J|A|C|S COMMUNICATIONS

Published on Web 07/28/2009

Combining Molecular Dynamics with Bayesian Analysis To Predict and Evaluate Ligand-Binding Mutations in Influenza Hemagglutinin

Peter M. Kasson, Daniel L. Ensign, and Vijay S. Pande*

Departments of Chemistry and Structural Biology, Stanford University, Stanford, California 94305

Received June 4, 2009; E-mail: pande@stanford.edu

Influenza attaches to host cells by binding cell-surface glycans via the viral hemagglutinin protein. Influenza hemagglutinin binds glycans in a species-specific manner: avian strains of influenza selectively bind glycans found in the avian upper respiratory tract, while human strains selectively bind human upper respiratory tract glycans.[1] Changes to this specificity are considered among the key factors required for efficient transmission of avian influenza between humans. In contrast, recent transmission between swine and humans is eased by the marked similarity between swine and human upper respiratory tract glycans.[2−5] Structural studies of H1N1 influenza from 1918 implicate specificity changes in the influenza pandemic of 1918−1919,[6,7] and retrospective characterizations of H5N1 avian influenza isolates from humans find mutations that shift H5N1 to an intermediate specificity between avian-type and human-type glycans.[8−10]

Despite the success of these retrospective analyses, prospective prediction of H5N1 mutations remains a much more difficult task. Influenza hemagglutinin is heavily glycosylated,[11] and the viral glycans can affect both affinity and specificity for host glycans.[12−14] Even using simplified ligands, large-scale expression and experimental screening of hemagglutinin glycoprotein mutants for specificity changes remain challenging due to biosafety issues and the difficulty of doing large-scale mutagenesis in cell-culture systems that will produce the relevant hemagglutinin glycosylation patterns. We have therefore designed an approach to large-scale computational screening of hemagglutinin mutants that will allow more directed experimental validation.

A number of experimental and computational methods have been developed to examine receptor-binding-domain mutants, but ligand-binding mutants from clinical isolates of influenza virus encompass *both receptor-binding-domain and allosteric sites*. Experimental data for allosteric sites are particularly sparse due to the challenges of high-throughput mutagenesis and screening of influenza hemagglutinin. We have designed a molecular-dynamics approach to score potential mutants with robust predictive power for both receptor-binding-domain and allosteric mutations. We perform thousands of simulations of 17 hemagglutinin mutants totaling >1 ms in length and employ a Bayesian model to rank mutations that disrupt hemagglutinin−ligand complex stability. Based on our analysis, we predict a significantly increased k_{off} for seven of these mutants. This means of analyzing molecular-dynamics data to make experimentally verifiable predictions offers a potentially general method to identify ligand-binding mutants, particularly allosteric ones. Our analysis provides a robust means to evaluate mutants prior to experimental mutagenesis and testing; these results also constitute an important step toward understanding the determinants of ligand binding by H5N1 influenza.

Dissociation rates were chosen as a means to evaluate predicted ligand-binding mutants because the association and dissociation rates (k_{on} and k_{off}) of ligands from wild-type hemagglutinins is relatively slow; data on monovalent k_{on} and k_{off} are not available, but X-31 hemagglutinin rosettes bind fetuin with multivalent rates of $k_{on} = 2 \times 10^3$ M^{-1} s^{-1} and $k_{off} = 2 \times 10^{-4}$ s^{-1}.[15] Experimental dissociation rates reported for hemagglutinin vary by up to 10 000-fold, however, based in part on the surface conjugation.[15,16] Depending on whether a ligand-binding mutation alters the transition-state free energy or only the free energy of the bound state, it would alter both k_{on} and k_{off} or k_{on} alone. We can sample and estimate fast processes more accurately via molecular dynamics than we can slow ones, so acceleration of k_{off} is a more accessible parameter than deceleration of k_{on}. Computational methods to predict free energies of binding under active development,[17] but predicting binding of charged, flexible ligands by a flexible protein is extremely challenging for current methods. Methods based on molecular mechanics-Generalized Born calculations have recently been applied to predict hemagglutinin glycan binding.[18] These show promise for predicting receptor-binding-domain mutations, while our molecular-dynamics-based calculations are designed also to detect allosteric mutants in a robust fashion.

