Detecting Patches of Protein Sites of Influenza A Viruses under Positive Selection

Christina Tusche,[1,2] Lars Steinbrück,[1,2] and Alice C. McHardy*,[1,2]

[1]Max Planck Research Group for Computational Genomics and Epidemiology, Max Planck Institute for Informatics, Saarbrücken, Germany

[2]Department of Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Institute for Computer Science, Düsseldorf, Germany

***Corresponding author:** E-mail: mchardy@mpi-inf.mpg.de.

Associate editor: Helen Piontkivska

Abstract

Influenza A viruses are single-stranded RNA viruses capable of evolving rapidly to adapt to environmental conditions. Examples include the establishment of a virus in a novel host or an adaptation to increasing immunity within the host population due to prior infection or vaccination against a circulating strain. Knowledge of the viral protein regions under positive selection is therefore crucial for surveillance. We have developed a method for detecting positively selected patches of sites on the surface of viral proteins, which we assume to be relevant for adaptive evolution. We measure positive selection based on dN/dS ratios of genetic changes inferred by considering the phylogenetic structure of the data and suggest a graph-cut algorithm to identify such regions. Our algorithm searches for dense and spatially distinct clusters of sites under positive selection on the protein surface. For the hemagglutinin protein of human influenza A viruses of the subtypes H3N2 and H1N1, our predicted sites significantly overlap with known antigenic and receptor-binding sites. From the structure and sequence data of the 2009 swine-origin influenza A/H1N1 hemagglutinin and PB2 protein, we identified regions that provide evidence of evolution under positive selection since introduction of the virus into the human population. The changes in PB2 overlap with sites reported to be associated with mammalian adaptation of the influenza A virus. Application of our technique to the protein structures of viruses of yet unknown adaptive behavior could identify further candidate regions that are important for host–virus interaction.

Key words: influenza, evolution, selection, adaptation, protein structure, pandemic.

Research article

Introduction

Influenza A viruses are single-stranded negative-sense RNA viruses typically causing short-term respiratory infections with considerable morbidity and mortality (WHO 2009). High mutation rates, swift spreading among individuals, and short replication times allow influenza A viruses to evolve and adapt rapidly to environmental conditions (Pybus and Rambaut 2009). Examples include the establishment of a virus in a novel host or an adaptation to escape increasing immunity of the host population to a circulating or a vaccine influenza strain (Dormitzer et al. 2011).

Past influenza pandemics resulted from the introduction into the human population of a transmissible virus with significantly different antigenicity from recent and currently circulating influenza strains. In all four pandemics that occurred within the last century, the respective influenza viruses carried hemagglutinin (HA) and several other genome segments of influenza A viruses from other host species, such as birds or swine (Webster et al. 1992; McHardy and Adams 2009). Configurational changes of multiple proteins of animal influenza A viruses are thought to be necessary to enable efficient replication and transmission in human hosts (Kuiken et al. 2006; Neumann and Kawaoka 2006). A region of particular importance for this process is the receptor-binding site of the viral hemagglutinin. It enables attachment to different types of host-specific glycosydic bonds on surface epithelial cells in the host respiratory and gastrointestinal tracts (Glaser et al. 2005; Neumann and Kawaoka 2006). Furthermore, certain areas of the viral polymerase complex determine host range (Neumann and Kawaoka 2006; Yamada et al. 2010). Following establishment of a virus within a novel host, additional adaptive changes are thought to optimize replication and dispersal rapidly within the population (Deem and Pan 2009; Hensley et al. 2009; Neumann et al. 2009; Smith et al. 2009).

Human influenza A viruses continuously change antigenically by accumulating changes in the antibody-binding sites of the viral surface proteins HA and neuraminidase (NA), (Bush et al. 1999; Smith et al. 2004; McHardy and Adams 2009; Weinstock and Zuccotti 2009). These changes allow reinfection of previously infected or vaccinated individuals. This requires the composition of the seasonal influenza A virus vaccine to be updated almost annually to ensure its continued effectiveness (Russell et al. 2008). Knowledge of the viral protein regions that are relevant for adaptation to a novel host or an increasingly immune population is therefore a crucial factor for the surveillance

and prevention of seasonal and pandemic influenza A virus infections.

Multiple methods allow identification of functional regions of proteins, for example, on the basis of evolutionary conservation ratios (Pupko et al. 2002; Glaser et al. 2003; Nimrod et al. 2005, 2008; Shazman et al. 2007; Ashkenazy et al. 2010). Regions under positive selection do not follow the assumption of strong conservation and can therefore not be detected by these methods. Other techniques predict the location of antibody-binding (epitope) sites based on structural and sequence information (Blythe and Flower 2005; El-Manzalawy et al. 2008; Rubinstein et al. 2008, 2009; Lacerda et al. 2010). However, besides epitope regions, receptor avidity–changing sites or host-specificity determinants can be subject to positive selection and might play a similarly important role for the adaptive evolution of influenza A viruses (Hensley et al. 2009). Furthermore, a part of the epitope regions is invariable due to functional and structural constraints.

Sites under positive selection indicate the relevance of a region within a protein for adaptation. Such sites can be identified based on the ratio of nonsynonymous to synonymous mutations (d*N*/d*S* ratio) (Bush et al. 1999). This has, for instance, identified regions of B- and T-cell epitopes which are under positive selection (Suzuki 2006). However, this measure is difficult to interpret directly when studying evolution within a population and lacks sensitivity when applied to individual sequence sites (Kryazhimskiy and Plotkin 2008). Other methods compare within-species with between-species substitution statistics or substitution rates at specific branches (Nei 2005; Nozawa et al. 2009). We have recently proposed how to identify individual alleles, or sets of mutations, instead of sites or genes, that might be under selection using a time series of sequence samples from human influenza A (H3N2) viruses (Steinbrück and McHardy 2011). Furthermore, maximum likelihood estimates of codon-based Markov models are used to detect sites under positive or directional selection (Yang 2000; Kosakovsky Pond et al. 2005, 2008) and can also consider the physiochemical properties of residues (Sainudiin et al. 2005). All these methods return statistics of positive selection for independent codons but do not consider protein structure and spatial information for sites. Other methods take the effects of solvent accessibility and pairwise interactions between amino acids into account in their evolutionary models (Robinson et al. 2003). In the method we describe here, we follow a similar approach but use a less complex evolutionary model and consider the spatial distribution of residues in a consecutive phase of our algorithm.

