Predicting influenza antigenicity from Hemagglutinin sequence data based on a joint random forest method

Yao, Yuhua; Li, Xianhong; Liao, Bo; Huang, Li; He, Pingan; Wang, Fayou; Yang, Jiasheng; Sun, Hailiang; Zhao, Yulong; Yang, Jialiang

Published in:
Scientific Reports

Published: 01/12/2017

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

License:
CC BY

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1038/s41598-017-01699-z

Publication details:
Yao, Y., Li, X., Liao, B., Huang, L., He, P., Wang, F., ... Yang, J. (2017). Predicting influenza antigenicity from Hemagglutinin sequence data based on a joint random forest method. Scientific Reports, 7(1), [1545]. DOI: 10.1038/s41598-017-01699-z

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.
Predicting influenza antigenicity from Hemagglutinin sequence data based on a joint random forest method

Yuhua Yao, Xianhong Li, Bo Liao, Li Huang, Pingan He, Fayou Wang, Jiasheng Yang, Hailiang Sun, Yulong Zhao & Jialiang Yang

Timely identification of emerging antigenic variants is critical to influenza vaccine design. The accuracy of a sequence-based antigenic prediction method relies on the choice of amino acids substitution matrices. In this study, we first compared a comprehensive 95 substitution matrices reflecting various amino acids properties in predicting the antigenicity of influenza viruses by a random forest model. We then proposed a novel algorithm called joint random forest regression (JRFR) to jointly consider top substitution matrices. We applied JRFR to human H3N2 seasonal influenza data from 1968 to 2003. A 10-fold cross-validation shows that JRFR outperforms other popular methods in predicting antigenic variants. In addition, our results suggest that structure features are most relevant to influenza antigenicity. By restricting the analysis to data involving two adjacent antigenic clusters, we inferred a few key amino acids mutation driving the 11 historical antigenic drift events, pointing to experimentally validated mutations. Finally, we constructed an antigenic cartography of all H3N2 viruses with hemagglutinin (the glycoprotein on the surface of the influenza virus responsible for its binding to host cells) sequence available from NCBI flu database, and showed an overall correspondence and local inconsistency between genetic and antigenic evolution of H3N2 influenza viruses.

Causing an estimated 500,000 deaths worldwide per year, influenza epidemics in humans seriously endanger population health and world economy. Vaccination is the primary option to reduce influenza outbreaks. The efficacy of a seasonal influenza vaccine depends largely on the selection of vaccine strains, i.e., the strains this vaccine is designed to prevent. A good vaccine recipe should target potential circulating strains able to escape from population immunity in the new flu season. However, influenza viruses are classic examples of antigenically variable pathogens and have a seemingly endless capacity to evade immune response, which makes vaccine design extremely challenging. According to reports from the center for disease control and prevention (CDC), flu shots fail half of the time. As such, timely identification of emerging antigenic variants is critical to influenza vaccine design, flu surveillance, and human health.

One of the most popular assays to evaluate the efficacy of a vaccine against an influenza virus is the hemagglutination inhibition (HI) assay, a binding assay measuring the ability of antiserum (vaccine) to block the hemagglutinin (HA) of the antigen (virus) from agglutinating red blood cells. However, HI assay is labor and cost intensive, which poses the need for efficient computational methods to estimate the antigenic similarity between antigens and antiserums. With the advances of sequencing techniques, influenza sequences have become more and more

1School of Mathematics and Statistics, Hainan Normal University, Haikou, 570100, P. R. China. 2College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou, 310018, P. R. China. 3College of Information Science and Engineering, Hunan University, Changsha, 410082, P. R. China. 4College of Sciences, Zhejiang Sci-Tech University, Hangzhou, 310018, P. R. China. 5School of Mathematics and Information Science, Henan Polytechnic University, Henan, 454000, P. R. China. 6Department of Civil and Environmental Engineering, National University of Singapore, Singapore, 119077, Singapore. 7College of Veterinary Medicine, Huanan Agricultural University, Guangzhou, 510000, P. R. China. 8Department of Mathematics, City University of Hong Kong, Hong Kong, P. R. China. 9Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, NY, 10029, USA. Correspondence and requests for materials should be addressed to Y.Z. (email: zhaoyulong@gmail.com) or J.Y. (email: jialiang.yang@mssm.edu)
available\(^2\), making them good candidates for predicting the antigenicity of new viruses and identifying antigenic variants.