Selecting Mutants for Simulation. We employ a combined approach that uses both molecular-dynamics simulation and sequence data to predict ligand-binding mutants of H5N1 influenza hemagglutinin. We first analyze dynamics of bound-state simulations to predict residues important to ligand binding. Covariance[19−21] and mutual-information methods[22,23] have previously been used to decompose protein motions in molecular simulations and identify important large-scale movements. Here, we score protein residues via a more targeted criterion: dynamic relationship to the ligand. We quantify this as excess mutual information between the residue α-carbon position and the ligand position and score according to this dynamic relationship. Our approach is designed to detect residues in the receptor-binding domain and allosteric sites as well as detect both interactions on a rapid time scale and ones that require slow conformational change.

We simulated the ligand-bound state of H5N1 hemagglutinin using the isolate VN1194 bound to α2,3-sialyllactose as previously crystallized.[10] The trimeric hemagglutinin complex was simulated for 100 ns, and excess mutual information was computed between each protein residue of each monomer and the corresponding bound ligand, using the average mutual information between the residue and all protein residues as an estimate of the "background" mutual information. The top 5% of residues scored via this method are rendered in Figure 1 (and listed in Table S1); they show substantial spatial overlap with ligand-binding specificity mutants identified by retrospective analysis of clinical isolates and confirmatory experimental mutagenesis.[8−10,24]

10.1021/ja904557w CCC: $40.75

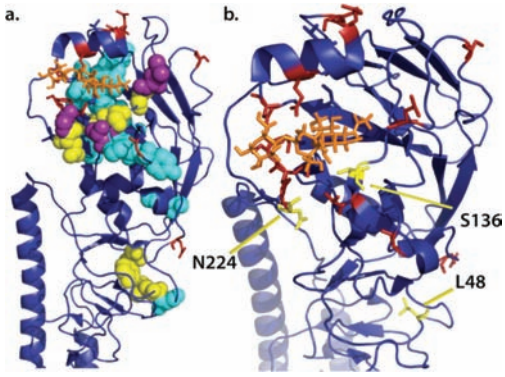

Figure 1. Mutation sites in hemagglutinin. Panel (a) shows a hemagglutinin monomer with experimentally identified ligand-binding mutations in red, the top 5% of residues by dynamics scoring in cyan (overlap of these two in magenta), and the six mutation sites identified by both dynamics and sequence analysis in yellow. Panel (b) shows the top three mutations from our ligand dissociation analyses in yellow. A modeled α2,3-sialyllactose is shown in orange.

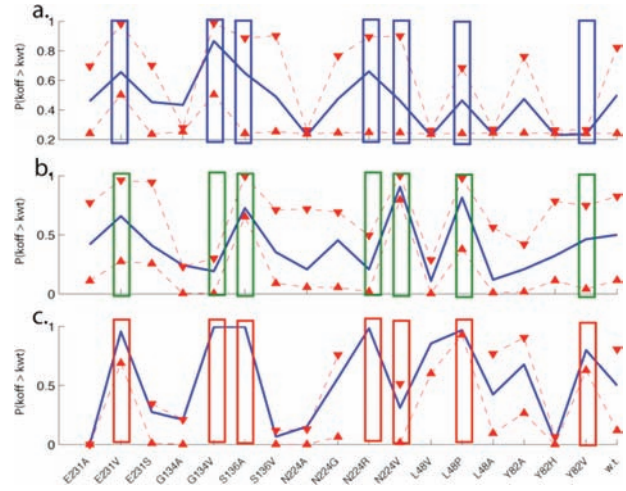

Figure 2. Estimated dissociation rate acceleration. For each starting monomer conformation from the crystal structure (a−c), plots show the probability that each mutant k_{off} is faster than wild-type VN1194. Triangles represent 90% bootstrap confidence intervals. Rectangles denote mutants with significantly increased k_{off} when conformations A−C are considered together ($p < 10^{-5}$, Kolmogorov−Smirnov test).

