Modeling mutations of influenza virus with IBM Blue Gene

Z. Xia
P. Das
T. Huynh
A. K. Royyuru
R. Zhou

Outbreaks of influenza cause considerable public health concerns and pose a potential threat of a global pandemic. In this paper, we describe some of our recent work on modeling of influenza virus with an IBM Blue Gene® supercomputer. The goal is to predict which mutations [on the viral glycoprotein hemagglutinin (HA)] are likely to occur in the next flu season, which mutations might escape antibody (Ab) neutralization, and which mutations might cause its receptor binding specificity to switch (e.g., from avian to human). We have analyzed more than 4,000 influenza A/H3N2 HA sequences from 1968 to 2010 to model the evolutionary path using a weighted mutual information method, which allows us to build a site transition network to predict antigenic drifts. We then used large-scale free energy perturbation calculations to study the mutation-induced effects on the antigen–Ab and antigen–receptor bindings. For example, we found that a single mutation T131I on H3N2 HA can decrease the HA–Ab binding affinity by 5.2 ± 0.9 kcal/mol, in excellent agreement with recent experimental results. We also found that a double mutation, i.e., V135S and A138S, could potentially switch the H5N1 HA binding specificity from avian to human, thus allowing the virus to gain a foothold in the human population. Detailed analyses also reveal a molecular picture of the influenza virus Ab and receptor binding mechanisms.

Introduction

The wide spread of influenza virus has become one of the most fatal diseases in humans and poultry [1–3]. The subtypes of A/H1N1, A/H3N2, and recent A/H5N1 have caused significant public health concerns due to the emergence of potential pandemic threats [4–10]. The viral surface glycoprotein hemagglutinin (HA) is the primary protein component of vaccines to provide protective immunity against influenza virus infection. However, the high mutation rates of HA ($\sim 5.7 \times 10^{-3}$ substitutions per site per year) make it difficult to effectively predict future mutations and develop appropriate vaccines/antibodies for potential emerging pandemics [11].

A number of studies aimed to understand the antigenic evolution of influenza caused by those mutations in HA [12–19]. Smith et al. [20] have mapped the antigenic drift and the site mutations of H3N2 using the hemagglutination inhibition (HI) assays, which were the first to directly relate the viral genotype and the inferred phenotype [20]. Shih et al. [21] and later Du et al. [22] have found that antigenic drifts might be enhanced not only by the accumulation of single-site mutations but also by the simultaneous multi-site mutations of HA. Meanwhile, some key mutations that can escape antibody (Ab) neutralization have been revealed by binding affinity studies of H3N2 HA/Ab complex [5]. However, the current lack of H5N1 HA/Ab complex structure and limited binding affinity data of H3N2 HA/Ab present barriers to rigorous computer modeling approaches [23, 24]. Glycan array experiments allow qualitative estimation of the antigen–receptor binding affinity and identify mutations that might switch the H5N1 HA receptor binding specificity from avian (α-2,3-linked sialylated glycan receptors) to human (α-2,6-linked sialylated glycan receptors) [4, 6, 9, 25, 26]. However, conflicting results from different groups with

Digital Object Identifier: 10.1147/JRD.2011.2163276

0018-8646/11/$5.00 © 2011 IBM

slightly different glycan arrays indicate the limits of these techniques. Therefore, a better understanding of the genetic evolution paths of influenza virus, as well as accurate molecular modeling of the antigen–Ab and antigen–receptor bindings, is critical for subsequent development of effective vaccines against future strains. Rigorous modeling that yields sufficient accuracy is computationally demanding. We perform molecular simulations of binding affinities with the free-energy perturbation (FEP) method, which, in turn, uses large-scale molecular dynamics (MD) simulations to sample the conformational space.

We use an IBM Blue Gene* supercomputer to perform these computationally demanding tasks. The Blue Gene Watson* supercomputer is a state-of-the-art high-performance computing facility located at the IBM Thomas J. Watson Research Center in Yorktown Heights, New York. It is the 80th fastest supercomputer in the world (see Top500** list: http://top500.org/list/2010/11/100). The IBM Blue Gene Watson supercomputer earned that distinction by demonstrating sustained performance of 91.29 teraflops on the linpack benchmark. The peak performance of IBM Blue Gene Watson is approximately 114 teraflops. Blue Gene Watson consists of 20 racks of hardware conforming to the IBM Blue Gene/L architecture. Each Blue Gene Watson rack consists of 1,024 nodes, and each node contains two 700-MHz IBM POWER* 440 processors and 512 MB of memory. The 20 racks are arranged as five rows of four racks each. The primary mission of Blue Gene Watson is to perform production science computations that could not be successfully undertaken on less powerful computers. Except for periodic maintenance, it runs 24 hours a day, seven days a week in production mode. For the binding affinity calculations, up to four racks of Blue Gene Watson (4,096 nodes, 8,192 processors) have been used to run the FEP calculations in parallel. To our knowledge, these are the largest FEP calculations for influenza viruses.

In this paper, we describe our recent work in order to provide a novel systematic approach to facilitate the development of influenza virus vaccine/Ab [17–19]. We begin with mapping the genetic evolution of influenza H3N2 virus by analyzing all available sequence data. A weighted mutual information (MI)-based machine-learning model is utilized to design a site transition network (STN) for each amino acid site of HA [18]. A novel five-step prediction algorithm based on this STN is used to predict the likely mutations that might occur in the next flu season. In the next step, we perform large-scale FEP calculations to quantitatively investigate the mutation-induced effects [17, 19, 27–31]. The current FEP calculations not only provide molecular mechanisms with detailed physical interactions but also identify potential mutations that might either escape Ab neutralization or switch the receptor binding specificity from avian to human [32, 33]. Identification of such key mutations could assist the development of vaccines in advance of the emergence of new strains. Detailed methods and results are reported and described in the following two sections.

Prediction of the antigenic variation of influenza virus

Network inference algorithm: Weighted MI

We use a weighted MI approach to predict the antigenic evolution of influenza A virus [18, 34–37]. The MI value for a pair of amino acid sites, i.e., x and y, can be defined as

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} w(x,y)p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \qquad (1)$$

where $I(X;Y)$ is the MI value; $w(x,y)$ is the weighting factor (mutated sites have double the weight as nonmutated sites); and $p(x)$, $p(y)$, and $p(x,y)$ in (1) are the probabilities associated with each discrete state, which are the probabilities of each residue to be mutated or not in our study.

We use A/H3N2 as an example to explain our model. The methodology developed here is equally applicable to H1N1 or H5N1 without any modification. All 4,311 HA1 full-length (>312 residue) sequences of A/H3N2 up to the release date of December 31, 2010, are aligned before the MI calculation [sequence data were downloaded from the National Center for Biotechnology Information Influenza Virus Resource] [38]. We then divided the sequences into 43 bins according to the year information (1968–2010). Ten sequences from each bin were randomly selected to generate one sample input sequence in each calculation. A total of 2,000 different samples are randomly selected to gather sufficient statistics. The output MI matrices from these samples were normalized (with mean 0 and standard deviation of 1) to generate the final "MI matrix."