Popular antigenic prediction methods include imputation-based methods\(^5\) and sequence-based methods. Sequence-based methods usually associate mutations in HA proteins with antigenic differences (among viruses) obtained from serological tests\(^5, 8, 9\). The antigenic differences are either quantified by antigenic distances\(^5\) or simply represented by a binary value to indicate if two viruses are antigenic variants\(^9, 10\). For example, Liao et al. tested four algorithms including iterative filtering, multiple regression, logistic regression, and support vector machine to predict antigenic variants from mutations in HA1, a sub-unit of HA forming globular domain\(^11\). They also explored six amino acids substitution models based on physiochemical grouping of 20 amino acids\(^10\). Sun et al. proposed Antigen-Bridges, a bootstrapped ridge regression model, to predict antigenic distances using amino acids substitutions quantified by pattern-induced multisequence alignment (PIMA) in HA1 protein sequences\(^2\). As H3N2 influenza viruses from 1968 to 2003 are grouped into 11 antigenic clusters, i.e., HK68, EN72, VI75, TX77, BK79, SI87, BE89, BE92, WU95, SY97, and FU02 in chronological order\(^5\), they also predicted HA mutations driving antigenic drift events between adjacent clusters and experimentally validated two predicted mutation sets (i.e., from BE92 to WU95 and from WU95 to SY97). Noticing that co-evolution in HA1 might contribute to antigenic evolution, Yang et al. developed a Lasso model incorporating both single and co-mutation features in HA1 protein sequences\(^12\). More recently, Qiu et al. developed a protein structure-based antigenic prediction model\(^13\), and Neher et al. developed an optimization model to interpret known antigenic data and evaluated its ability in predicting composition of future influenza virus populations\(^14\). There are also many other methods in this hot topic, e.g., Huang et al.\(^15\) and Ren et al.\(^16\) to name but a few.

Despite the fact that these methods are greatly helpful in selecting antigenic variants and optimizing vaccine strains, there are still a few points to be improved. First, it is known that the scoring matrices to quantify amino acids substitutions are critical to the accuracy of prediction algorithms\(^10, 13\). However, only a few matrices reflecting partial protein attributes have been tested, e.g., binary substitution matrix\(^17\), physiochemical models\(^10\), PAM250, BLOSUM62, PIMA\(^12\), and structure model\(^13\). A systematic study of the relationship between amino acids attributes and influenza antigenicity is largely missing. AAindex, a database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids, provides an opportunity to fix this gap\(^18\). The predictive powers of the 94 physicochemical and biochemical properties of amino acids in AAindex could be helpful to elucidate their contribution to influenza antigenic evolution. Second, it is unclear if the combination of a few important amino acids attributes can better predict influenza antigenicity. Third, most previous models either predict antigenic variants or adopt linear models. There might be some advantages in predicting continuous antigenic distances using nonlinear models like random forest\(^19\) since antigenic distances have higher resolution than binary values and the relationship among antigenic sites might be nonlinear.

In this study, we propose and test Joint Random Forest Regression (JRFR), a novel algorithm that combines multiple substitution matrices into the random forest algorithm to predict antigenic distances from HA1 protein sequences. We also systematically compare 95 amino acids substitution matrices in predicting the antigenicity of H3N2 influenza viruses. These substitution matrices reflect a comprehensive list of amino acids attributes including structural, physicochemical, and biochemical information. Finally, we explore the relationship between genetic and antigenic evolution of H3N2 influenza viruses based on the prediction results from JRFR.

**Results**

**JRFR: a random forest model to predict antigenic distances.** We illustrated our computational framework in Fig. 1. We trained a predictive model by taking the amino acids’ changes among virus pairs at each protein site as a feature and the pairwise antigenic distances among viruses as the response (see Materials and
then downloaded the HA1 protein sequences (of length 329) for the 253 viruses from NCBI flu database and in AAindex and binary substitution one) by applying the random forest algorithm (Fig. 1a). We then proposed first evaluated the prediction powers of a comprehensive 95 amino acids substitution matrices (i.e, 94 matrices EN72, VI75, TX77, BK79, SI87, BE89, BE92, WU95, SY97, and FU02, respectively. The pairwise antigenic distances among the viruses were calculated based on Metric MDS and were used as responses in our models. We a mutation to antigenic change and antigenic drift events.

acy matrices. Based on the importance score of each protein locus in JRFR, we also evaluated the contribution of selected for predicting the antigenicity of new influenza viruses. For simplicity, we only allowed at most 2 second-the model in a greedy manor to improve prediction accuracy (see Materials and Methods). The top model was matrix. The 94 secondary matrices (matrices other than the major matrix) were then selected one by one to join the model in a greedy manor to improve prediction accuracy (see Materials and Methods). The top model was selected for predicting the antigenicity of new influenza viruses. For simplicity, we only allowed at most 2 secondary-matrices. Based on the importance score of each protein locus in JRFR, we also evaluated the contribution of a mutation to antigenic change and antigenic drift events.

