Putative amino acid determinants of the emergence of the 2009 influenza A (H1N1) virus in the human population.

Daphna Meroz, Sun-Woo Yoon, Mariette Ducatez, Thomas P Fabrizio, Richard J Webby, Tomer Hertz, Nir Ben-Tal

Putative amino acid determinants of the emergence of the 2009 influenza A (H1N1) virus in the human population

Daphna Meroz[a], Sun-Woo Yoon[b], Mariette F. Ducatez[b], Thomas P. Fabrizio[b], Richard J. Webby[b], Tomer Hertz[c,1], and Nir Ben-Tal[a,1]

[a]Department of Biochemistry and Molecular Biology, The George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel-Aviv, Israel 69978; [b]Department of Infectious Diseases, St. Jude Children's Research Hospital, Memphis, TN 38105; and [c]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

Edited by Barry Honig, Columbia University Howard Hughes Medical Institute, New York, NY, and approved July 8, 2011 (received for review October 6, 2010)

The emergence of the unique H1N1 influenza A virus in 2009 resulted in a pandemic that has spread to over 200 countries. The constellation of molecular factors leading to the emergence of this strain is still unclear. Using a computational approach, we identified molecular determinants that may discriminate the hemagglutinin protein of the 2009 human pandemic H1N1 (pH1N1) strain from that of other H1N1 strains. As expected, positions discriminating the pH1N1 from seasonal human strains were located in or near known H1N1 antigenic sites, thus camouflaging the pH1N1 strain from immune recognition. For example, the alteration S145K (an antigenic position) was found as a characteristic of the pH1N1 strain. We also detected positions in the hemagglutinin protein differentiating classical swine viruses from pH1N1. These positions were mostly located in and around the receptor-binding pocket, possibly influencing binding affinity to the human cell. Such alterations may be liable in part for the virus's efficient infection and adaptation to humans. For instance, 133_A and 149 were identified as discriminative positions. Significantly, we showed that the substitutions $R133_AK$ and R149K, predicted to be pH1N1 characteristics, each altered virus binding to erythrocytes and conferred virulence to A/swine/NC/18161/02 in mice, reinforcing the computational findings. Our findings provide a structural explanation for the deficient immunity of humans to the pH1N1 strain. Moreover, our analysis points to unique molecular factors that may have facilitated the emergence of this swine variant in humans, in contrast to other swine variants that failed.

A pandemic H1N1 (pH1N1) human influenza virus was identified in April 2009 (1) and has since spread to over 200 countries and caused over 18,000 deaths (2). Evolutionary analysis of the pH1N1 strain indicates that its HA belongs to the classical swine lineage. HAs in this lineage are more similar to those of historical human strains such as the 1918 H1N1 strain than to those of circulating H1N1 strains from recent years (3). Prior to the emergence of the pH1N1 strain, only sporadic cases of human infection by swine influenza viruses had been reported (4–8). The molecular basis enabling this recent strain to efficiently infect and be transmitted between humans remains obscure (9–12).

Herein, we used a computational approach to uncover signature residues in the receptor-binding domain (RBD) of the pH1N1 HA protein. Using a machine-learning algorithm of alternating decision trees (ADTs) (13), we predicted positions that differentiated between HA protein sequences of the pH1N1 strain and HA sequences of other H1N1 strains. The algorithm combines moderately successful rules to produce highly accurate predictions (13). Here the rules represent correlations between the presence of specific amino acid type(s) in selected positions and the sequence annotation, e.g., pandemic versus seasonal strain. This is done by iteratively reweighting training examples (i.e., HA sequences), thus, concentrating on the sequences that are more "difficult" to classify. We subsequently ranked each predicted position according to its discriminative potency (see *SI Text S5*). First, we detected the alterations distinguishing the antigenicity of the pH1N1 strain from that of prior seasonal human strains. We used all human H1N1 sequences (seasonal H1N1 and pH1N1) available in the National Center for Biotechnology Information (NCBI) database (14) and trained the classification algorithm to detect residues in the HA sequence discriminating pH1N1 from those strains. Next, we carried out the same analysis to identify molecular features that distinguish this swine-origin virus from classical swine strains. The detected positions, which were situated mainly in and around the receptor-binding pocket, may reveal why the pandemic virus was able to emerge in the human population whereas other swine viruses were unsuccessful. Notably, we experimentally validated two of our predictions: We inserted either mutation $R133_AK$ or mutation R149K —predicted pH1N1 characteristics—into the HA of the swine strain A/swine/NC/18161/02. Each substitution altered virus binding to erythrocytes and resulted in a strain that was more virulent than the parental strain in DBA/2J mice. These results substantiate the computational method presented here.

Results

To identify signature sites we trained the ADT algorithm on the RBD of the HA sequence. We chose to focus our analysis on the RBD, because previous studies have shown that minor changes in this region can account for differences in virus transmissibility, which is a prerequisite for emergence (15, 16). Furthermore, the H1N1 HA includes four known antigenic sites (17) in the RBD region (Fig. 1), namely, Sa (residues 128, 129, 156–160, and 162–167), Sb (residues 187–198), Ca (residues 140–145, 169–173, 206–208, 224, 225, and 238–240), and Cb (residues 79–84) [numbering according to Protein Data Bank (PDB) ID code 3lzg].

pH1N1 Versus Seasonal H1N1 Human Strains. We compared pH1N1 isolates with isolates from human-seasonal H1N1. We selected the 10 most discriminative positions (see *SI Text S5*), whose combination provided near-perfect classification accuracy of 98%, on average. Mapping these positions on the HA structure, we discovered that all residues were situated in and around known H1N1 antigenic sites (17) (Fig. 1 and Table 1). Specifically, five residues were located in the known Ca antigenic site, one was in direct contact with the Sb site, and four were in the vicinity of these sites. These results are in agreement with a recent study showing that

Author contributions: D.M., S.-W.Y., M.F.D., T.H., and N.B.-T. designed research; D.M., S.-W.Y., M.F.D., T.P.F., and T.H. performed research; D.M., S.-W.Y., M.F.D., R.J.W., T.H., and N.B.-T. analyzed data; and D.M., S.-W.Y., M.F.D., R.J.W., T.H., and N.B.-T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. E-mail: thertz@fhcrc.org or NirB@tauex.tau.ac.il.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1014854108/-/DCSupplemental.

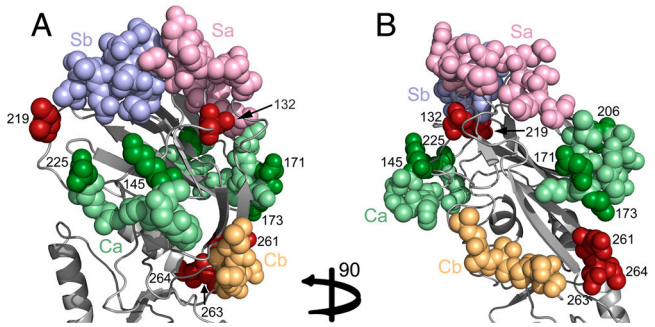

Fig. 1. *Detected positions, discriminating between prior human H1N1 circulating strains and the pH1N1 strain, correspond to known antigenic sites.* The RBD of the HA protein from the human A/California/04/2009 H1N1 strain (PDB ID code 3lzg) is shown in cartoon representation. Front (*A*) and side (*B*) views of the RBD with the identified positions and antigenic sites presented as all-atom spheres. The Sa antigenic site is in pink, Sb site in blue, Ca site in green, and the Cb antigenic site in orange. The identified specificity determinants that overlap with H1N1 antigenic sites are shown using a dark variant of the site color (e.g., a position identified in Ca is colored in dark green, whereas the rest of the site is in pale green). Predictions that do not overlap with a known antigenic site are in dark red. All newly detected positions are located in and around the recognition hot spots of the immune system.

the majority of antigenic differences between pH1N1 and historical strains appear in the Ca antigenic site, whereas the Sa site has remained relatively conserved (18).