The 2-D MI matrix obtained from the MI calculation includes the co-mutation correlation (MI value) for any pair of sites in HA1. A higher value between a pair of sites means a higher likelihood of co-mutation. An STN for all 312 sites can then be generated by the MI matrix. The nodes in the STN represent individual amino acid sites in HA1, and the edges between the two nodes represent the co-mutation correlation. An example of STN is shown in **Figure 1(a)**, in which 63 positive selection sites were selected and the edge with the normalized MI value lower than 0.5 was removed from the network [21]. The positive selection sites are thought to be responsible for the most antigenic variations. These are mainly distributed in the receptor binding domain and the five known epitopes. The effectiveness of the STN was evaluated by comparing STN with the "antigenic maps" from Smith et al. [20]. The antigenic map utilizes data from serum HI assays to measure

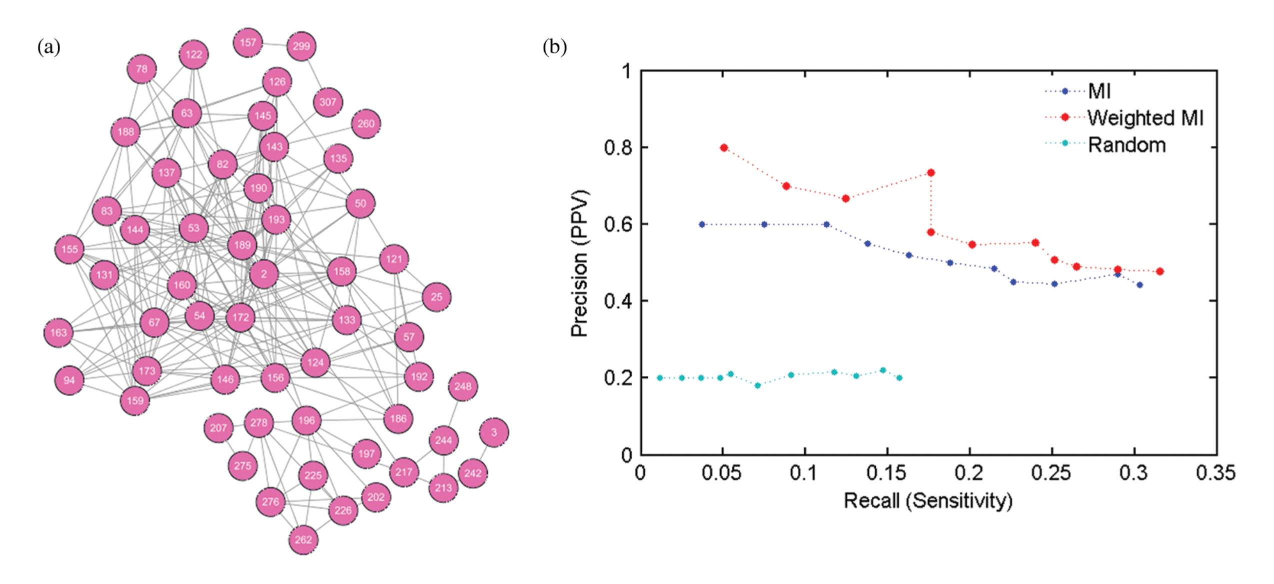


Figure 1

(a) Example of an STN from years 1968–2008. A total of 63 positive selection sites out of 312 were found, which were then chosen to plot the network (other sites were omitted since they are mostly conserved with little interactions with other nodes and will appear as isolated nodes in the network if plotted). Each node represents a site with its residue number marked on top, and each edge represents the interaction between a pair of mutation sites if its normalized MI score is higher than 0.5. (b) Prediction accuracy for several different network inference methods applied to influenza antigenic variation predictions. The weighted MI method, the regular MI method, and a random selection method are shown.

the cross-immunity "distance" between each strain of influenza to every other strain in an "antigenic space." We found that the antigenic drift in A/H3N2 occurs more smoothly at a sequence level, which indicates that the mutation on antigenic sites of HA occurs all of the time and certain positive substitutions might result in a partial structural change in the antigenic regions. It seems that co-mutations and structural changes in a specific group of these sites may confer sufficient advantage to induce an antigenic change. To confirm this hypothesis, we performed further cluster analysis of the MI matrix (see below).

Co-mutation sites responsible for the antigenic drift

We applied a hierarchical clustering analysis to cluster mutation sites of HA1 in the MI matrix. The rationale behind this is that if any two sites i and j each have high MI values with some other common sites, they will all show high correlations and thus appear as a cluster. A total of five clusters were obtained from the cluster analysis, with each cluster representing roughly one or several antigenic transitions (see **Table 1**). For example, the mutations at sites 122, 207, and 188 responsible for the HK68-EN72 antigenic drift appear as a cluster. Such cluster analyses suggest that the sequence data alone, particularly for H3N2 where historical data is abundant, is adequate to uncover most of the antigenic variations of influenza virus. This finding implies that a cluster of co-mutating sites on HA1 might create large-enough structural change on the protein surface at antigenic sites to escape Ab neutralization and induce antigenic drift to a new strain.

STN-guided prediction of antigenic variation

The STN shows that antigenic drifts can be enhanced by cumulative co-occurring multi-site mutations in the epitope regions on the HA protein surface. These co-occurring mutations can be exploited to predict future mutations from the present network. Similarly, mutation hot spots can be identified by calculating the mutation frequency for each site of HA1. In addition, we observe nonrandom probabilities for different amino acid types at each mutation site. These strong preferences are a reflection of evolutionary selection.

A five-step algorithm has thus been designed to predict future antigenic variations (see **Figure 2**). Before we start, we first define the year that we want to predict as the target year, i.e., N, and the years $N - 1$ and $N - 2$ as induction years. The steps in the prediction algorithm are listed as follows.

Step 1: Calculate all sites that are under positive selection in HA1 before year N. Here, the "positive selection site" is defined as a site that has mutated in successive years and then remained fixed in the

Table 1 Amino acid sites responsible for co-occurring mutations. A total of five clusters were obtained from the cluster analysis, with each cluster representing roughly one or several antigenic transitions. The sites shown in red in rows 2 and 3 are the identical sites obtained from both the real historical data and our clustering analysis (and the ones in gray are missed ones).