Evaluation of amino acids attributes in predicting influenza antigenicity. We applied our computational framework into the H3N2 human influenza data from 1968 to 2003, a curated HI table of 253 viruses and 79 vaccines. According to Smith et al. the viruses are classified into 11 antigenic clusters namely HK68, EN72, VI75, TX77, BK79, SI87, BE89, BE92, WU95, SY97, and FU02, respectively. The pairwise antigenic distances among the viruses were calculated based on Metric MDS and were used as responses in our models. We then downloaded the HA1 protein sequences (of length 329) for the 253 viruses from NCBI flu database, and aligned them using MUSCLE. The alignment was then transformed into 95 feature matrices based on the 94 amino acids substitution models in AAindex and the binary substitution model. The 10-fold cross-validation root-mean-square errors (RMSEs) for all 95 models were summarized in Supplementary Table S1, among which we listed a few top or well studied models in Table 1.

Methods for details). There are many amino acids substitution matrices reflecting different attributes of amino acids and the choice of substitution model has been proven to be critical in prediction accuracy. Thus, we first evaluated the prediction powers of a comprehensive 95 amino acids substitution matrices (i.e, 94 matrices in AAindex and binary substitution one) by applying the random forest algorithm (Fig. 1a). We then proposed the JRFR algorithm by jointly considering 2 or more substitution matrices. Specifically, we first selected the top 15 substitution matrices according to 10-fold cross-validation prediction accuracy, each was considered as a major matrix. The 94 secondary matrices (matrices other than the major matrix) were then selected one by one to join the model in a greedy manor to improve prediction accuracy (see Materials and Methods). The top model was selected for predicting the antigenicity of new influenza viruses. For simplicity, we only allowed at most 2 secondary-matrices. Based on the importance score of each protein locus in JRFR, we also evaluated the contribution of a mutation to antigenic change and antigenic drift events.

Table 1. The top 12 amino acids substitution matrices in predicting influenza antigenicity.

| Accession No | Description | RMSE  |
|--------------|-------------|-------|
| NIEK910102   | Structure-derived correlation matrix 2 | 0.965 |
| NIEK910101   | Structure-derived correlation matrix 1 | 0.967 |
| RIER950101   | Hydrophobicity scoring matrix | 0.968 |
| MIYS930101   | Base-substitution-protein-stability matrix | 0.968 |
| DOSZO10102   | Normalised version of SM_SAUSAGE | 0.970 |
| LUTR910102   | Structure-based comparison table for inside other class | 0.970 |
| AZAE970102   | The substitution matrix derived from spatially conserved motifs | 0.971 |
| BENS940101   | Log-odds scoring matrix collected in 6.4-8.7 PAM | 0.972 |
| TUDE900101   | Isomorphism of replacements | 0.973 |
| AZAE970101   | The single residue substitution matrix from interchanges of spatially neighbouring residues | 0.973 |
| HENS920102   | BLOSUM62 substitution matrix | 0.986 |
| DAYM780301   | Log-odds matrix for 250 PAMs | 1.000 |
| Binary       | 1 for substitution and 0 for match | 1.098 |

Evaluation of amino acids attributes in predicting influenza antigenicity. We applied our computational framework into the H3N2 human influenza data from 1968 to 2003, a curated HI table of 253 viruses and 79 vaccines. According to Smith et al. the viruses are classified into 11 antigenic clusters namely HK68, EN72, VI75, TX77, BK79, SI87, BE89, BE92, WU95, SY97, and FU02, respectively. The pairwise antigenic distances among the viruses were calculated based on Metric MDS and were used as responses in our models. We then downloaded the HA1 protein sequences (of length 329) for the 253 viruses from NCBI flu database, and aligned them using MUSCLE. The alignment was then transformed into 95 feature matrices based on the 94 amino acids substitution models in AAindex and the binary substitution model. The 10-fold cross-validation root-mean-square errors (RMSEs) for all 95 models were summarized in Supplementary Table S1, among which we listed a few top or well studied models in Table 1.