In contrast to this type of methods, we assume that not only mutations at individual sites but also of multiple sites within a certain region of a gene can cause adaptive protein conformation changes. Shape and charge modifications within larger patches of residues on the protein surface are important for viral adaptation to structural changes in the interacting proteins of the host (see e.g., Yamada et al. 2010). We therefore devised a method to detect dense patches showing a high average positive selection, using d*N*/d*S* estimates of positive selection for individual sites and information on the spatial distances between them. With this approach, we also included sites with a large, but not exceptionally large, d*N*/d*S* ratio. Such residues would be discarded by methods that rank sites based on a measure of selection and then cut the list below a certain threshold. With our method, such residues were included if their spatial position supported the continuity of a patch. By searching for clusters of sites that are close to each other in the protein structure and consistently exhibit elevated d*N*/d*S* values, one might have greater statistical power to detect adaptive evolution in genes compared to methods that test for elevated d*N*/d*S* ratios at individual sites.

As mentioned above, more advanced techniques can be used for estimating positive selection. We here rely on the d*N*/d*S* statistic to allow an easy understanding of the principles of our method. The d*N*/d*S* statistic used for clustering can easily be exchanged with other measures.

There are similar methods that search for clusters of positively selected sites (Suzuki 2004; Berglund et al. 2005; Zhou et al. 2008). These differ from ours in that they use a sliding window–based search for sphere-shaped clusters on the surface of the tertiary structure. Our approach does not require specification of a cluster radius nor does it restrict the geometrical form of the inferred clusters. We evaluated our method by applying it to HA data for human influenza A viruses of the subtypes H3N2 and H1N1. These are particularly suited for evaluation as large numbers of sequences are available and their interaction with the human host is very well studied. Additionally, we applied the method to HA and polymerase basic protein 2 (PB2) of swine-origin influenza virus (S-OIV) A/H1N1 to study the more recent development of the virus.

Materials and Methods

We implemented a graph-cut algorithm to cluster protein residues based on structural and evolutionary protein information. Our goal was to identify dense patches of spatially close residues on the protein surface that show significant signs of positive selection. Generally speaking, our algorithm includes residues in a patch if they show evidence for positive selection and are close to other patch residues. A patch is rated both by its average *P* value and the density of sites under selection. Individual sites can compensate for a weaker signal of positive selection by being close to neighbors with a strong signal. Structural protein models were used to identify the spatial coordinates of individual residues. To measure positive selection for individual sites, ancestral character states were inferred from phylogenetic trees constructed from available genetic sequences for a particular protein. Subsequently, d*N*/d*S* statistics for each site were calculated, according to the ratio of the number of synonymous and nonsynonymous changes mapping to the tree edges (Bush et al. 1999; Suzuki 2006). After clustering, the identified patches were visualized on the protein structure. The complete process is shown in figure 1.

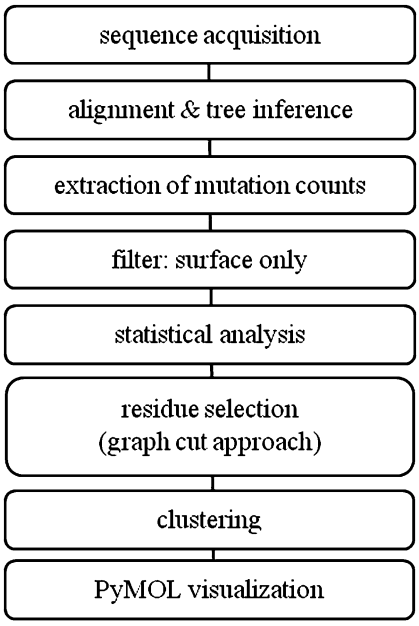

FIG. 1. Workflow for predicting patches under positive selection.

Structural Models

HA structures of the human influenza A/H3N2 virus, the human influenza A/H1N1, and S-OIV A/H1N1 were downloaded from the RSCB Protein Data Bank (PDB) (http://www.rcsb.org/) (for identifier codes of structures, sequences, and templates, see table 1). The analysis process was restricted to residues annotated in the PDB structure file and to sites found to be on the protein surface using the NetSurf software (Petersen et al. 2009). Structural models were generated for PB2 of the S-OIV isolate A/California/14/2009 (H1N1) based on the PB2 structures of PDB. To this end, the S-OIV PB2 sequence was compared with sequences of PB2 proteins with experimentally determined structure using Blast (Altschul et al. 1990). For PB2, there was no single structural template that covered all protein domains. Therefore, two models were generated from two templates, one for the PB2cap and one for the PB2c domain. The highest sequence identity, the largest coverage of the S-OIV protein, and the quality according to resolution and free R-factor values were used as criteria to select the best matching structural templates for the PB2cap and PB2c domains. The S-OIV sequences were aligned to the templates with MODELLER (version 9v6) (Sali and Blundell 1993). The alignments are expected to be reliable, given a sequence identity of 94% and a lack of insertions and deletions. Subsequently, the structural models were generated with MODELLER.

