JVI Journals.ASM.org

Quantitative Modeling of Virus Evolutionary Dynamics and Adaptation in Serial Passages Using Empirically Inferred Fitness Landscapes

Hyung Jun Woo, Jaques Reifman

Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, Fort Detrick, Maryland, USA

We describe a stochastic virus evolution model representing genomic diversification and within-host selection during experimental serial passages under cell culture or live-host conditions. The model incorporates realistic descriptions of the virus genotypes in nucleotide and amino acid sequence spaces, as well as their diversification from error-prone replications. It quantitatively considers factors such as target cell number, bottleneck size, passage period, infection and cell death rates, and the replication rate of different genotypes, allowing for systematic examinations of how their changes affect the evolutionary dynamics of viruses during passages. The relative probability for a viral population to achieve adaptation under a new host environment, quantified by the rate with which a target sequence frequency rises above 50%, was found to be most sensitive to factors related to sequence structure (distance from the wild type to the target) and selection strength (host cell number and bottleneck size). For parameter values representative of RNA viruses, the likelihood of observing adaptations during passages became negligible as the required number of mutations rose above two amino acid sites. We modeled the specific adaptation process of influenza A H5N1 viruses in mammalian hosts by simulating the evolutionary dynamics of H5 strains under the fitness landscape inferred from multiple sequence alignments of H3 proteins. In light of comparisons with experimental findings, we observed that the evolutionary dynamics of adaptation is strongly affected not only by the tendency toward increasing fitness values but also by the accessibility of pathways between genotypes constrained by the genetic code.

Viruses with RNA genomes evolve rapidly, evading selective pressure from the host immune response and adapting to changing environments (1–4). In particular, their capacity to switch host species, emerge into new vulnerable populations, and cause outbreaks has significant implications for viral disease control (5). Many such outbreaks in recent history have been attributed to viral species jumps: influenza A virus has jumped from birds and pigs to humans multiple times (6–9), the severe acute respiratory syndrome (SARS) epidemic was caused by a species jump of coronavirus from bats and palm civets to humans (10–13), and human immunodeficiency virus type 1 (HIV-1) is believed to have switched hosts from primates (14).

Characterizing the complex factors affecting the evolution of viruses in natural settings among diverse groups of interacting hosts is a challenging task. Valuable insights have been gained by evolutionary experiments under more controlled conditions, particularly within the context of serial-passage experiments (15). In a serial-passage experiment, a cell culture or live host is inoculated by viral (or other) pathogens, usually already well adapted to different cell types or hosts. A pathogen's growth under the restrictive host environment leads to within-host selection for advantageous variants, either present as a minority in the founder population or generated from error-prone replications. After a certain amount of time (approximately days) of such growths, a small subset of the resulting pathogen population is sampled and used to inoculate a fresh new medium or host, initiating a subsequent round of the passage. Generally, rapid adaptation to the new host environment in the form of increased fitness and virulence is observed (typically within ~10 passages) along with attenuation of adaptation to the former host (15). In addition to revealing key adaptation strategies a pathogen can exhibit, an important application of serial passages is the production of attenuated vaccine strains capable of eliciting immune responses without virulence (16).

The recent availability of rapid and inexpensive deep-sequencing techniques (17) has the potential to significantly improve our understanding of how key factors affect virus evolutionary dynamics, including species jumps, especially when combined with controlled experiments such as serial passages. A major obstacle in leveraging the growing sequence data, however, is the lack of quantitative connections between such genotype data and experimentally measurable viral phenotypes, including growth properties, infectivity, virulence, and tissue tropism. The main objective of this study was to develop, validate, and apply stochastic models of viral evolutionary dynamics that can bridge this gap by realistically modeling genomic diversification and adaptation processes during serial passages.

Mathematical models of viral dynamics have long been used to obtain insights into how viruses interact with and adapt to different hosts. The mainstay among such models is the continuum description of population dynamics formulated with ordinary differential equations (18–20). One of their drawbacks, however, is that they are not readily extendable to descriptions of extensive

Received 9 October 2013 **Accepted** 30 October 2013
Published ahead of print 6 November 2013
Address correspondence to Hyung Jun Woo, hwoo@bhsai.org, or Jaques Reifman, jaques.reifman.civ@mail.mil.

doi:10.1128/JVI.02958-13

genomic diversification processes. Lee et al. have modeled HIV-1 intrahost sequence evolution and divergence (21). Lim et al. studied the design of growth-attenuated viruses using computational models for the intracellular growth of vesicular stomatitis viruses (22, 23). Evolutionary bottleneck effects associated with serial passages have been modeled using a different mathematical approach (24, 25). Explicit descriptions of within-host dynamics with mutations have been used by Russell et al. (26) to examine the potential for emergence of airborne-transmissible influenza A H5N1 virus variants (6, 27), but without consideration of serial passages used in the experiment.

Recently, we introduced a stochastic dynamics-based modeling framework of virus dynamics under immune selective pressure (28). It combines the classic quasispecies theory description of mutation-selection balance (1, 29, 30) and the discrete-state stochastic simulation algorithm widely used for chemical reactions (31). The model was successfully used to describe HIV-1 within-host dynamics in both short (immune escape dynamics) and long (disease progression) time scales. In this study, we developed an analogous stochastic model suitable for describing the within-host adaptation of viruses in serial passages. We demonstrated that the approach allows one to quantitatively probe the effects of key factors of evolutionary dynamics on adaptation, including the fitness distribution in sequence space, bottleneck size, and mutation rate. In particular, we examined the adaptation process by considering the relative accessibility of genotypes well adapted to a new host environment from the starting stock during passages. We found that the number of passage rounds necessary to achieve this adaptation increases exponentially with the required number of amino acid mutations, rendering triple mutants practically inaccessible. This observation is consistent with experimental studies of the adaptation of influenza A H5N1 viruses (6) and SARS coronavirus (CoV) (12).

