Bioinformatics, 35(1), 2019, 77–87
doi: 10.1093/bioinformatics/bty457
Advance Access Publication Date: 7 June 2018
Original Paper

OXFORD

Systems biology

Graph-guided multi-task sparse learning model: a method for identifying antigenic variants of influenza A(H3N2) virus

Lei Han[1,2], Lei Li[1], Feng Wen[1], Lei Zhong[1], Tong Zhang[2] and Xiu-Feng Wan[1,*]

[1]Department of Basic Science, College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39759, USA and [2]Tencent AI Lab, Tencent, Shenzhen 518052, China

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on November 7, 2017; revised on March 30, 2018; editorial decision on June 1, 2018; accepted on June 6, 2018

Abstract

Motivation: Influenza virus antigenic variants continue to emerge and cause disease outbreaks. Time-consuming, costly and middle-throughput serologic methods using virus isolates are routinely used to identify influenza antigenic variants for vaccine strain selection. However, the resulting data are notoriously noisy and difficult to interpret and integrate because of variations in reagents, supplies and protocol implementation. A novel method without such limitations is needed for antigenic variant identification.

Results: We developed a Graph-Guided Multi-Task Sparse Learning (GG-MTSL) model that uses multi-sourced serologic data to learn antigenicity-associated mutations and infer antigenic variants. By applying GG-MTSL to influenza H3N2 hemagglutinin sequences, we showed the method enables rapid characterization of antigenic profiles and identification of antigenic variants in real time and on a large scale. Furthermore, sequences can be generated directly by using clinical samples, thus minimizing biases due to culture-adapted mutation during virus isolation.

Availability and implementation: MATLAB source codes developed for GG-MTSL are available through http://sysbio.cvm.msstate.edu/files/GG-MTSL/.

Contact: wan@cvm.msstate.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Each year in the United States, influenza causes >200 000 hospitalizations and ~23 000 deaths, and many more hospitalizations and deaths occur globally (Thompson *et al.*, 2010, 2004). Vaccination is the primary strategy for reducing the impact of influenza outbreaks (Harper *et al.*, 1984). However, antigenic changes caused by antigenic drift or shift at virus surface glycoproteins, especially hemagglutinin (HA), allow influenza viruses to evade the herd immunity acquired by a population from prior infections or vaccination. The key to a successful influenza vaccination program is to select a vaccine candidate that antigenically matches the viruses that will be circulating during the coming influenza season.

Serologic assays, such as hemagglutination inhibition (HI) and neutralization inhibition assays, are routinely used during the influenza vaccine strain selection process to identify influenza antigenic variants. However, these serologic assays are labor intensive, costly and middle-throughput, and they require the isolation of virus. Thus, a genomic sequence-based strategy for antigenic variant identification would be ideal because the genomic sequences can be obtained directly from clinical samples, which is efficient and economic. Such a method must be able to quantify antigenicity directly using genomic sequences. That is, a quantitative function between the virus genetic information (i.e. mutations in protein sequences) to the virus antigenic properties (i.e. changes in antigenicity) should be

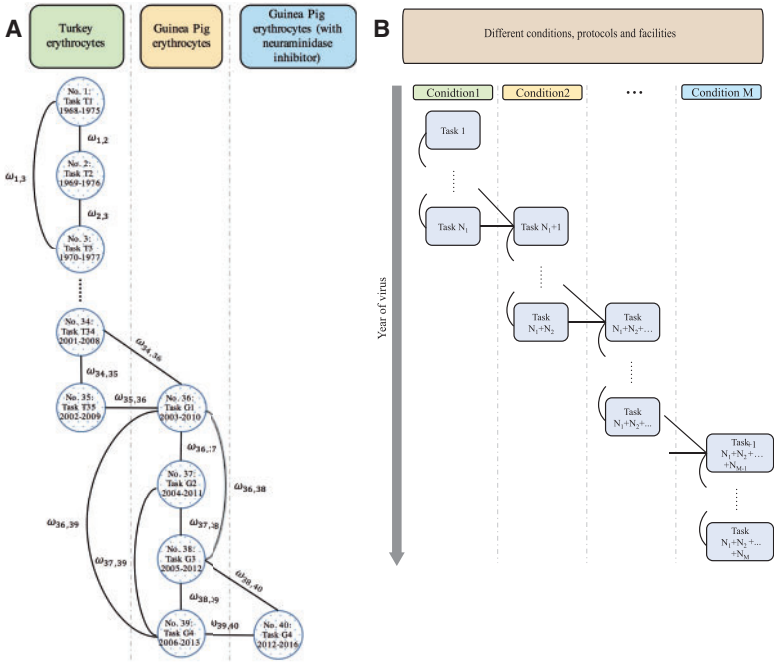

Fig. 1. Graph structure of the multi-task sparse learning model. (**A**) The tasks are first divided into three groups according to different data sources (i.e. HI datasets generated using turkey erythrocytes without neuraminidase inhibitor, guinea pig erythrocytes without neuraminidase inhibitor or guinea pig erythrocytes with neuraminidase inhibitor). Then, in each group, the tasks are formulated by sliding windows, denoted by circles. The edges indicate the information sharing among tasks with task similarity weight $\omega_{i,j}$. (**B**) A general graph structure of the multi-task learning concept

developed. In the past decade, there have been a few attempts on these efforts. For instance, a simple statistical analysis of the correlation between the HI titer and the count of mutations was used by (Lee and Chen, 2004); Regression and Bayesian models were introduced by treating the mutations as features and the HI values or antigenic similarities between sequences as responses (Harvey *et al.*, 2016; Liao *et al.*, 2008; Mansfield, 2007; Ren *et al.*, 2015); more recently, sparse learning techniques (Cai *et al.*, 2012; Neher *et al.*, 2016; Sun *et al.*, 2013; Yang *et al.*, 2014) have been proposed to reduce the affine relationship to sparse structure with concentration on a few key mutated residues. Nevertheless, these prior studies have demonstrated that only a small number of residues on influenza surface glycoproteins, especially antibody binding sites at HA, are associated with antigenic drift of influenza viruses (Smith *et al.*, 2004; Sun *et al.*, 2013), providing rationales for applying sparse learning based methods in developing an effective sequence-based antigenic variant predictor using serologic data.

One necessary condition for developing a robust sequence-based antigenicity inference methods is the availability of large scale of both genomic and serologic data, which must be derived from influenza viruses with much diversity on both genetic mutations and antigenic characteristics. Fortunately, a large set of serologic data for H3N2 influenza viruses have been generated during past decades, providing an opportunity to develop, validate and apply machine learning in antigenic analyses. However, such serologic data are notoriously noisy and difficult to interpret because of inherent variations in reagents, supplies and protocol implementation by laboratory personnel (Yuan *et al.*, 2013) and because of human error (Ampofo *et al.*, 2012). In addition, over time, the protocols have been updated to minimize the effects of changing biologic attributes during virus evolution. For example, erythrocytes from various hosts were used to improve hemagglutination (Ampofo *et al.*, 2012), and neuraminidase inhibitor was used to pretreat influenza viruses before HI assays to minimize the effects of neuraminidase-mediated hemagglutination (Lin *et al.*, 2010). Thus, integrating such data for machine learning is not a trivial task (Yuan *et al.*, 2013). To the best of our knowledge, none of the existing methods can perform sequence-based antigenicity inference effectively with direct usage of these large amount of diverse data. In this study, we developed a novel a Graph-Guided Multi-Task Sparse Learning (GG-MTSL) model to learn antigenicity-associated residues from multi-sourced serologic data. A quantitative model was developed to determine antigenic distances between any two viruses given their HA protein sequences, and this model was further applied to illustrate the antigenic drift patterns of these human A(H3N2) influenza viruses.