The physicochemical characteristics of many of the predicted alterations (Table 1) may influence the structure of the antigenic sites and consequently the antibody binding. For instance, our analysis suggests that in seasonal strains, serine appears in position 145, whereas in the pH1N1 strain lysine appears in this position. Position 145 is located in the Ca antigenic site (17) (Fig. 2), and the substitution of serine with lysine may alter the geometry and physicochemical characteristics of the site. Interestingly, this position is located in the site that binds the sialic acid of the host receptor (Fig. 1). As binding of the HA to the host's negatively charged sialic acid mediates viral infection, the predicted substitution to the positively charged lysine may increase affinity to the host receptor in the pH1N1 strain.

Recent studies have shown that the 2009 pH1N1 strain lacks two glycosylation sites in the RBD that are conserved in human-seasonal H1N1 strains (18, 19). Specifically, in human-seasonal strains, threonine or serine at positions 131 and 165 introduce glycosylation sites at asparagine 129 and 163, respectively. The algorithm does not specifically account for changes that alter potential glycosylation patterns, and yet three of these four positions—129, 131, and 165—were detected as discriminative positions between pH1N1 and seasonal H1N1 human strains. However, these positions were assigned relatively low ranks (51, 42, and 50, respectively).

pH1N1 Versus Classical Swine Strains. The molecular characteristics giving rise to a human-adapted swine virus are of major interest and concern. Therefore, we looked for positions in the HA sequence separating the pH1N1 strain from previous swine strains.

We selected the 13 most discriminative positions (see *SI Text S5* and Table 2); their combination resulted in an average of approximately 90% classification accuracy. Almost all detected positions are located in the receptor-binding pocket, suggesting that they have some role in receptor-binding efficacy.

Some of the positions were particularly interesting. Residues 131, 132, and 133_A form a structural cluster (Fig. 3) located in the receptor-binding pocket. These positions may have mutated simultaneously, which in turn may have affected the structural conformation of the receptor-binding pocket, ultimately influencing the host cell binding. Furthermore, this cluster is near the Sb antigenic site, and the mutations might also be involved in antigenic variation.

Residues 206 and 208 (Fig. 3) are both located in antigenic site Ca and can affect antigenicity. More interestingly, these residues interact with positions 220, 221, and 229 of the receptor-binding pocket in the adjacent monomer (Fig. 4). Therefore, mutations in residues 206 and 208 may result in a structural change to the adjacent receptor-binding pocket or influence the interaction between the monomers.

Position 186 binds the receptor directly and is in direct contact with the Sb antigenic site (Fig. 3). Our analysis suggests that serine in this position is a characteristic of the pH1N1 strain. Indeed, a previous study has found serine in this position to be a human H1N1 characteristic, whereas proline in this position characterizes swine strains (20). Moreover, this position has been recently reported to affect growth in cell culture and eggs (21). Thus, the appearance of serine in pH1N1 may be one of the characteristics leading it to become a more "human-like" virus. Additionally, residues 188 and 189 are also located in the receptor-binding pocket (Fig. 3) and are both in antigenic site Sb. Mutations in these positions may affect the receptor binding and antigenicity as well.

Residues 225 and 226 both bind the host cell receptor (Fig. 3); therefore, alterations in these positions may induce a variation in the binding specificity to the receptor. Moreover, position 225

Table 1. Highly ranked residues that discriminate the pH1N1 strain from circulating human H1N1 strains in the hemagglutinin RBD

Residue number in structure 3lzg	Rank*	Circulating human strain characteristics†	pH1N1 characteristics‡	Structural comments
145	1	S	K	In antigenic site Ca. A change from the (positively charged) K to (polar) S can change the stereochemistry of the antigenic site and affect receptor specificity.
171	2	N	D	In antigenic site Ca. May change the physicochemical characteristics of Ca.
225	3	—	—	In antigenic site Ca. Known specificity determinant for altering avian to human H1N1 sequences.
219	4	E	I	Close to antigenic site Sb. The E-to-I mutation may affect the antigenic site's structure.
261	5	S	E	Residues 261 + 263 + 264 form a cluster in the structure. Residue 261 is very close to antigenic site Ca. From the other side, residues 263 and 264 are close to antigenic site Cb. These positions may be a continuous part of the antigenic sites. The alterations may affect the protein structure or the approximate antigenic sites.
132	6	I	P,S	In the vicinity of the Sa antigenic site.
206	7	—	—	In antigenic site Ca.
173	8	—	G	In antigenic site Ca. The mutation may alter the antibody binding.
264	9	F	A	See comments for position 261.
263	10	G	N	

Positions are numbered as in the A/California/04/2009 H1N1 strain (PDB ID code 3lzg) structure.
*Rank of contribution to discrimination given to the detected position (see *Computational Methods*).
†The amino acid suggested to characterize circulating human H1N1 strains by the algorithm.
‡The amino acid suggested to characterize the pH1N1 strain. The symbol "—" indicates that no characteristic amino acid was identified.

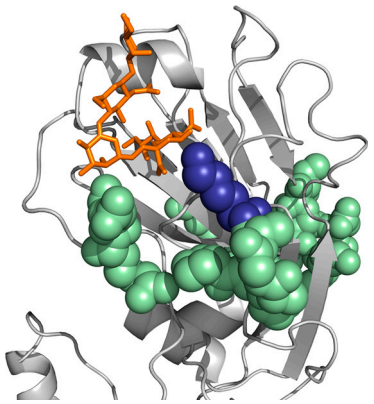

Fig. 2. Residue 145 (shown in blue spheres) has been detected as a discriminative position in our analysis. The HA receptor-binding domain of the H1 A/California/04/2009 H1N1 strain (PDB ID code 3lzg) is shown in gray cartoon representation, and the human receptor analogue is shown in orange. Interestingly, this amino acid is in the known antigenic Ca (in pale green spheres) and is located near the human receptor as well. The human receptor analogue was modeled into the RBD by superimposing the structures of A/California/04/2009 H1N1 and A/Brevig Mission/1/1918 H1N1 HAs (PDB ID codes 3lzg and 2wrg) with the α2-6 analogue bound.

is already known to affect receptor-binding specificity of avian and human H1N1 viruses (22). In pH1N1, the amino acid in this position is characteristic of the human virus (23), which may account for its efficient adaption and transmission in humans.

We note that although a small number of the positions that were highly ranked by our computational approach have been reported to affect adaptation or transmissibility in humans, the majority of the detected positions in our analysis have not been previously described. However, the strategic locations of these predicted alterations, together with their characteristics, suggest a role for these features in the emergence of pH1N1 as an effective human pathogen.

We also conducted the same analyses (i.e., pH1N1 versus seasonal and pH1N1 versus swine) on the whole HA sequence (instead of the RBD alone). Reassuringly, most of the highly ranked positions in the former analyses appeared in the full HA analyses as well (see *SI Text S1* and Tables S1 and S2).

In addition, we were interested to see whether positions identified in our analyses were located in or around T-cell and B-cell epitopes, which would provide additional support for the role of these positions in antigenicity. However, we found that 78% of the RBD sequence is covered by one or more epitopes reported in the Immune Epitope Database (www.immuneepitope.org) (see Fig. S1), rendering this analysis uninformative.

Experimental Validation. In order to evaluate whether our computational findings had identified residues with phenotypic relevance, we altered an H1N1 virus with a classical swine-lineage HA, with the goal of rendering it more "pH1N1-like." The experimental methods are described in *SI Text S2*. Residues 133_A and 149 were detected in our analysis as discriminating between classical swine H1N1 and pH1N1 strains (Table 2). Accordingly, we generated two single amino acid HA mutants by inserting mutation $R133_AK$ or mutation R149K into the HA of A/swine/NC/18161/02, an endemic H1N1 swine virus. Both mutant viruses, rg-swine/NC/18161/02-$HA133_A$ and rg-swine/NC/18161/02-HA149, respectively, were successfully rescued. "Swine-like" pandemic reverse mutants rg-TN/560-1/09-$HA133_A$ and rg-TN/560-1/09-HA149 were also successfully generated.