Since influenza antigenicity is known to associate with many attributes of HA protein, e.g., structure and hydrophobicity, we conjectured that it might be helpful to combine a few amino acids characteristics. Thus, we also applied our JRFR model by combining multiple substitution matrices. The prediction accuracy is indeed improved by JRFR (see Table 2 for a few top combinations), confirming that influenza antigenicity is affected by many protein attributes. The best combination uses NIEK910101 mutation matrix ("Structure-derived correlation matrix 1") as the main feature matrix and RUSR970101 ("Substitution matrix based on structural alignments of analogous proteins") and KOS950107 ("Context-dependent optimal substitution matrices for buried turn") as auxiliary feature matrices. The 10-fold cross-validation RMSE could be reduced to 0.941 (with prediction accuracy 0.963, sensitivity 0.980, specificity 0.775 and the Mathew’s correlation coefficient (MCC) 0.756) by this model. It is of note that the prediction accuracy could be further improved by adding new secondary matrices, however, the improvement is marginal. As such, we adopt the best model with 3 feature matrices for all subsequent analyses.
A template for scientific text, including tables and figures.
Mutations driving H3N2 influenza antigenic drift events. There are 10 antigenic drift events for H3N2 influenza viruses occurred between 1968 and 2003. We also applied JRFR on viruses in chronologically adjacent antigenic clusters to infer amino acids mutations driving specific drift event based on importance scores. The combination of mutations responsible for the 10 drift events were listed in Table 4. Interestingly, we observed a few antigenic sites critical for multiple antigenic drift events. For example, site 193 drove the early two antigenic drift events, i.e., HK68-EN72 and EN72-VI75. Site 145 was involved in 4 antigenic drifts, i.e., EN72-EN75, SI87-BE89, BE89-BE92, and BE92-WU95.

To validate the power of these sites in driving antigenic drift events, we manually selected one virus at the center of one antigenic group, and mutated its amino acids, by which we obtained a few artificial variants of the selected virus. We then predicted the antigenicity of the variants (based on its HA1 sequence) and test if it has similar antigenicity with viruses in the later antigenic group. For a better view, we plotted the antigenic cartography of the viruses in two adjacent antigenic clusters, the selected virus, and its artificial variants (see Fig. 5). We only plotted the late 4 clusters since they have more viruses. As one can see, in most cases the predicted mutation combinations are capable of driving the antigenic drifts. Sun et al. experimentally validated that N145K can drive BE92-WU95, which is consistent with our predictions.

Discussion
As we know, the antigenicity of influenza viruses changes rapidly and vaccines should be updated accordingly to avoid influenza outbreaks. The vaccines have been updated at least 27 times for H3N2 viruses since 1968, 9 times for H1N1 virus from 1977 to 2009, and 15 times for influenza B virus from 1972 to 2011. However, the selection of vaccine strains is non-trivial due to the exceptional large number of new influenza viruses in the new season and the burden in testing antigenicity of these virus against known vaccines. As a result, flu shots from CDC fail half the time. Thus, timely surveillance of the antigenic evolution of these viruses is critical. However, traditional experimental methods such as HI assay for evaluating the antigenicity of new viruses face a multitude of issues.
For example, the reduction seen in H3N2 virus binding to red blood cells\textsuperscript{26,27} can lead to problems performing and interpreting HI assays. Furthermore, because antigenic characterization is relatively labor-intensive, only a small portion (generally, fewer than 20\%) of the influenza isolates sequenced will be antigenically characterized. With the development of sequencing techniques, influenza protein sequencing has become a routine for influenza studies. Many newly sequenced influenza protein sequences have been stored in influenza databases like NCBI influenza database\textsuperscript{15}. Much of the burden for influenza surveillance could be avoided if reliable sequence-based antigenic prediction methods can be established. JRFR produces relative low prediction error and reasonable antigenic cartography, which could serve as an initial screen of antigenic variants for further experimental analyses.

In the JRFR framework, we systematically evaluated the powers of 95 amino acids substitution matrices in predicting influenza antigenicity, which reflecting various physicochemical and biochemical properties of amino acids. We found that structure-based features outperformed all other features, followed by amino acids hydrophobicity. The importance of HA structure in influenza antigenicity has long been identified\textsuperscript{28}. Interestingly, commonly used substitution matrices including binary matrix, PAM250 and BLOSUM62 are not performing well, suggesting the necessity of combining protein structure information in studies to predict phenotypes using protein sequences. By combining a few feature matrices, the prediction accuracy could be further improved, indicating that influenza antigenicity might be determined by multiple factors, though HA structure might play dominant roles.

In addition, we inferred 34 antigenically important protein sites, most of which are located at the 5 epitopes, especially in epitope A and B. Ndifon \textit{et al.}\textsuperscript{29} presented a competitive model to predict antibody escape and proposed that antigenic drift events would be associated with amino acid changes that occur in epitopes with high neutralization efficiencies (i.e., epitopes A, B, and D) rather than in those with low neutralization efficiencies (i.e., epitopes C and E). The results are consistent with our findings. Meanwhile, we found that a small subset of amino

