhibition and drug resistance of influenza A M2 channel. Proc Natl Acad Sci U S A 106: 7379-7384. doi:10.1073/pnas.0902548106. PubMed: 19383794.
83. Gubareva LV (2004) Molecular mechanisms of influenza virus resistance to neuraminidase inhibitors. Virus Res 103: 199-203. doi:10.1016/j.virusres.2004.02.034. PubMed: 15163510.
84. Chen GL, Subbarao K (2009) Attacking the flu: neutralizing antibodies may lead to 'universal'. Vaccine - Nature Medicine 15: 1251-1252. doi:10.1038/nm1109-1251.
85. Blok V, Cianci C, Tibbles KW, Inglis SC, Krystal M et al. (1996) Inhibition of the influenza virus RNA-dependent RNA polymerase by antisera directed against the carboxy-terminal region of the PB2 subunit. J Gen Virol 77 (5): 1025-1033. doi:10.1099/0022-1317-77-5-1025. PubMed: 8609468.
86. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059-3066. doi:10.1093/nar/gkf436. PubMed: 12136088.
87. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234: 779-815. doi:10.1006/jmbi.1993.1626. PubMed: 8254673.
88. Koradi R, Billeter M, Wüthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 14: 29-55. 8744573.
89. Valkenburg SA, Gras S, Guillonneau C, La Gruta NL, Thomas PG et al. (2010) Protective efficacy of cross-reactive CD8+ T cells recognising mutant viral epitopes depends on peptide-MHC-I structural interactions and T cell activation threshold. PLoS Pathog 6: e1001039. PubMed: 20711359.
90. Pond SL, Frost SD, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676-679. doi:10.1093/bioinformatics/bti079. PubMed: 15509596.
91. Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10: 1396-1401. PubMed: 8277861.
92. Conant GC, Stadler PF (2009) Solvent exposure imparts similar selective pressures across a range of yeast proteins. Mol Biol Evol 26: 1155-1161. doi:10.1093/molbev/msp031. PubMed: 19233963.
93. Conant GC (2009) Neutral evolution on mammalian protein surfaces. Trends Genet 25: 377-381. doi:10.1016/j.tig.2009.07.004. PubMed: 19716195.
94. Sokal R, Rohlf F (1995) Biometry. WH Freeman. New York. p. 887.
95. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, et al. (2006) Comparative protein structure modeling using Modeller. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] Chapter 5: Unit 5 6