Antigenic changes	*HK68-EN72*		*EN72-VI75*		*VI75-TX77*		*TX77-BK79*	*BK79-SI87*	*SI87-BE89*	*BE89-BE92*		*BE92-WU95*		*WU95-SY97*	*SY97-FU02*
Observed mutant sites in HA1 during the antigenic transitions	78 122 188 207	242 144 155 275	53 137 213 145	189 217 278	2 137 213 260	50 82 158 193	54 133 143 156 160 172 197 217 53 146 162 244	124 155 189	145	145 157 189	190 276 156	135 145 226 262	172 197 278	276 196 226 62 156 158	25 57 75 83 131 142 144 155 186 202 222 225 227 50 156 159 189
Sites in different groups in cluster analysis	78 122 188 207 242 276		2 53 137 213 260 244 50				54 133 143 156 160 172 197 217 121	124 135 145 157 189 190 196 226 262 276							25 57 75 83 131 142 144 155 186 202 222 225 227 63 82 94 126 192299

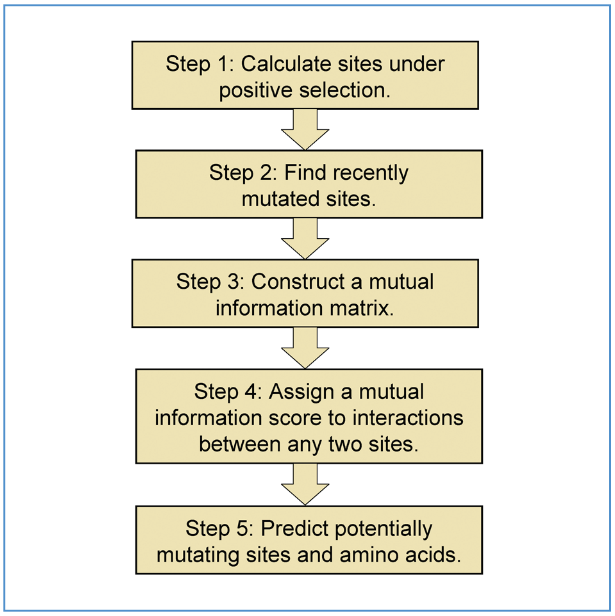

Step 1: Calculate sites under positive selection.
Step 2: Find recently mutated sites.
Step 3: Construct a mutual information matrix.
Step 4: Assign a mutual information score to interactions between any two sites.
Step 5: Predict potentially mutating sites and amino acids.

Figure 2

Five-step prediction algorithm based on STN.

population for at least one year, similar to the definition used by Shih et al. [21].

Step 2: Among the positive selection sites, find the sites that have just mutated in any of the induction years. Such sites are considered as the initial state of the present network.

Step 3: Use all of the available sequences before year N as a data source to construct the aforementioned sample sequence input file and calculate the MI matrix.

Step 4: Since the MI matrix quantifies the interaction between any two sites by a MI score, for each site X in HA1, we sum up the scores between the site X and all the sites found in Step 2 (i.e., newly mutated positive selection sites in induction years). The sites with high MI scores are chosen as predicted sites.

Step 5: Find the most probable amino acid type for each predicted mutation site from Step 4 by searching the historical amino acid type database for each site. Historical data suggests that there is a strong preference for each residue site to have some specific amino acid types (see the section "Results"). Therefore, we use the most probable amino acid type other than the current one as the final mutated type.

The five-step predictive method based on STN was then validated by testing the prediction accuracy of known years. The prediction of mutating sites was tested for every year from 1999 to 2008 using only the sequence data of prior years (thus, a blind test). We found that the accuracy of prediction was reasonably good and fairly stable, approximately 70% for most of the years tested, which means that the network-guided method can be a reliable tool to predict the antigenic variations. **Figure 1(b)** summarizes the statistical results and comparisons among various methods, i.e., the weighted MI method, the normal MI (unweighted) method, and a random selection method. Obviously, the weighted MI method performs best, whereas the random selection method is not predictive, as one would expect.

Following the same procedure for the above validations, we used all the sequence data available up to year 2010 to predict likely mutations in year 2011. Our method predicted six possible mutation sites, i.e., N6I, N121I,

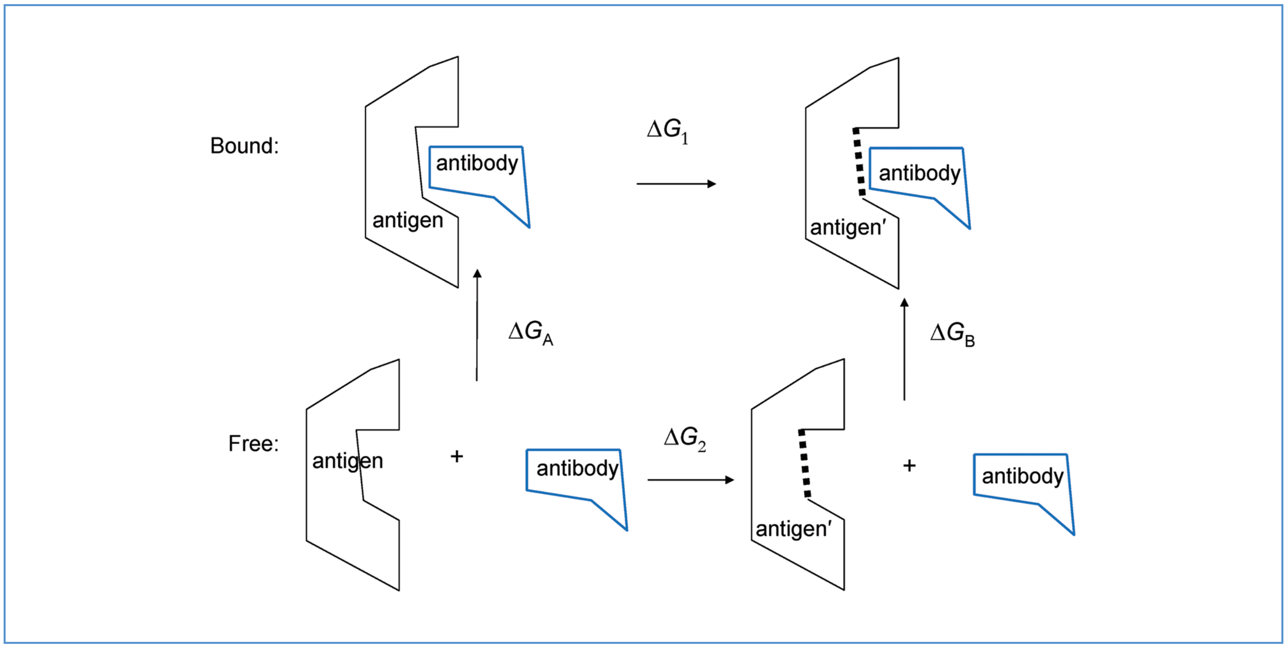


Figure 3

Scheme of the thermodynamic cycle for the calculation of the antigen–Ab binding affinity change due to a mutation on the antigen (represented by viral surface glycoprotein HA; *antigen′* stands for the mutant).

R142G, N144V, I192T, and R261H, that may appear in the next flu season. These predictions will be validated against actual strains in due course, as 2011 strains are sequenced and the sequence information is released. These predictions can also be used as starting points in structural modeling of antigen–Ab and antigen–receptor binding affinities in order to computationally forecast the consequences of these mutations.