Sequence Data, Alignments, and Phylogenetic Tree Construction

Available HA sequences of the seasonal influenza A virus, subtypes H1N1 and H3N2, were downloaded from the GISAID EpiFlu database (http://platform.gisaid.org). Only sequences longer than 1,500 bp were selected, resulting in 1,734 and 3,221 sequences for H1 and H3, respectively (supplementary table S2, Supplementary Material online). Alignments of DNA and protein sequences were computed with MUSCLE (Edgar 2004), and manually curated. Phylogenetic trees were inferred with PhyML v3.0 (Guindon and Gascuel 2003) under the general time reversible (GTR) + I + Γ4 model, with the frequency of each substitution type, the proportion of invariant sites (I), and the gamma distribution of among-site rate variation with four rate categories (Γ4) estimated from the data. Subsequently, the tree topology and branch lengths of the maximum likelihood tree inferred with PhyML were optimized for 200,000 generations with Garli v0.96b8 (Zwickl 2006). Substitution events were inferred for the genome segment tree topologies from intermediates reconstructed with accelerated transformation (AccTran; Felsenstein 2004). The total number of substitutions occurring on all reconstructed internal branches was then calculated for each site independently. These numbers were used to compute the d*N*/d*S* ratio for each codon site (Bush et al. 1999; Suzuki 2006). The ratios were transformed to *P* values by a one-sided Fisher test for independence of the d*N* and d*S* values at an individual site and the mean values of the protein. *P* values were corrected for the ranking comparison with the false discovery rate (Benjamini and Yekutieli 2001) and used as a measure of selection for individual sites. Furthermore, 3,419 sequences of the PB2 protein and 7,373 sequences of the HA protein of the 2009 S-OIV A/H1N1 strains were downloaded from the GISAID EpiFlu database (supplementary table S2, Supplementary Material online). Phylogenetic trees were inferred using neighbor joining with PAUP (Swofford 2003) under the GTR model. Sequence alignment and residue statistics were inferred as described above.

Structural Clustering

Before clustering, all spatial coordinates were normalized to fit the protein structure into a hypercube of size 1. For

Table 1. Sequence Codes and PDB Codes of Selected Templates.

Protein	Query S-OIV Sequence	Template PDB Code and Chain	Template PDB Sequence	Query/Template Sequence Identity (%)
H1 (seas)	—	2wrgH,I	A/Brevig Mission/1/1918[1]	—
H3 (seas)	—	3hmgA,B	A/Aichi/2/1968[2]	—
H1 (swl)	—	3al4A,B	A/California/04/2009[3]	—
PB2cap	A/California/14/2009[3]	2vqzA	A/Victoria/3/1975[2]	94.00
PB2c	A/California/14/2009[3]	2vy6A	A/Victoria/3/1975[2]	94.00

NOTE.—Strains are of the subtypes [1]H1N1, [2]H3N2, and [3]H1N1swl.

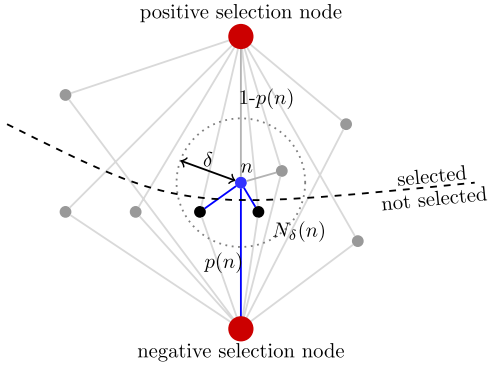

Fig. 2. Schematic drawing of the graph-cut approach. The minimum cut minimizes the sum of weights of all edges cut by the line separating the positive and negative selection nodes. For a single node *n*, these are the lines shown in blue: the scaled distances to the nonselected neighbors in $N_\delta(n)$ and the connection to the other side (i.e., the negative) selection node with the weight $P(n)$.

clustering with a graph-cut algorithm (Boykov et al. 2002), we constructed a graph in which each node represents a residue in the protein. Edges were added between all pairs of residues *m* and *n* for which the Euclidean distance dist(*m*, *n*) was below a threshold δ, and these edges were weighted according to their spatial distance (fig. 2). Weights were set to be in inverse exponential proportion to the Euclidean distance dist(*m*, *n*), that is, the closer the residues were located relative to each other on the protein structure, the larger the weight of the corresponding edge. Therefore, nodes that are close to each other have a strong connection to each other. We then augmented the graph with two additional nodes, which we call the "positive selection node" and the "negative selection node," corresponding to "source" and "sink" nodes in a standard graph-cut formulation. These two special nodes are connected to each residue node, with the weights equal to the *P* value $P(n)$ of the residue *n* in the case of the negative selection node or $1 - P(n)$ in the case of the positive selection node. Thus, residues that have high dN/dS ratios (large $1 - P(n)$) have a strong connection with the positive selection node, whereas nodes with low dN/dS values (large $P(n)$) have a strong connection with the negative selection node. The two types of edges and edge weights were added to the graph to represent the spatial information for each residue (by adding distances to close neighbors) and the evolutionary evidence for selection (by encoding the *P* value of the dN/dS ratios).