In particular, for influenza A H5N1 viruses, which are normally restricted to avian hosts but can occasionally spill over into mammals, specific hemagglutinin (HA) mutations that allow for binding to mammalian receptors have been suggested as the key factor for potential species jumps (6, 27). Upon serial passaging in ferrets of an engineered strain containing three mutations (two in HA) known to facilitate receptor binding, viral strains capable of respiratory infection containing two additional mutations were found. Modeling the evolutionary adaptation of such specific systems requires an empirical fitness landscape (the correspondence relationship between genotypes and replication rate), which could be inferred from a sufficiently large set of multiple-sequence alignments (MSAs). For this fitness inference procedure, we adopted the direct coupling analysis (32, 33), a method intended for extracting the degree of coevolution between sites within a protein due mostly to direct spatial interactions that are obscured in frequency correlation data. It allows for the determination of a maximum-entropy fitness function (a "disordered Potts model") that can be used to estimate fitness values of arbitrary sequences, while taking into account interresidue correlations. Ferguson et al. have recently used a similar approach in their study of HIV fitness landscapes, but employing a reduced description of a two-letter alphabet (wild type [WT] and mutant) and Monte Carlo simulations for inference (34) rather than the mean field approximation developed by Morcos et al. (33). We used this inferred landscape to simulate the adaptation of H5N1: an H5 protein segment evolves under the fitness landscape inferred from sequences (H3) natural to mammalian hosts, e.g., the H3N2 subtype endemic in human populations. Our simulations revealed that both the fitness values of genotypes and their relative accessibilities dictated by the genetic code strongly affect the adaptation dynamics.

MATERIALS AND METHODS

Stochastic model and simulation. The basic model of virus evolution can be described by the following set of events:

$$U + V_n \xrightarrow[a]{} I_n \text{ (infection)} \quad (1)$$

$$I_n \xrightarrow[Q_{mn} r_n]{} I_n + V_m \text{ (replication)} \quad (2)$$

$$I_n \xrightarrow[b]{} 0 \text{ (death)} \quad (3)$$

$$V_n \xrightarrow[b]{} 0 \text{ (clearance)} \quad (4)$$

where U is an uninfected host cell and V_n and I_n are a virion with genotype n and a cell infected with it, respectively. Equation 1 represents the infection of a host cell with rate a, and equation 2 corresponds to a replication event in which an infected cell of genotype n sheds a virion with genotype m with replication rate r_n and mutation probability (29, 30),

$$Q_{mn} = (1 - \mu)^{L - d_{mn}} (\mu/3)^{d_{mn}} \quad (5)$$

where μ is the mutation rate (probability for substitution per nucleotide site per replication), L is the total length of the genomic nucleotide segment, and d_{mn} is the Hamming distance (total number of nucleotides that differ) between genotypes m and n. Equation 5 gives the joint probability of mutating d_{mn} sites with probability $\mu/3$ (one out of three possibilities from n to m), while leaving $L - d_{mn}$ sites alone with probability $1 - \mu$. In the actual simulations, equation 5 is a consequence of the algorithm rather than an input, because the target genotype m is determined by randomly mutating each site of genotype n with probability $\mu/3$ into one of three possibilities. The genome length is a multiple of 3 such that $L = 3L_a$, where L_a is the number of amino acids. A given genotype n is translated into the corresponding amino acid sequence $\alpha(n)$ using the standard genetic code. The replication rate r_n is a function of amino acid sequence only: $r_n = r_{\alpha(n)}$. Equations 3 and 4 represent the death of an infected cell and the clearance of a virion, respectively, which are assumed to have the same rate b.

The dynamic model was simulated stochastically by the Gillespie algorithm (31), using as the initial condition a given number of uninfected cells (U_0), founder viruses (V_0), and no infected cells. The founder group consisted of virions with a monoclonal (WT) genotype. Simulations proceeded by dynamically updating the list of genotypes and corresponding amino acid sequences encountered as the viral population became more diverse (28), instead of enumerating all possible sequences, which become exponentially numerous even for small L_a.

For comparison, we also considered an extended model for adaptation in animal hosts with immune response for which equation 3 is replaced by

$$I_n \xrightarrow[s]{} I_n + T_n \text{ (stimulation)} \quad (6)$$

$$I_n + T_n \xrightarrow[c]{} T_n \text{ (killing)} \quad (7)$$

$$T_n \xrightarrow[b]{} 0 \text{ (clearance)} \quad (8)$$

where T_n denotes a cytotoxic T lymphocyte specific to genotype n, stimulated with rate s by the presence of infected cells and capable of killing them with rate c.

Serial-passage protocol. To model serial passages, we implemented a stochastic simulation algorithm where the monoclonal founder group underwent a growth process for a fixed amount of time (τ), and the resulting quasispecies was randomly sampled to form a new polyclonal founder group for the next round of passage. All free virions were sampled, while infected cells were excluded. A given virion has a probability $f = V_0/V < 1$ to be sampled, such that one round of growth producing population size V is followed by the next round with an initial number fV of viruses. This procedure leaves the set of frequencies of genotypes pres-

ent before and after the bottleneck event roughly the same. Strains with small frequencies, however, often disappear after a bottleneck event ("extinctions").

Simulation under random landscapes. We first performed two different classes of simulations: those in which the founder viruses are already well adapted to the host and those where the founder group faces a novel host environment. In both cases, a key input to the model was the fitness function r_α, for which we first adopted random fitness landscapes: the fitness function was modeled as a Gaussian random variable with mean $\langle r_\alpha \rangle = r_0 \exp(-d_\alpha/\xi)$ and standard deviation σ, where d_α is the distance of sequence α (amino acid sites that are different) from a reference with fitness r_0, and ξ is a characteristic distance of the fitness decay. We used $\xi = 1$ amino acid (aa) and $\sigma = 0.1$ unless otherwise specified. The choice of reference sequence ($d_\alpha = 0$) distinguishes the two scenarios without and with adaptations: in the former, the reference sequence coincides with the founder WT sequence, while in the latter, they are distinct. The reference (i.e., the "most fit" [MF] sequence) was taken as a random nucleotide sequence of a certain length, translated into the amino acid sequence, and assigned the fitness r_0. The results shown were obtained typically by averaging >100 different realizations, each simulated with different random sequences under the given parameter set and length/distance specifications. Dependence on the sample size of realizations indicated rapid convergence within less than ~100 realizations, with consistently large variances (data not shown).

Simulations of passages without adaptation were started with these MF sequences as the founder group. For simulations of passages with adaptation, WT and MF sequences in each realization were generated by first assigning the WT randomly and then subsequently mutating it with a μ of 0.1 until a sequence with the desired amino acid distance was obtained. The latter sequence was designated the MF sequence.