Modeling of receptor specificity and escape from Ab neutralization

Large-scale FEP with Blue Gene Watson

The FEP method has been widely used to calculate binding affinities for a variety of biophysical phenomena such as solvation free energy values, enzyme catalysis, redox, pKa, ion conductance, ligand–receptor binding, protein–protein interaction, and protein–nucleic acid binding [27–29, 31–33, 39–47]. A thermodynamic cycle is often employed to estimate the relative free energy change ($\Delta\Delta G$) caused by a mutation ($A \rightarrow B : \Delta G_B - \Delta G_A$), since the absolute binding free energy is usually very difficult to directly calculate in FEP simulations. We follow a similar approach as in previous studies and design a thermodynamic cycle, as shown in **Figure 3**, to calculate the relative binding affinities of the bound (ΔG_1 : antigen + antibody) and free states (ΔG_2: antigen only) (for details, see our previous publications, i.e., [17] and [19]). The total free energy change should be zero in any thermodynamic cycle, i.e.,

$$\Delta G_A + \Delta G_1 - \Delta G_B - \Delta G_2 = 0, \quad (2)$$

which gives the binding affinity change due to the mutation from A to B as

$$\Delta\Delta G = \Delta G_B - \Delta G_A = \Delta G_1 - \Delta G_2, \quad (3)$$

where ΔG_1, ΔG_2, ΔG_A, and ΔG_B are the free energy changes defined above.

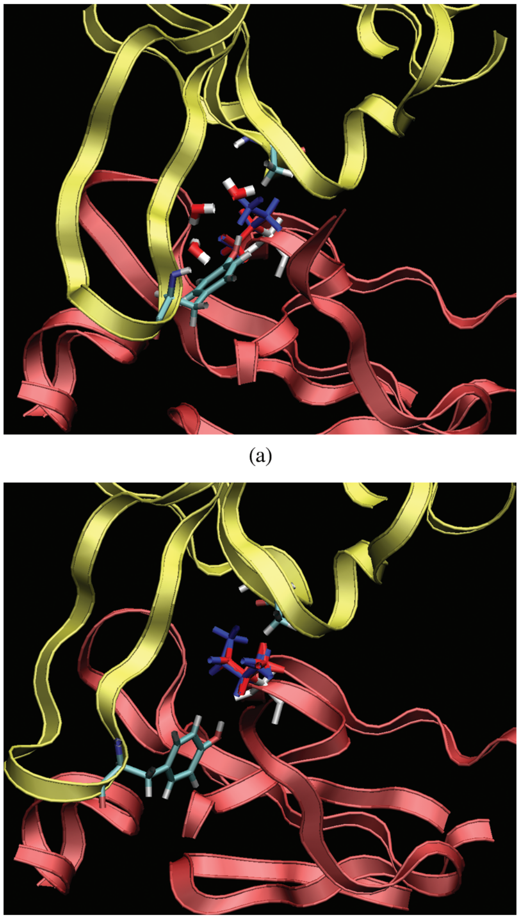
Among several available computational methods developed in the past years, FEP using an all-atom explicit solvent model serves as the most accurate approach for our current needs in estimating the relative antigen–Ab and antigen–receptor binding affinities [45, 47]. Such simulations for realistically sized biological systems often require large computational resources [48–54]. Here, we have utilized the massively parallel MD software developed on IBM Blue Gene to perform these FEP calculations [48–53].

Escape from Ab neutralization

The accuracy of FEP calculations was first validated by introducing an experimentally known mutation T131I [5]. This T131I single mutation in H3N2 HA can increase the Ab–antigen dissociation constant K_d by a factor of approximately 4,000 (equivalent to a binding affinity decrease of -5 kcal/mol), thus causing an escape of the

Ab neutralization [5]. Our FEP calculation using the IBM Blue Gene supercomputer estimated the HA–Ab binding affinity decrease of 5.2 ± 0.9 kcal/mol, which is in excellent agreement with the experimental results. The decomposition of the total binding affinity revealed that the electrostatic interactions dominate the free energy change, with about 70% from electrostatic and 30% from van der Waals interactions. The physicochemical factors behind the FEP results were also investigated in atomic detail. Two or more bridging water molecules were constantly found between the wild-type HA and the Ab near residue T131 in the simulation [see **Figure 4(a)**]. These bridge water molecules form hydrogen bonds with the HA T131 side chain's hydroxyl group and the Ab fragment (Fab) heavy chain residues (hydroxyl group or backbone), such as sites Y107, S31, and A53. These water molecules appear to be lubricants at the Ab–antigen interface; the water molecules were not fixed in space but fairly mobile with hydrogen bonds forming and breaking all the time. During the FEP simulation, the T131 was gradually mutated to I (Ile); the much bulkier hydrophobic side chain of I131 displaces these two or three bridging water molecules [see **Figure 4(b)**]. Therefore, the displacement of bridge water molecules from the binding site has contributed to the loss of the binding affinity in the T131I mutation [17].

In addition to the T131I mutation, another 14 neutral mutations and 4 charged ones were also performed at the same site 131. All of the binding affinity changes with standard deviations are listed in **Table 2**. For the neutral residue mutations, T131W, T131Y, and T131F are found to have even larger binding affinity decreases, with $\Delta\Delta G$ values of 7.46 ± 1.91, 6.01 ± 1.31, and 5.68 ± 1.48 kcal/mol, respectively. These mutations also reveal a significant displacement of bridging water molecules. Other residues such as T131H (3.84 kcal/mol), T131L (3.15 kcal/mol), T131N (2.92 kcal/mol), T131V (2.58 kcal/mol), and T131Q (1.22 kcal/mol) also show a decrease in the binding affinity. Meanwhile, a few others show an increase in the binding energy, such as T131G (−3.72 kcal/mol), T131A (−2.81 kcal/mol), and T131S (−0.48 kcal/mol). The smaller sizes of mutated residues Gly (G) and Ala (A) seem to have accommodated more bridging water molecules (three to four water molecules) in the active site, as well as less "steric repulsion" in the relatively flat binding pocket, resulting in a more favorable binding. Interestingly, the T131A mutation was observed in 1990, and then, a back-mutation A131T was observed in 1994, indicating that it is not a favorable mutation for the virus [21]. Although we are not suggesting that this binding affinity increase is the only reason for this back A131T mutation, our FEP simulation is consistent with the real-world influenza A/H3N2 virus evolution in this instance. All of the mutations to charged residues show a decrease in binding affinity. The negatively charged residues (T131D and T131E) have a significantly larger decrease than the basic residues (T131K and T131R). This is because the nearby acidic residue D98 from the Ab heavy chain contributes favorable electrostatic interaction with K131 and R131.


(a)

(b)

Figure 4

Bridging water between T131 and the Ab: (a) bridging water in native T131–Ab complex; (b) bridging water disappeared at the end of the T131I mutation. The bridge water and hydrogen-bonding residues A53 and Y107 from Ab are shown with sticks. The T131I residue is colored by dual topology (red: T131; blue: I131). (Reproduced from [17], with permission.)

Receptor binding specificity modeling

The high pathogenicity and human mortality rates of H5N1 virus infections have raised serious public health concerns that this virus may seed the next pandemic. The determining factor to the host specificity of influenza viruses is thought

Table 2 FEP simulation results for the H3N2 HA/Ab binding free energy change due to various mutations of T131. (*Data were obtained by incorporating no counter ions when T131 was mutated to charged residues.)