A "graph cut" will divide this graph in two halves, one containing the positive selection node and the other containing the negative one (fig. 2). A "minimum graph cut" is a graph cut that minimizes the sum *E* of the weights of the edges connecting these two halves:

$$E = \sum_{n \in \mathrm{Pos}} P(n) + \alpha \sum_{n \in \mathrm{Neg}} \bar{P}(n) + \beta \sum_{n \in \mathrm{Pos}} \sum_{\substack{m \in \mathrm{Neg}, \\ m \in N_{\delta(n)}}} e^{-\mathrm{dist}(m,n)},$$

where $\bar{P}(n) = 1 - P(n)$, pos represents all nodes assigned to the positive selection half, Neg represents all nodes assigned to the negative one, and $N_\delta(n)$ represents all neighbors of residue *n* within a distance less than δ. This means that the minimum cut will select residues to be in Pos if they show strong signs of positive selection (i.e., a low *P* value) and if they separate well spatially from the residues in Neg. The distance δ defines how many sites of a single residue are considered to be neighbors. We set δ such that a residue has, on average, ten close neighbors. The factor β weighs this distance statistic. The smaller the β, the more likely the method is to balance the residue evenly between the positive and the negative selection set halves according to the ratio $1{:}\alpha$ (we set $\alpha = 1$). The larger the β, the more expensive an even distribution becomes, and the more stringently the method searches for a small exclusive set of residues that spatially separate well from the rest. Since the total distance statistic is dependent on the number of residues in the protein, β has to be set manually (see supplemental text S1, Supplementary Material online). Finally, the selected residues were grouped into patches by merging all residues within a spatial distance *d* of each other into a set. The parameter *d* was set to represent the first quartile of all pairwise distances in the protein. Finally, we excluded outliers by filtering out all patches that contained two or less residues. Patches were identified for the H1 and H3 proteins of human influenza A viruses of the subtypes H1N1 and H3N2, respectively, and for the HA and PB2 proteins of the 2009 S-OIV of subtype H1N1. Subsequently, we analyzed their enrichment with known epitope sites (Caton et al. 1982; Wiley and Skehel 1987) and receptor avidity–changing sites (Hensley et al. 2009).

Evaluation and Visualization

For evaluation, we calculated the precision (ratio of selected epitope sites to all selected residues) and recall (ratio of selected epitope sites to all epitope sites) of the inferred patches based on the epitope regions defined for subtypes H1 (Caton et al. 1982) and H3 (Wiley et al. 1981; Wiley and Skehel 1987; Suzuki 2006). For a list of epitope sites used as a reference for evaluation, see supplementary table S1 (Supplementary Material online). The identified patches of all proteins were visualized with PyMOL software (Schrödinger 2012).

Results

We analyzed the merits of a clustering technique based on a graph-cut formalization for identification of patches of sites under selection on the surface of the HA and PB2 proteins of several influenza A viruses. Our goal was to rediscover regions known to play an important role in the interaction of the virus with the host's immune system and that comprise many important sites for adaptation. We therefore first considered known antigenic site regions on the HA of the human influenza A virus (Caton et al. 1982; Wiley and Skehel 1987) as our approximate reference for evaluation. The clustering algorithm identified dense patches of residues, which mostly consisted of sites with substantial deviation from the expected value of the protein-wide dN/dS. In comparison, a site ranking based

Table 2. Precision and Recall of Different Settings and Approaches When Put to the Task of Detecting Influenza Epitope Sites.

Setting	Recall (H1)	Precision (H1)	Recall (H3)	Precision (H3)
Graph cut	0.53	0.49	0.25	0.94
PV 0.05	0.19	0.4	0.15	0.86
PV 0.1	0.19	0.4	0.17	0.81

on P value alone resulted only in a low sensitivity for discovering relevant sites, with only 6 of 32 (H1) or 19 of 131 (H3) known antigenic or receptor avidity–changing sites exhibiting a significant ($P < 0.05$) signal. To compare this approach with our method, we calculated the precision and recall for sites selected by setting a P value ranking at $\theta = 0.05$ (PV 0.05) or $\theta = 0.1$ (PV 0.1) as a threshold as well as calculating these characteristics for the sites in patches identified with our graph-cut approach. Our evaluation (table 2) showed that including information on the spatial proximity of residues under selection and applying our clustering algorithm resulted in a significant improvement in recall (i.e., a larger number of epitope sites being identified) while maintaining similar or better precision (meaning that a similar or lower number of nonepitope sites were inferred). In the light of a recently proposed hypothesis on the relevance of receptor avidity–changing sites (Hensley et al. 2009), as opposed to the epitope sites of hemagglutinin in subtype H1 antigenic evolution, we also tested the value of these sites as a reference and compared these with the inferred patches of sites. The currently available data do not allow discrimination between these two hypotheses, as the reference sites of known receptor avidity–changing sites and antigenic sites overlap greatly (fig. 3). Still, residues 156 and 158, found to play the most significant role in receptor avidity, are included in the second patch identified for subtype H1.

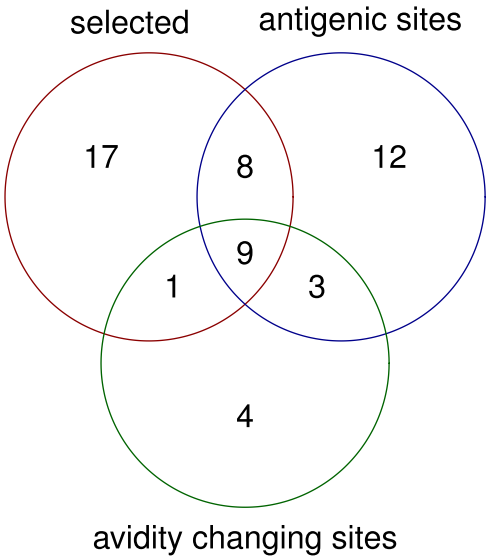

Fig. 3. Overlap between selected epitope and avidity-changing sites. Venn diagram showing the overlap between subtype H1 residues in patches selected by the dN/dS graph-cut approach (red), the influenza A H1 epitope sites according to Caton et al. (1982) (blue), and avidity-changing sites according to Hensley et al. (2009) (green).