| Site | Importance Score | Antigenic Domain | Site | Importance Score | Antigenic Domain |
|------|-----------------|-----------------|------|-----------------|-----------------|
| 189  | 5.273           | B               | 262  | 3.678           | E               |
| 2    | 4.651           | other           | 157  | 3.664           | B               |
| 159  | 4.645           | B               | 193  | 3.646           | B               |
| 163  | 4.541           | B               | 190  | 3.545           | B               |
| 278  | 4.421           | C               | 131  | 3.460           | A               |
| 156  | 4.417           | B               | 62   | 3.453           | E               |
| 133  | 4.196           | A               | 226  | 3.408           | D               |
| 158  | 4.167           | B               | 299  | 3.406           | C               |
| 145  | 4.133           | A               | 137  | 3.383           | A               |
| 276  | 4.116           | C               | 248  | 3.326           | D               |
| 135  | 3.915           | A               | 144  | 3.270           | A               |
| 196  | 3.899           | B               | 197  | 3.189           | B               |
| 307  | 3.890           | C               | 172  | 3.154           | D               |
| 173  | 3.862           | D               | 142  | 3.091           | A               |
| 94   | 3.859           | E               | 217  | 2.998           | D               |
| 83   | 3.775           | E               | 143  | 2.946           | A               |
| 155  | 3.712           | B               | 121  | 2.936           | D               |

Table 3. The top 34 antigenic importance sites according to importance scoring in JRFR for H3N2 influenza data.

| Antigenic drift events | Combination of mutations |
|------------------------|--------------------------|
| HK68-EN72              | S193N-G144D              |
| EN72-VI75              | N53D-S193D-S145N         |
| VI75-TX77              | S137G                    |
| TX77-BK79              | D144V-N2K                |
| BK79-S87               | Y155H-K189R              |
| S87-BE89               | N145K,N145K-G135E,N145K-N193S |
| BE89-BE92              | K145N-E156K-R189S,K145N-E156K-T262N-R189S,K145N-E156K-S133D-R189S |
| BE92-WU95              | K135T-N145K-L226V,K135T-N145K-N262S |
| WU95-SY97              | V196A-N276K-E158K-K156Q-V196A-N276K-E158K-K156Q-K62E |
| SY97-FU02              | A131T-H155T-V202I        |

Table 4. Multiple mutations driving 10 antigenic drifts for H3N2 influenza viruses inferred by JRFR.
acids mutations could drive most antigenic drift events for H3N2 influenza viruses, confirming the result from a few previous studies. In the end, we would like to point out that our framework works generally for all influenza sub-types though we used H3N2 influenza as an illustration. In the future, we plan to predict the antigenicity of other sub-types like H1N1, H5N1 and H7N9, and compare their antigenically associated amino acid mutations or antigenic determinant regions. In addition, our method could also be applied to investigate other problems, such as disease and drug prediction, DNA-binding protein prediction, protein fold recognition, detection of tubule boundary, and other related problems etc. However, it is out of the scope of this study.

Materials and Methods

**Influenza data.** In this study, we adopted the H3N2 influenza data from Smith et al. which contains a partially revealed HI table consisting of 253 viruses and 79 vaccines from 1968 to 2003. There are 11 major antigenic clusters including HK68, EN72, VI75, TX77, BK79, SI87, BE89, BE92, WU95, SY97 and FU02, named by vaccine strain in the respective cluster. We also downloaded the HA protein sequences of the 253 viruses and those of a total 1638 viruses available from NCBI flu database from 1968 to 2014. The HA protein sequences were aligned by MUSCLE, and we only kept the 329 sites belong to HA1 protein for further analyses.

**The random forest algorithm.** We applied the random forest algorithm to predict antigenic distances using 95 single amino acids substitution models. It is a nonlinear ensemble algorithm taking each non-conservative protein site as a feature and pair-wise antigenic distances among viruses as responses. Let $n$ be the number of viruses ($n = 253$ for H3N2 data) and $S$ be the alignment of their HA1 sequences. Since conservative sites are non-informative, we removed them from $S$. Let $m$ be the number of non-conservative sites ($m = 154$ for H3N2 data). Thus after removing conservative sites, $S$ is an $n \times m$ matrix with each entry being an amino acid.

**Figure 5.** Antigenic cartographies to illustrate the key mutations driving the 4 antigenic drift events including (a) SI87-BE89, (b) BE89-BE92, (c) BE92-WU95, and (d) WU95-SY97.
Feature construction. We constructed the feature matrix $X$ as follows: Let $S_{ij}$ be the amino acid at position $(i, j)$ in $S$ and $X_j$ be the $j^{th}$ column of $X$. Then $X_j = \{D_{S_{ij}, S_{ik}}\}$, a vector traversing all ordered pairs of $(k, l)$ with $1 \leq k \leq l \leq n$. Clearly, $X_j$ is a vector of length $N = \binom{n}{2}$ and $X$ is an $N \times m$ matrix. Here, $D_{ab}$ denotes the dissimilarity between amino acids $a$ and $b$. It is of note that AAIndex only provides similarity matrix $A$ among amino acids. We transformed $A$ into dissimilarity matrix $D$ using the following formula,

$$D_{a,b} = (A_{a,a} + A_{b,b}) - 2A_{a,b}. \tag{1}$$

Response. We adopted antigenic distance proposed by Smith et al.\textsuperscript{5} as our response $y$, where $y$ consists of all distances between ordered pairs of viruses. Clearly, $y$ is also of length $N$. Given an HI table of $n$ viruses, $y$ could be calculated directly from the website provided in ref.\textsuperscript{5}.