The adaptation process studied in this work was defined by the exploration of sequence space by the viral population during passages starting from a founder population with a suboptimal fitness (WT), leading to the discovery and dominance of the MF sequence. Simulations were performed with the genomic segment length $L_a = d$, where d is the distance between WT and MF sequences. As a characteristic measure of the efficiency of adaptation, we probed the time required ("jumping time") for the MF sequence to be discovered and its frequency (calculated by summing the numbers of free virions and infected cells) to reach 50% of all sequences. The inverse of this time was defined as the "jumping rate" (J), a stochastic quantity with a wide distribution including $J = 0$, which corresponds to cases where the trajectory never reached the MF sequence. The maximum simulation time was set such that it was much longer than the typical jumping times for trajectories that did reach MF. Realizations that became stuck with a nonzero distance to MF exhibited stationary MF frequencies of <50% (often zero) beyond the time maximum and were counted as having a zero jumping rate.

Inference of fitness landscapes. To model specific evolutionary adaptations more directly linked to experimental studies, we derived empirical fitness landscapes of HA sequences for influenza H5N1 viruses and used them as input to our stochastic model. For the fitness inference, we adopted a modified version of the direct coupling analysis algorithm (32, 33) and used it to infer the fitness landscape of a model influenza HA protein segment. In this inference procedure, one employs an empirical functional form of the stationary distribution of genotypes that maximizes its entropy and infers the parameters in this expression using MSAs. We downloaded influenza virus sequence sets from the NCBI (35): for H5, we used 2,085 full-length HA sequences of H5N1, excluding laboratory strains, while for H3, we used 4,113 full-length HA sequences of H3 strains.

For inference, the MSAs were used to derive the empirical single-site and two-site amino acid distributions, $f_i(\alpha)$ and $f_{ij}(\alpha,\beta)$, each giving the frequency of amino acid α on site i and the joint frequency of α and β on sites i and j, respectively. A prior distribution $p_0(\alpha)$ with a count n_0 was used in frequency calculations, such that the total effective number of sequences was $n_0 + N_s$, where N_s is the actual sequence count. For the prior distribution $p_0(\alpha)$, we used the frequency values from the Jones-Taylor-Thornton matrix (36), augmented by $p_0 = 0.01$ for the gap, and renormalized the 21 values. The prior count n_0 values used were 3,000 and 4,000 for H5 and H3, respectively. The set of empirical distributions, $f_i(\alpha)$ and $f_{ij}(\alpha,\beta)$, were then matched to the maximum-entropy expression of the probability of a sequence $q = (\alpha_1, \ldots, \alpha_{La})$, $p(q) \sim \exp(r_q/r_s)$, where

$$r_q = \sum_i h_i(\alpha_i) + \sum_{i<j} C_{ij}(\alpha_i, \alpha_j) \quad (9)$$

was obtained using the mean-field approximation (33), where the matrix inverse of $f_{ij}(\alpha,\beta)$ gives the parameters $C_{ij}(\alpha,\beta)$ after excluding a reference amino acid, which we took as Ala, from the argument list. In our application, the frequency data for the $L_a = 4$ minigenome (103, 156, 222, and 224 [H5 numbering], identified as the locations of key mutations in HA during H5N1 adaptation in ferrets [6]) were used to infer the parameter set $C_{ij}(\alpha,\beta)$ and $h_i(\alpha)$. The inference only determined fitness values up to an additive constant (in addition to the reference fitness value r_s), which we chose such that genotypes neighboring the MF sequence had suitable replication rates within the stochastic model.

Simulation of influenza virus adaptation. We used the inferred H3 fitness landscape as an input to the stochastic simulations of H5 WT adaptation in mammalian host environments. Each simulation was started with the WT genotype CATACACAAGGA (HTQG) of H5N1 with a different random number seed, and each mutant generated in the simulations was dynamically assigned the fitness calculated by the inferred fitness landscape parameters. The reference fitness value was set as $r_s = 1$ day^{-1}. The WT and MF (SAQG) fitness values were 4.9 and 21.7 (in units of r_s), respectively. This setup mimics the adaptation of H5 under the selective environments native to H3 strains.

RESULTS

Stochastic model. The basic model of stochastic virus evolution used in this study involved the following set of species: uninfected host cell U, a virion with genotype n (denoted by V_n), and a cell infected with the virion I_n. They interacted according to the following set of events: an infection of a host cell with rate a, a replication event where I_n produced V_m with replication rate r_n and mutation probability Q_{mn} (a function of the mutation rate μ, defined as the probability of making a substitution error per nucleotide site per replication), and two additional events in which an infected cell dies and a virion is cleared with rate b (see Materials and Methods). We adopted the same rate for the last two processes for simplicity. The clearance of viruses may be viewed as reflecting the effect of any innate and/or acquired immune responses present if passages occur in live animals or unspecified decay mechanisms in cell cultures.

If the mutation rate μ is zero ($Q_{mn} = 1$ if $m = n$ and 0 otherwise), the continuum approximation to equations 1 to 4 becomes equivalent to the target cell-limited model used by Baccam et al. (20) and Pawelek et al. (37) to study the kinetics of influenza A virus infection. We compared the full discrete stochastic simulation with the outcomes of the continuum differential equation representation and found that the latter severely overestimated the speed of adaptation (explored in detail in the following subsections) because it assigned nonzero frequencies to all genotypes via equation 5 at all times t of >0 (data not shown). Only for a simplified model with $L = 3$ nt (64 possible genotypes) and an artificially high mutation rate μ of ~0.1 was the difference between the two methods relatively small. As μ decreased (or conversely with increasing L, which increases the total number of genotypes exponentially while the population size remains small), the population dynamics quickly became dominated by the finite-