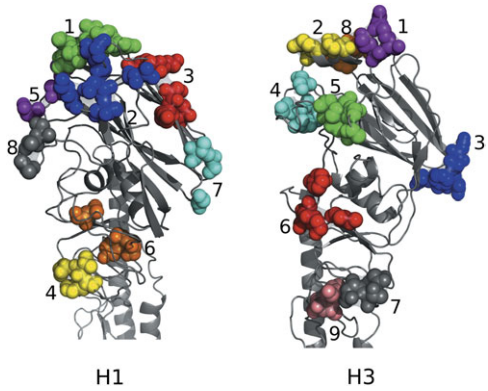

Fig. 4. Patches under positive selection on HA. Patches on the HA protein structure of subtype H1 and H3 selected by the graph-cut algorithm. Patches are numbered according to tables 3 and 4.

Of the detected patches on the HA protein surface (fig. 4 and tables 3 and 4), several include known epitope or receptor avidity–changing sites up to a fraction of 100%. The patches contain many sites that are relevant for antigenic evolution (Matrosovich et al. 1997; Hay et al. 2003; Lin et al. 2004; Yamada et al. 2010), including position 145, which has been shown experimentally to have a high antigenic impact (Smith et al. 2004).

We also compared our results with similar techniques for predicting the properties of sites under positive selection or relevant for adaptive evolution. Our predictions match 7 of 13 sites inferred to be under positive selection by a maximum likelihood approach (Yang 2000). However, 10 of these 13 sites are at least direct neighbors of those listed by our method, confirming its ability to find positively selected regions on the tertiary structure. Similar observations can be made for sites identified in Fitch et al. (1997), where five of six are matches or direct neighbors and the sites discussed in Bush et al. (1999) and Yang (2000) (10 of 13). Furthermore, several techniques combine biochemical and phylogenetic information to gain insights into the adaptive evolution of influenza A. It has recently been suggested that HA evolves by increasing the number of charged amino acids in regions recognized by the immune system, particularly in the dominant epitope (i.e., the one with the highest proportion of amino acid mutations, see Pan et al. 2011). We therefore compared the number of charged and uncharged amino acids in

Table 3. Patches and Residues Selected for the Influenza A Hemagglutinin Protein, Subtype H1.

Patch	Residues
1	187, 188, 189, 190, 192, 193, 196, 197, 198
2	131, 132, 133, 158, 156, 129
3	163, 165, 166, 244, 248
4	274, 275, 276
5	227, 225, 219
6	82, 81, 56
7	240, 169, 173
8	142, 144, 145

Note.—Underlined numbers refer to known epitope sites according to Caton et al. (1982) and supplementary table S1 (Supplementary Material online). All positions are given in H3 numbering (Aoyama et al. 1991).

Table 4. Patches and Residues Selected for the Influenza A Hemagglutinin Protein, Subtype H3.

Patch	Residues
1	156, 157, 158, 159
2	188, 189, 192, 193
3	171, 172, 173, 174, 175
4	186, 220, 229
5	137, 140, 142, 144, 145
6	62, 91, 92, 94
7	53, 275, 276
8	196, 197, 198, 199
9	47, 48, 50

NOTE.—Underlined numbers refer to known epitope sites (Wiley et al. 1981; Wiley and Skehel 1987; Suzuki 2006; see supplementary table S1, Supplementary Material online). All positions are given in H3 numbering.

the H1 and H3 consensus sequences for selected sites in the patches and sites lying outside the patches. Indeed, we found that the percentage of charged amino acids is much higher within patches (H1: 67%, H3: 67%) than outside patches (H1: 27%, H3: 28%). Finally, other authors suggest statistics based on rates of substitutions toward specific residues (Kosakovsky Pond et al. 2008; Kryazhimskiy and Plotkin 2008) or based on epistatic effects between pairs of sites (Kryazhimskiy et al. 2011). The overlap between the predictions by both methods and ours is not large, possibly due to the different nature of the measured quantities and statistics, and because, as Kryazhimskiy et al. 2011 discuss, hitchhiking changes without selective impact might comprise a fraction of identified epistatic pairs, particularly among the trailing change of a pair. However, our simple criterion for positive selection can easily be exchanged for more advanced estimates for adaptive evolution, allowing a search for clusters of residues that show significantly elevated statistics of such properties.

Additionally, we identified one patch in H1 without known epitope sites, but with similar evidence for positive selection as the other patches, which indicates its potential importance for antigenic evolution (table 3 and fig. 4, patch 4). For both subtypes, one patch in HA overlaps with the receptor-binding site of the protein. This could be due to the overlap of the antigenic and receptor-binding regions. However, the receptor-binding site, particularly position 189, is also known to be relevant for adaptation to avian and human hosts (Matrosovich et al. 1997; Sorrell et al. 2009). Both the H1 and H3 of human influenza A viruses show evidence of selection acting upon the receptor-binding region when grown in eggs, due to the effects of egg adaptation (Robertson et al. 1987; Gambaryan et al. 1999). Therefore, part of the signal in the receptor-binding sites could also be due to the effects of egg cultivation.

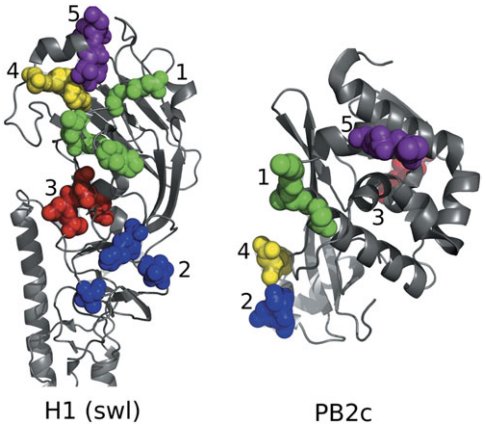

FIG. 5. Patches under positive selection on the HA and PB2 proteins of 2009 S-OIV. Patches on the 2009 swine-origin influenza A protein structures of HA and the c-terminal region of PB2, selected by the graph-cut algorithm. Patches are numbered according to tables 5 and 6.