2 Materials and methods

2.1 GG-MTSL model

2.1.1 Problem formulation

The overall goal of this study was to develop a genomic sequence-based antigenicity inference method. Our serologic datasets were composed of data generated by using three different protocols: turkey erythrocytes with untreated viruses, guinea pig erythrocytes with untreated viruses, and guinea pig erythrocytes with neuraminidase inhibitor-pretreated viruses. Viruses involved in these datasets span a long time period and may not react with each other; the datasets included a large quantity of low reactors, had missing HI titers, and had a unique distribution of data (Cai *et al.*, 2010), presenting a challenge in matrix completion and determination of accurate distances for the low reactors. We formulated the problem of dealing with different protocols and with viruses spanning a long time period into a multi-task problem by separating the data into multiple temporal tasks (Fig. 1); within each task, the HI data was generated by using the same protocol and with a minimal number of

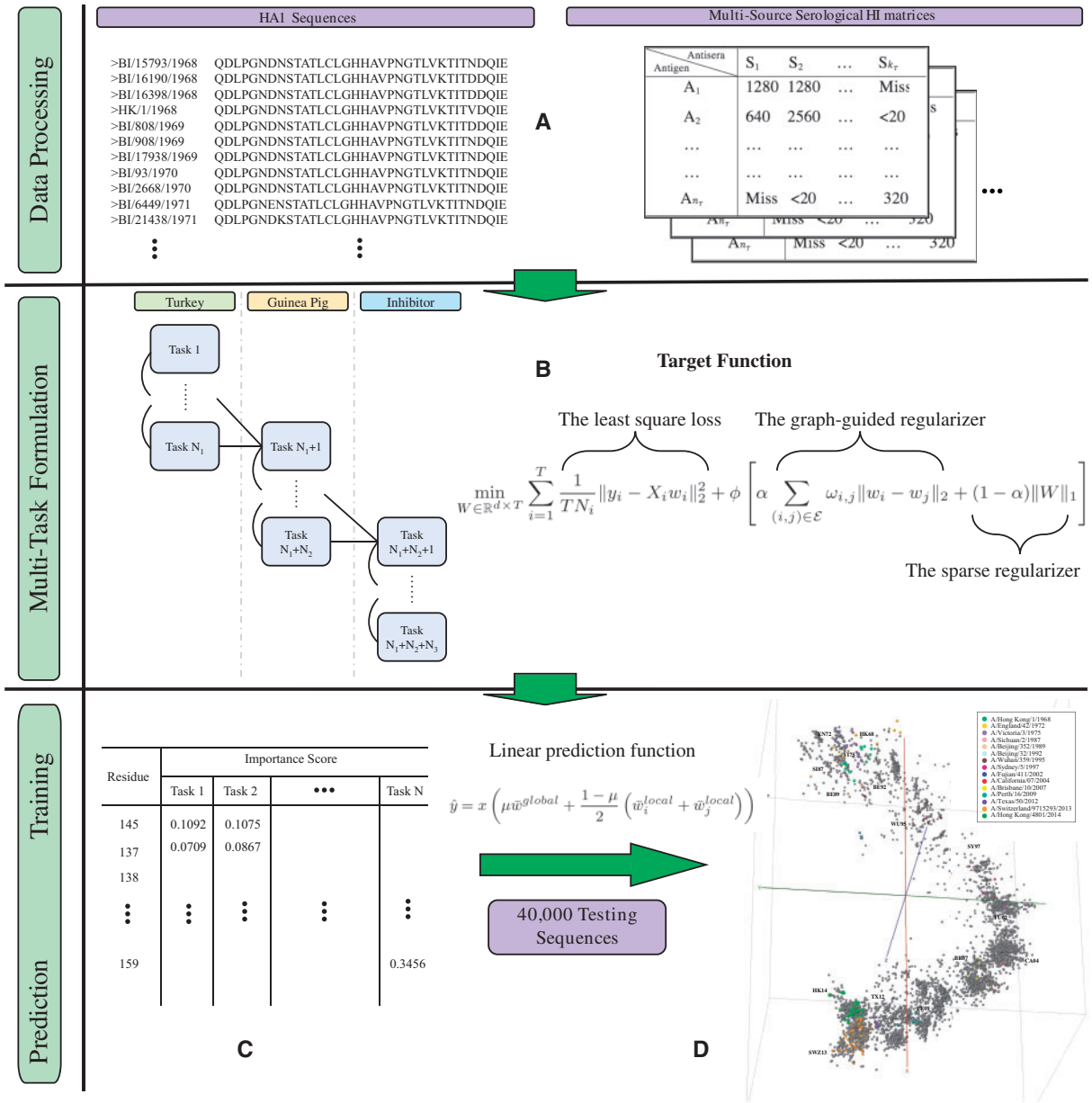

Fig. 2. Workflow of the multi-task learning system. (**A**) The data processing integrates two types of data: sequence data (e.g. HA1 sequences shown in the left panel) and serologic data (e.g. HI data in the right panel). (**B**) Multiple tasks are formulated and integrated via a graph (Fig. 1). Specifically, in this study, the serologic data from multiple sources (e.g. data generated in a different time or using different protocols [i.e. HI datasets generated using turkey erythrocytes without neuraminidase inhibitor, guinea pig erythrocytes without neuraminidase inhibitor or guinea pig erythrocytes with neuraminidase inhibitor]) were separated into >50 individual tasks and processed by the multi-task matrix completion model. (**C**) Graph-based multi-task feature learning is conducted to identify and integrate influenza virus antigenicity-associated sites and their weights for each individual task. The finalized residues and associated weights are used to develop an ensemble prediction model to quantify antigenic distances given protein sequences. (**D**) Large-scale, sequence-based antigenic maps are constructed, and antigenic evolution of influenza viruses is studied by using data mining and machine learning (e.g. spectral clustering to antigenic drift events and Bayesian modeling to identify temporal and spatial origins for influenza antigenic variants)

low reactors (Supplementary Fig. S1). Of note, the problem formulation can be conveniently generalized to any multi-sourced dataset by splitting tasks according to protocols or specific settings, and then within each source the tasks can be further decentralized along the temporal (Fig. 1A) or other dimensions (Fig. 1B). The key is that as long as the connections inter- or intra-sources can be clearly represented as a general graph as the one shown in Figure 1, and then our learning framework can adopt any general graph structure to learn multiple tasks simultaneously. Following the protocol in the literature (Cai *et al.*, 2010), we sorted the viruses and serum samples by time and then, after evaluating the results of temporal task generation obtained by using window sizes of 4, 6, 8, 12, 14 and 16 years, we chose 12 years as the window size to generate temporal tasks. This window size is the same as that used in other studies, suggesting that a window size of 12 years achieved the best performance in minimizing the effects of low-reactor viruses (Cai *et al.*, 2010; Sun *et al.*, 2013). A GG-MTSL method was then developed to identify key features associated with viral antigenicity, and a quantitative function was developed to measure antigenic distances between influenza A viruses on the basis of their HA protein sequences. As shown in

Figure 2, multi-task learning consisted of three integrated steps: multi-task matrix completion; dynamic multi-source multi-task feature learning; and proposal of an ensemble antigenicity prediction model.

2.1.2 Multi-task matrix completion

Serologic data can typically be classified into three types of information: high-reactor data, low-reactor data or data with missing values. The assessment of these three types of data can be naturally formulated as a low-rank matrix completion problem (Cai *et al.*, 2010). In this study, we proposed a multi-task matrix completion method by separating matrix completion into multiple tasks. To minimize the effects of the protocols on HI data for data integration, we ensured that the HI data in each individual task were generated by the same protocol. One major challenge in the multi-task matrix completion method is that the optimal rank for each individual task may not be the same; it is not practical to optimize a universal rank for all tasks. To overcome this challenge, we adopted the nuclear norm-based regularization technique to optimize ranks for each individual tasks by penalizing the small eigenvalues in the matrix to be zeros (Han and Zhang, 2016; Jaggi *et al.*, 2010). By solving the nuclear norm regularized problem, the optimal rank for each individual matrix completion task can be automatically identified.

Formally, given a $m \times n$ sub-matrix A and the set of regular entries and low reactors in A (denoted as a set Ω), the considered matrix completion problem is to infer the missing values conditioning on the regular entries and low reactors while completing these low reactors with a more confident value, by solving the optimization problem:

$$\min_{H} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}(H_{i,j}^{\Omega} - A_{i,j}^{\Omega})^2 \mathbb{I}(H_{i,j}^{\Omega} \geq \theta) + \lambda||H||_*, \tag{1}$$

where the matrix H is the estimated completed matrix of A; $H_{i,j}$ denotes the (i, j)-th element of H; $H_{i,j}^{\Omega}$ denotes the projection of $H_{i,j}$ on the set Ω (i.e. $(i, j) \in \Omega$); $\mathbb{I}$ is the indicator function; θ is a predefined threshold for identifying the low-reactor value, where we set $\theta = \log_2 20$ and 20 is the signal of the low-reactor value in the HI titers; $||H||_* = \sum_{i=1}^{\min(m,n)} \sigma_i$ is the nuclear norm, which is the sum of all the singular values σ_is of H; and λ is a regularization parameter to trade off between data fitting and the regularization of the matrix rank. In formulation (1), the first term is the least square loss defined on the regular values that are larger than θ (by noting the indicator function), because these entries have true values in A; the second term is the nuclear norm of H that forcing the small eigenvalues of the estimated matrix H to be zeros and thus H will have low rank with the rank detected automatically via this penalization. The algorithm for solving the nuclear norm-regularized matrix completion problem in (1) is straightforward by following the existing approaches (Jaggi *et al.*, 2010; Yuan *et al.*, 2013). The final HI matrix is calculated by averaging the overlapped entries from multiple sub-matrices from each learning task.