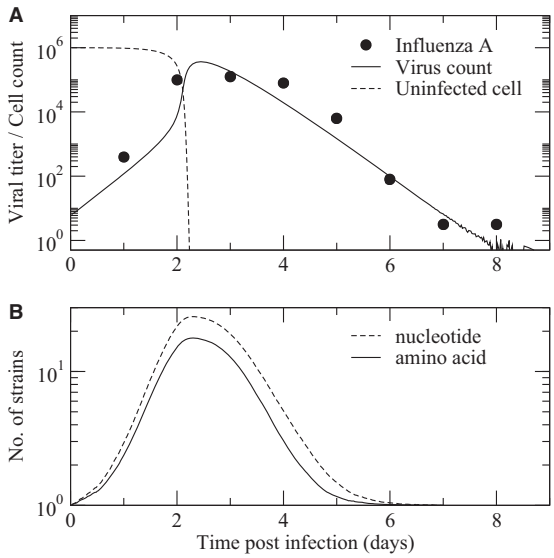

FIG 1 Viral growth, host cell infection, and within-host diversification dynamics predicted by the stochastic model for the following parameter values: $L_a = 3$ aa, $V_0 = 1 \times 10^3$, $a = 1.0 \times 10^{-3}$ day^{-1}, $b = 3.0$ day^{-1}, $r_0 = 6.0$ day^{-1}, $\mu = 1.0 \times 10^{-5}$, and $U_0 = 1 \times 10^6$. (A) Viral titer and uninfected cell number U_0. Symbols represent the experimental data from references 20 and 38, plotted assuming a scaling factor of 1 viral count per 1 $TCID_{50}$/ml. (B) Total numbers of genotypes (nucleotide and amino acid sequences) as a function of time.

size effects: stochastic simulations of the full discrete model predicted long waiting times for the discovery of MF genotype, while the continuum approximation yielded a rapid rise in MF frequency. Therefore, to realistically model the time-dependent adaptation behavior of virus populations, it is important to take into account the discrete nature of population sizes.

Viral growth curve. The typical time-dependent behavior of virus growth and uninfected cell depletion for a single cell culture or single host is shown in Fig. 1A, along with the experimental data of influenza A virus H1N1 infection kinetics from references 20 and 38. The total virus count V in our model output was assumed to be proportional to the experimentally measured viral titers. For comparison, we assumed an empirical scaling factor such that 1 virus count corresponded to 1 50% tissue culture infective dose ($TCID_{50}$)/ml. During simulations, the initial clonal founder virus group (V_0 virions) underwent a short period of clearance before the target cells (initially U_0) became infected appreciably. The viral titer began to rise as infection and replication occurred, whereas the number of uninfected cells decreased. The growth stopped when the target cells became depleted and was followed by gradual cell death and virus clearance processes. The parameter values were chosen to obtain a good agreement in virus count between the simulation and the experimental data. The value of a primarily affected the shape of the initial rise in virus count, with larger values leading to a flatter curve up to ~2 days (initial virions rapidly consumed by infection) followed by a sharper increase (infected cells start producing copied virions). The value of b affected the decay rate after saturation, with larger values leading to a faster decrease in virus count.

This overall behavior of total species count (uninfected cells and viruses of all genotypes) shown in Fig. 1 was qualitatively similar to those obtained with a kinetic model by Baccam et al. (20). However, the discrete stochastic model provided previously unavailable and realistic views of the genomic diversification process accompanying this growth (Fig. 1B). The results shown in Fig. 1 were obtained by averaging over many explicit representations of $L_a = 3$ aa ($L = 9$ nt) genomic segments with random sequences and decreasing fitness distributions of mutants around the WT with increasing distance (see Materials and Methods), a choice that has been empirically validated using experimental data (28). The numbers of distinct nucleotide and amino acid sequences increased with the viral titer growth, reaching a peak together with the virus count. The cell death/clearance phase (time, 3 to 8 days) led to a restoration of the monoclonal population, which was due to the relatively low mutation rate μ of 1×10^{-5}, close to typical values estimated for RNA viruses (39, 40). The maximum numbers of genotypes shown in Fig. 1B, ~25 nt sequences (and ~17 aa sequences), in fact, are close to the number of all possible single mutants plus the WT, $1 + 3L = 28$ nt sequences, which suggests that with the given mutation rate value, single mutants dominated the within-host quasispecies for small genomic segments of L_a on the order of 1 aa. In summary, in addition to illustrating qualitative behavior of viral growths in agreement with experimental trends, our model also revealed that genomic diversification is maximal when the population size grows rapidly.

Serial passages. We extended the basic stochastic model given by equations 1 to 4 to simulate the typical growth and evolutionary dynamics encountered in experimental serial passages. In this protocol, simulations of single-host growth as shown in Fig. 1 were terminated at a certain time (passage interval τ, chosen near the viral count maximum), and the quasispecies population was sampled with equal probability for all virions present to produce a founder group of virions to infect fresh new target cells (with the same U_0) for the next round of the passage. The "bottleneck factor" $f < 1$, defined as the ratio of the initial viral population size of a round to the final viral population size of the previous round (and controlled by varying the founder group size V_0), quantifies the strength of selection pressure exerted on viruses: smaller f values cause only the most dominant genotypes (higher frequencies) to survive these bottleneck events. The survival probability therefore depends on how efficiently a given strain infects the limited number of target cells to replicate in competition with others.

We first examined the simpler case in which the viral population was already well adapted to the target cell environment, and no systematic evolutionary drift in sequence space was expected. This situation was modeled by using a fitness function centered around the WT (the genotype corresponding to the founder group of the first passage). Figure 2 shows the typical time-dependent behavior of the virus population size and genomic diversity. The total viral count repeated the growth pattern after a bottleneck event with the passage period τ of 3 days, with the initial value at the beginning of a round kept constant as V_0. The smaller value of death/clearance rate b in simulations shown in Fig. 2 compared to Fig. 1 resulted in flatter plateau regions near the viral count peaks, where the genomic diversity was maintained near the single-mutant maximum level of ~28 nt sequences. Although the bottleneck selection cuts this diversity down to ~1 (WT) in this particular case, we found the initial number of strains at each round to vary depending on f, approaching the maximum for larger f values.

The combined frequency of mutants shown at the bottom part

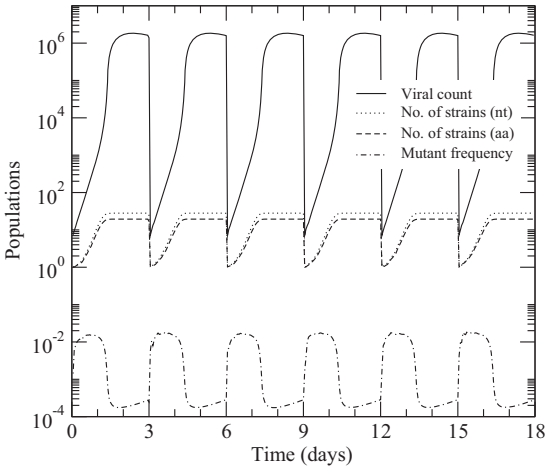

FIG 2 Time dependence of viral population size and diversity over serial passage simulations without adaptation. The WT and MF coincide in the sequence space. The parameter values were the same as in Fig. 1 except that $b = 1.0\ \text{day}^{-1}$ and $\tau = 3$ days.