Mutation	*Calculated ΔΔ*G (kcal/mol)
T131I	5.20 ± 0.94
T131G	−3.72 ± 0.69
T131A	−2.81 ± 0.91
T131C	0.117 ± 1.24
T131V	2.58 ± 0.89
T131M	0.57 ± 1.63
T131Q	1.22 ± 1.20
T131S	−0.48 ± 1.57
T131F	5.68 ± 1.48
T131W	7.46 ± 1.91
T131L	3.15 ± 1.19
T131H	3.84 ± 1.17
T131Y	6.01 ± 1.31
T131N	2.92 ± 1.16
T131K	2.38 ± 1.07*
T131R	3.86 ± 1.37*
T131D	12.69 ± 2.18*
T131E	8.20 ± 1.95*

Table 3 Receptor binding free energy changes of avian H5 HA on a number of mutations at V135 and A138. The results for V135S and A138S double mutations are in red.

H5 HA mutation	*Calculated ΔΔ*G (kcal/mol)	
	α-2,3-linked (avian like)	α-2,6-linked (human like)
V135T	−0.83 ± 0.99	0.28 ± 0.92
V135K	2.09 ± 0.89	1.57 ± 0.52
V135N	−0.29 ± 0.88	0.71 ± 0.94
V138I	−1.31 ± 0.51	−0.60 ± 0.92
A138I	2.03 ± 0.8	3.63 ± 0.98
A138V	1.13 ± 0.41	0.17 ± 0.69
A138N	0.39 ± 0.6	3.05 ± 1.35
A138K	−1.68 ± 1.17	−0.43 ± 1.64
A138T	0.52 ± 0.4	0.45 ± 0.59
V135T + A138T	0.48 ± 0.9	−0.3 ± 1.00
V135S	1.12 ± 0.7	−0.60 ± 0.19
A138S	−0.13 ± 0.59	−0.41 ± 0.32
V135S + A138S	0.84 ± 1.02	−2.56 ± 0.73

to be the linkage between sialic acids (SAs) and the penultimate sugar in the host cell receptors [55–58]. A switch in the receptor specificity from α-2,3-linked to α-2,6-linked sialylated glycans is believed to facilitate bird-to-human, as well as human-to-human, transmission of influenza viruses [56, 57, 59–61]. Unfortunately, the HA–SA binding affinity at each HA monomer level is only in millimolars (i.e., a few kilocalories per mole), which poses a significant challenge for computer modeling, since the thermal fluctuations can be easily seen at *kT* level (∼0.6 kcal/mol at body temperature). In this paper, we again use the IBM Blue Gene supercomputer to carry out large-scale FEP calculations to investigate the HA–glycan binding at the atomic level and identify likely mutations in currently circulating H5N1 influenza viruses that may be critical for binding to human receptors. The goal is to characterize the effect of mutations on HA–glycan binding specificity [17, 27–31].

Similar to the FEP simulation for Ab–HA, the simulation protocol was also validated by comparing the simulated binding affinities to experimentally available data [6, 9, 25]. The simulated binding affinities agree fairly well with currently available glycan array data for several H5 HA mutants. Then, a number of single/double amino acid mutations were performed at position 135 and/or 138 to find mutations causing such a switch in the H5 HA receptor specificity from α-2,3-glycan (avian) to α-2,6-glycan (human) [19]. Overall, the binding affinities for those mutations suggest that the majority of them results in either no change or a decrease in binding affinity to α-2,6-glycan over α-2,3-glycan. However, V135S and A138S single mutations result in a small preference for α-2,6-glycan, with a ΔΔ*G* of −0.6 ± 0.19 kcal/mol for V135S and −0.41 ± 0.32 kcal/mol for A138S (see **Table 3**).

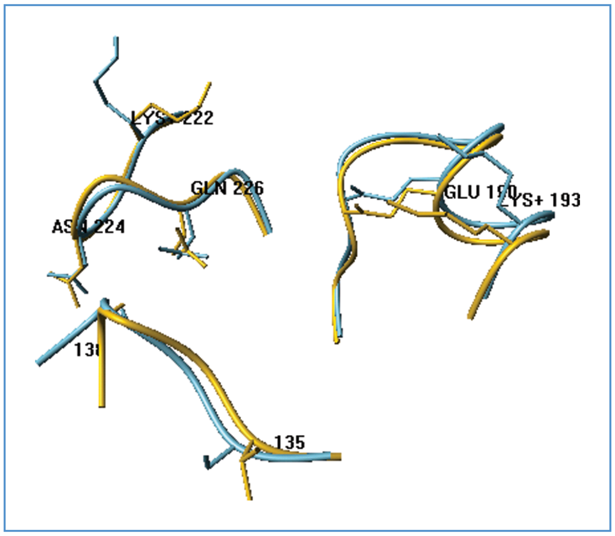


Figure 5

Conformational change observed in the free state of modeled avian H5 HA on the (V135S and A138S) double mutation. The backbone of the wild-type protein is shown in yellow, whereas the backbone of the double mutant protein is shown in cyan. Side chains are shown for the two mutation sites, i.e., 135 and 138, in addition to the residues E190, K193, K222, N224, and Q226. Clearly, introduction of the double mutation alters the conformation of the receptor binding pocket of H5 HA, which significantly facilitates human receptor binding. (Reproduced from [19], with permission.)

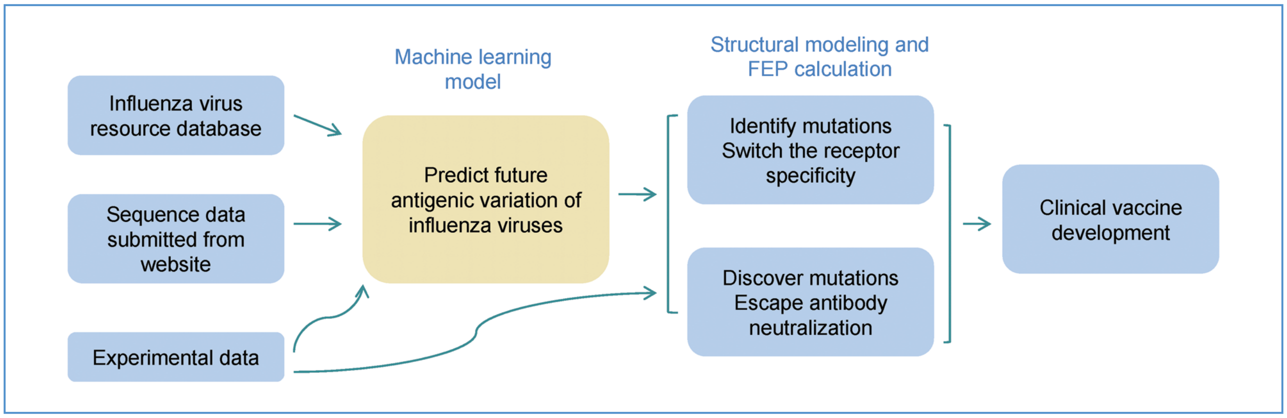

Machine learning model
Structural modeling and FEP calculation
Influenza virus resource database
Sequence data submitted from website
Experimental data
Predict future antigenic variation of influenza viruses
Identify mutations Switch the receptor specificity
Discover mutations Escape antibody neutralization
Clinical vaccine development

Figure 6

Flowchart illustrating the suggested systematic approach to facilitate the development of new influenza virus vaccines. (FEP: free energy perturbation.)