After the feature matrix $X$ and the response vector $y$ were constructed, we applied the random forest function in R package ‘randomForest’ to construct the prediction model\textsuperscript{46}. We set the bootstrapping (tree) number to be 500 and the number of features to be 500 and the number of features to be $\frac{m}{4}$ for each tree. We then tested and compared the 95 dissimilarity matrices based on a 10-fold cross-validation process and used root-mean-square error (RMSE) to evaluate their performances, which are described as follows.

RMSE. Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_N)$ be two prediction vectors, then the RMSE between $\alpha$ and $\beta$ is defined as

$$\text{RMSE}(\alpha, \beta) = \sqrt{\frac{\sum_{i=1}^{N} (\alpha_i - \beta_i)^2}{N}} \tag{2}$$

Accuracy and specificity. For a pair of viruses with underlying antigenic distance $y$, we define them to be true antigenic variants if $y \geq 2$, and false otherwise. Similar they are called positive antigenic variants if the predicted antigenic distance $\hat{y} \geq 2$ by JRFR, and negative otherwise. By this way, we can define true positive (TP) pairs (i.e., $y \geq 2$ and $\hat{y} \geq 2$), false positive (FP) ($y < 2$ and $\hat{y} \geq 2$), true negative (TN) ($y < 2$ and $\hat{y} < 2$), and false negative (FN) ($y \geq 2$ and $\hat{y} < 2$). The accuracy of a method is defined as

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

and the specificity is defined as

$$\text{specificity} = \frac{TN}{TN + FP} \tag{4}$$

10-fold cross-validation. We divided $y$ randomly into 10 equal parts, 9 parts of which (with the corresponding sub-matrices of $X$) are used for training and the remaining one part for prediction. The process is repeated for 10 times until each part is used as the prediction set once. By merging the prediction results for each part, we obtained a prediction vector $\hat{y}$. The RMSE between $y$ and $\hat{y}$ was used to tune model parameters and compare the performances of different models.

The JRFR algorithm. In the JRFR algorithm, we combined multiple feature matrices (derived from different amino acid similarity matrices in AAIndex\textsuperscript{39}) to construct decision trees in random forest algorithm. Specifically, let $X^{(1)}, X^{(2)}, \ldots, X^{(k)}$ be $k$ feature matrices, each consisting of $m$ features corresponding to $m$ non-conservative sites in HA1. We set one of the feature matrices, say $X^{(1)}$ to be the main matrix and other to be auxiliary ones. For each auxiliary matrix $X^{(i)}$ with $2 \leq i \leq k$, we then applied an linear regression process to remove its overlapping information from the main matrix as follows: let $X^{(j)}$ and $X^{(1)}$ be the $j^{th}$ feature vectors of $X^{(i)}$ and $X^{(1)}$ respectively. There exist constants $\alpha, \beta, \rho$ and vector $\varepsilon_j$ such that

$$X^{(i)}_j = \alpha_j + \beta_j X^{(1)}_j + \varepsilon_j \tag{5}$$

where $\alpha_j$ is the intercept, $\beta_j$ is the regression coefficient and $\varepsilon_j$ is the residual vector. We constructed a new feature matrix with each feature vector being $\varepsilon_j$, for $1 \leq j \leq m$. For brevity, we still used $X^{(i)}$ to represent this matrix for $2 \leq i \leq k$. By this way, we obtained new $k - 1$ auxiliary matrices independent of $X^{(1)}$.

We then combined the main feature matrix and new auxiliary ones to predict antigenic distance using random forest algorithm. That is, we extended the number of features to $k \times m$, and applied the usual random forest process introduced above. In practice, we only used the top 15 feature matrices (as ranked in Supplementary Table S1) as main feature matrices and restricted the number of auxiliary feature matrices to be less than 2. Similar to single feature matrix analysis, we set the bootstrapping (tree) number to be 500 and the number of features to be $\frac{k \times m}{4}$ for each tree.