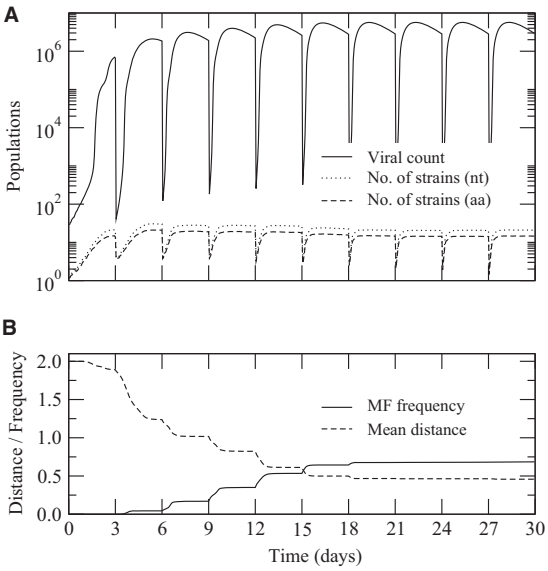

FIG 3 Time dependence of population size and diversity (A) as well as mean distance to the MF sequence and the frequency of the MF (B). Parameter values were as follows: $L_a = d = 2$ aa, $V_0 = 1 \times 10^4$, $U_0 = 1 \times 10^6$, $a = 1.0 \times 10^{-3}\ \text{day}^{-1}$, $b = 1.0\ \text{day}^{-1}$, $r_0 = 20\ \text{day}^{-1}$, $\mu = 1.0 \times 10^{-5}$, and $\tau = 3$ days.

of Fig. 2 indicates that during the initial growth process in a given passage round, the rapid population growth with mutations led to a sharp increase in diversity. Then, as the target cells were depleted, due to an increase in viral population, the mutant frequency decreased again after ~1.5 days. Because in this case the fitness distribution was centered at the WT, the viral population reverted to the WT-dominant structures after each moderate diversification stage under selection. We examined the trend in the total number of genotypes present after the end of each passage round as a function of passage number up to 30 passages and confirmed that under this condition, each passage round was statistically independent of the previous round (data not shown). In summary, our model allows for a simple description of repeated bottleneck events under serial passages. If the initial population is already well adapted to the host environment, the selection events did not affect the quasispecies structure appreciably.

Adaptation. More typical in experimental serial passages is the situation in which the WT is exposed to a new host environment in which its fitness is not optimal. Evolutionary dynamics then exhibit an adaptation process where systematic drifts toward more fit genotypes occur within a characteristic timescale, which we refer to as the (species) "jumping time." In our modeling, the adaptation process was defined as the exploration of sequences near the initial WT during passages and the discovery of more fit genotypes leading to the increase in their frequencies. We adopted fitness distributions where a WT was chosen at a certain distance d from the MF genotype. We found this distance to MF to be the dominant factor influencing the relative probability of adaptation. The jumping time was defined as the time required since the initiation of passages for the MF genotype to be discovered and for its frequency to reach 50% of the viral population.

Figure 3A shows typical evolutionary dynamics of such an adaptation process from our simulations averaged over 100 realizations, where $d = 2$ aa and other parameter values were chosen to best illustrate the typical behavior in simulations under these settings. During the earlier rounds where the population was dominated by the WT with a lower fitness, the total virus count reached was relatively smaller. It gradually increased with the increasing number of passages as mutants with higher fitness values appeared and displaced the WT. Figure 3B shows variations of the mean distance to the MF and the frequency of the MF. The mean distance roughly decreased monotonically with the number of passages, whereas the MF frequency did not increase appreciably from zero until the third passage round: the adaptation at the early stage primarily arose from mutants other than the MF, which only appeared after a considerable period of "search" in the sequence space. Both the approach to the MF and the increase in MF frequency occurred in graduated "steps" with most changes concentrated in the early phase of each passage round, where active growth occurred and competition among strains was most severe.

The late-stage limiting values of the mean distance and frequency after many rounds of passages were fairly distant from zero and one, respectively. We found that this feature was primarily due to the sizeable presence of realizations in which evolutionary dynamics got "stuck" in mutants distinct from the MF but with fitness values larger than that of WT, rather than a consistent convergence to these average limiting values. This feature can be seen more clearly in Fig. 4, which shows the distribution of the jumping rate J, defined as the inverse of the jumping time. The distribution is multimodal, with broad peaks at larger, finite rates and a nonzero probability of zero rate. The latter, more pronounced for $d = 2$ aa (Fig. 4B), can be attributed to the topological structure of the genetic code: transitions via substitution errors between a significant number of pairs of amino acids require multiple mutations, making them highly unlikely. A population can become dominated by a genotype close to the MF in distance but not directly accessible. The fine structure of the distributions shown in Fig. 4 reflects the characteristics of passage dynamics: the rightmost cluster of peaks in Fig. 4A ($d = 1$ aa) corresponds to cases in which the MF frequency exceeded 50% during the first passage cycle, the cluster to its left the second cycle, and so on. For a d of 2 aa, the probability of jumping during the first passage cycle was virtually zero (Fig. 4B). The gaps between these clusters ap-

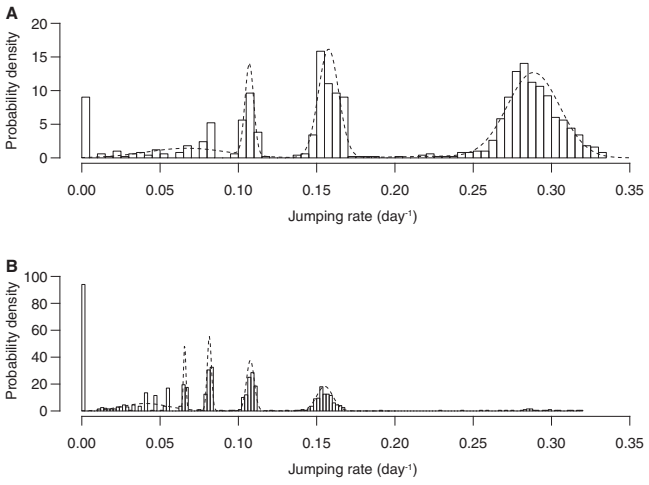

FIG 4 Jumping rate (J) distributions from the sets of 10^3 realizations using parameter values as in Fig. 3 for two different distances: $d = 1$ aa (A) and $d = 2$ aa (B). The dashed lines show maximum-likelihood fits to linear combinations of normal distributions. The cluster of peaks in panel A centered at a J of $\approx$0.3 day^{-1}, 0.16 day^{-1}, 0.11 day^{-1}, and 0.07 day^{-1} correspond to jumping events (MF frequency exceeding 50%) during the first, second, third, and fourth passage rounds, respectively. The peaks at a J of 0 arise from trajectories trapped at sequences without direct single substitution route to the MF sequence.