We first compared the binding specificities of A/swine/NC/18161/02 with those of the two HA mutants. As a surrogate measure of receptor specificity, we carried out hemagglutination assays using erythrocytes from different species. Both swine HA mutant viruses showed decreased binding to chicken, goose, guinea pig, and human (type O) erythrocytes compared with the parental strain. Pandemic viruses containing the reverse mutations (i.e., mutation to the corresponding swine virus residue) were unaltered in their erythrocyte binding patterns as compared to the parental strain (Tables S3 and S4). The three viruses showed similar binding to turkey erythrocytes, and none of the viruses bound to horse erythrocytes (Table S3). Whereas horse erythrocytes contain α2-3-linked sialic acids (α2-3SAL, "avian-like" receptors) but no α2-6SAL (human-like receptors), chicken and goose erythrocytes contain more α2-3SAL than α2-6SAL, and human O, guinea pig, and turkey erythrocytes have more α2-6SAL than α2-3SAL (24, 25). It is important to note that these erythrocyte receptor generalizations are very much oversimplifications as highlighted by the lack of difference between the three viruses in binding to turkey erythrocytes. Therefore, these find-

Table 2. Highly ranked residues that discriminate between the pH1N1 and H1N1 swine strains

Residue number in structure 3lzg	Rank*	Swine characteristics†	pH1N1 characteristics‡	Structural comments
149	1	R	K	In close proximity to the Ca and Cb antigenic sites.
171	2	N	D	In antigenic site Ca. The alteration may interfere with antibody binding.
225	3	G, E	D	This position is in direct contact with the host cell receptor and known to affect avian to human receptor specificity. The alteration may alter receptor binding. Also in antigenic site Ca.
132	4	T	S	Residues $131 + 132 + 133_A$ form a cluster in the structure, situated in the receptor-binding pocket and
133_A	5	R	—	may affect binding. Also in vicinity to the Sa antigenic site and may affect this site.
188	6	T	S	In contact with residue 189, which is in direct contact with the host receptor. An alteration in this position may indirectly affect receptor specificity and binding. Also in antigenic site Sb.
226	7	—	—	In direct contact with the sialic acid of the host cell receptor.
206	8	—	T	Residue 206 is in direct contact with residue 221 of the adjacent RBD, which is in the receptor-binding pocket. The alteration may indirectly affect the receptor binding. Additionally, this residue is located in antigenic site Ca; a mutation may affect antibody binding.
208	9	K	R	Binds position 206, which is in direct contact with the adjacent RBD, therefore, can affect binding. Located in antigenic site Ca and mutation here may alter antibody binding.
200	10	T	A	In direct contact with the Sb antigenic site; its alteration may affect antigenicity.
131	11	E	D	Residues $131 + 132 + 133_A$ form a cluster in the structure (described above).
189	12	—	A	Binds the receptor; alterations may affect binding. Also in the Sb antigenic site.
186	13	P	S	Binds the receptor; a mutation may alter the binding.

Positions are numbered as in the A/California/04/2009 H1N1 strain (PDB ID code 3lzg) structure sequence.
*Rank of contribution to discrimination given to the detected position (see *Computational Methods*).
†The amino acid suggested to characterize circulating swine H1N1 strains by the algorithm.
‡The amino acid suggested to characterize the pandemic H1N1 human strain. The symbol "—" indicates that no characteristic amino acid was identified.

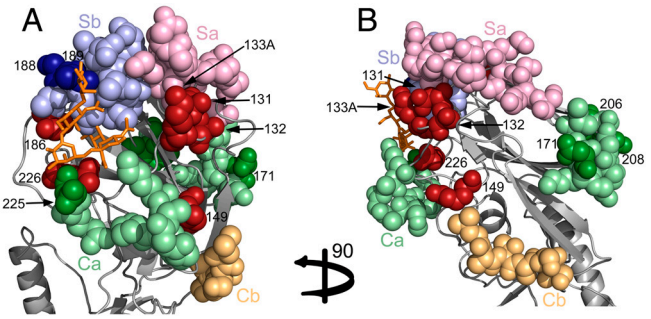

Fig. 3. Detected positions discriminating between swine H1N1 strains and the pH1N1 strain RBD of the HA protein from the human A/California/04/2009 H1N1 strain (PDB ID code 3lzg) is shown in cartoon representation. Front (*A*) and side (*B*) views of the RBD with the identified positions and antigenic sites presented as all-atom spheres (the orientations are similar to these in Fig. 1 *A* and *B*). The Sa antigenic site is in pink, Sb site in blue, Ca site in green, and the Cb antigenic site in orange. The identified specificity determinants that overlap with H1N1 antigenic sites are shown using a dark variant of the site color (e.g., a position identified in Ca is colored in dark green, whereas the rest of the site is in pale green). Predictions that do not overlap with a known antigenic site are in dark red.

ings indicate that indeed the mutations presented here affect the binding to different erythrocytes; however, they do not point to a preference of a specific sialic acid receptor. Because HA residue 133_A is situated in the receptor-binding pocket (Table 2 and Fig. 3), we anticipated that it would play a role in receptor binding; thus, the differences observed between the binding of the 133_A mutant and that of the parental virus were in line with our expectations. Interestingly, mutation of residue 149 produced similar results in the hemagglutination assay, even though this residue is not located directly in the receptor-binding pocket.

We further wished to investigate the effect of these mutations on viral virulence. We used the DBA/2J mouse strain, as it has been shown to be highly susceptible to most influenza viruses (26). To measure the lethal dose of each virus, we infected mice with either 5×10^1, 5×10^2, 5×10^3, or 5×10^4 egg 50% infective dose (eID_{50}) of A/swine/NC/18161/02, rg-A/swine/NC/18161/02-HA133_A, or rg-A/swine/NC/18161/02-HA149. Both swine HA mutants were more virulent than their parental strain, with 50% mouse-lethal doses (MLD_{50}s) of <1.5, <1.5, and 2.45 for rg-A/swine/NC/18161/02-HA133_A, rg-A/swine/NC/18161/02-HA149, and rg-A/swine/NC/18161/02, respectively. DBA/2J mice were also infected with 10^1, 10^2, 10^3, and 10^4 eID_{50} of rg-TN/560-1/09-HA133_A, rg-TN/560-1/09-HA149 (i.e., pandemic viruses with swine virus signature mutations), and rg-A/TN/560-1/09. These swine-like pandemic mutants were less pathogenic than their parental strain with MLD_{50}s of <3.4, 3.5, and 2.4 for rg-TN/560-1/09-HA133_A, rg-TN/560-1/09-HA149, and rg-TN/560-1/09, respectively (Table S4). The critical biological roles of HA positions 133_A and 149 were therefore confirmed in vivo.

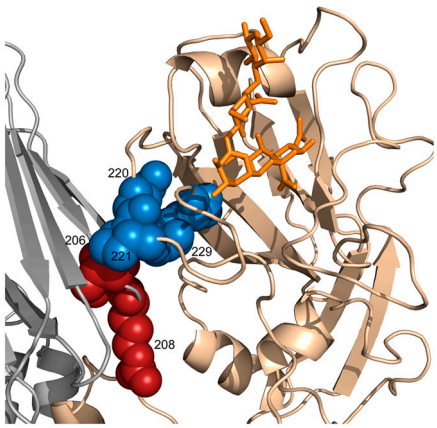

Fig. 4. *Residues 206 and 208*. Positions 206 and 208 were identified as discriminative between the swine and pH1N1 strains. Two adjacent monomers of the H1 A/California/04/2009 H1N1 (PDB ID code 3lzg) trimer are shown in cartoon representation (the third monomer is not shown for clarity). Each monomer is in a different color. The human receptor analogue is shown in an orange stick representation. Positions 206 and 208 are shown in red spheres, and positions 220, 221, and 229 from the adjacent monomer, which are in direct contact with the receptor, are shown in blue spheres. An alteration in residues 206 or 208 may affect the structural conformation of the receptor-binding pocket from the adjacent monomer.

Discussion

In this study we identified molecular features differentiating the HA protein of the pH1N1 from the HAs of other H1N1 strains. We detected alterations in amino acid positions in the receptor-binding pocket and antigenic sites of the pH1N1 HA that could be responsible in part for the successful transmission of the virus into humans. Moreover, we experimentally confirmed that two of the highly ranked positions indeed had a phenotypic effect in vivo.