As a second application, we analyzed data of 2009 S-OIV A/H1N1. The molecular basis of the successful establishment of the triple reassortant swine virus, which contains several recently acquired avian segments (Smith et al. 2009), in the human host is not fully understood. It has, in particular, been argued that lysine at position 627 of the PB2 protein of the viral polymerase complex, instead of the avian-like glutamic acid, is required for successful transmission and replication within mammals (Gabriel et al. 2005). However, the 2009 H1N1 virus still has lysine at position 627 in PB2, which it has maintained since its descent from an originally avian lineage. A change at residue 591 has been proposed to compensate for the lack of lysine in 627, allowing its efficient replication in mammals (Yamada et al. 2010). We searched for regions with evidence for positive selection and relevance for adaptation of PB2 since the introduction of the 2009 S-OIV into the human population. The virus might have acquired changes in PB2 to further optimize replication and transmission in the novel host. We identified five patches. The first (fig. 5 and table 5) is localized in a region around residue 591, which lends support to its relevance for mammalian and, in particular, human adaptation. To gain more insight, we allowed the method to also report patches containing only two residues. The resulting second patch surrounds residue 714, which is known to increase polymerase activity in mammals (Gabriel et al. 2005).

We furthermore analyzed the genetic sequences and protein structure of the HA protein of 2009 S-OIV A/H1N1. We identified five patches of sites under positive

Table 5. Patches and Residues Selected for the PB2 Protein of the 2009 Swine-Origin Influenza A/H1N1 Virus.

Patch	Residues
1	586, 588, 590
2	714, 715
3	660, 661
4	709, 711
5	575, 578

Table 6. Patches and Residues Selected for the HA Protein of the 2009 Swine-Origin Influenza A/H1N1 Virus.

Patch	Residues
1	135, 137, 140, 141, 142, 144, 145
2	53, 54, 56, 57, 276
3	63, 91, 92, 93, 94
4	186, 188, 189, 218
5	197, 198, 199, 200

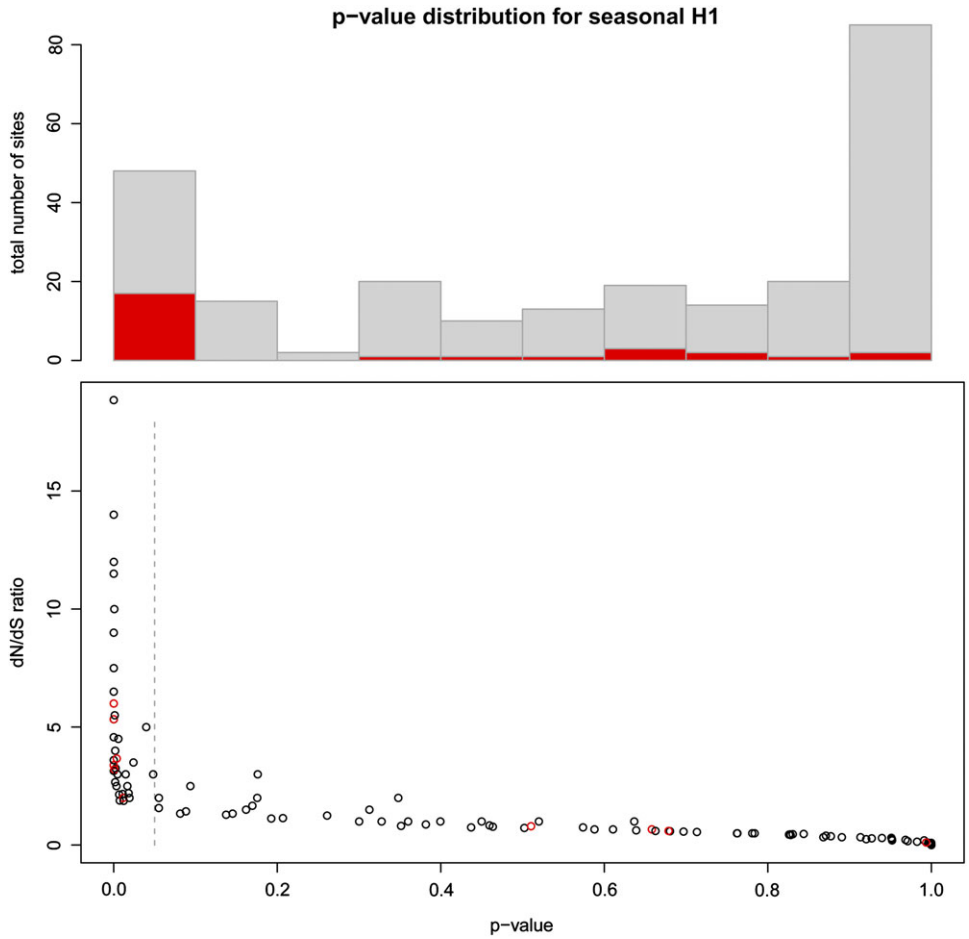

FIG. 6. Epitope sites not under positive selection. The histogram displays the ratio of residues within the corresponding *P* value intervals and demonstrates that many epitope sites feature insignificant *P* values resulting from an average d*N*/d*S* ratio. Epitopic sites are marked in red. The lower plot shows the distribution of the *P* values versus the d*N*/d*S* ratios for all residues of the H1 subtype.

selection. The first one (fig. 5 and table 6) overlaps with the Ca2 epitope site of seasonal H1 (Caton et al. 1982). The remaining ones cluster densely at the head of the protein, indicating emerging areas of relevance for adaptation and antigenic evolution of the 2009 H1N1 virus.

Our software, AdaPatch, is available online (http://www.cs.uni-duesseldorf.de/AG/AlgBio) and can also be applied to analyze other viral proteins.