2.1.3 GG-MTSL

In this study, we assume the variations in serologic data with similar temporal information would be determined with similar variations (i.e. genetic features, such as residues) in genetic data, regardless of the sources of the serologic data. Thus, we can logically represent the relationships among individual tasks by using graphs based on temporal orders (Fig. 1). Next, we explain how to apply the graph structure to establish the multi-task learning framework.

Formally, let T be the number of tasks and d be the number of residues (the feature dimensionality). d can either be the sum of the number of residues and the number of co-mutations if we consider the synergetic effects among multiple residues. The input data matrix $X_i \in \mathbb{R}^{N_i \times d}$ for the i-th task contains the pairwise genetic distances for all the viruses in the corresponding window, where N_i is the number of pairs. The response $y_i \in \mathbb{R}^{N_i \times 1}$ indicates the pairwise antigenic distance calculated from the HI matrix. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote the graph, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ indicates the edge set. If we encode each node as a task, then an edge in the graph implies that the connected tasks are close to each other, and the weight on the edge indicates the strength of their similarity. Now, the graph-based multi-task model aims to make the tasks connected by edges share similar parameters and it can be formulated as an optimization problem:

$$\begin{aligned}\min_{W \in \mathbb{R}^{d \times T}} &\sum_{i=1}^{T} \frac{1}{TN_i}||y_i - X_i w_i||_2^2 + \\ &\phi\left[\alpha \sum_{(i,j)\in\mathcal{E}} \omega_{i,j}||w_i - w_j||_2 + (1-\alpha)||W||_1\right],\end{aligned} \tag{2}$$

where (i, j) denotes an edge between the i-th task and j-th task; $w_i \in \mathbb{R}^{d \times 1}$ is the model parameter of the i-th task; $||\cdot||_2$ and $||\cdot||_1$ indicate the ℓ_2 and ℓ_1 norms of vector and matrix, respectively; ϕ and α ($0 \leq \alpha \leq 1$) are regularization parameters that control the overall sparseness and the trade-off between the task similarity and sparsity, respectively. In problem (2), the first term is the averaged square loss defined on the linear function mapping the genetic variations to the antigenic variations; the second term penalizes the ℓ_2 norm of the difference between the parameters of any pair of connected tasks, and the effect of this ℓ_2 norm is to make the parameters from two tasks to be similar and hence share common patterns in the affine relationship; the ℓ_1 term is employed to make the solution sparse and force the solution to select the important residues for each task, and this term is regardless of the graph structure.

Solving problem (2) is not trivial because the second term in (2) is non-smooth and general sub-gradient based optimization algorithms are inefficient. We proposed to employ the smoothing proximal gradient (SPG) method (Chen *et al.*, 2012) to solve it. The problem considered by the SPG method takes the form $\min_Z f(W) + r(Z)$, where $f(\cdot)$ is convex and Lipschitz continuous and $r(\cdot)$ is convex but non-smooth. In order to employ the SPG method, we used $f(\cdot)$ to represent the first term and $r(\cdot)$ to represent the second term in (2). We could then rewrite $r(\cdot)$ as

Algorithm 1 SPG algorithm for solving the GG-MTSL models.

Require: X, Y, μ, ω, λ and $\widehat{W}^{(0)}$.
Ensure: W.
Initialize $t = 0$ and $\tau_0 = 1$;
repeat
Compute $\nabla_W \tilde{f}(\widehat{W}^{(t)})$ as in (6);
Solve the proximal step:

$$\begin{aligned}W^{(t+1)} = \arg\min_{W} &\tilde{f}(\widehat{W}^{(t)}) + \langle W - \widehat{W}^{(t)}, \nabla_W \tilde{f}(\widehat{W}^{(t)})\rangle \\ &+ \frac{L}{2}||W - \widehat{W}^{(t)}||_F^2 + \lambda||W||_1.\end{aligned} \tag{3}$$

$\tau_{t+1} = \frac{2}{t+3}$;
$\widehat{W}^{(t+1)} = W^{(t+1)} + \frac{1-\tau_t}{\tau_t}\tau_{t+1}(W^{(t+1)} - \widehat{W}^{(t)})$;
$t = t + 1$;
until some convergence criterion is satisfied.

$$r(W) = ||CW^\top||_{1,2} + \lambda||W||_1,$$

where $\lambda = \phi(1-\alpha)$ and $C \in \mathbb{R}^{E\times m}$ ($E = |\mathcal{E}|$ is the number of edges) is a sparse matrix with each row containing only two non-zero entries, 1 and –1, in two corresponding positions, denoting an edge in the graph $\mathcal{G}$. For example, when the graph is a chain, the matrix C is

$$C = \phi\alpha\begin{bmatrix} \omega_{i,j} & -\omega_{i,j} & 0 & \cdots & \cdots \\ 0 & \omega_{i,j} & -\omega_{i,j} & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \omega_{i,j} & -\omega_{i,j} \end{bmatrix}.$$

Based on the definition of the dual norm, $r(W)$ can be reformulated as

$$r(W) = \max_{A\in Q}\langle CW^\top, A\rangle + \lambda||W||_1, \tag{4}$$

where α_i is a vector of auxiliary variables corresponding to the i-th row of $CW^\top$, $A = (\alpha_1, \ldots, \alpha_E)^\top$ is the auxiliary matrix variable, and $Q = \{A|||\alpha_i||_2 \le 1, \forall i\}$ is the domain of A. Then the smooth approximation of the first term in (4) is given by

$$g_\mu(W) = \max_{A\in Q}\langle CW^\top, A\rangle - \mu d(A), \tag{5}$$

where $\langle CW^\top, A\rangle$ is the inner product of the two matrices, $d(A) = \frac{1}{2}||A||_F^2$, and $||\cdot||_F$ is the matrix Frobenius norm. (5) is convex and smooth with gradient $\nabla g_\mu(W) = A^{*\top}C$, where A^* is the optimal solution to (5) (Han and Zhang, 2015). The computation of A^* is depicted as follows (Chen *et al.*, 2012; Han and Zhang, 2015):

Proposition 1. By denoting by $A^* = \left(\alpha_1, \ldots, \alpha_E^*\right)^\top$ the optimal solution to (5), for any i, we have

$$\alpha_i^* = S\left(\frac{[CW^\top]^i}{\mu}\right),$$

where $[CW^\top]^i$ denotes the i-th row of the matrix $CW^\top$, and S(x) is the projection operator to project vector x on the ℓ_2 ball as

$$S = \begin{cases} \dfrac{x}{||x||_2}, & ||x||_2 > 1, \\ x, & ||x||_2 \le 1. \end{cases}$$

Then, instead of directly solving (2), we solve its approximation as

$$\min_W \tilde{f}(W) + \lambda||W||_1 = f(W) + g_\mu(W) + \lambda||W||_1.$$

The gradient of $\tilde{f}(W)$ with respect to W can be computed as

$$\nabla_W\tilde{f}(W) = \nabla_W f(W) + A^{*\top}C. \tag{6}$$

By using the square loss, the i-th column of $\nabla_W f(W)$ can be easily obtained as

$$\frac{2}{TN_i}X_i^\top(X_i w_i - y_i).$$

Moreover, it is easy to prove that $\tilde{f}(W)$ is L-Lipschitz continuous where L can be determined by numerical approaches (Chen *et al.*, 2012). The SPG algorithm is depicted in Algorithm 1, where (3) has a closed-form solution as

$$W^{(t+1)} = H_\lambda\left(W^{(t)} - \frac{1}{L}\nabla_W\tilde{f}\left(\widehat{W}^{(t)}\right)\right),$$

where $H_\lambda(x) = \text{sign}(x)\max(x-\lambda, 0)$ is the soft-thresholding operator used in solving the Lasso problem (Beck and Teboulle, 2009).