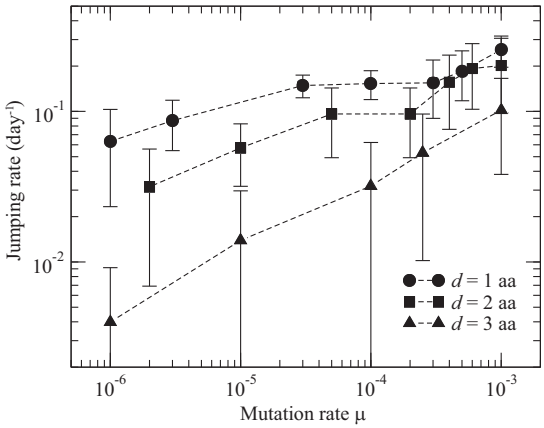

FIG 5 Dependence of jumping rate on the mutation rate for the WT-to-MF distance from 1 to 3 aa. Symbols represent averages and error bars represent 1 standard deviation over 100 realizations. Parameter values were as follows: $V_0 = 1 \times 10^5$, $U_0 = 1 \times 10^6$, $a = 1.0 \times 10^{-3}$ day^{-1}, $b = 0.1$ day^{-1}, $r_0 = 10$ day^{-1}, and $\tau = 3$ days.

peared because large-scale changes to frequency distribution occur mostly during the early growth stage within each round (Fig. 3B). The jumping rate correlates with the general likelihood of observing a given adaptation during passages. To support this conclusion quantitatively, we fitted the distributions in Fig. 4 using maximum-likelihood estimation to the functional form of a sum of normal distributions with the amplitude, mean, and standard deviations as parameters, shown in Fig. 4 as dashed lines. The jumping time in units of passage numbers for the modes with significant amplitudes (the peaks in Fig. 4 from right to left) were 1.2, 2.1, and 3.1 for Fig. 4A and 2.1, 3.1, 4.1, and 5.1 for Fig. 4B, confirming that jumping events mostly occur near the beginning of each passage cycle. In summary, the model allowed for quantitative characterizations of the speed of adaptation in terms of the time required for a highly fit genotype to appear and dominate the quasispecies population under passages.

Variations of jumping rate. We examined the dependence of jumping rate J, which is a measure of the relative probability of observing a certain adaptation from a WT to a MF genotype during passages, on mutation rate μ (Fig. 5). For a given distance d between the WT and MF, the jumping rate monotonically increased with increasing μ. The large variance in jumping rate showed little sample-size dependence and reflected the contribution of pathways inaccessible to the MF genotype in sequence space, as well as the multimodal structure of its distribution over the initial phase of many passage cycles (Fig. 4). The relative amplitudes of normal distributions fitted to the data in Fig. 4A and B were distributed evenly, with proportions of 0.5, 0.3, and 0.1 and 0.2, 0.2, and 0.2, respectively, for the first 3 main peaks from the right. To examine the effects of heterogeneity in substitution rates at different sites, which would become more important for longer sequences, we also considered the case where the mutation rate μ at each nucleotide site was sampled from a gamma distribution (41). For an L of 6 nt, the site heterogeneity did not affect the jumping rate appreciably (data not shown). Its effect, however, may become stronger for longer genomic segments, for which deep-sequencing data of viral genomes (42) could be used to assign site and nucleotide-dependent substitution rates.

While jumping rate values became statistically similar for different distances near μ of $\sim 10^{-3}$ (Fig. 5), they differed widely near the experimental range of mutation rates of $\sim 10^{-5}$, where J decreased exponentially with increasing d, reaching J of $\sim 10^{-2}$ day^{-1} for a d of 3 aa, or a jumping time of $\sim$100 days (33 passages). From the results shown in Fig. 5, we may infer estimates for the typical number of passages (of 3 days for each round) that would be required under a typical mutation rate of $\sim 10^{-5}$ to arrive at and jump to an MF sequence: 3, 6, and 33 passages for 1-, 2-, and 3-aa distances, respectively.

The dependences of the jumping rate with other physical parameters are shown in Fig. 6. The host cell number U_0 determines the viral population size reached within each passage round. The jumping rate monotonically increased with increasing U_0, saturating near $\sim$0.1 day^{-1} (Fig. 6A). The bottleneck size V_0 for a given U_0 represents the strength of the selection pressure, determining the bottleneck factor f. Although at first the jumping rate increased with increasing V_0 for small V_0, it reached a maximum and began to decrease as V_0 approached U_0 (Fig. 6B). For small V_0, more fit mutants found during a passage round frequently were lost because of their low frequencies, whereas for large enough V_0, they usually survived the bottleneck selections. The increase of J with increasing V_0 for small V_0 therefore is a gradual attenuation of the finite-size effect. For a V_0 of $>10^4$ (or f of $>10^{-2}$ for V of 10^6), the starting viral population at each round was sufficiently large, making a smaller bottleneck size more favorable for adaptation.