Importance score. To calculate the importance of a feature, the random forest algorithm permutes the values of this feature. It then calculates and normalizes the difference of out-of-bag error before and after permutation\textsuperscript{35}. By applying the random forest function in R, we obtained importance scores for features in both main feature matrix
and auxiliary ones. Let $I^{(i)}_j$ be the importance score of the $j^{th}$ feature in the main feature matrix and $I^{(i)}_j$ with $2 \leq i \leq k$ be its importance scores in auxiliary feature matrices. Since the $j^{th}$ feature in both main matrix and auxiliary ones represents information from the $j^{th}$ non-conservative site in HA1 sequence, we defined the importance of the $j^{th}$ non-conservative site to be $\sum_{i=2}^{k} I^{(i)}_j$.

**Building the JRFR prediction model.** We sorted all non-conservative sites in descending order based on their importance scores, and selected site one by one from top to join the JRFR algorithm, until the best 10-fold cross-validation RMSE is obtained. For H3N2 data, we selected the top 85 sites to building the final JRFR model by this process.

**Construction of antigenic and genetic cartography.** The antigenic map was directly constructed from the predicted antigenic distance matrix among viruses using classical multidimensional scaling (MDS). To construct the genetic map, we first calculated the p-distance matrix between pair of viruses and then applied classical MDS.