Comparison of aligned sequences of viruses from various sources as a means to detect altered positions is a commonly used approach to guide experimental studies (e.g., refs. 10 and 23). A thorough investigation of all the thousands of available HA sequences would be desirable. As manual inspection of this volume of data is virtually impossible; it is necessary to use computational methods. Entropy (mutual-information)-based methods that use the amino acid frequency in each position are commonly used to this effect (27–29). These methods consider each position in the alignment separately, disregarding the possible relations between the sequence positions. Allen et al. (30) used a classification algorithm that did take such relations into account. They used it to detect host-specific alterations in the influenza A proteins, but none of the positions they detected were in the HA, the main protein responsible for the host cell binding. Our approach is similar to theirs in that it also takes into account the relations between the sequence positions. It selects a set of positions whose combination provides optimal discrimination between two groups (e.g., pH1N1 and swine-H1N1 sequences). For comparison we analyzed the same datasets using a mutual-information-based method (29). Reassuringly, the majority of the highly ranked positions presented here were also obtained in that analysis (*SI Text S3* and Tables S5 and S6). The mutual-information method identified 49 positions distinguishing between the human-seasonal H1N1 and the pH1N1 dataset, and our methodology showed that 10 residues were sufficient to provide a very high classification accuracy of 98%. Eight of the residues we identified were also identified in the mutual-information analysis. For the swine H1N1 and pH1N1 dataset, the mutual-information analysis produced 14 discriminative amino acid positions, whereas our method produced 13, 6 of which were predicted by both methods. Intriguingly, position 133_A, which our method highlighted, and whose subsequent substitution in an H1N1 virus led to a phenotypic effect (see *Results*), did not come up in the mutual-information analysis.

Recently, probabilistic approaches that explicitly take into account the phylogenetic relations between the taxa were used to identify host, clade, and drug-resistance signature residues in the various influenza proteins (31–33). This approach may solve problems caused due to uneven sampling in sequence space. However, because of the large quantity of available HA sequences and their high similarity, we do not believe that uneven sampling poses a major concern in our case. With these data we obtained an average accuracy of 90% for the discrimination of swine versus pH1N1 sequences (using only the 13 highly ranked positions), which is very high given the low diversification within the swine and pH1N1 groups and the relatively close evolutionary relationship between pH1N1 and other swine strains. The analy-

sis of sequence datasets with more distant evolutionary relationships, such as human seasonal versus pH1N1 (see *Results*) and human seasonal versus swine, resulted in even better accuracy of 98% (*SI Text S4*).

An obvious concern regarding our approach would be that because antigenic sites evolve rapidly (5), they are prone to be detected by the algorithm. We therefore examined the correlation between the variability of selected positions and their discriminative rank (see Fig. S2). Reassuringly, we found a very weak negative correlation ($r = -0.26$; $p = 0.04$). Thus, the positions detected here were identified based on their discriminative power and not their variability.

Another caveat to our approach relates to the low diversity in the HA sequences of the pH1N1 strain, which results from the fact that the unique influenza strain has been circulating in the human population for a relatively short time, probably insufficient to mutate and substantially diverge. Presumably, some of the unique amino acid characteristics this virus acquired are not crucial for its functionality, but are merely founder effects, i.e., mutations that were randomly "fixed" during the transition to humans and have not mutated because they are not yet under antigenic pressure (32). Indeed, over time, some of these positions will diverge and may prove to be functionally insignificant, but others might provide a supportive genetic background for other functionally important mutations (31, 34). Bearing in mind these difficulties, it is encouraging that essentially all the positions identified in our analysis were located in functionally important regions.

We studied two different, albeit related, questions. First, we sought the distinctive properties of pH1N1 in comparison with prior human H1N1 strains. Vaccines for seasonal influenza strains induced little or no neutralizing antibody response to pH1N1 among children and adults under 64 y of age (35), thus indicating that the antigenic properties of pH1N1 differ from those of previous seasonal strains. Indeed, our analysis showed that all positions detected as discriminating between the pandemic and seasonal human strains are located in and around the known H1N1 antigenic sites. Considering that many of the characteristic substitutions of the pH1N1 viruses substantially affected the physicochemical nature of the amino acids in these positions (Table 1), these alterations could have resulted in variation of the stereochemistry of the antibody-binding sites. Thus, a host immune to either the swine or human virus lineages might not be able to recognize the pH1N1 virus, a phenomenon known as "antigenic drift" (36).

Second, we studied the molecular adaptations that enabled the virus to emerge in humans. The efficient adaptation and vast transmission of the swine-origin virus, pH1N1, into the human population has caused substantial concern. Prior to this pandemic, only isolated cases of infection by a swine virus had been seen in humans (4–8), and limited human-to-human transmission had been documented. Although the impact of this pandemic has been mild in terms of overall mortality, discerning the molecular factors enabling this virus's success is crucial for future detection of potential pandemic threats. To this end, we identified positions in the HA protein sequence separating pH1N1 from classical swine strains. Our study revealed residues in the receptor-binding pocket that may have altered binding to the host cell receptor.

Swine hosts were shown to be susceptible to infection by both avian and human influenza viruses (37), owing to the presence of both avian and human receptors in the pig trachea (38). Therefore, reassortment between avian and human viruses may occur upon coinfection (39). Alternatively, viruses with preferential binding specificity to avian receptors may mutate (as a result of the human receptors' presence in the pig trachea) and eventually acquire the ability to efficiently bind the human receptor. Indeed, our analysis showed that the residues detected as discriminating between the swine and pH1N1 strains are located at a mean distance of 10.2 Å from the receptor, and the residues discriminating between the human-seasonal and pH1N1 strains are located at a mean distance of 19.4 Å from the receptor. An independent t test with a confidence level of 0.95 showed that the difference between these means was indeed significant ($p = 0.04$). These results emphasize and support the probable role of the former identified positions in increasing binding to the human host. Therefore, these sites may represent milestones in the adaptation of the swine virus to humans. In this context it is noteworthy that four residues appeared in both analyses (residues 132, 171, 206, and 225, Fig. 5).

We validated the significance of residues 133_A and 149, identified in our analysis. A hemagglutination assay showed that the introduction of mutation $R133_AK$ or of mutation R149K into the HA of A/swine/NC/18161/02 viruses had the effect of decreasing binding to chicken, goose, guinea pig, and human (type O) erythrocytes as compared with the parental strain (Table S3). Furthermore, we compared the pathogenesis of our two swine and two pandemic mutants with that of their parental strain in the DBA/2J mouse model. Both mutants were more virulent than their parental strain, whereas the mutated pandemic strains were less. Furthermore, two mutants were successfully generated for positions 171 and 132 (ranks 2 and 4) but did not alter erythrocyte binding patterns or pathogenicity in mice (Table S4). The consequences of such mutations in the human host, however, are still unclear, and further studies are needed to better address the question. Nevertheless the fact that such mild substitutions (i.e., lysine to arginine) caused significant phenotypic effects confirm the biological significance of HA positions 133_A and 149 and validate the computational method presented here.

After this paper was submitted for publication, Ye et al. reported that the insertion of mutations D131E and S186P into the pandemic strain A/California/04/ 09/(H1N1) increased the virus's pathogenicity in BALB/c mice (40). Indeed, our analysis detected these positions as discriminating between swine H1N1 and pH1N1 (Table 2), and D131 and S186 were identified as characteristic of the pandemic strain. Interestingly, the amino acids we predicted to be swine H1N1 characteristics (131E and 186P) were shown to generate the pathogenic strain studied by Ye et al. (40). An additional study published after the submission of this paper (41) disclosed two positions in the HA, namely, 200 and 227, as affecting receptor binding of the pandemic strain. Residue 200 is indeed a highly ranked position in our study (Table 2), and 227 was detected by our algorithm, albeit with a lower rank.