Analysis of Mutational Data and Prediction of Mutants. We combine these results with sequence analysis of H5N1 mutational data to predict clusters of residues that undergo coordinated mutation. Such residues have some capacity to vary but are subject to selective pressure relating mutation in some residue *i* to mutation in residue *j*. We hypothesize that these may be richer targets to change ligand specificity than residues that are absolutely conserved (and may be required for hemagglutinin function) or residues that display uncorrelated mutations and may be involved in immune escape. We use pairwise sequence mutual information[25,26] as a robust nonlinear measure of relatedness.

Residues in H5N1 hemagglutinin were identified as previously described[26] by computing pairwise mutual information on a multiple sequence alignment of all available human and avian H5N1 hemagglutinin sequences (see Supporting Information (SI)). All residues scoring in the top 0.1% of pairs were included. The intersection of residues selected via this sequence-based method and those in the top 5% via dynamics-based analysis is shown in Figure 1: the residues identified are L48, Y82, G134, S136, N224, and E231. We hypothesized that point mutations at coordinated residues might disrupt interactions important to ligand binding; we therefore selected mutants for further analysis by mutating each residue to Ala or Val and to residues found in the multiple sequence alignment. The 17 point mutants thus identified were further evaluated via computational mutagenesis as described below.

Simulation of Ligand-Binding Mutants. Predicted ligand-binding affinity mutants were evaluated by simulating each point mutant in complex with α2,3-sialyllactose and assessing changes to complex stability. We have developed Bayesian analysis methods to predict dissociation rates based on extensive simulation of each mutant and evaluate whether a mutant has a faster dissociation rate than the influenza clinical isolate that we use as a wild-type reference. This method uses the stochastic nature of physical kinetics to predict rare events by combining many trajectories each less than the mean time for the process. Three monomers of each mutant were simulated (in the trimeric complex) for one run of 100 ns and >200 runs of at least 50 ns; these simulations were used to estimate the dissociation rate for each mutant. We have analyzed the expected number of dissociation events as a function of $\Delta\Delta G^{\ddagger}$; with our degree of sampling we expect to detect mutants with $\Delta\Delta G^{\ddagger} > \sim 5$ kcal/mol (Figure S1).

The estimated dissociation rates incorporate both positive (dissociation observed) and negative simulation data of varying length; see the SI for k_{off} probability density functions of each mutant and starting conformation. We predict the probability that each mutation accelerates k_{off}, destabilizing the bound complex (Figure S2). We then perform a bootstrap analysis to identify mutations that significantly speed k_{off} compared to the wild-type VN1194 complex.

We predict that seven mutations significantly ($p < 10^{-5}$) perturb binding of α2,3-sialyllactose by hemagglutinin: E231V, G134V, S136A, N224R, N224V, L48P, and Y82V (Figure 2). More dissociation events were observed from starting conformation C than A or B, suggesting that the three protein−ligand complexes derived from the crystal structure differ in baseline stability. The effects of each mutation were relatively consistent across starting conformations; however, all but one of these mutants (Y82V) significantly perturb binding in at least two of the three conformations tested. This conformation dependence emphasizes the need for extensive molecular dynamics sampling; we have performed additional simulations of six mutants using 30 additional starting states (Figure S3).

The three mutations most strongly predicted to destabilize ligand binding by hemagglutinin are S136A, N224V, and L48P. As shown in Figure 1, S136 lies within the binding pocket, N224 is a loop residue at the edge of the ligand-binding pocket, and L48 is distant from the binding pocket and does not directly contact the bound ligand. The mutations we predict to affect the stability of ligand binding thus include both binding-pocket and allosteric sites. S136A and N224V lie adjacent to experimentally identified ligand-specificity mutation sites 138, 225, 226, 227, and 228. The E231V mutant, which scored fifth in our analysis, lies within the monomer−monomer interface; it is possible that this mutation acts to alter interactions between monomers.

