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.

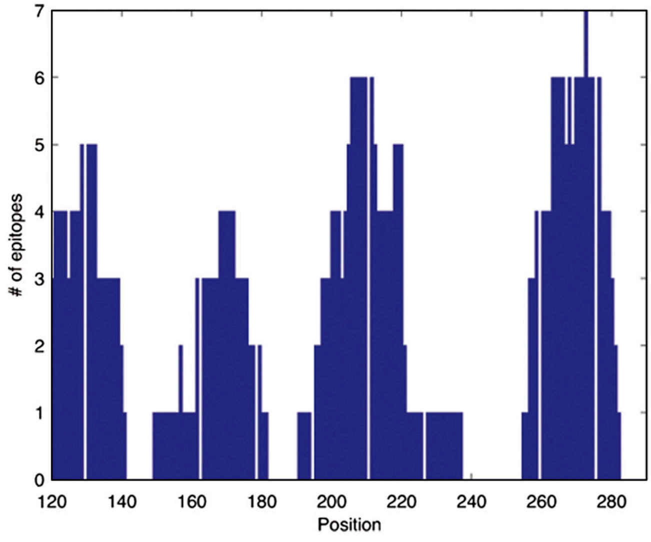

Fig. S1. A histogram of all T- and B-cell epitopes reported for the RBD of influenza A H1N1, or influenza A (unspecified) in the Immune Epitope Database. Seventy-eight percent of the RBD sequence is covered by one or more epitopes.

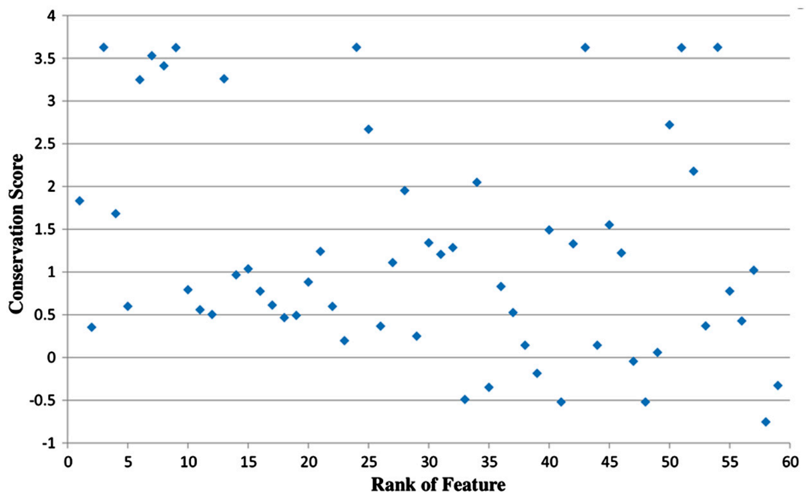

Fig. S2. *The discriminating positions are not necessarily conserved.* Scatter plot of the evolutionary conservation scores versus our calculated rank for the positions that were detected as discriminative between the pH1N1 and the circulating human strains. The evolutionary conservation scores were 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.

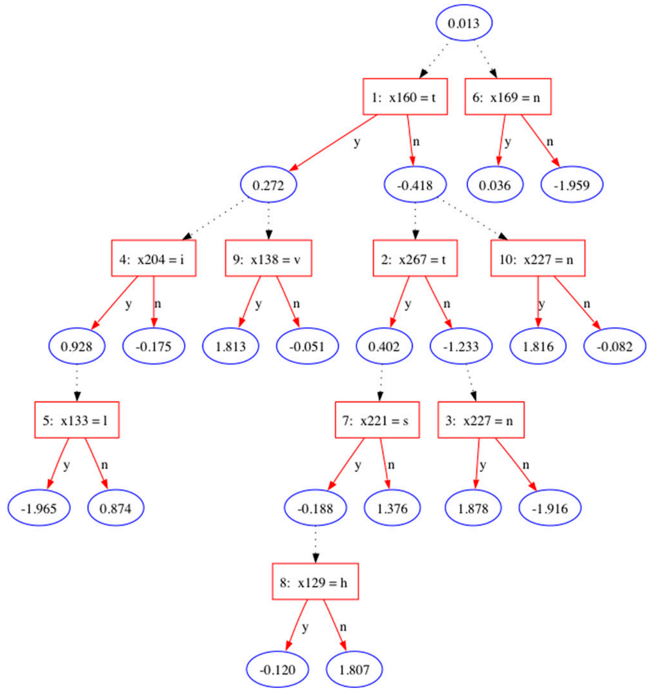

Fig. S3. *Representative ADT obtained after 10 iterations.* The ovals in the decision tree are the prediction nodes, and the rectangles represent the splitter nodes. The final prediction score is obtained by starting from the score of the top prediction node and summing the scores of the relevant prediction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