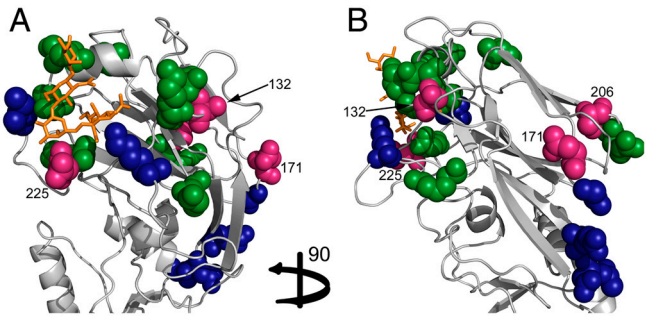

Fig. 5. *Positions detected as discriminating both between human circulating H1N1 and pH1N1 strains and between swine and pH1N1 strains.* The receptor-binding domain of the HA protein from the human A/California/04/2009 H1N1 strain (PDB ID code 3lzg) is shown in cartoon representation. Front (*A*) and side (*B*) views of the RBD with the identified positions presented as all-atom spheres. Positions detected as discriminating between human circulating and pH1N1 are in blue, and those detected as discriminating between swine and pH1N1 are in green. Residues identified in both analyses are colored in pink. For clarity, only the overlapping positions are numbered. It is evident that residues discriminating between swine and pH1N1 strains are mostly around the receptor-binding pocket.

We suggest that the basis for the difference in antigenicity between the pH1N1 and seasonal human H1N1 strains is associated with the detected sites presented in our study. We suggest signature sites in the HA protein that may enable the efficient host cell binding, infection, and transmission of the strain in the human population. Bearing in mind that HA is the main antigen on the viral surface and is responsible for the first step in the viral infection (42), our study is a significant step toward providing testable predictions for positions that may contribute to the elusiveness of this previously undescribed viral strain.

Computational Methods

An elaborate description of the methodology is provided in *SI Text S5*. Briefly, H1N1 HA sequences from swine and human hosts were collected from the NCBI Influenza Database (14). Duplicate sequences and partial sequences (less than 80% of full length) were removed from the data. We aligned sequences using the MUSCLE program (43) and visually inspected alignments to verify their quality. The main analysis was limited to the RBD (positions 114–268, H3 numbering) (44). Two datasets were created from these sequences: pH1N1 sequences versus prior circulating human strains, and pH1N1 sequences versus classical swine strains. The first dataset consisted of 706 pH1N1 sequences and 852 prior circulating human strains, excluding the 1918 historical strain. The second dataset consisted of 245 swine sequences and 782 pH1N1 sequences. In each dataset, sequences were labeled according to the group they belonged to ("pH1N1," "swine," or "human circulating" strains).

We used JBoost (http://jboost.sourceforge.net/) and the above datasets to identify positions in HA that distinguish pH1N1 isolates from human circulating H1N1 isolates, as well as positions that distinguish pH1N1 from swine H1N1 isolates (Fig. S3).

ACKNOWLEDGMENTS. We acknowledge Aaron Arvey and Maya Schushan for fruitful discussions. We also thank the St. Jude Hartwell Center for Biotechnology and Bioinformatics. This work was supported by the Specific Targeted Research Project project "EuroFlu," funded by the FP6 European Commission Program (SP5B-CT-2007-044098 to N.B.-T.) R.W. was supported in part by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract HHSN266200700005C. D.M. was supported by the Edmond J. Safra Bioinformatics program at Tel Aviv University.

1. Dawood FS, et al. (2009) Emergence of a novel swine-origin influenza A (H1N1) virus in humans: Novel swine-origin influenza A (H1N1) virus investigation team. *N Engl J Med* 360:2605–2615.
2. World Health Organization (2010) Pandemic (H1N1) 2009—update 101 (http://www.who.int/csr/don/2010_05_21..
3. Sinha NK, Roy A, Das B, Das S, Basak S (2009) Evolutionary complexities of swine flu H1N1 gene sequences of 2009. *Biochem Biophys Res Commun* 390:349–351.
4. Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y (1992) Evolution and ecology of influenza A viruses. *Microbiol Rev* 56:152–179.
5. Hay AJ, Gregory V, Douglas AR, Lin YP (2001) The evolution of human influenza viruses. *Philos Trans R Soc Lond B Biol Sci* 356:1861–1870.
6. Kendal AP, Goldfield M, Noble GR, Dowdle WR (1977) Identification and preliminary antigenic analysis of swine influenza-like viruses isolated during an influenza outbreak at Fort Dix, New Jersey. *J Infect Dis* 136(Suppl):S381–S385.
7. Myers KP, Olsen CW, Gray GC (2007) Cases of swine influenza in humans: A review of the literature. *Clin Infect Dis* 44:1084–1088.
8. Shinde V, et al. (2009) Triple-reassortant swine influenza A (H1) in humans in the United States, 2005–2009. *N Engl J Med* 360:2616–2625.
9. Childs RA, et al. (2009) Receptor-binding specificity of pandemic influenza A (H1N1) 2009 virus determined by carbohydrate microarray. *Nat Biotechnol* 27:797–799.
10. Maines TR, et al. (2009) Transmission and pathogenesis of swine-origin 2009 A(H1N1) influenza viruses in ferrets and mice. *Science* 325:484–487.
11. Garten RJ, et al. (2009) Antigenic and genetic characteristics of swine-origin 2009 A (H1N1) influenza viruses circulating in humans. *Science* 325:197–201.
12. Itoh Y, et al. (2009) In vitro and in vivo characterization of new swine-origin H1N1 influenza viruses. *Nature* 460:1021–1025.
13. Freund Y, Mason L (1999) The alternating decision tree algorithm. *Proceedings of the 16th International Conference on Machine Learning* (Morgan Kaufmann Publishers, San Francisco), pp 124–133.
14. Bao YM, et al. (2008) The influenza virus resource at the National Center for Biotechnology Information. *J Virol* 82:596–601.
15. 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 USA* 106:7565–7570.
16. Tumpey TM, et al. (2007) A two-amino acid change in the hemagglutinin of the 1918 influenza virus abolishes transmission. *Science* 315:655–659.
17. 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.
18. Xu R, et al. (2010) Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. *Science* 328:357–360.
19. Wei CJ, 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.
20. Matrosovich MN, et al. (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.
21. Suphaphiphat P, et al. (2010) Mutations at positions 186 and 194 in the HA gene of the 2009 H1N1 pandemic influenza virus improve replication in cell culture and eggs. *Virol J* 7:157.
22. Matrosovich M, et al. (2000) Early alterations of the receptor-binding properties of H1, H2, and H3 avian influenza virus hemagglutinins after their introduction into mammals. *J Virol* 74:8502–8512.
23. Neumann G, Noda T, Kawaoka Y (2009) Emergence and pandemic potential of swine-origin H1N1 influenza virus. *Nature* 459:931–939.
24. Medeiros R, Escriou N, Naffakh N, Manuguerra JC, van der Werf S (2001) Hemagglutinin residues of recent human A(H3N2) influenza viruses that contribute to the inability to agglutinate chicken erythrocytes. *Virology* 289:74–85.
25. Ito T, et al. (1997) Receptor specificity of influenza A viruses correlates with the agglutination of erythrocytes from different animal species. *Virology* 227:493–499.
26. Boon AC, et al. (2010) Cross-reactive neutralizing antibodies directed against pandemic H1N1 2009 virus are protective in a highly sensitive DBA/2 mouse influenza model. *J Virol* 84:7662–7667.
27. Chen GW, et al. (2006) Genomic signatures of human versus avian influenza A viruses. *Emerg Infect Dis* 12:1353–1360.
28. Finkelstein DB, et al. (2007) Persistent host markers in pandemic and H5N1 influenza viruses. *J Virol* 81:10292–10299.
29. Miotto O, Heiny A, Tan TW, August JT, Brusic V (2008) Identification of human-to-human transmissibility factors in PB2 proteins of influenza A by large-scale mutual information analysis. *BMC Bioinformatics* 9(Suppl 1):S18.
30. Allen JE, Gardner SN, Vitalis EA, Slezak TR (2009) Conserved amino acid markers from past influenza pandemic strains. *BMC Microbiol* 9:77.
31. Bloom JD, Gong LI, Baltimore D (2010) Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* 328:1272–1275.
32. Tamuri AU, Dos Reis M, Hay AJ, Goldstein RA (2009) Identifying changes in selective constraints: Host shifts in influenza. *PLoS Comput Biol* 5:e1000564.
33. Forsberg R, Christiansen FB (2003) A codon-based model of host-specific selection in parasites, with an application to the influenza A virus. *Mol Biol Evol* 20:1252–1259.
34. Holmes EC (2010) Virology. Helping the resistance. *Science* 328:1243–1244.
35. Katz J, et al. (2009) Serum cross-reactive antibody response to a novel influenza A (H1N1) virus after vaccination with seasonal influenza vaccine. *MMWR Morb Mortal Wkly Rep* 58:521–524.
36. Treanor J (2004) Influenza vaccine—Outmaneuvering antigenic shift and drift. *N Engl J Med* 350:218–220.
37. Gambaryan AS, et al. (2005) Receptor-binding properties of swine influenza viruses isolated and propagated in MDCK cells. *Virus Res* 114:15–22.
38. Ito T, et al. (1998) Molecular basis for the generation in pigs of influenza A viruses with pandemic potential. *J Virol* 72:7367–7373.
39. Ma W, Kahn RE, Richt JA (2008) The pig as a mixing vessel for influenza viruses: Human and veterinary implications. *J Mol Genet Med* 3:158–166.
40. Ye J, et al. (2010) Variations in the hemagglutinin of the 2009 H1N1 pandemic virus: Potential for strains with altered virulence phenotype? *PLoS Pathog* 6:e1001145.
41. de Vries RP, et al. (2011) Only two residues are responsible for the dramatic difference in receptor binding between swine and new pandemic H1 hemagglutinin. *J Biol Chem* 286:5868–5875.
42. Ha Y, Stevens DJ, Skehel JJ, Wiley DC (2002) H5 avian and H9 swine influenza virus haemagglutinin structures: Possible origin of influenza subtypes. *EMBO J* 21:865–875.
43. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
44. Lin T, et al. (2009) The hemagglutinin structure of an avian H1N1 influenza A virus. *Virology* 392:73–81.