The double mutation (V135S and A138S) in H5 HA significantly enhances the human receptor binding by $\Delta\Delta G = -2.56 \pm 0.73$ kcal/mol over the avian one ($\Delta\Delta G = 0.84 \pm 1.02$ kcal/mol). Simulating the same double mutation in another H5 HA–receptor complex, the Sing97 virus has reproduced this effect, where this double mutation also shows a substantial increase in the human receptor binding ($\Delta\Delta G = -1.18 \pm 0.57$ kcal/mol) over the avian one ($\Delta\Delta G = -0.15 \pm 0.99$ kcal/mol). Considering that the HA–glycan binding is weak in general, a -2.56 kcal/mol relative binding free energy change per monomer of H5 HA should be regarded as a strong indication for a substantial increase in HA receptor binding affinity. A free energy component analysis indicates that the electrostatic interactions dominate the contribution to the free energy change associated with human receptor binding of the double V135S and A138S mutant ($\sim$80% from electrostatic and $\sim$20% from van der Waals interactions). A structural analysis offers further explanation of why the double mutant prefers the human receptor (see **Figure 5**). The HA–glycan hydrogen-bonding network is rearranged in the double V135S and A138S mutant. Residues Y95, S135, S136, S137, E190, K222, G225, and Q226 of this altered receptor binding domain form favorable hydrogen-bonding interactions with the human-like receptor, similar to the human-adapted HAs. In addition, the crucial presence of S138 in human H1 HA further emphasizes the significance of this predicted double mutant of avian H5 HA in gaining pandemic potential.

Conclusion

In summary, we have proposed a novel systematic approach to facilitate the development of vaccines against influenza virus (see **Figure 6**) [17–19]. We first use machine-learning techniques to predict likely mutations that might occur in the next flu season. A weighted MI algorithm is developed to capture the nonlinear correlation for each amino acid site of HA. Then, the variations at the predicted sites of mutation could be structurally modeled in order to investigate the consequence of these mutations. A few examples have been provided for the antigen–Ab and antigen–receptor binding affinity calculations. These advanced structural modeling and simulations provide detailed atomistic mechanisms and help identify potential mutations that could escape Ab neutralization or switch the receptor specificity from avian to human. Overall, our comprehensive approach could provide potent leads for the design of future vaccines against influenza virus.

Acknowledgments

The authors would like to thank Isidore Rigoutsos, Alice McHardy, Gulei Jin, Jun Zhu, Bruce Berne, Steven Turner, and David Topham for many useful discussions, as well as Ian Wilson, Jim Paulson, and Peter Palese for helpful comments at the beginning of this project.

*Trademark, service mark, or registered trademark of International Business Machines Corporation in the United States, other countries, or both.

**Trademark, service mark, or registered trademark of Top500.org, in the United States, other countries, or both.

References

1. N. J. Cox and K. Subbarao, "Global epidemiology of influenza: Past and present," *Annu. Rev. Med.*, vol. 51, pp. 407–421, 2000.
2. M. R. Hilleman, "Realities and enigmas of human viral influenza: Pathogenesis, epidemiology and control," *Vaccine*, vol. 20, no. 25/26, pp. 3068–3087, Aug. 19, 2002.
3. T. Horimoto and Y. Kawaoka, "Influenza: Lessons from past pandemics, warnings from current incidents," *Nat. Rev. Microbiol.*, vol. 3, no. 8, pp. 591–600, Aug. 2005.