2.1.4 Ensemble prediction model proposal

After solving (2), we can obtain the coefficient vector w_i for each task i, indicating the importance of each residue in task i. Now, given the sequences of a pair of viruses, i and j, we need a scoring function to predict the antigenic distance between them. Suppose virus i is from year a_i, then, we define our prediction model as

$$\widehat{y} = x\left(\mu\bar{w}^{global} + \frac{1-\mu}{2}(\bar{w}_i^{local} + \bar{w}_j^{local})\right) \tag{7}$$

where x is the genetic distance vector based on the sequences; $\widehat{y}$ is the predicted antigenic distance between the two viruses; $\bar{w}^{global} = \frac{\sum_{t=1}^T w_t}{T}$ and $\bar{w}_i^{local} = \frac{\sum_{t\in\mathcal{A}(a_i)} w_t}{|\mathcal{A}(a_i)|}$; $\mathcal{A}(a_i)$ denotes the set of tasks that the year a_i is covered by any task in $\mathcal{A}(a_i)$; $|\mathcal{A}(a_i)|$ denotes the cardinality of $\mathcal{A}(a_i)$; and μ is a parameter trading off between the global coefficient and the local coefficient. Note that the proposed prediction model in (7) is an ensemble of two parts, the global coefficient and the local coefficient, under a trade-off parameter $0 \le \mu \le 1$. Because our GG-MTSL model in (7) captures the distinct antigenically associated residues with respect to each task, the local coefficient, $\bar{w}_i^{local}$, reveals the important residues in a certain local time period, while the global coefficient, $\bar{w}_i^{global}$, captures the information of the important residues in the entire H3N2 influenza virus history.

2.1.5 Parameter tuning and performance evaluation

In the problem of (1), a regularization parameter λ needs to be tuned to obtain the best performance of the matrix completion. We chose λ from a candidate set $[0.1, 0.2, \ldots, 1.0]$, which was found to be a reasonable range to effectively achieve low rank estimations. The performance of different choices of λ was evaluated under 10-fold cross-validation, in which, during each fold, we randomly chose 90% of the known values (i.e. the high reactors) for training and use the remaining 10% of values as the testing set. We used the relative mean square error (ReMSE) for performance assessment

$$ReMSE = \frac{\sum_{(i,j)\in S}\left(H_{i,j} - A_{i,j}\right)^2}{\sum_{(i,j)\in S}A_{i,j}^2},$$

where S denotes the testing set of elements.

In the problem of (2), for simplicity, we set $\omega_{i,j} = 1$ if task i and task j are from the neighbored windows; otherwise $\omega_{i,j} = 0$. Since different HI sources have overlapped time windows, the tasks from different data sources can also be connected in the graph. In addition to $\omega_{i,j}$, there are two regularization parameters, ϕ and α, that need to be tuned to obtain the best performance of the multi-task feature learning. ϕ controls the overall sparseness of the solution, and α trades off between the task similarity and the element-wise sparsity. We proposed to choose ϕ from a candidate set $\left[10^4, 10^3, \ldots, 10^{-2}\right]$ which are common practice used by sparse learning methods (Han and Zhang, 2015, 2016; Han *et al.*, 2016), and α from $[0.1, 0.2, \ldots, 0.9]$ (because α is a value in the interval $[0, 1]$) via 10-fold cross-validation. A larger ϕ will induce a sparser solution, and a

larger α will lead to more similar tasks in the solution. We used the average rooted mean square error (RMSE) for performance assessment, which is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}.$$

2.2 Antigenic cartography and identification of antigenic clusters

The antigenic maps were constructed by using AntigenMap (Barnett *et al.*, 2012; Cai *et al.*, 2010), which is based on the antigenic distance matrix derived from serologic data or the GG-MTSL model described above. To identify the antigenic clusters in antigenic cartography, we used a spectral cluster method (Ng *et al.*, 2002). Our spectral clustering method does not require prior knowledge of the number of clusters because the number can be determined through a nuclear norm regularization algorithm.

2.3 Genetic and antigenic distance generation

The sequence-based antigenic inference in our machine learning system explores genetic features that determine the connections between genetic and antigenic variations. The pairwise genetic distances were measured by using a binary coding function or a pattern-induced multi-sequence alignment (PIMA) scoring function (Smith and Smmith, 1992). When considering the synergetic effects among multiple residues, the co-mutation features were represented by the product among the according single genetic features. After we thoroughly compared the learning performances derived from these two scoring systems, we adopted the PIMA scoring system for conducting all analyses in this study.

The pairwise antigenic distances between viruses and antigens were derived from antigenic cartography (Cai *et al.*, 2010). Each unit in the antigenic map corresponds to a $2\log_2(\text{HI})$; in antigenic maps, 2 units of antigenic distance represent a 4-fold change in HI titers, which, as described elsewhere (Smith *et al.*, 1999), defines whether one virus is an antigenic variant of the other.

2.4 Sequence and serologic data

The HI data used in antigenic cartography and machine learning were collected from the literature (Smith *et al.*, 2004; Sun *et al.*, 2013) and from the annual reports for vaccine strain selection by the World Health Organization influenza collaborative centers, including data for 1528 viruses and 303 serum samples. HI data were obtained by using assays based on turkey erythrocytes (samples from 1968 to 2009), guinea pig erythrocytes (samples from 2006 to 2013) or guinea pig erythrocytes with neuraminidase inhibitor-pretreated viruses (samples from 2012 to 2016) (Supplementary Fig. S1). Because of the multiple variations in the way these multi-sourced serologic data were collected, they presented a challenge to data integration (Yuan *et al.*, 2013) and, thus, provided a rationale for applying multi-task learning methods.

The full length of the HA protein sequences for 39 370 human influenza A(H3N2) viruses collected during 1968–2016 were obtained from public databases (Supplementary Fig. S1). The sequence data are downloaded from public databases, including Influenza Virus Resource (Bao *et al.*, 2008), Influenza Research Database (Squires *et al.*, 2012) and GISAID (a global initiative on sharing all influenza data) (Shu and McCauley, 2017). Sequence and serologic data are available for download through http://sysbio.cvm.msstate.edu/files/GG-MTSL/.

2.5 Reverse genetics and serologic assays

The full-length cDNA for HA and neuraminidase genes of influenza A/Texas/50/2012(H3N2) virus were amplified by using SuperScript One-Step RT-PCR (Invitrogen, Grand Island, NY), and the 6: 2 recombinant viruses with six internal genes of influenza A/PR/8/1934(H1N1) virus were generated by using reverse genetics. Site mutagenesis was performed using a QuikChange II Site-Directed Mutagenesis Kit (Stratagene, La Jolla, CA). Serologic assays were performed using 0.5% turkey erythrocytes.

3 Results

3.1 Graph-guided multi-task sparse learning model can predict the antigenic variant using sequence data

The aim of this study was to develop a genomic sequence-based antigenicity inference method and then to understand antigenic evolution of subtype H3N2 influenza A viruses by using large-scale antigenic profiles derived from this method. To achieve these goals, we proposed a novel GG-MTSL method and multi-source serologic data to identify key features associated with viral antigenicity. A quantitative function was then developed to measure antigenic distances between influenza A viruses on the basis of their HA protein sequences. We formulated the problem (i.e. dealing with different protocols and viruses spanning a long time period) into a multi-task problem by separating the data into multiple temporal tasks (Fig. 1). As shown in Figure 2, multi-task learning consisted of three integrated steps: multi-task matrix completion; dynamic multi-source, multi-task feature learning; and proposal of an ensemble antigenicity prediction model.

During multi-task matrix completion, we optimized $\lambda = 0.3$ by selecting the best average performance for each individual task (Supplementary Fig. S2A). To evaluate the overall performance of multi-task matrix completion, we used cross validation by randomly blinding 10% of the high reactors in the matrices for testing because there is no ground truth for the missing values and low reactors. Results showed that multi-task matrix achieved the best ReMSE (0.0413), which indicates only a 4.14% error rate of the true values in the original HI matrices. The training time of completing the matrices associated with all tasks was 1.2 min.

During multi-task feature learning, we optimized various parameters in the model by selecting the average performance for each individual task (Supplementary Fig. S2B–E). Results showed that $\phi = 10^2$ serves as a cutoff for the best trading-off between the number of selected residues and model performance. For α, we observed that larger α always achieved lower average RMSE, implying that the multi-task sharing is important for boosting model performance. Hence, we set $\alpha = 0.9$. With these parameter settings, we finally achieved an average training RMSE of 0.78 (units) with around 24 selected residues per task. To evaluate the overall prediction performance of the GG-MTSL system procedure and benchmark single-task learning method (Lasso model), we numerically validated our method by using historical HI data for training to predict future virus antigenicity given their sequences. Following a published protocol (Sun *et al.*, 2013), we used the HI data from $[1996, k]$ for training and predicted the antigenic distance between any pair of viruses in the consequent years $[k, k+1]$, where $k \in [2009, 2016]$. We reported the performance in terms of antigenic distance prediction errors on the antigenic drift identification accuracy. Here antigenic distances >4-fold (2 units of antigenic distance) were treated as antigenic drift, and we used this value as the threshold to partition each pair of antigens into either non-variant or variant.