Supporting Information

Meroz et al. 10.1073/pnas.1014854108

SI Text

Text S1—Analysis of the Whole HA Protein. In order to further verify that our approach is capable of identifying functionally important sites, we conducted a second set of experiments in which the algorithm was provided with full HA sequences rather than the receptor-binding domain (RBD) alone. The HA sequences of the human pandemic and circulating human H1N1 strains were collected from the National Center for Biotechnology Information (NCBI) influenza database (1) following the same method described for the RBD analysis. The dataset consisted of 821 circulating human H1N1 and 673 pandemic H1N1 (pH1N1) sequences.

We hypothesized that a significant number of the detected sites would overlap with the sites selected when analyzing the RBD, and that in general, most discriminative sites would be in the RBD, taking into account that it consists of approximately 27% of the whole HA sequence (the whole HA sequence is approximately 560 amino acids long). Indeed, for the pH1N1 versus human seasonal H1N1 strains, 9 of the 18 most highly ranked positions of the whole HA analysis (i.e., 50%; Table S1) were in the RBD. Out of 10 highly ranked positions from the RBD analysis (Table 1), 7 appeared in the highly ranked set from the analysis of the entire HA. For the swine versus pH1N1 strains, 15 of the 32 (approximately 47%, Table S2) highly ranked positions in the full HA analysis were from the RBD sequence. Additionally, 11 out of the 13 (approximately 85%, Table 2) highly ranked positions from the RBD analysis were ranked highly in the analysis of the whole HA. These results demonstrate the power of the approach and its ability to identify the known functional regions and residues, even when provided with a very large set of features. Moreover, the analysis reinforces the importance of the highly ranked residues selected.

Text S2—Experimental Methods. ***Generation of viruses.*** The eight genes of the A/swine/NC/18161/02 (H1N1) virus were cloned into a dual-promoter plasmid, pHW2000. The HA of A/swine/NC/18161/02 was mutated with the QuikChange mutagenesis kit (Stratagene) following the instructions of the manufacturer. Reverse genetics (rg) viruses were generated by DNA transfection as described previously (2). Each viral HA segment was sequenced to confirm the identity of the virus.

Hemagglutination assay. Hemagglutination assays were performed as previously described (3). Six types of packed erythrocytes (Rockland) were used in different concentrations: 0.5% for turkey, chicken, and goose RBCs; 0.75% for guinea pig and human (group O) RBCs; and 1% for horse RBCs (4). We added 0.5% bovine serum albumin (Sigma) to the horse RBCs. Virus titers were normalized to $10^{6.25}$ egg 50% infective does (eID_{50}) per milliliter prior to the hemagglutination assay. Turkey red blood cells were used to measure the eID_{50}s.

Mouse experiments. Six- to 8-wk-old female DBA/2J mice (Jackson Laboratory) were housed at St. Jude Children's Research Hospital according to the institution's Animal Care and Use Committee guidelines. The experiments were performed in compliance with relevant institutional policies of the National Institutes of Health and the Animal Welfare Act. Mice were sedated with 2,2,2-tribromoethanol (Avertin; Sigma) and intranasally inoculated with 30 μL of virus diluted in phosphate buffer saline ($n = 5$ mice per group). The mice were monitored daily for survival and body weight loss over a period of 14 d. Any mouse showing more than 30% of body weight loss was considered to have reached the experimental end point and was humanely euthanized. The mouse-lethal dose (MLD_{50}) was calculated using the method of Reed and Muench (5).

Text S3—Mutual Information Analysis with AVANA. We applied the AVANA (*Antigenic Variability Analyzer*) method (6), a software program that calculates entropy profiles from multiple sequence alignments, to the same input datasets used in our study (see *Computational Methods*). Specifically, we carried out two analyses with AVANA, comparing seasonal human H1N1 versus pH1N1, and swine H1N1 versus pH1N1 strains. For the human H1N1 versus pH1N1 dataset, AVANA selected 49 positions, which included 8 of the 10 highly ranked positions detected in our study (see *Results* in the main text and Table S5). When applied to the pH1N1 and swine H1N1 dataset, AVANA detected 14 positions, 6 of which overlapped with the 13 highly ranked positions from our approach (see *Results* in the main text and Table S6). Remarkably, position 133_A, which was detected as discriminative by our method and was shown to have a phenotypic effect in vivo (see *Results*), was not identified by AVANA, reinforcing the advantage of our method.

Text S4—Seasonal Human H1N1 Versus Swine H1N1 Strains. Swine and human seasonal H1N1 sequences were collected from the NCBI database (1), and a dataset was built as described in *Computational Methods* (main text). The resulting dataset consisted of 195 swine H1N1 and 525 human seasonal H1N1 sequences. We applied our computational approach to this set and obtained an overall mean test accuracy of 98% (with 50 runs of 10-fold cross-validation).

Text S5—Computational Methods. Two datasets were created as described in the main text (*Computational Methods*): pH1N1 sequences versus prior circulating human strains, and pH1N1 sequences versus classical swine strains. These datasets were analyzed using JBoost (http://jboost.sourceforge.net/) to identify positions in HA that distinguish "pH1N1" isolates from "human circulating" H1N1 isolates, as well as positions that distinguish pH1N1 from "swine" H1N1 isolates. JBoost is an open-source Java implementation of the Adaboost (7) machine-learning algorithm. This discriminative learning approach tries to identify the features that best distinguish between different data categories. Ultimately, classifiers in the form of decision trees called alternating decision trees (ADTs) (8) are generated. The ADT algorithm is an easily interpretable, boosting-based algorithm that is a generalization of decision trees and boosting using decision stumps. This algorithm also provides a measure of confidence, called a classification margin, for each prediction. An example of a decision tree created by the ADT method is presented in Fig. S3. The rectangles in the decision tree are the decision (or splitter) nodes, and the ovals are the prediction nodes; the values in each oval correspond to the contribution of that node to the prediction score. The number in each decision node represents the number of the iteration in which that feature was selected. In order to predict the label of a given example, we begin at the root of the decision tree and traverse the tree, using the decision nodes and summing the scores in the prediction nodes along the selected path.

In our setting each data instance is an influenza HA sequence, so the dimensionality of each data point is $N = 155$ for the receptor-binding site of the HA dataset. Each data instance consists

of the amino acid sequence alone, without taking into account functional annotations of the protein (e.g., glycosylation sites). The data labels are the host species from which each isolate was obtained: pH1N1, human circulating or swine. The algorithm uses the data and labels to learn an ADT that can then be used to predict which strain a certain sequence belongs to.