As negative controls, we have tested two mutants, S127P and N197K, where the EC50 in a plate-binding assay using live whole virus is within 10-fold of wild-type VN1194,[10] and we

thus expect the monomeric k_{off} to be minimally perturbed. Our analysis does not predict acceleration of either k_{off} (Figure S4). In addition, one of our three top-scoring mutants, S136A, has been tested in H3N2 influenza. S136 is conserved between VN1194 and human H3N2 strains; the S136A mutant of A/Aichi/68 has been studied experimentally and has 30% of wild-type activity in an erythrocyte binding assay.[27]

Here, we present a new means to predict ligand-binding affinity mutations in influenza hemagglutinin. Our method combines molecular dynamics analysis with sequence data and employs information-theoretical methods to score mutation sites based on their relation to ligand binding in a robust manner. We combine this scoring with a sequence-based analysis of residue covariation to produce a focused set of mutation sites. In this report, we predict mutants with altered binding affinity, but the underlying method is designed to be applicable to predicting altered binding specificity as well.

We also demonstrate a new statistical methodology for computational evaluation of ligand-binding mutants by estimating changes to k_{off}. Dissociation is a more accessible process than association, and it may also be more relevant, as studies of binding of small flexible ligands by MHC molecules[28] and RNaseS[29] show dissociation rates to be sensitive to mutation while association rates remain relatively constant. This occurs when the transition state is dominated by nonspecific interactions. The bound state free energy is then more mutation-sensitive than the transition state, so k_{off} is affected much more than k_{on}. Though plausible, it is unknown if hemagglutinin-ligand binding has such a transition state.

We model the dissociation reaction as approximately two-state kinetically, and a Bayesian framework allows rate estimation based on both positive and negative observations of dissociation in a heterogeneous data set of many molecular dynamics simulations. This approach is particularly helpful in comparing two rates, as one can utilize the entire probability distribution rather than only the maximum-likelihood estimate. To account for small-sample-size effects, we encapsulate the Bayesian analysis in a bootstrap error analysis to give robust estimates of statistical significance. We have tested this analytical procedure by retrospective validation against mutants that bind similarly to wild-type hemagglutinin in experimental assays and comparing our top-scoring mutant to experimental data on the analogous mutation in H3N2 influenza.

The mutation sites predicted by analysis of the molecular dynamics data include both residues immediately contacting the bound glycan and residues located farther away on the globular head of the hemagglutinin molecule. The spatial patterning of these residues is particularly provocative, especially since the two allosteric mutation sites we predict are located adjacent to the E79 residue implicated in ligand-binding specificity by Yamada et al.[10] Any relationship must be considered speculative at this stage until more experimental testing and computational analyses are completed, but the potential for an allosteric regulatory "locus" in that region is extremely intriguing. Perhaps even more so is the potential for a new method to predict such sites in a general manner.

Acknowledgment. The authors thank R. Brandman, B. Daigle, T. Fenn, O. Troyanskaya, and V. Voelz for many helpful discussions. P.K. was supported by a Berry Foundation fellowship. Computational resources were provided by Folding@Home donors worldwide, by NSF Award CNS-0619926, by E. Lindahl and the Swedish Royal Institute of Technology, and by award NIH R01-GM062868 to V.S.P.