Table 1. Residue sites identified to be associated with influenza A(H3N2) virus antigenicity

Site	ABS†	$\bar{w}$‡	Site	ABS	$\bar{w}$	Site	ABS	$\bar{w}$	Site	ABS	$\bar{w}$	Co-mutation	ABS	$\bar{w}$
25	–	0.0476	131	A	0.0126	174	D	0.0104	219	D	0.0075	⟨193, 196⟩	⟨B, B⟩	0.0042
31	–	0.0366	133	A	0.0206	186	B	0.0325	223	—	0.0494	⟨142, 196⟩	⟨A, B⟩	0.0025
45	C	0.0369	135	A	0.0209	188	B	0.023	225	—	0.0151	⟨50, 196⟩	⟨C, B⟩	0.0025
50	C	0.0278	137	A	0.0163	189	B	0.0511	226	D	0.0468	⟨196, 225⟩	⟨B,-⟩	0.0020
53	C	0.0878	138	A	0.0791	190	B	0.0097	230	D	0.0176	⟨193, 225⟩	⟨B,–⟩	0.0018
57	E	0.0529	140	A	0.0457	192	B	0.0262	242	D	0.0329	⟨157, 189⟩	⟨–, B⟩	0.0017
62	E	0.0431	142	A	0.0306	193	B	0.073	260	E	0.0332	⟨140, 196⟩	⟨A, B⟩	0.0016
67	E	0.0354	144	A	0.071	196	B	0.1357	262	E	0.0214	⟨145, 173⟩	⟨A, D⟩	0.0015
75	E	0.0057	145	A	0.0625	198	B	0.0294	275	C	0.0083	⟨188, 196⟩	⟨B, B⟩	0.0015
78	E	0.0123	155	B	0.0092	199	—	0.0576	276	C	0.0252	⟨189, 196⟩	⟨B, B⟩	0.0015
82	E	0.0159	156	B	0.0741	201	D	0.006	278	C	0.0635	⟨158, 189⟩	⟨B, B⟩	0.0009
83	E	0.0347	158	B	0.1373	202	—	0.026	280	C	0.0152	⟨145, 225⟩	⟨A,–⟩	0.0007
88	E	0.053	159	B	0.0335	207	D	0.0193	299	C	0.0098	⟨144, 225⟩	⟨A,–⟩	0.0006
112	—	0.0169	160	B	0.0211	212	D	0.0063	311	C	0.0085	⟨145, 159⟩	⟨A, B⟩	0.0006
121	D	0.018	163	B	0.0301	213	D	0.0148	312	C	0.0122	—	—	—
122	A	0.0275	172	D	0.0045	214	D	0.0382	—	—	—	—	—	—
124	A	0.032	173	D	0.102	217	D	0.0105	—	—	—	—	—	—

Note: †ABS, antibody binding site; '–', residue that is not in ABS; ‡66 selected residues with global coefficient, $\bar{w}^{global}$, are included in the table, and the local coefficients, $\bar{w}^{local}$, for different tasks are provided in Supplementary Table S2.

Then, we could define classification tasks to measure the prediction accuracy.

We compared the GG-MTSL method with two other multi-task learning methods, the $\ell_{1,2}$ norm regularized MTL and $\ell_{1,\infty}$ norm regularized MTL (Liu *et al.*, 2009), and two single task learning methods, the Lasso and Ridge regressions. Results showed that the average RMSE of the GG-MTSL system was 0.9154 (units) and its average accuracy for identifying antigenic variants was 85.55% for $k \in [2009, 2016]$. Such results outperformed all the other four methods which we compared with (Supplementary Table S1). Moreover, the training time of the GG-MTSL on the entire feature learning task was 2.5 min, which were much faster compared with the other multi-task learning methods and slightly slower compared with the single-task methods, indicating that our optimization algorithm for solving the GG-MTSL model is efficient (Supplementary Table S1). These results demonstrated the effectiveness of the GG-MTSL model for inferring antigenicity.

For the ensemble prediction model, we need to optimize μ, which leverages the ratio of the component of local coefficients (Supplementary Table S2) and global coefficients (Table 1). However, optimizing μ over the available HI data will lead to $\mu \to 0$ because we barely have the ground truth value of the antigenic distance between two viruses lying outside a window (due to the band-matrix shaped structure that off-diagonal values are generally missing or due to low reactors). That is, the available testing data generally lie within the same task and will tend to emphasize the local weights, making them dominant. However, by evaluating μ over a candidate set $[0.1, 0.2, \ldots, 0.9]$, we clearly identified the one re-emerged event (i.e. some H3N2 variant (H3N2v)-like viruses were predicted to be antigenically similar to the A/Beijing/32/92(H3N2) (BE92) cluster, when $\mu = 0.2$). Hence, we set $\mu = 0.2$ in our study.

In summary, study results suggested that GG-MTSL could not only predict the antigenic variants, but could also achieve much better prediction performance than a single-task learning system by overcoming difficulties associated with integrating serologic data derived by using different protocols and obtained from multiple time periods and sources.

3.2 Mutations on HA protein drive antigenic drifts for H3N2 viruses in human

When we used optimized settings, each of 50 learning tasks in the model selected an average of 24 residues to be associated with antigenicity of H3N2 viruses, and a total of 66 unique residues were obtained from all 50 learning tasks (Table 1). These residues were mapped onto a 3D structure of the HA protein (Supplementary Fig. S3). Among these 66 residues, 59 were located in reported antibody binding sites A-E. To test the synergetic effects of multiple residues in the learning, by following (Yang *et al.*, 2014), we also incorporated all the pairwise co-mutations among the resides locating on the surface of the protein structure into the GG-MTSL procedure. After training, the GG-MTSL system identified 186 co-mutation pairs and the top-10 pairwise co-mutations were ⟨193, 196⟩, ⟨142, 196⟩, ⟨50, 196⟩, ⟨196, 225⟩, ⟨193, 225⟩, ⟨157, 189⟩, ⟨140, 196⟩, ⟨145, 173⟩, ⟨188, 196⟩ and ⟨189, 196⟩. Interestingly, four co-mutation pairs with non-zero weights: ⟨158, 189⟩, ⟨145, 225⟩, ⟨144, 225⟩ and ⟨145, 159⟩ were shown to have caused antigenic drifts of H3N2 viruses (Tables 1 and 2). Overall, the multi-task learning identified that mutations N145S-N225D-A138S-F159S and N145S-N225D-N144S-F159Y-Q311H drove the emergence of influenza viruses A/Switzerland/9715293/2013 (SWZ13) and A/Hong Kong/5738/2014 (HK14) from influenza virus A/Texas/50/2012 (TX12) (Table 2). These mutations are located in antibody binding sites A or B. It was probable that viruses with mutation N145S-N225D served as intermediate precursor viruses for SWZ13 and HK14, a suggestion supported by phylogenic analyses and antigenic cartography (Fig. 3A and B).

To confirm this hypothesis, we used the HA and neuraminidase genes of TX12 as template to generate six mutant viruses (bold letters indicate mutations against TX12-like viruses): 145N-225N-138A-144N-159F-311Q (TX12-like), **145S-225D**-138A-144N-159F-311Q (intermediate-like), **145S-225D-138S**-144N-**159S**-311Q (SWZ13-like), **145S-225D**-138A-**144S-159Y-311H** (HK14-like), 145N-225N-**138S**-144N-**159S**-311Q (TX12-like) and 145N-225N-138A-144N-**159Y-311H** (TX12-like). Serologic testing showed that these reassortants antigenically matched our predicted results (Table 3 and Fig. 3D). Specifically, results showed that mutant

145S-225D-138A-144N-159F-311Q is located among the center position of TX12, SWZ13 and HK14 viruses; that **145S-225D-138S**-144N-**159S**-311Q is close to SWZ13 virus; and that **145S-225D**-138A-**144S-159Y-311H** is close to HK14-like virus. Such results indicated that mutations N145S-N225D-A138S-F159S caused antigenic drift from TX12 to SWZ13 and that mutations N145S-N225D-N144S-F159Y-Q311H caused antigenic drift from TX12 to HK14 (Table 2). Furthermore, the antigenic variants of SWZ13 and HK14 were derived from the same intermediate variant bearing residues **145S-225D**-138A-144N-159F-311Q in their HA protein sequences.