Discussion

We have developed a technique for identification of candidate regions under positive selection in viral proteins. Our method utilizes a common measure of selection, and state-of-the-art techniques for phylogenetic tree inference, ancestral state reconstruction, and clustering or separation techniques. It requires only sequence information and a PDB structure file as input. We identified clusters of sites under positive selection based on information on the spatial proximity of sites. Although other methods search for functional importance, that is, conserved regions, or focus specifically on the detection of epitope sites, we aim to provide a fast and easy solution for identification of patches of arbitrary shape and size whose combined evolutionary signature indicates their importance for viral adaptive evolution. In addition to d*N*/d*S* statistics, other methods for evaluating positive selective pressure (e.g., Kosakovsky Pond et al. 2005) can easily be included.

Focusing on the HA of two subtypes of the seasonal influenza A virus and the HA and PB2 proteins of 2009 S-OIV A/H1N1, we searched for patches of sites under positive selection on their protein structures. The patches we identified for the HA of the seasonal influenza A viruses largely map to known epitope sites and sites associated with receptor binding. Among the patch sites, we identified for the PB2 protein of the 2009 S-OIV are sites with known relevance for successful replication in mammalian hosts. Our analysis showed that our approach increases the predictive accuracy relative to the commonly used approach of searching for individual sites with significantly deviating d*N*/d*S* statistics. This indicates that focusing on evolutionary change in larger regions, instead of individual sites, is helpful for revealing patches of residues that are important for adaptation, which together show a stronger signal of positive selection.

The precision and recall values for detecting known epitope sites based on patches under positive selection are rather low overall, mostly at or below 50%, indicating that not all sites in the epitope regions are under positive

selection and contributing to adaptation of the viral HA. Influenza A epitopes seem to be variable only in part (fig. 6) and probably change over time, thus diluting the overall signal of positive selection. Furthermore, receptor avidity–changing sites or host-specificity determinants may play a similarly important role in adaptive evolution, which lowers precision if one considers only the epitope sites that are predicted to be evolving under positive selection.

We evaluated our method using the influenza A viruses as they are very well studied and much is already known about the relevant sites for adaptive evolution. Still, our inferred patches might be more informative than individual sites for monitoring circulating viral strains for adaptive changes with relevance for transmission and spread in the human population. Our analyses of HA and PB2 identified many sites known to be relevant for antigenic drift or for the adaptation of influenza A to its host, improving its ability for infection, replication, and immune evasion. We therefore suggest analysis of the new patches identified in this study to determine the underlying causes of their consistent variability. We also suggest applying the method to other protein structures of rapidly evolving viruses with as yet unknown adaptive behavior in order to identify candidate regions that are important for virus–host interaction.

Supplementary Material

Supplementary text S1 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Dr Francisco Domingues for the provision of structural models of the PB2 protein of the A/California/04/2009 strain. We gratefully acknowledge funding by Max Planck Society and Heinrich Heine University Düsseldorf.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.

Aoyama T, Nobusawa E, Kato H. 1991. Comparison of complete amino acid sequences among 13 serotypes of hemagglutinins and receptor-binding properties of influenza A viruses indirect immunofluorescence. *Mutagenesis* 182:475–485.

Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* 38:1–5.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann Stat.* 29:1165–1188.

Berglund AC, Wallner B, Elofsson A, Liberles DA. 2005. Tertiary windowing to detect positive diversifying selection. *J Mol Biol.* 60:499–504.

Blythe M, Flower D. 2005. Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci.* 14:246–248.

Boykov Y, Veksler O, Zabih R. 2002. Fast approximate energy minimization via graph cuts. *Pattern Anal Mach Learn.* 23: 1222–1239.

Bush RM, Fitch WM, Bender CA, Cox NJ. 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol.* 16:1457–1465.

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.

Deem MW, Pan K. 2009. The epitope regions of H1-subtype influenza A, with application to vaccine efficacy. *Protein Eng Des Sel.* 22:543–546.

Dormitzer PR, Galli G, Castellino F, Golding H, Khurana S, Del Giudice G, Rappuoli R. 2011. Influenza vaccine immunology. *Immunol Rev.* 239:167–177.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

El-Manzalawy Y, Dobbs D, Honavar V. 2008. Predicting linear B-cell epitopes using string kernels. *J Mol Recognit.* 21:243–255.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates, Inc.

Fitch WM, Bush RM, Bender CA, Cox NJ. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc Natl Acad Sci U S A.* 94:7712–7718.

Gabriel G, Dauber B, Wolff T, Planz O, Klenk HD, Stech J. 2005. The viral polymerase mediates adaptation of an avian influenza virus to a mammalian host. *Proc Natl Acad Sci U S A.* 102: 18590–18595.

Gambaryan AS, Robertson JS, Matrosovich MN. 1999. Effects of egg-adaptation on the receptor-binding properties of human influenza A and B viruses. *Virology* 258:232–239.

Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. 2003. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 19:163–164.

Glaser L, Stevens J, Zamarin D, Wilson IA, García-Sastre A, Tumpey TM, Basler CF, Taubenberger JK, Palese P. 2005. A single amino acid substitution in 1918 influenza virus hemagglutinin changes receptor binding specificity. *Virology* 79:11533–11536.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.

Hay AJ, Lin Y, Gregory V, Bennet M. 2003. WHO collaborating centre for reference and research on influenza, Annual Report. Tech. Rep. London: National Institute for Medical Research.

Hensley SE, Das SR, Bailey AL, et al. (11 co-authors). 2009. Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. *Science* 326:734–736.

Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.

Kosakovsky Pond SL, Poon AFY, Leigh Brown AJ, Frost SDW. 2008. A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Mol Biol Evol.* 25:1809–1824.

Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB. 2011. Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS Genet.* 7:e1001301.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. *PLoS Genet.* 4:e1000304.

Kuiken T, Holmes EC, McCauley J, Rimmelzwaan GF, Williams CS, Grenfell BT. 2006. Host species barriers to influenza virus infections. *Science* 312:394–397.