4. J. Stevens, O. Blixt, L. Glaser, J. K. Taubenberger, P. Palese, J. C. Paulson, and I. A. Wilson, "Glycan microarray analysis of the hemagglutinins from modern and pandemic influenza viruses reveals different receptor specificities," *J. Mol. Biol.*, vol. 355, no. 5, pp. 1143–1155, Feb. 2006.
5. D. Fleury, S. A. Wharton, J. J. Skehel, M. Knossow, and T. Bizebard, "Antigen distortion allows influenza virus to escape neutralization," *Nat. Struct. Biol.*, vol. 5, no. 2, pp. 119–123, Feb. 1998.
6. Z. Y. Yang, C. J. Wei, W. P. Kong, L. Wu, L. Xu, D. F. Smith, and G. J. Nabel, "Immunization by avian H5 influenza hemagglutinin mutants with altered receptor binding specificity," *Science*, vol. 317, no. 5839, pp. 825–828, Aug. 2007.
7. T. M. Tumpey, T. R. Maines, N. Van Hoeven, L. Glaser, A. Solorzano, C. Pappas, N. J. Cox, D. E. Swayne, P. Palese, J. M. Katz, and A. Garcia-Sastre, "A two-amino acid change in the hemagglutinin of the 1918 influenza virus abolishes transmission," *Science*, vol. 315, no. 5812, pp. 655–659, Feb. 2007.
8. J. Stevens, A. L. Corper, C. F. Basler, J. K. Taubenberger, P. Palese, and I. A. Wilson, "Structure of the uncleaved human H1 hemagglutinin from the extinct 1918 influenza virus," *Science*, vol. 303, no. 5665, pp. 1866–1870, Mar. 19, 2004.
9. J. Stevens, O. Blixt, T. M. Tumpey, J. K. Taubenberger, J. C. Paulson, and I. A. Wilson, "Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus," *Science*, vol. 312, no. 5772, pp. 404–410, Apr. 21, 2006.
10. S. J. Gamblin, L. F. Haire, R. J. Russell, D. J. Stevens, B. Xiao, Y. Ha, N. Vasisht, D. A. Steinhauer, R. S. Daniels, A. Elliot, D. C. Wiley, and J. J. Skehel, "The structure and receptor binding properties of the 1918 influenza hemagglutinin," *Science*, vol. 303, no. 5665, pp. 1838–1842, Mar. 19, 2004.
11. R. Chen and E. C. Holmes, "Avian influenza virus exhibits rapid evolutionary dynamics," *Mol. Biol. Evol.*, vol. 23, pp. 2336–2341, Dec. 2006.
12. R. Nielsen and Z. Yang, "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene," *Genetics*, vol. 148, no. 3, pp. 929–936, Mar. 1998.
13. J. P. Huelsenbeck and K. A. Dyer, "Bayesian estimation of positively selected sites," *J. Mol. Evol.*, vol. 58, no. 6, pp. 661–672, Jun. 2004.
14. Y. Suzuki, "New methods for detecting positive selection at single amino acid sites," *J. Mol. Evol.*, vol. 59, no. 1, pp. 11–19, Jul. 2004.
15. Z. Yang and W. J. Swanson, "Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes," *Mol. Biol. Evol.*, vol. 19, no. 1, pp. 49–57, Jan. 2002.
16. Z. Yang, R. Nielsen, N. Goldman, and A. M. Pedersen, "Codon-substitution models for heterogeneous selection pressure at amino acid sites," *Genetics*, vol. 155, no. 1, pp. 431–449, May 2000.
17. R. Zhou, P. Das, and A. K. Royyuru, "Single mutation induced H3N2 hemagglutinin antibody neutralization: A free energy perturbation study," *J. Phys. Chem. B*, vol. 112, no. 49, pp. 15 813–15 820, Dec. 2008.
18. Z. Xia, G. Jin, J. Zhu, and R. Zhou, "Using a mutual information-based site transition network to map the genetic evolution of influenza A/H3N2 virus," *Bioinformatics*, vol. 25, no. 18, pp. 2309–2317, Sep. 15, 2009.
19. P. Das, J. Li, A. K. Royyuru, and R. Zhou, "Free energy simulations reveal a double mutant avian H5N1 virus hemagglutinin with altered receptor binding specificity," *J. Comput. Chem.*, vol. 30, no. 11, pp. 1654–1663, Aug. 2009.
20. D. J. Smith, A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. Osterhaus, and R. A. Fouchier, "Mapping the antigenic and genetic evolution of influenza virus," *Science*, vol. 305, no. 5682, pp. 371–376, Jul. 16, 2004.
21. A. C. Shih, T. C. Hsiao, M. S. Ho, and W. H. Li, "Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 104, no. 15, pp. 6283–6288, Apr. 10, 2007.
22. X. Du, Z. Wang, A. Wu, L. Song, Y. Cao, H. Hang, and T. Jiang, "Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution," *Genome Res.*, vol. 18, no. 1, pp. 178–187, Jan. 2008.
23. A. Apisarnthanarak, S. Erb, I. Stephenson, J. M. Katz, M. Chittaganpitch, S. Sangkitporn, R. Kitphati, P. Thawatsupha, S. Waicharoen, U. Pinitchai, P. Apisarnthanarak, V. J. Fraser, and L. M. Mundy, "Seroprevalence of anti-H5 antibody among Thai health care workers after exposure to avian influenza (H5N1) in a tertiary care center," *Clin. Infect. Dis.*, vol. 40, no. 2, pp. e16–e18, Jan. 15, 2005.
24. T. Sawai, Y. Itoh, H. Ozaki, N. Isoda, K. Okamoto, Y. Kashima, Y. Kawaoka, Y. Takeuchi, H. Kida, and K. Ogasawara, "Induction of cytotoxic T-lymphocyte and antibody responses against highly pathogenic avian influenza virus infection in mice by inoculation of apathogenic H5N1 influenza virus particles inactivated with formalin," *Immunology*, vol. 124, no. 2, pp. 155–165, Jun. 2008, DOI: 10.1111/j.1365-2567.2007.02745.x.
25. S. Yamada, Y. Suzuki, T. Suzuki, M. Q. Le, C. A. Nidom, Y. Sakai-Tagawa, Y. Muramoto, M. Ito, M. Kiso, T. Horimoto, K. Shinya, T. Sawada, M. Kiso, T. Usui, T. Murata, Y. Lin, A. Hay, L. F. Haire, D. J. Stevens, R. J. Russell, S. J. Gamblin, J. J. Skehel, and Y. Kawaoka, "Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors," *Nature*, vol. 444, no. 7117, pp. 378–382, Nov. 2006.
26. O. Blixt, S. Head, T. Mondala, C. Scanlan, M. E. Huflejt, R. Alvarez, M. C. Bryan, F. Fazio, D. Calarese, J. Stevens, N. Razi, D. J. Stevens, J. J. Skehel, I. van Die, D. R. Burton, I. A. Wilson, R. Cummings, N. Bovin, C. H. Wong, and J. C. Paulson, "Printed covalent glycan array for ligand profiling of diverse glycan binding proteins," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 101, no. 49, pp. 17 033–17 038, Dec. 2004.
27. T. Simonson, G. Archontis, and M. Karplus, "Free energy simulations come of age: Protein–ligand recognition," *Acc. Chem. Res.*, vol. 35, no. 6, pp. 430–437, Jun. 2002.
28. B. L. Tembre and J. A. McCammon, "Ligand receptor interactions," *Comput. Chem.*, vol. 8, no. 4, pp. 281–283, 1984.
29. W. L. Jorgensen, "Free-energy calculations—A breakthrough for modeling organic-chemistry in solution," *Acc. Chem. Res.*, vol. 22, no. 5, pp. 184–189, May 1989.
30. A. Warshel, F. Sussman, and G. King, "Free energy of charges in solvated proteins: Microscopic calculations using a reversible charging process," *Biochemistry*, vol. 25, no. 26, pp. 8368–8372, Dec. 1986.
31. P. Kollman, "Free-energy calculations—Applications to chemical and biochemical phenomena," *Chem. Rev.*, vol. 93, no. 7, pp. 2395–2417, 1993.
32. A. Pathiaseril and R. J. Woods, "Relative energies of binding for antibody–carbohydrate–antigen complexes computed from free-energy simulations," *J. Amer. Chem. Soc.*, vol. 122, no. 2, pp. 331–338, Jan. 2000.
33. A. Laederach and P. J. Reilly, "Modeling protein recognition of carbohydrates," *Proteins—Struct. Function Bioinformatics*, vol. 60, no. 4, pp. 591–597, Sep. 2005.
34. A. J. Butte and I. S. Kohane, "Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements," in *Proc. Pacific Symp. Biocomputing*, 2000, vol. 5, pp. 418–429.
35. A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane, "Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 97, no. 22, pp. 12 182–12 186, Oct. 2000.
36. A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, "ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, no. Suppl 1, p. S7, 2006.
37. J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale mapping and validation of Escherichia coli transcriptional

regulation from a compendium of expression profiles," *PLoS Biol.*, vol. 5, no. 1, p. e8, Jan. 2007.

38. Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman, "The influenza virus resource at the National Center for Biotechnology Information," *J. Virol.*, vol. 82, no. 2, pp. 596–601, Jan. 2008.
39. W. Wang, O. Donini, C. M. Reyes, and P. A. Kollman, "Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein–ligand, protein–protein, and protein–nucleic acid noncovalent interactions," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 30, pp. 211–243, 2001.
40. A. Warshel, "Simulating the energetics and dynamics of enzymatic reactions in specificity in biological interactions," *Pontificiae Academiae Scientiarum Scripta Varia*, vol. 55, pp. 60–81, 1984.
41. A. Warshel, P. K. Sharma, M. Kato, and W. W. Parson, "Modeling electrostatic effects in proteins," *Biochim. Biophys. Acta*, vol. 1764, no. 11, pp. 1647–1676, Nov. 2006.
42. J. Tirado-Rives and W. L. Jorgensen, "Contribution of conformer focusing to the uncertainty in predicting free energies for protein–ligand binding," *J. Med. Chem.*, vol. 49, no. 20, pp. 5880–5884, Oct. 2006.
43. Y. Deng and B. Roux, "Calculation of standard binding free energies: Aromatic molecules in the T4 lysozyme L99A mutant," *J. Chem. Theory Comput.*, vol. 2, pp. 1255–1273, 2006.
44. G. Jayachandran, M. R. Shirts, S. Park, and V. S. Pande, "Parallelized-over-parts computation of absolute binding free energy with docking and molecular dynamics," *J. Chem. Phys.*, vol. 125, no. 8, p. 084 901, Aug. 2006.
45. M. Almlof, J. Aqvist, A. O. Smalas, and B. O. Brandsdal, "Probing the effect of point mutations at protein–protein interfaces with free energy calculations," *Biophys. J.*, vol. 90, no. 2, pp. 433–442, Jan. 2006.
46. B. O. Brandsdal, F. Osterberg, M. Almlof, I. Feierberg, V. B. Luzhkov, and J. Aqvist, "Free energy calculations and ligand binding," *Adv. Protein Chem.*, vol. 66, pp. 123–158, 2003.
47. B. O. Brandsdal and A. O. Smalas, "Evaluation of protein–protein association energies by free energy perturbation calculations," *Protein Eng.*, vol. 13, no. 4, pp. 239–245, Apr. 2000.
48. S. Kumar, C. Huang, G. Almasi, and L. V. Kale, "Achieving strong scaling with NAMD on Blue Gene/L," in *Proc. IEEE Int. Parallel Distrib. Process. Symp.*, 2006, p. 10, DOI: 10.1109/IPDPS.2006.1639298.
49. M. Eleftheriou, R. S. Germain, A. K. Royyuru, and R. Zhou, "Thermal denaturing of mutant lysozyme with both the OPLSAA and the CHARMM force fields," *J. Amer. Chem. Soc.*, vol. 128, no. 41, pp. 13 388–13 395, Oct. 2006.
50. P. Liu, X. Huang, R. Zhou, and B. J. Berne, "Observation of a dewetting transition in the collapse of the melittin tetramer," *Nature*, vol. 437, no. 7055, pp. 159–162, Sep. 2005.
51. R. Zhou, M. Eleftheriou, A. K. Royyuru, and B. J. Berne, "Destruction of long-range interactions by a single mutation in lysozyme," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 104, no. 14, pp. 5824–5829, Apr. 2007.
52. R. Zhou, X. Huang, C. J. Margulis, and B. J. Berne, "Hydrophobic collapse in multidomain protein folding," *Science*, vol. 305, no. 5690, pp. 1605–1609, Sep. 2004.
53. S. Kumar, C. Huang, G. Zheng, E. Bohm, A. Bhatele, J. C. Phillips, H. Yu, and L. V. Kale, "Scalable molecular dynamics with NAMD on the IBM Blue Gene/L system," *IBM J. Res. & Dev.*, vol. 52, no. 1/2, pp. 177–188, Jan. 2008.
54. R. Zhou, "Exploring the protein folding free energy landscape: Coupling replica exchange method with P3ME/RESPA algorithm," *J. Mol. Graph. Model.*, vol. 22, no. 5, pp. 451–463, May 2004.
55. T. Kuiken, E. C. Holmes, J. McCauley, G. F. Rimmelzwaan, C. S. Williams, and B. T. Grenfell, "Host species barriers to influenza virus infections," *Science*, vol. 312, no. 5772, pp. 394–397, Apr. 2006.
56. R. J. Russell, D. J. Stevens, L. F. Haire, S. J. Gamblin, and J. J. Skehel, "Avian and human receptor binding by hemagglutinins of influenza A viruses," *Glycoconj. J.*, vol. 23, no. 1/2, pp. 85–92, Feb. 2006.
57. J. J. Skehel and D. C. Wiley, "Receptor binding and membrane fusion in virus entry: The influenza hemagglutinin," *Annu. Rev. Biochem.*, vol. 69, pp. 531–569, 2000.
58. D. van Riel, V. J. Munster, E. de Wit, G. F. Rimmelzwaan, R. A. M. Fouchier, A. Osterhaus, and T. Kuiken, "H5N1 virus attachment to lower respiratory tract," *Science*, vol. 312, no. 5772, p. 399, Apr. 2006.
59. K. Shinya, M. Ebina, S. Yamada, M. Ono, N. Kasai, and Y. Kawaoka, "Influenza virus receptors in the human airway," *Nature*, vol. 440, no. 7083, pp. 435–436, Mar. 2006.
60. D. van Riel, V. J. Munster, E. de Wit, G. F. Rimmelzwaan, R. A. M. Fouchier, A. Osterhaus, and T. Kuiken, "Human and avian influenza viruses target different cells in the lower respiratory tract of humans and other mammals," *Amer. J. Pathol.*, vol. 171, no. 4, pp. 1215–1223, Oct. 2007.
61. J. M. Nicholls, A. J. Bourne, H. Chen, Y. Guan, and J. S. M. Peiris, "Sialic acid receptor detection in the human respiratory tract: Evidence for widespread distribution of potential binding sites for human and avian influenza viruses," *Respir. Res.*, vol. 8, no. 1, p. 73, 2007.

Received January 15, 2011; accepted for publication January 31, 2011

Zhen Xia *Department of Biomedical Engineering, The University of Texas at Austin, Austin, TX 78712 USA (zhenxia@mail.utexas.edu).* Mr. Xia is currently a Ph.D. student and a Research Assistant in the Biomedical Engineering Department at the University of Texas at Austin. He received two B.S. degrees in automation and bioinformatics from Zhejiang University, Hangzhou, China, in 2005. He performed the current work at IBM Computational Biology Center at the Thomas J. Watson Research Center as a research intern in 2010, where he has worked on protein folding/recognition and structure-function prediction in the Protein Science Group. He is interested in understanding and engineering biological molecular systems for pharmaceutical and biomedical applications.

Payel Das *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (daspa@us.ibm.com).* Dr. Das is a Research Staff Scientist in the Protein Science Group of Computational Biology Center at IBM Thomas J. Watson Research Center, Yorktown Heights, NY. She received her Ph.D. degree in chemistry from Rice University, Houston, TX, in 2007 and then performed postdoctoral work at IBM Thomas J. Research Center. Her current research interests are large-scale simulations for protein folding, misfolding, and aggregation, protein-protein interaction, and multiscale modeling.

Tien Huynh *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (gdesktop@us.ibm.com).* Mr. Huynh is a Senior Software Engineer in the Protein Science Group of Computational Biology Center at the Thomas J. Watson Research Center. He received B.S. and M.S. degrees in computer science from Kent State University, Kent, OH, in 1983 and 1985, respectively. He joined IBM at the Thomas J. Watson Research Center in 1984 and has worked on various projects in programming language design, parallel programming, biometrics, and bioinformatics.

Ajay K. Royyuru *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (ajayr@us.ibm.com).* Dr. Royyuru is Senior Manager of the Computational Biology Center at IBM Research. He obtained his Ph.D. degree in molecular biology from the Tata Institute of Fundamental Research, Mumbai, in 1993 and then performed postdoctoral work in structural biology at Memorial Sloan-Kettering Cancer Center, New York. Currently his work focuses on collaborative research at the interface of information technology and biology. Working with biologists and institutions around the world, he is engaged in research that will advance personalized, information-based medicine.

Ruhong Zhou *IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (ruhongz@us.ibm.com).* Dr. Zhou is Manager of the Protein Science Group, Computational Biology Center at IBM Thomas J. Watson Research Center, and an Adjunct Professor at Department of Chemistry, Columbia University, New York, NY. He received his Ph.D. degree from Columbia University in 1997. His current research interests include development of novel methods and algorithms for computational biology and bioinformatics, and large-scale simulations for protein folding, misfolding, and aggregation, protein-protein interaction, ligand-receptor binding, and protein nanoparticle interactions (nanotoxicity).