The ADT algorithm selects the set of positions that best discriminate between the requested groups. In order to measure the predictive power of our proposed method over test data, we performed 50 runs of 5-fold cross-validation experiments over 100 iterations, producing 50 different runs altogether.

Stopping criteria. While boosting algorithms have been shown to be empirically robust to overfitting, some simple criteria for choosing the number of iterations have been suggested. Here we used a stopping criterion based on the convergence of the distribution of margins over all training points. Specifically, let us denote by m_t^i the margin of the ith data point in iteration t, and by S_t the average margin over all data points in iteration t: $S_t = \frac{1}{N}\sum_{i=1}^{N} m_t^i$. Our stopping criterion was defined by $(S_{t+1} - S_t)^2 < \varepsilon$, where $\varepsilon = 10^{-5}$.

Adjusting for biases in training set size. In order to balance the sizes of the different sets of HA sequences (number of swine, pH1N1, and circulating human sequences), we used a standard technique in boosting to account for biases in the label distribution and to reweight the data such that each label had equal weight. This is easily done in boosting algorithms, where each point i is associated with a weight w_t^i in each iteration, by tweaking $W_1 = (w_1^1, w_1^2, \cdots, w_1^N)$ to be such that $\sum_{l(i)=\mathrm{pH1N1}} w_1^i = \sum_{l(i)=\mathrm{seasonal/swine}} w_1^i$. This forces the algorithm to focus equally on the different HA sequence sets in the initial rounds of training.

Measuring the informativeness of selected features. In order to assess the importance of the selected features over the different decision trees created, we developed a scoring function to rank positions selected by the algorithm. Our scoring function is an extension of the one suggested by Creamer et al. (9). Intuitively, given a set of decision trees generated using many different partitions of the data into training and test data, a feature is more important if it appears in many of the trees and is selected in earlier boosting iterations. Moreover, because our main concern is predicting mutations that characterize the pH1N1 strain, our scoring function also takes into account the relative contribution of a given feature in assigning a sequence to the pH1N1 class. More formally, the score of a given feature i is given by $S(i) = n_i{}^* m_{\mathrm{iter}}{}^* \max_{d(i)}(\mathrm{p_{H1N1}})$, where n_i is the number of appearances of feature i in the set of trees, m_{iter} is the mean iteration in which feature i appears, and $\max_{d(i)}(\mathrm{p_{H1N1}})$ is the maximal value of the pH1N1 label prediction nodes taken over all of the decision nodes that contain feature i. A larger contribution score implies a greater importance of the feature for predictions related to the pH1N1 strain.

Decision of the cutoff for top-ranked positions. In order to choose a cutoff for a smaller subset from the list of ranked positions, we looked for a set of positions that would cover 70% of the cumulative distribution of the computed ranking scores. That is to say, the sum of the scores of the positions that we chose for further analysis consisted of 70% of the total ranking scores for all detected positions.

1. Bao YM, et al. (2008) The influenza virus resource at the national center for biotechnology information. *J Virol* 82:596–601.
2. Hoffmann E, Neumann G, Kawaoka Y, Hobom G, Webster RG (2000) A DNA transfection system for generation of influenza A virus from eight plasmids. *Proc Natl Acad Sci USA* 97:6108–6113.
3. WHO (2002) WHO manual on animal diagnosis and surveillance. Available at http://www.who.int/csr/resources/publications/influenza/en/whocdscsrncs20025rev.pdf (accessed on January 31, 2011).
4. Wiriyarat W, et al. (2010) Erythrocyte binding preference of 16 subtypes of low pathogenic avian influenza and 2009 pandemic influenza A (H1N1) viruses. *Vet Microbiol* 146:346–349.
5. Reed LJ, Muench H (1938) A simple method for estimating fifty percent endpoints. *Am J Hyg* 27:493–497.
6. Miotto O, Heiny A, Tan TW, August JT, Brusic V (2008) Identification of human-to-human transmissibility factors in PB2 proteins of influenza A by large-scale mutual information analysis. *BMC Bioinformatics* 9:S18.
7. Freund Y, Schapire RE (1999) A short introduction to boosting. *J Jap Soc Artif Intell* 14:771–780.
8. Freund Y, Mason L (1999) The alternating decision tree algorithm. *Proceedings of the 16th International Conference on Machine Learning*, eds (Morgan Kaufmann Publishers, San Francisco), pp 124–133.
9. Creamer G, Freund Y, Moore M (2005) Using Adaboost for equity investment scorecards. http://papers.ssrn.com.

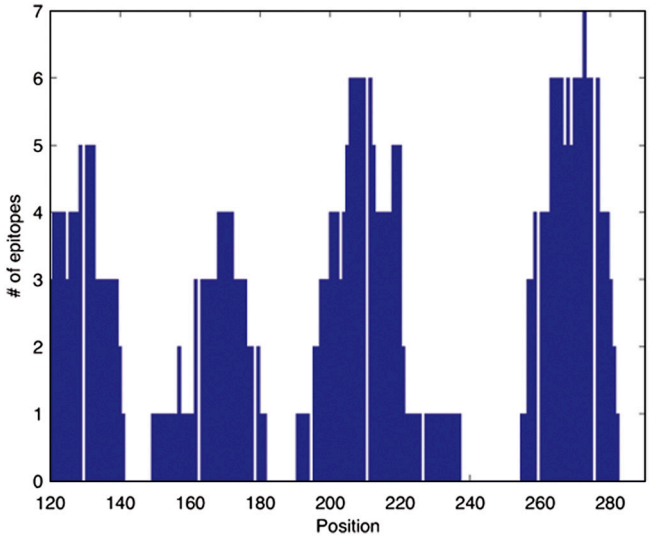

Fig. S1. A histogram of all T- and B-cell epitopes reported for the RBD of influenza A H1N1, or influenza A (unspecified) in the Immune Epitope Database. Seventy-eight percent of the RBD sequence is covered by one or more epitopes.

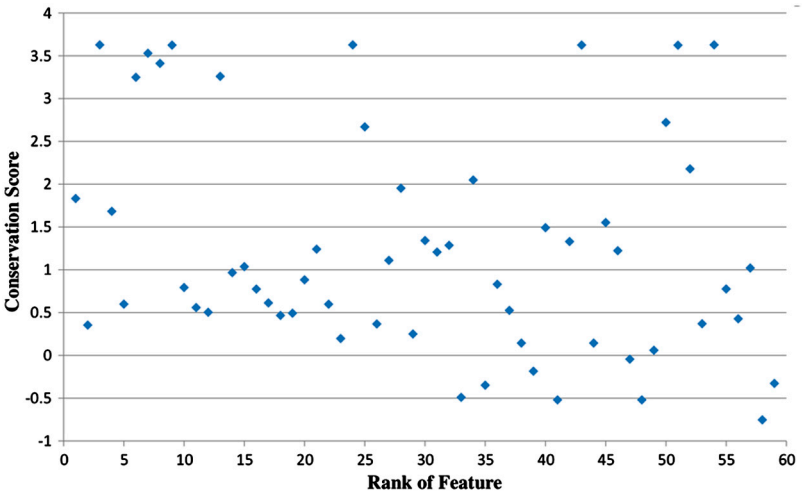

Fig. S2. *The discriminating positions are not necessarily conserved*. Scatter plot of the evolutionary conservation scores versus our calculated rank for the positions that were detected as discriminative between the pH1N1 and the circulating human strains. The evolutionary conservation scores were calculated using the ConSurf web server (http://consurf.tau.ac.il) (1). Higher conservation scores are given to evolutionarily variable sites. Evidently there is no correlation between the conservation score and our rank.

1. 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:W529–W533.