Table 2. Antigenic drift events for seasonal influenza A(H3N2) viruses, in order of occurrence (2007–2016), and the residues determining the drift events

Antigenic drift event†	Predominant mutations
BR07 → PE09	K158N-N189K
PE09 → TX12	N278K-S45N
TX12 → SWZ13	N145S-N225D-A138S-F159S
TX12 → HK14	N145S-N225D-N144S-F159Y-Q311H

Note: †BR07, A/Brisbane/59/2007; HK14, A/Hong Kong/4801/2014; PE09, A/Perth/16/2009; SWZ13, A/Switzerland/9715293/2013; TX12, A/Texas/50/2012.

In addition, we used our scoring function from the GG-MTSL ensemble model to predict the pairwise antigenic distance of the mutant viruses. The predicted antigenic distances had a correlation coefficient of 0.75 compared with the HI assay-based antigenic distances, which suggested a high correspondence between real antigenic distances (HI-based) and predicted antigenic distances (sequence-based). Machine learning results suggested that 1–5 mutations led to the antigenic changes in the four antigenic drift events since 2007.

3.3 Large-scale sequence-based prediction infers antigenic profile of H3N2 seasonal influenza viruses

The quantitative function using these features described in Table 1 was developed and then applied to quantify antigenic distances among 39 370 H3N2 viruses recovered from influenza virus-infected humans during 1968–2016 (Fig. 4). Antigenic cartography was constructed, and by using a spectral clustering algorithm (which does not require a predetermined cluster number) 16 antigenic clusters (HK68, EN72, VI75, TX77, SI87, BE89, BE92, WU95, SY97, FU02, CA04, BR07, PE09, TX12, SWZ13 and HK14) were identified with an average Silhouette index of 0.7486; the Silhouette index is a value ranging from −1 to 1, with higher values indicating better clustering performance. A total of 15 antigenic drift events were identified as leading to 16 antigenic variants; the most recent drift from TX12 during the 2013–2014 influenza season led to two

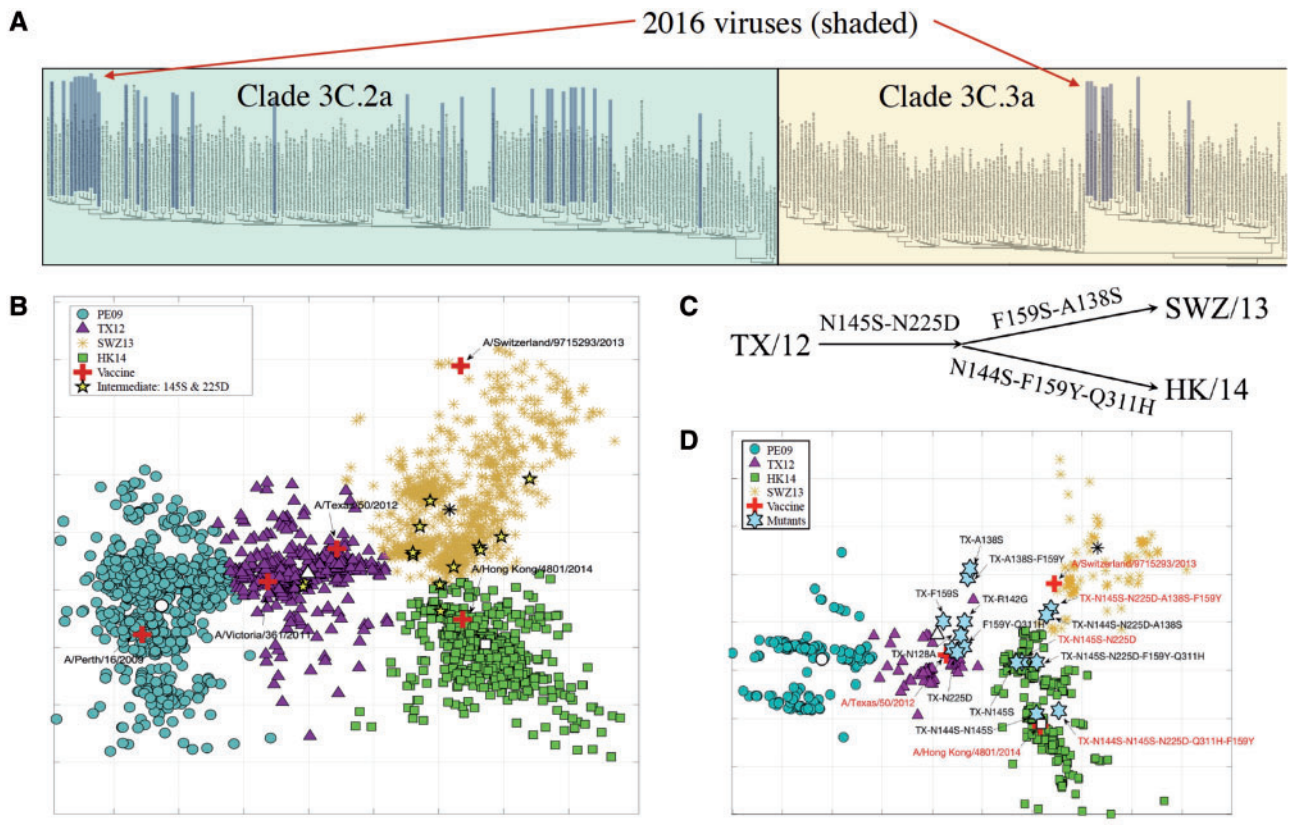

Fig. 3. Co-circulation of two influenza A H3N2 virus antigenic variants, SWZ13-like and HK14-like viruses. (**A**) Phylogenic analyses demonstrating genetic diversity of H3N2 viruses during the 2015–2016 influenza season. Shaded samples represent viruses that emerged in 2015–2016, implying that the two clades were still co-circulating as of 2016. (**B**) Antigenic map demonstrating that SWZ13-like and HK14-like viruses are co-circulating along two different directions. Some of the viruses that emerged in 2015–2016 are labeled. Markers with white face and black edge indicate the estimated centers for each antigenic cluster. (**C**) Estimated mutations leading to antigenic drift events TX12 →SWZ13 and TX12 →HK14. An intermediate mutation (i.e. a double-mutation from TX12 vaccine strain) was identified. (**D**) Bench hemagglutination inhibition value-based antigenic cartography. The three key mutants illustrated in panel C, are demonstrated in this bench validation. Markers with white face and black edge indicate the estimated centers for each antigenic cluster

Table 3. Bench serologic results based on ferret serum for validating the predicted antigenically associated residues

Virus†	Ferret Serum											
	Br/07	Perth/09	Vic/11	TX/12	SWZ/13	Utah/13	CR/13	HK/14	Palau/14	FJ/15	Vic/15	Br/15
TX-A138S-F159Y	<10	40	160	160	160	80	<10	40	160	160	<10	<10
TX-N145S-N225D	<10	640	640	640	160	640	160	320	320	640	80	160
TX-F159Y-Q311H	40	160	320	320	80	320	80	640	80	640	160	320
TX-N145S-N225D-A138S	<10	320	640	640	160	320	160	160	160	640	80	160
TX-N145S-N225D-A138S-F159Y	<10	<10	160	160	160	160	<10	80	160	160	<10	80
TX-N145S-N225D-F159Y-Q311H	<10	320	640	640	320	640	80	640	640	1280	160	640
TX-N144S-N145S-N225D-Q311H-F159Y	10	320	640	640	320	640	160	1280	160	2560	320	1280
TX WT	<10	640	10	1280	160	320	320	160	160	640	640	160
TX-N128A	<10	1280	10	1280	320	640	320	640	320	1280	1280	320
TX-A138S	<10	640	10	640	80	320	160	160	160	640	640	80
TX-R142G	<10	1280	10	1280	320	640	320	320	160	640	1280	160
TX-N145S	<10	640	10	1280	320	640	160	320	320	640	640	320
TX-F159S	<10	10	10	320	160	10	10	80	160	320	160	40
TX-N225D	<10	1280	10	2560	320	640	320	320	320	640	1280	320
TX-N144S-N145S	<10	640	10	1280	320	640	320	640	320	1280	1280	320
Br/07	1280	<10	<10	<10	<10	<10	<10	<10	<10	<10	<10	<10
CR/13	<10	160	160	80	20	40	160	320	<10	320	80	80
FJ/15	<10	80	160	160	<10	<10	80	640	<10	640	20	160
HK/14	<10	160	320	160	<10	<10	160	1280	<10	1280	40	320
Palau/14	<10	<10	<10	160	320	320	80	80	1280	320	80	80
Perth/09	<10	640	<10	160	<10	80	80	80	<10	320	80	40
SWZ/13	<10	160	320	320	640	640	40	320	640	1280	<10	160
TX/12	<10	640	<10	1280	160	320	160	160	80	640	640	640
Utah/13	<10	320	<10	640	160	1280	80	320	160	640	1280	160
Vic/11	<10	320	640	640	80	160	80	160	80	320	<10	80
Vic/15	20	160	40	80	<10	<10	160	1280	<10	640	640	320

Note: †Viruses propagated in Madin–Darby canine kidney cells. The results confirmed the antigenic difference among viruses in TX12, SWZ13 and HK14 clusters and the intermediate virus, TX-N145S-N225D (shown in bold).

co-circulating antigenic variants, SWZ13-like and HK14-like viruses (Fig. 3).