Supporting Information Available: Simulation details and additional analyses, complete refs 8 and 10. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Matrosovich, M. N.; Matrosovich, T. Y.; Gray, T.; Roberts, N. A.; Klenk, H. D. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4620–4.
(2) Nicholls, J. M.; Chan, R. W.; Russell, R. J.; Air, G. M.; Peiris, J. S. *Trends Microbiol.* **2008**, *16*, 149–57.
(3) Dawood, F. S.; Jain, S.; Finelli, L.; Shaw, M. W.; Lindstrom, S.; Garten, R. J.; Gubareva, L. V.; Xu, X.; Bridges, C. B.; Uyeki, T. M. *N. Engl. J. Med.* **2009**, *360*, 2605–15.
(4) Rota, P. A.; Rocha, E. P.; Harmon, M. W.; Hinshaw, V. S.; Sheerar, M. G.; Kawaoka, Y.; Cox, N. J.; Smith, T. F. *J. Clin. Microbiol.* **1989**, *27*, 1413–6.
(5) Gambaryan, A. S.; Karasin, A. I.; Tuzikov, A. B.; Chinarev, A. A.; Pazynina, G. V.; Bovin, N. V.; Matrosovich, M. N.; Olsen, C. W.; Klimov, A. I. *Virus Res.* **2005**, *114*, 15–22.
(6) Gamblin, S. J.; Haire, L. F.; Russell, R. J.; Stevens, D. J.; Xiao, B.; Ha, Y.; Vasisht, N.; Steinhauer, D. A.; Daniels, R. S.; Elliot, A.; Wiley, D. C.; Skehel, J. J. *Science* **2004**, *303*, 1838–42.
(7) Stevens, J.; Corper, A. L.; Basler, C. F.; Taubenberger, J. K.; Palese, P.; Wilson, I. A. *Science* **2004**, *303*, 1866–70.
(8) Auewarakul, P.; et al. *Virol.* **2007**, *81*, 9950–5.
(9) Stevens, J.; Blixt, O.; Tumpey, T. M.; Taubenberger, J. K.; Paulson, J. C.; Wilson, I. A. *Science* **2006**, *312*, 404–10.
(10) Yamada, S.; et al. *Nature* **2006**, *444*, 378–82.
(11) Wilson, I. A.; Skehel, J. J.; Wiley, D. C. *Nature* **1981**, *289*, 366–73.
(12) Gambaryan, A. S.; Marinina, V. P.; Tuzikov, A. B.; Bovin, N. V.; Rudneva, I. A.; Sinitsyn, B. V.; Shilov, A. A.; Matrosovich, M. N. *Virology* **1998**, *247*, 170–7.
(13) Marinina, V. P.; Gambarian, A. S.; Bovin, N. V.; Tuzikov, A. B.; Shilov, A. A.; Sinitsyn, B. V.; Matrosovich, M. N. *Mol. Biol. (Mosk)* **2003**, *37*, 550–5.
(14) Kasson, P. M.; Pande, V. S. *Biophys. J.* **2008**, *95*, L48–50.
(15) Takemoto, D. K.; Skehel, J. J.; Wiley, D. C. *Virology* **1996**, *217*, 452–8.
(16) Mandenius, C. F.; Wang, R. H.; Alden, A.; Bergstrom, G.; Thebault, S.; Lutsch, C.; Ohlson, S. *Anal. Chim. Acta* **2008**, *623*, 66–75.
(17) Mobley, D. L.; Dill, K. A. *Structure* **2009**, *17*, 489–98.
(18) Xu, D.; Newhouse, E. I.; Amaro, R. E.; Pao, H. C.; Cheng, L. S.; Markwick, P. R.; McCammon, J. A.; Li, W. W.; Arzberger, P. W. *J. Mol. Biol.* **2009**, *387*, 465–91.
(19) Balsera, M. A.; Wriggers, W.; Oono, Y.; Schulten, K. *J. Phys. Chem.* **1996**, *100*, 2567–2572.
(20) Ichiye, T.; Karplus, M. *Proteins* **1991**, *11*, 205–17.
(21) Hunenberger, P. H.; Mark, A. E.; van Gunsteren, W. F. *J. Mol. Biol.* **1995**, *252*, 492–503.
(22) Lange, O. F.; Grubmüller, H. *Proteins* **2006**, *62*, 1053–61.
(23) Lange, O. F.; Grubmüller, H. *Proteins* **2008**, *70*, 1294–312.
(24) Stevens, J.; Blixt, O.; Chen, L. M.; Donis, R. O.; Paulson, J. C.; Wilson, I. A. *J. Mol. Biol.* **2008**, *381*, 1382–94.
(25) Martin, L. C.; Gloor, G. B.; Dunn, S. D.; Wahl, L. M. *Bioinformatics* **2005**, *21*, 4116–24.
(26) Kasson, P. M.; Pande, V. S. *Pac. Symp. Biocomput.* **2009**, 492–503.
(27) Martín, J.; Wharton, S. A.; Lin, Y. P.; Takemoto, D. K.; Skehel, J. J.; Wiley, D. C.; Steinhauer, D. A. *Virology* **1998**, *241*, 101–11.
(28) Kasson, P. M.; Rabinowitz, J. D.; Schmitt, L.; Davis, M. M.; McConnell, H. M. *Biochemistry* **2000**, *39*, 1048–58.
(29) Goldberg, J. M.; Baldwin, R. L. *Biochemistry* **1998**, *37*, 2556–63.

JA904557W