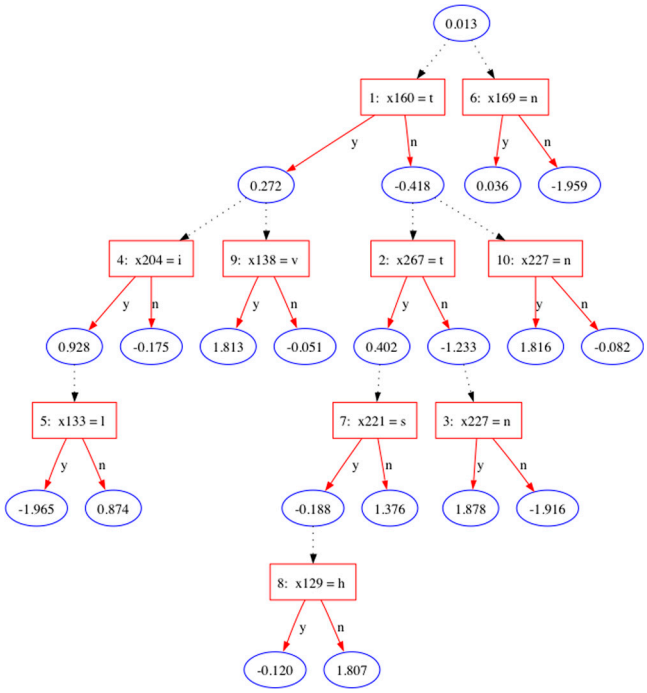

Fig. S3. *Representative ADT obtained after 10 iterations*. The ovals in the decision tree are the prediction nodes, and the rectangles represent the splitter nodes. The final prediction score is obtained by starting from the score of the top prediction node and summing the scores of the relevant prediction nodes that meet the conditions of the splitter nodes.

Table S1. Highly ranked residues detected as discriminating between the pH1N1 and human circulating H1N1 strains in the analysis of the entire HA

Position in structure 3lzg	Rank*	Antigenic site?	In RBD?†	Detected in RBD analysis?‡
145	1	in Ca antigenic site	yes	yes
242	2		yes	yes
317	3		no	no
219	4		yes	yes
206	5	in Ca antigenic site	yes	yes
171	6	in Ca antigenic site	yes	yes
261	7		yes	yes
296	8		no	no
225	9	in Ca antigenic site	yes	yes
−5	10		no	no
55	11		no	no
−11	12		no	no
132	13		yes	yes
305	14		no	no
211	15		yes	yes
301	16		no	no
36	17		no	no
275	18		no	no

*Rank refers to the rank for contribution to discrimination (according to the ranking function, see *Methods*). *Antigenic site?* refers to whether the position is in a known antigenic site.

†*In RBD?* refers to whether the position is part of the RBD sequence (positions 114–268).

‡*Detected in RBD analysis?* refers to whether the position was detected in the analysis of the HA RBD sequences of the pH1N1 versus the circulating human H1N1 strains. Positions are numbered as in the A/California/04/2009 H1N1 strain (PDB ID code 3lzg) structure sequence; therefore, residues appearing before the first position of the structure sequence are numbered with a minus sign (e.g., −5).

Table S2. Highly ranked residues detected as discriminating between the pH1N1 and swine H1N1 strains in the analysis of the entire HA

Position in structure 3lzg	Rank*	Antigenic site?	In RBD?†	Detected in RBD analysis?‡
149	1		yes	yes
225	2	in Ca antigenic site	yes	yes
132	3		yes	yes
171	4	in Ca antigenic site	yes	yes
186	5		yes	yes
188	6	in Sb antigenic site	yes	yes
226	7		yes	yes
−1	8		no	no
324	9		no	no
233	10		yes	yes
71	11		no	no
−5	12		no	no
206	13	in Ca antigenic site	yes	yes
318	14		no	no
530	15		no	no
−12	16		no	no
263	17		yes	yes
131	18		yes	yes
426	19		no	no
51	20		no	no
75	21		no	no
300	22		no	no
414	23		no	no
−3	24		no	no
527	25		no	no
120	26		yes	yes
−2	27		no	no
377	28		no	no
557	29		no	no
200	30		yes	yes
189	31	in Sb antigenic site	yes	yes
88	32		no	no
−7	33		no	no
146	34		yes	yes

*_Rank_ refers to the rank for contribution to discrimination (according to the ranking function, see _Computational Methods_). _Antigenic site?_ refers to whether the position is in a known antigenic site.

†_In RBD?_ refers to whether the position is part of the RBD sequence (positions 114–268).

‡_Detected in RBD analysis?_ refers to whether the position was detected in the analysis of the HA RBD sequences of the pH1N1 versus the classical swine strains. Positions are numbered as in the structure of the A/California/04/2009 H1N1 strain (PDB ID code 3lzg); therefore, residues appearing before the first position of the structure are numbered with a minus sign (e.g., −1).

Table S3. Differential binding of reverse genetics A/swine/NC/18062/02, A/swine/NC/18062/02-HA133$_A$, A/swine/NC/18062/02-HA149, A/TN/560-1/09, A/TN/560-1/09-HA133$_A$, and A/TN/560-1/09-HA149 with different erythrocytes as measured by hemagglutination assay

Erythrocytes type	rg-SW/NC/02	rg-sw/NC/18062/02-HA133$_A$	rg-sw/NC/18062/02-HA149	rg-TN/560-1/09	rg-TN/560-1/09-HA133$_A$	rg-TN/560-1/09-HA149
Turkey	32	32	32	32	32	32
Chicken	128	32	32	16	16	16
Goose	64	16	16	32	32	32
Guinea pig	24	4	8	32	32	32
Horse	<1	<1	<1	<1	<1	<1
Human (type O)	24	4	4	16	16	16

Table S4. Experimental validation of residues discriminating classical swine H1N1 and pH1N1 strains

Rank	Residue number in structure 3lzg	Virus rescued*	HA assay*	MLD_{50} (log_{10})†
1	149	rg-Sw/NC/02-HA-R149K	D	<1.5
		rg-TN/560-HA-K149R	S	3.5
2	171	rg-NC/02-HA-N171D did not rescue	ND	—
		rg-TN/560-HA-D171N	S	2.17
3	225	ND	ND	—
		ND	ND	—
4	132	rg-Sw/NC/02-HA-T132S	S	2.53
		ND	ND	—
5	133_A	rg-Sw/NC/02-HA-R133_AK	D	<1.5
		rg-TN/560 HA-K133_AR	S	3.38

*ND: not done, D: different from parental strain, S: same as parental strain.
†Sw/NC/02 MLD_{50}: $10^{2.45}$, TN/560 MLD_{50}: $10^{2.4}$.

Table S5. Highly ranked residues detected as discriminating between the pH1N1 and human seasonal H1N1 strains by AVANA (6) and the method presented here

Position in structure 3lzg	Appears in AVANA analysis	Appears in our highly ranked set
124	yes	no
131	yes	no
132	**yes**	**yes**
133	yes	no
136	yes	no
138	yes	no
140	yes	no
142	yes	no
144	yes	no
145	**yes**	**yes**
149	yes	no
152	yes	no
155	yes	no
156	yes	no
158	yes	no
159	yes	no
160	yes	no
163	yes	no
169	yes	no
171	**yes**	**yes**
173	**yes**	**yes**
182	yes	no
186	yes	no
187	yes	no
188	yes	no
189	yes	no
192	yes	no
193	yes	no
196	yes	no
197	yes	no
198	yes	no
199	yes	no
203	yes	no
205	yes	no
206	no	yes
208	yes	no
211	yes	no
214	yes	no
219	**yes**	**yes**
225	no	yes
230	yes	no
237	yes	no
242	yes	no
244	yes	no
248	yes	no
252	yes	no
253	yes	no
260	yes	no
261	**yes**	**yes**
263	**yes**	**yes**
264	**yes**	**yes**

Positions appearing in both analyses are marked in bold.

Table S6. Highly ranked residues detected as discriminating between the pH1N1 and swine H1N1 strains by AVANA (6) and the method presented here

Position in structure 3lzg	Appears in AVANA analysis	Appears in our highly ranked set
131	no	yes
132	**yes**	**yes**
133_A	no	yes
145	yes	no
149	**yes**	**yes**
171	**yes**	**yes**
186	**yes**	**yes**
188	no	yes
189	**yes**	**yes**
200	no	yes
206	no	yes
208	**yes**	**yes**
210	yes	no
219	yes	no
225	no	yes
226	no	yes
227	yes	no
242	yes	no
261	yes	no
263	yes	no
264	yes	no

Positions appearing in both analyses are marked in bold.