The passage period τ affects the jumping rate primarily via the size and frequency of viral populations at which bottleneck events occur. The approximate increase of J with decreasing τ (Fig. 6C) can be explained by the increase in the fraction of time a population experienced strong competition via rapid growth during each passage round (Fig. 3). If the period is too short ($\tau = 1$ day), this growth phase begins to be interrupted by bottlenecks, decreasing J again. We found the jumping rate J to be insensitive to variations

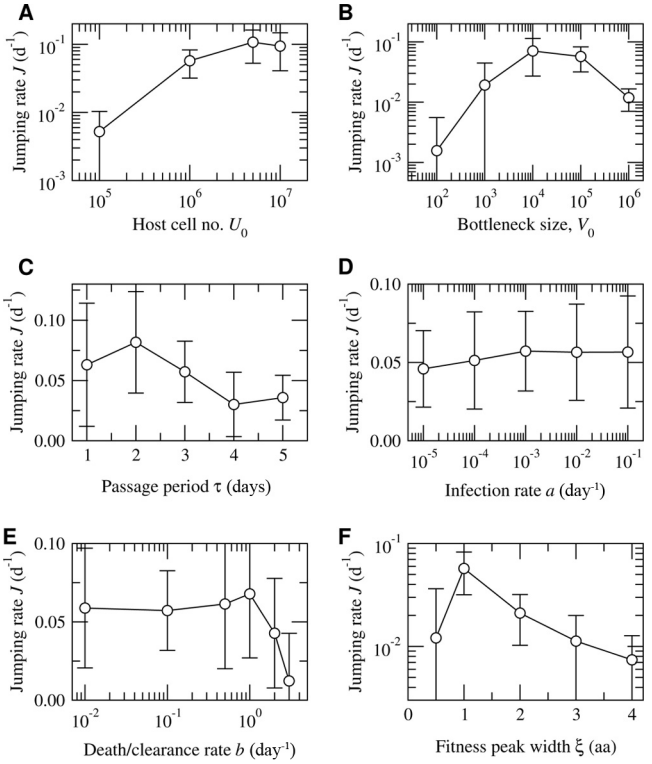

FIG 6 Dependence of jumping rate on other parameters. (A) Number of host cells U_0. (B) Bottleneck size V_0, or the number of virions selected at each passage round. (C) Passage period τ, or the time duration of each passage round. (D) Infection rate a. (E) Rate b of infected cell death and virus clearance. (F) Fitness landscape peak width parameter ξ of the MF sequence. Default values of parameters other than those varied in each case were the same as in Fig. 5, with $d = 2$ aa, $\mu = 1.0 \times 10^{-5}$, and $\xi = 1$ aa.

in infection rate a (Fig. 6D). The dependence on death/clearance rate b was similar (Fig. 6E), except when b increased beyond 1 day^{-1}, where cell death and clearance began to annihilate the growth phase. Finally, we varied the characteristic width parameter ξ of the fitness peak (Fig. 6F), where increasing ξ corresponded to the fitness peak at the MF sequence becoming broader, which sharply reduced the jumping rate: the probability for adaptation was reduced when the associated fitness gain decreased. For a ξ of <1 aa, fitness values of the WT and its neighbors are too small to support adaptation. Overall, the global picture shown in Fig. 6 of the dependence of the jumping rate J on different variables other than distance suggests that the efficiency of adaptation is much more sensitive to the characteristics of bottleneck events than host cell dynamics.

Effects of immune response. For serial passages in animal hosts (43, 44), the action of innate and adaptive host responses may affect the dynamics of adaptation. To examine such effects, we considered an extended model in which host cells infected with a genotype stimulate an immune cell (cytotoxic T lymphocyte) specific to the genotype, which can kill infected cells with the same genotype (equation 3 replaced by equations 6 to 8). A similar implementation is possible with the inclusion of antibodies interacting with viruses rather than infected cells. We found the adaptation dynamics of the extended model (Fig. 7) to be qualitatively similar to Fig. 3, which suggests that the basic model defined by

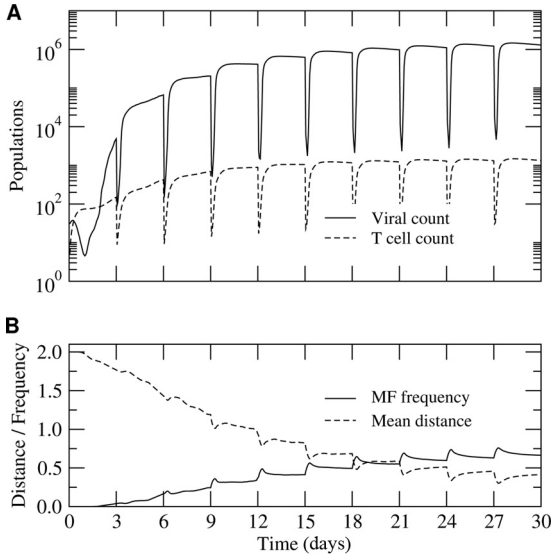

FIG 7 Adaptation dynamics of the extended model given by equations 1, 2, and 4 and 6 to 8 explicitly including host immune response. Parameter values were the same as in Fig. 3 except for $b = c = 0.1$ day^{-1} and $s = 0.01$ day^{-1}.

equations 1 to 4 may be regarded as an effective model indirectly reflecting the selection pressure of immune response.

Influenza A virus H5N1 adaptation. To gain insight into how the generic behavior of adaptation statistics examined as described above for random sequences manifest themselves in specific virus evolution, we have modeled the avian influenza virus H5N1 evolutionary adaptation investigated in recent experimental studies (6, 27). We obtained fitness landscapes specific to influenza virus HA sequences by adopting the direct coupling analysis (32, 33) for fitness inference. The inferred fitness parameters allow one to score fitness values of arbitrary genotypes that might be encountered during evolutionary adaptation. We focused on the segments of H5 HA protein identified as allowing the H5N1 viruses to acquire the capacity to infect via respiratory routes: His103, Thr156, Gln222, and Gly224 (H5 numbering; we focused on the HA protein only, excluding Glu627 of the PB2 protein reported in the experiment) (6). Upon serial passaging in ferrets with preintroduced mutations Q222L and G224S, which had previously been shown to enhance the binding affinity of H5 HA to human influenza A virus receptors (45), two additional mutations were found among strains exhibiting the capability of spreading via airborne routes: H103Y and T156A.