Prediction performance of the GG-MTSL system procedure could also be validated by comparing the correlations between antigenic maps generated from sequence data (prediction) and serologic data (real data). Antigenic cartography shown in Figure 4 and Supplementary Figure S4 were generated from HA sequences and serologic data (HI), respectively. In sequence-based prediction cartography, all serologically tested antigenic clusters showing a clear evaluation pattern of those clusters could be observed, and the pattern matched well with the patterns for serologically tested sequences from each major antigenic cluster.

3.4 Large-scale, sequence-based prediction infers re-emerging H3N2v antigenic variant

Based on key mutations identified from GG-MTSL, a sequence-based prediction could not only predict/infer the antigenic distances and the relationships among all historical H3N2 human viruses, but also could identify re-emergence events in history. Specifically, H3N2v-like viruses were predicted to be antigenically similar to the BE92 cluster, and such results had been confirmed in previous studies (Sun *et al.*, 2013). In addition, antigenic cartography identified an H3N2v-like variant that was antigenically similar to viruses in clusters BE92-SY97 (Fig. 4). The H3N2v-like variant was identified in the summer of 2011 at agricultural fairs and caused 2055 infections among humans in the United States during August 2011–April 2012 (Biggerstaff *et al.*, 2013). This H3N2v virus was possibly transmitted from humans to swine in the mid-1990s and then re-emerged in humans in 2011 (Feng *et al.*, 2013).

4 Discussion

This study presents a robust genomic sequence-based method for quantifying antigenic distances. This method enables the rapid characterization of antigenic profiles and identification of antigenic variants for influenza viruses in real time and on a large scale. In addition, since sequences can be generated directly by using clinical samples, this method can help minimize biases due to culture-adapted mutation during virus isolation (Stevens *et al.*, 2010). This method also allows for the inclusion of uncultivable virus samples into the analyses. Furthermore, multi-task learning allows for the independent characterization of serologic datasets from multiple sources, which are usually difficult to integrate due to various factors, such as types and batches of biologic materials (e.g. reference antiserum and erythrocytes) and supplies (e.g. plates) and variations in the protocol implementation by personnel (Yuan *et al.*, 2013). Another advantage to using multi-task learning is that it makes it possible to use all available data and could help avoid local optimization (referred to as 'overfitting') and false positive results.

In the past years, a few attempts at influenza antigenic variant prediction based on HI data were reported. For example, Lee and Chen (2004) developed a simple correlation method between HA titer and the number of mutations between test viral HA and reference viral HA. Liao *et al.* (2008) applied multiple regression and logistic regression between mutations and HI values; Huang *et al.* (2009) developed a decision tree algorithm in drift variant prediction by deriving association rules from HI data based on information theory. In our previous work, we developed a sparse learning method to identify antigenicity-associated residues by using serologic data and formulated this sparse learning problem as an

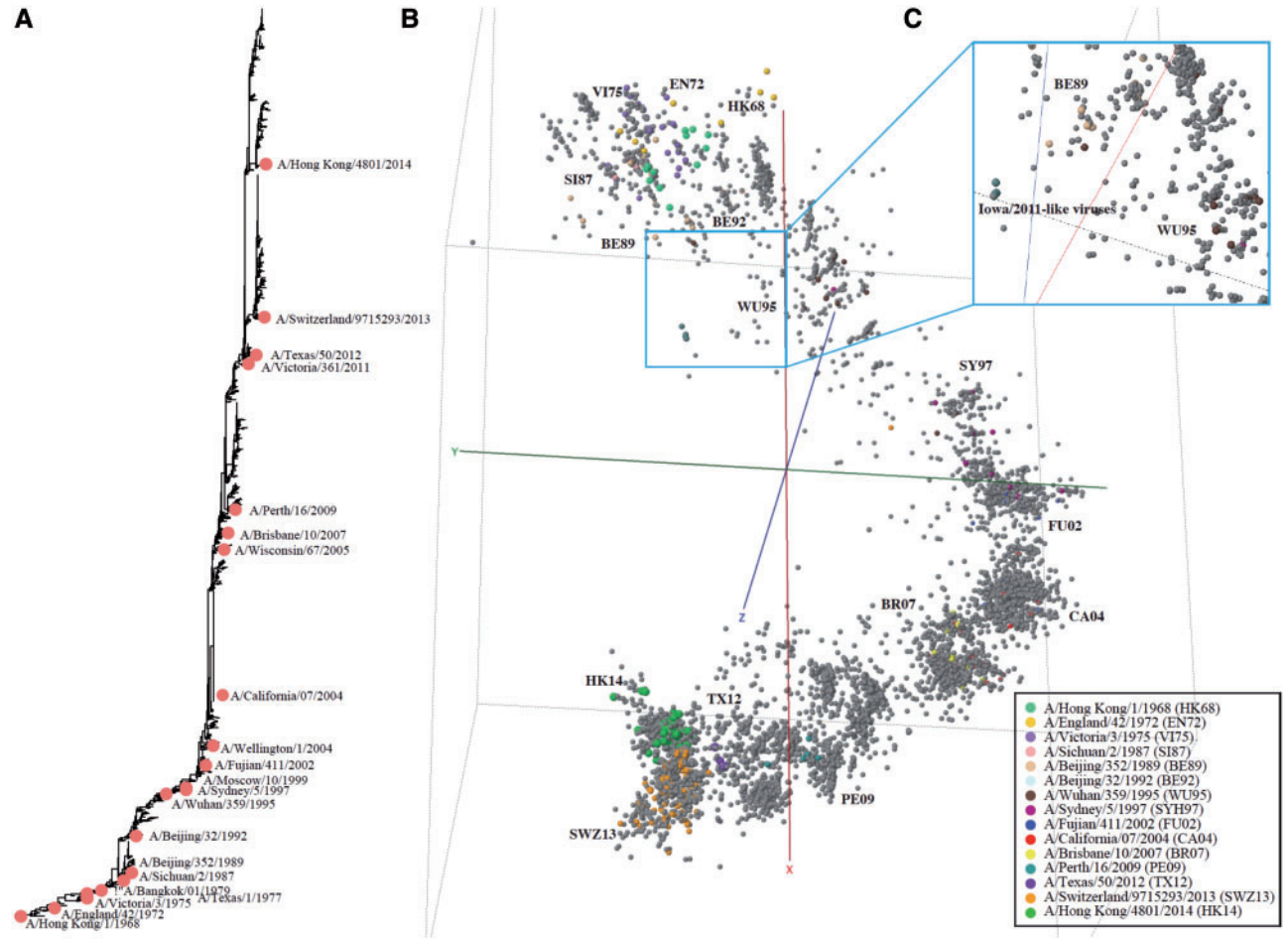

Fig. 4. Genetic and antigenic drift of seasonal influenza A(H3N2) viruses (1968–2016). (**A**) Phylogenetic tree of HA genes for H3N2 viruses showed a continuous natural selection leading to a major truck structure. (**B**) Antigenic map of 39 370 viruses demonstrating zig-zag 'S' shape for antigenic relationship among H3N2 viruses. A total of 16 antigenic clusters were identified by using a spectrometry clustering program. (**C**) Detection of one potential antigenic variant H3N2v-like viruses. H3N2v-like viruses, which emerged in 2011, were antigenically similar to WU95

optimization problem that measures the correlation between the antigenic distance changes in serologic data and antigenic profiling by using a scoring function that characterizes the magnitude of mutations in protein sequences (Cai *et al.*, 2012; Han *et al.*, 2016; Sun *et al.*, 2013; Yang *et al.*, 2014). Instead of predicting antigenic distances among viruses, Neher *et al.* developed a sparse learning method to predict the HI titers for pairs of antigen and sera (Neher *et al.*, 2016). Nevertheless, these methods treat the data analyses or learning as a single task and require data integration; therefore, these methods face the associated challenges previously described. Thus, the GG-MTSL method presented in this study is unique from other available methods, and our findings show that GG-MTSL performed superiorly over a single task-sparse learning method, indicating the effectiveness of the multi-task strategy (Supplementary Table S1).