Lacerda M, Scheffler K, Seoighe C. 2010. Epitope discovery with phylogenetic hidden Markov models. *Mol Biol Evol.* 27: 1212–1220.

Lin YP, Gregory V, Bennett M, Hay A. 2004. Recent changes among human influenza viruses. *Virus Res.* 103:47–52.

Matrosovich MN, Gambaryan AS, Teneberg S, Piskarev VE, Yamnikova SS, Lvov DK, Robertson JS, Karlsson KA. 1997. Avian influenza A viruses differ from human viruses by recognition of sialyloligosaccharides and gangliosides and by

a higher conservation of the HA receptor-binding site. *Virology* 233:224–234.
McHardy AC, Adams B. 2009. The role of genomics in tracking the evolution of influenza A virus. *PLoS Pathog.* 5:e1000566.
Nei M. 2005. Selectionism and neutralism in molecular evolution. *Mol Biol Evol.* 22:2318–2342.
Neumann G, Kawaoka Y. 2006. Host range restriction and pathogenicity in the context of influenza pandemic. *Emerg Infect Dis.* 12:881–886.
Neumann G, Noda T, Kawaoka Y. 2009. Emergence and pandemic potential of swine-origin H1N1 influenza virus. *Nature* 459:931–939.
Nimrod G, Glaser F, Steinberg D, Ben-Tal N, Pupko T. 2005. In silico identification of functional regions in proteins. *Bioinformatics* 21:i328–i337.
Nimrod G, Schushan M, Steinberg DM, Ben-Tal N. 2008. Detection of functionally important regions in "hypothetical proteins" of known structure. *Structure* 16:1755–1763.
Nozawa M, Suzuki Y, Nei M. 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci U S A.* 106:6700–6705.
Pan K, Long J, Sun H, Tobin GJ, Nara PL, Deem MW. 2011. Selective pressure to increase charge in immunodominant epitopes of the H3 hemagglutinin influenza protein. *J Mol Biol.* 72:90–103.
Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C. 2009. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol.* 9:51.
Pupko T, Bell RE, Mayrose I, Glaser F, Ben-tal N. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 18(Suppl 1):S71–S77.
Pybus OG, Rambaut A. 2009. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet.* 10:540–550.
Robertson JS, Bootman JS, Newman R, Oxford JS, Daniels RS, Webster RG, Schild GC. 1987. Structural changes in the haemagglutinin which accompany egg adaptation of an influenza A(H1N1) virus. *Virology* 160:31–37.
Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol.* 20:1692–1704.
Rubinstein ND, Mayrose I, Halperin D, Yekutieli D, Gershoni JM, Pupko T. 2008. Computational characterization of B-cell epitopes. *Mol Immunol.* 45:3477–3489.
Rubinstein ND, Mayrose I, Pupko T. 2009. A machine-learning approach for predicting B-cell epitopes. *Mol Immunol.* 46: 840–847.
Russell CA, Jones TC, Barr IG, et al. (11 co-authors). 2008. The global circulation of seasonal influenza A (H3N2) viruses. *Science* 320:340–346.
Sainudiin R, Wong W, Yogeeswaran K, Nasrallah J, Yang Z, Nielsen R. 2005. Detecting site-specific physicochemical selective pressures: applications to the Class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *J Mol Biol.* 60:315–326.
Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol.* 234:779–815.
Schrödinger. 2012. The PyMOL molecular graphics system, Version 1.4. New York: Schrödinger, LLC.
Shazman S, Celniker G, Haber O, Glaser F, Mandel-Gutfreund Y. 2007. Patch Finder Plus (PFplus): a web server for extracting and displaying positive electrostatic patches on protein surfaces. *Nucleic Acids Res.* 35:W526–W530.
Smith D, Lapedes A, de Jong J, Bestebroer T, Rimmelzwaan G, Osterhaus A, Fouchier R. 2004. Mapping the antigenic and genetic evolution of influenza virus. *Science* 305:371.
Smith GJD, Vijaykrishna D, Bahl J, et al. (11 co-authors). 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 459:1122–1125.
Sorrell EM, Wan H, Araya Y, Song H, Perez DR. 2009. Minimal molecular constraints for respiratory droplet transmission of an avian-human H9N2 influenza A virus. *Proc Natl Acad Sci U S A.* 106:7565–7570.
Steinbrück L, McHardy AC. 2011. Allele dynamics plots for the study of evolutionary dynamics in viral populations. *Nucleic Acids Res.* 39:e4.
Suzuki Y. 2004. Three-dimensional window analysis for detecting positive selection at structural regions of proteins. *Mol Biol Evol.* 21:2352–2359.
Suzuki Y. 2006. Natural selection on the influenza virus genome. *Mol Biol Evol.* 23:1902–1911.
Swofford D. 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.
Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. 1992. Evolution and ecology of influenza A viruses. *Microbiol Rev.* 56:152–179.
Weinstock DM, Zuccotti G. 2009. The evolution of influenza resistance and treatment. *JAMA* 301:1066–1069.
WHO. 2009. Influenza Fact Sheet No 211. [cited 2011 Apr 12]. Available from: http://www.who.int/mediacentre/factsheets/fs211/en/index.html
Wiley D, Skehel J. 1987. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annu Rev Biochem.* 56:365–394.
Wiley D, Wilson I, Skehel J. 1981. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature* 289:373.
Yamada S, Hatta M, Staker BL, et al. (11 co-authors). 2010. Biological and structural characterization of a host-adapting amino acid in influenza virus. *PLoS Pathog.* 6:e1001034.
Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J Mol Biol.* 51:423–432.
Zhou T, Enyeart PJ, Wilke CO. 2008. Detecting clusters of mutations. *PLoS One* 3:e3765.
Zwickl D. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion [dissertation]. [Austin (TX)]: University of Texas.