**References**

1. WHO. Influenza(seasonal). Fact sheet No 211 (2014).
2. Sun, H. et al. Using sequence data to infer the antigenicity of influenza virus. MBio 4(4), e00230-13, doi:10.1128/mBio.00230-13 (2013).
3. Blackburne, B. P., Hay, A. J. & Goldstein, R. A. Changing selective pressure during antigenic changes in human influenza h3. PLoS Pathogens 4(5), e1000558, doi:10.1371/journal.ppat.1000558 (2008).
4. CDC. CDC admits flu shots fail half the time. http://www.nvic.org/nvic–vaccine–news/april–2016/cdc–admits–flu–shots–fail–half–the–time.aspx.
5. Smith, D. J. et al. Mapping the antigenic and genetic evolution of influenza virus. Science 305(5682), 371–376, doi:10.1126/science.1097211 (2004).
6. Hirst, G. K. Studies of antigenic diversity among strains of influenza a by means of read cell agglutination. The Journal of Experimental Medicine 78(5), 407–423, doi:10.1084/jem.78.5.407 (1943).
7. Bao, Y. et al. The influenza virus gene resource at the national center for biotechnology information. Journal of Virology 82(2), 596–601, doi:10.1128/JVI.02005-07 (2008).
8. Barnett, J. L., Yang, J., Cai, Z., Zhang, T. & Wan, X. F. Antigenic map 3d: an online antigenic cartography resource. Bioinformatics 28, 1292–1293, doi:10.1093/bioinformatics/bts105 (2012).
9. Lee, M. S. & Chen, J. S. Predicting antigenic variants of influenza a/h3n2 viruses. Emerging Infectious Diseases 10(8), 1385–1390, doi:10.3202/eed.100640107 (2004).
10. Liao, Y. C., Lee, M. S., Ko, C. Y. & Hsiung, C. A. Bioinformatics models for predicting antigenic variants of influenza a/h3n2 virus. Bioinformatics 24, 505–512, doi:10.1093/bioinformatics/btm638 (2008).
11. Wang, W., DeFeo, C. J., Alvarado-Facundo, E., Vassell, R. & Weiss, C. D. Intermonomer interactions in hemagglutinin subunits h1 and h2 affecting hemagglutinin stability and influenza virus infectivity. Journal of Virology 89(20), 10602–10611, doi:10.1128/JVI.00899-15 (2015).
12. Yang, J., Zhang, T. & Wan, X. F. Sequence-based antigenic change prediction by a sparse learning method incorporating co-evolutionary information. PLoS One 20, 317–330 (2009).
13. Qiu, J., Qiu, T., Yang, Y., Wu, D. & Cao, Z. Incorporating structure context of ha protein to improve antigenicity calculation for influenza virus/a/h3n2. Scientific Reports 6, 31156, doi:10.1038/srep31156 (2016).
14. Neher, R. A., Bedford, T., Daniels, R. S., Russell, C. A. & Shraiman, B. I. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proceedings of the National Academy of Sciences of the United States of America 113, E1701–1709, doi:10.1073/pnas.1525578113 (2016).
15. Huang, J. W., King, C. C. & Yang, J. M. Co-evolution positions and rules for antigenic variants of human influenza a/h3n2 viruses. BMC Bioinformatics 10(Suppl. 1), S41, doi:10.1186/1471-2105-10-S1-S41 (2009).
16. Ren, X. et al. Computational identification of antigenicity-associated sites in the hemagglutinin protein of a/h1n1 seasonal influenza virus. PLoS One 10(5), e0126742, doi:10.1371/journal.pone.0126742 (2015).
17. Cai, Z. et al. Identifying antigenicity-associated sites in highly pathogenic h5n1 influenza virus hemagglutinin by using sparse learning. Journal of Molecular Biology 422(1), 145–55, doi:10.1016/j.jmb.2012.05.011 (2012).
18. Kawashima, S. et al. Aaindex: amino acid index database, progress report 2008. Nucleic Acids Research 36 (Database issue), D202euc2025 (2008).
19. Tourn, W. G. et al. Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? Briefings in Bioinformatics 14(3), 315–26, doi:10.1093/bib/bbs034 (2013).
20. Edgar, R. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32(5), 1792–1797, doi:10.1093/nar/gkh340 (2004).
21. Wiley, D. C., Wilson, I. A. & Skelh, J. J. Structural identification of the antibody-binding sites of hong kong influenza haemagglutinin and their involvement in antigenic variation. Nature 298, 373–378, doi:10.1038/35073397 (1981).
22. Wilson, I. A. & Cox, N. Structural basis of immune recognition of influenza virus hemagglutinin. Annual Review of Immunology 8, 737–787, doi:10.1146/annurev.immunol.080403.160353 (1990).
23. Kilbourne, E. D. Future influenza vaccines and the use of genetic recombinants. Bulletin of the World Health Organisation 41(3), 643–645 (1969).
24. Chambers, B., Parkhouse, K., Ross, T., Alby, K. & Hensley, S. Identification of hemagglutinin residues responsible for h3n2 antigenic drift during the 2014–015 influenza season. Cell Reports 12(1), 1–6, doi:10.1016/j.celrep.2015.06.005 (2015).
25. Koel, B. F., Burke, D. F., Bestebroer, T. M. & Vlet, S. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. Science 342(6161), 976–979, doi:10.1126/science.1244730 (2013).
26. Morishita, T., Nobsuwa, E. & Nakajima, S. Studies on the molecular basis for loss of the ability of recent influenza a (h1n1) virus strains to agglutinate chicken erythrocytes. Journal of General Virology 77, 2499–2506, doi:10.1099/0022-1317-77-10-2499 (1996).
27. Nobsuwa, E. et al. Change in receptor-binding specificity of recent human influenza a viruses (h3n2): A single amino acid change in hemagglutinin altered its recognition of sialyloligosaccharides. Virology 278(2), 587–596, doi:10.1006/viro.2000.0679 (2000).
28. Wilson, I. et al. The structure of an antigenic determinant in a protein. Cell 37(3), 767–778 (1984).
29. Nötf, W., Wingreen, N. S. & Levin, S. A. Differential neutralization efficiency of hemagglutinins epitopes, antibody interference, and the design of influenza vaccines. Proceedings of the National Academy of Sciences of the United States of America 106, 8701–8706, doi:10.1073/pnas.0909342710 (2009).
30. Wei, L., Tang, J. & Zou, Q. Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information. Information Sciences 384, 135–144, doi:10.1016/j.ins.2016.06.026 (2017).
31. Wei, L. & Zou, Q. Recent Progress in Machine Learning-Based Methods for Protein Fold Recognition. *International Journal of Molecular Sciences* 17(12), 2118. doi:10.3390/ijms17122118 (2016).
32. Su, R. et al. Detection of tubule boundaries based on circular shortest path and polar-transformation of arbitrary shapes. *Journal of microscopy* 264(2), 127–142. doi:10.1111/jmi.2016.264.issue-2 (2016).
33. Wei, L., Xing, P., Shi, G., Ji, Z. L. & Zou, Q. Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, doi:10.1109/TCBB.2017.2670558 (2017).
34. Wei, L., Xing, P., Tang, J. & Zou, Q. PhosPred-RF: a novel sequence-based predictor for phosphorylation sites using sequential information only. *IEEE Transactions on Nanobioscience*, doi:10.1109/TNB.2017.2661756 (2017).
35. Liaw, A. & Wiener, M. Classification and regression by randomforest. *R News* 20, 317–330 (2009).
36. Liaw, A. & Wiener, M. Package randomForest. https://www.stat.berkeley.edu/~breiman/RandomForests/.

**Acknowledgements**
This work was partially supported by the National Natural Science Foundation of China (No. 61272312) and the 521 Talent Cultivation Plan of Zhejiang Sci-Tech University.

**Author Contributions**
J.L.Y. and Y.Z. conceived and designed the experiments. Y.Y., X.L. and B.L. performed the experiments and analysed the data. J.L.Y., X.L. and Y.Z. wrote the paper. Y.Y., B.L., L.H., F.W., J.S.Y., H.S. contribute to the discussion, and helped to revise the paper.

**Additional Information**
Supplementary information accompanies this paper at doi:10.1038/s41598-017-01699-z

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2017