By using the antigenic characterization results of 39 370 H3N2 viruses recovered from patients during 1968–2016, we showed that the GG-MTSL system proposed in this study identified 16 antigenic clusters of subtype H3N2 influenza virus (Fig. 4 and Supplementary Fig. S4) and showed the dynamics of antigenic evolution of these viruses. The results of our large-scale and sequence-based antigenic cartography suggest that antigenic evolution of H3N2 viruses is much less punctuated than it used to be (Shih *et al.*, 2007), as confirmed by antigenic maps derived from serologic assays (Russell *et al.*, 2008). The continuity of antigenic variations presents great challenges for identifying and defining a virus as an antigenic variant. Although the large set of serologic and genetic data reflects the complete picture of viral evolution, it also complicates identification of antigenic variants during the vaccine strain selection process.

Of interest, this study suggested that two variants (genetic clade C3.2a and clade C3.3a; antigenic clusters HK14 and SWZ13) emerged in 2013 and then co-circulated during the subsequent three influenza seasons, with one variant predominating in some regions and the other predominating in other regions. These two genetic variants are antigenically distinct (Fig. 3), and the extent to which an SWZ13-like vaccine would be effective against a HK14-like viruses, and vice-versa, is not known. It is also not known how long these two antigenic variants will continue to co-circulate among humans. Co-circulation of multiple antigenic variants presents great challenges in vaccine strain selection (Ampofo *et al.*, 2012).

In this study, we detected one re-emerging H3N2 variant; this finding creates another challenge in influenza surveillance by adding another layer of complexity in antigenic variant detection. For example, among the swine population in North America, the current predominant influenza A(H3N2) virus is associated with a spillover of human seasonal H3N2 viruses to pigs in the 1990s (Zhou *et al.*, 1999). In the past 2 decades, genomic analyses suggested at least 22 introductions of influenza A viruses from humans to swine, eight of

which were human seasonal subtype H3N2 viruses (Shu *et al.*, 2012). Uncertainty surrounding the emergence of such variants at the human-swine interface increases the need for surveillance coverage beyond urban areas with dense human populations.

Acknowledgements

This project was supported by the National Institutes of Health [grant number R01AI116744]. We thank Dr. Scott Hensley for providing the HA plasmids of influenza TX12 virus and Dr. Hang Xie for providing the HA plasmids of influenza SWZ13 and HK14 viruses.

Conflict of Interest: none declared.

References

Ampofo,W.K. *et al.* (2012) Improving influenza vaccine virus selectionreport of a WHO informal consultation held at WHO headquarters, Geneva, Switzerland, 14–16 June 2010. *Influenza Other Respir. Viruses*, **6**, 142–152.

Bao,Y. *et al.* (2008) The influenza virus resource at the national center for biotechnology information. *J. Virol.*, **82**, 596–601.

Barnett,J.L. *et al.* (2012) Antigenmap 3d: an online antigenic cartography resource. *Bioinformatics*, **28**, 1292–1293.

Beck,A. and Teboulle,M. (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, **2**, 183–202.

Biggerstaff,M. *et al.* (2013) Estimates of the number of human infections with influenza A (H3N2) variant virus, United States, August 2011–April 2012. *Clin. Infect. Dis.*, **57**, S12–S15.

Cai,Z. *et al.* (2012) Identifying antigenicity-associated sites in highly pathogenic H5N1 influenza virus hemagglutinin by using sparse learning. *J. Mol. Biol.*, **422**, 145–155.

Cai,Z. *et al.* (2010) A computational framework for influenza antigenic cartography. *PLoS Comput. Biol.*, **6**, e1000949.

Chen,X. *et al.* (2012) Smoothing proximal gradient method for general structured sparse regression. *Ann. Appl. Stat.*, **6**, 719–752.

Feng,Z. *et al.* (2013) Antigenic characterization of H3N2 influenza A viruses from Ohio agricultural fairs. *J. Virol.*, **87**, 7655–7667.

Han,L. and Zhang,Y. (2015) Learning multi-level task groups in multi-task learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, pp. 2638–2644.

Han,L. and Zhang,Y. (2016) Multi-stage multi-task learning with reduced rank. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1638–1644.

Han,L. *et al.* (2016). Generalized hierarchical sparse model for arbitrary-order interactive antigenic sites identification in flu virus data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 865–874.

Harper,S.A. *et al.* (1984) Prevention and control of influenza.

Harvey,W.T. *et al.* (2016) Identification of low-and high-impact hemagglutinin amino acid substitutions that drive antigenic drift of influenza A (H1N1) viruses. *PLoS Pathog.*, **12**, e1005526.

Huang,J.-W. *et al.* (2009) Co-evolution positions and rules for antigenic variants of human influenza A/H3N2 viruses. *BMC Bioinformatics*, **10**, S41.

Jaggi,M. *et al.* (2010). A simple algorithm for nuclear norm regularized problems. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 471–478.

Lee,M.-S. and Chen,J.S.-E. (2004) Predicting antigenic variants of influenza A/H3N2 viruses. *Emerg. Infect. Dis.*, **10**, 1385.

Liao,Y.-C. *et al.* (2008) Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. *Bioinformatics*, **24**, 505–512.

Lin,Y.P. *et al.* (2010) Neuraminidase receptor binding variants of human influenza A (H3N2) viruses resulting from substitution of aspartic acid 151 in the catalytic site: a role in virus attachment? *J. Virol.*, **84**, 6769–6781.

Liu,J. *et al.* (2009). Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 339–348.

Mansfield,K. (2007) Viral tropism and the pathogenesis of influenza in the mammalian host. *Am. J. Pathol.*, **171**, 1089–1092.

Neher,R.A. *et al.* (2016) Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proc. Natl. Acad. Sci. USA*, **113**, E1701–E1709.

Ng,A.Y. *et al.* (2002) On spectral clustering: analysis and an algorithm. *Adv. Neural Inf. Process. Syst.*, **2**, 849–856.

Ren,X. *et al.* (2015) Computational identification of antigenicity-associated sites in the hemagglutinin protein of A/H1N1 seasonal influenza virus. *PLoS One*, **10**, e0126742.

Russell,C.A. *et al.* (2008) The global circulation of seasonal influenza A (H3N2) viruses. *Science*, **320**, 340–346.

Shih,A.C.-C. *et al.* (2007) Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proc. Natl. Acad. Sci. USA*, **104**, 6283–6288.

Shu,B. *et al.* (2012) Genetic analysis and antigenic characterization of swine origin influenza viruses isolated from humans in the united states, 1990–2010. *Virology*, **422**, 151–160.

Shu,Y. and McCauley,J. (2017) Gisaid: global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, **22**, pii: 30494.

Smith,D.J. *et al.* (1999) Variable efficacy of repeated annual influenza vaccination. *Proc. Natl. Acad. Sci. USA*, **96**, 14001–14006.

Smith,D.J. *et al.* (2004) Mapping the antigenic and genetic evolution of influenza virus. *Science*, **305**, 371–376.

Smith,R.F. and Smmith,T.F. (1992) Pattern-induced multi-sequence alignment (PUMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Eng. Des. Sel.*, **5**, 35–41.

Squires,R.B. *et al.* (2012) Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza Other Respir. Viruses*, **6**, 404–416.

Stevens,J. *et al.* (2010) Receptor specificity of influenza A H3N2 viruses isolated in mammalian cells and embryonated chicken eggs. *J. Virol.*, **84**, 8287–8299.

Sun,H. *et al.* (2013) Using sequence data to infer the antigenicity of influenza virus. *MBio*, **4**, e00230-13–e00213.

Thompson,M. *et al.* (2010) Estimates of deaths associated with seasonal influenza-United States, 1976–2007. *Morb. Mortal. Wkly Rep.*, **59**, 1057–1062.

Thompson,W.W. *et al.* (2004) Influenza-associated hospitalizations in the United States. *JAMA*, **292**, 1333–1340.

Yang,J. *et al.* (2014) Sequence-based antigenic change prediction by a sparse learning method incorporating co-evolutionary information. *PLoS One*, **9**, e106660.

Yuan,X.-T. *et al.* (2013) A joint matrix completion and filtering model for influenza serological data integration. *PLoS One*, **8**, e69842.

Zhou,N.N. *et al.* (1999) Genetic reassortment of avian, swine, and human influenza A viruses in american pigs. *J. Virol.*, **73**, 8851–8856.