We took the four key residues at positions 103, 156, 222, and 224 to form the model protein segment of $L_a = 4$ aa, for which the H5 WT sequence is HTQG. We used an HA protein MSA of H5N1 sequences to infer the fitness landscape of the model segment. Figure 8A shows the first 20 genotypes ranked by their fitness values within all possible sequences. The sequence HAQG has a fitness comparable (slightly larger) to that of the WT, reflecting the prevalence of Ala156 within the sequence set. Most of the high-fitness mutants involve a single substitution at position 156. This landscape corresponds to the natural environment to which H5 proteins are already well adapted: avian hosts. To obtain the analogous fitness landscape of HA in mammalian hosts, we used an MSA of H3 sequences. Figure 8B shows the highest-fitness genotypes of the corresponding positions (110, 160, 226, and 228 in

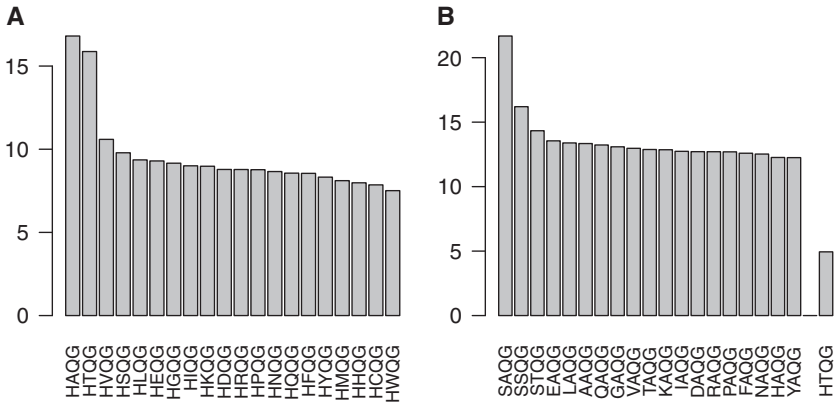

FIG 8 Fitness landscapes of an HA protein 4-aa segment inferred from MSA. (A) H5 landscape derived from H5N1 sequences. (B) H3 landscape from H3 sequences. The height of the bars represents the fitness values in units of $r_s = 1.0\ \text{day}^{-1}$. The abscissa shows the list of genotypes sorted with decreasing order in fitness. In panel B, the rightmost bar shows the H5 WT, which is 251st in rank.

H3 numbering), in addition to the fitness of H5 WT, HTQG, which is less than 25% of the fitness value of the MF sequence, SAQG. The last two positions on the segment are still dominated by QG as in H5: the enhanced fitness of H5 Q222L/G224S mutants in H3 environments, which a recent structural study linked to the *cis-trans* change of receptor analogs (46), is not reflected strongly in the natural H3 populations when coupled mutations are taken into account. We therefore modeled the adaptation of the first two HA positions by taking the sequence HTQG as the WT and SAQG as the MF sequence and simulating the adaptation of the WT under the H3 fitness landscape in Fig. 8B. In these serial-passage simulations, the WT nucleotide sequence at time zero was fixed (CATACACAAGGA) and the fitness values of the appearing mutants were obtained from the computationally inferred landscape parameters instead of being randomly assigned from a distribution.

Typical viral count evolutionary patterns were similar to Fig. 3, while the outcomes of the evolutionary trajectories in the sequence space varied greatly between different realizations, suggesting that the "ruggedness" of the empirically derived fitness landscape with multiple peaks and valleys significantly affected the adaptation process. The majority of the trajectories did not reach the MF sequence within $\sim 10^3$ days ($\sim$300 cycles), getting trapped in some of the genotypes listed in Fig. 8B. One trajectory that reached the MF sequence is illustrated in Fig. 9: the simulation starts with a monoclonal WT (HTQG) population at time zero. The T156A mutation already appears at 3 days, reflecting the large jump in fitness (Fig. 8B) as well as the simplicity of the nucleotide substitution required (A $\rightarrow$ G). The Ala156 is already dominant after 9 days. After a fairly long period of explorations, a new mutant (RAQG) grows in frequency ($t = 99$ days), eventually taking over at around 150 days. It is closely followed by LAQG, which replaces RAQG at 162 days, and finally by SAQG, the MF sequence, at 174 days. Afterwards, the MF frequency grows to around 0.99 and remains dominant, with the subpopulation of small minority mutants continually drifting around. The time-dependent progression of the most frequent genotypes (master sequences; HTQG $\rightarrow$ HAQG $\rightarrow$ RAQG $\rightarrow$ LAQG $\rightarrow$ SAQG) covers the fitness landscape (Fig. 8B) with a sequential increase in rank (251 $\rightarrow$ 18 $\rightarrow$ 14 $\rightarrow$ 5 $\rightarrow$ 1) and fitness (4.9 $\rightarrow$ 12.3 $\rightarrow$ 12.7 $\rightarrow$ 13.4 $\rightarrow$ 21.7), while reflecting the accessibility of the required mutations (Fig. 10). We calculated the mean frequencies of dominant genotypes (HAQG, QAQG, LAQG, and SAQG) as functions of time by averaging over many trajectories (Fig. 11). The mean frequency of the MF sequence SAQG of $\sim$10% with the timescale of $\sim$300 days reflects the probability of jumping rather than the actual composition in individual trajectories: if it occurs, the genotype eventually dominates the population (Fig. 9). In summary, our results show that the basic stochastic model of adaptation can be combined with sequence-based empirical data to yield detailed and experimentally relevant predictions of virus adaptation dynamics.

DISCUSSION

Our stochastic model of viral evolutionary dynamics during serial passages allows for quantitative characterizations of the probability of observing systematic drifts in population structure and adaptations to a new host environment. The discrete stochastic representation of the viral evolutionary dynamics is essential to realistically capture adaptations in finite-size populations dominated by extinctions of minor variants during bottleneck events, which cannot be described using viral copy numbers as continuous variables.

We may view the serial passage as an idealized experimental setup with well-controlled variables deemed important for evolutionary dynamics. The ability to model passages quantitatively, therefore, would provide insights into analogous events occurring in natural settings, including species jumps and outbreaks (5). By calibrating the model parameters using certain serial passage experimental data, one could apply the model to interpretations of experiments as well as generation of predictions and testable hypotheses. The host cell number U_0 would depend on the plating density for cell cultures, typically U_0 of $\sim 10^6$ or less, whereas for passages using live host animals it would be larger and less clear-cut. The bottleneck size V_0, representing the number of virions sampled after each passage round for initiating the next round, could be readily estimated based on the sampling protocol of the experiment. Appropriate values for the infection and death rates, a and b, could be determined by fitting growth curves as shown in Fig. 1, whereas the mutation rate μ is known empirically for many viral pathogens (39, 40, 47).

In addition to these physical parameters, the fitness landscape

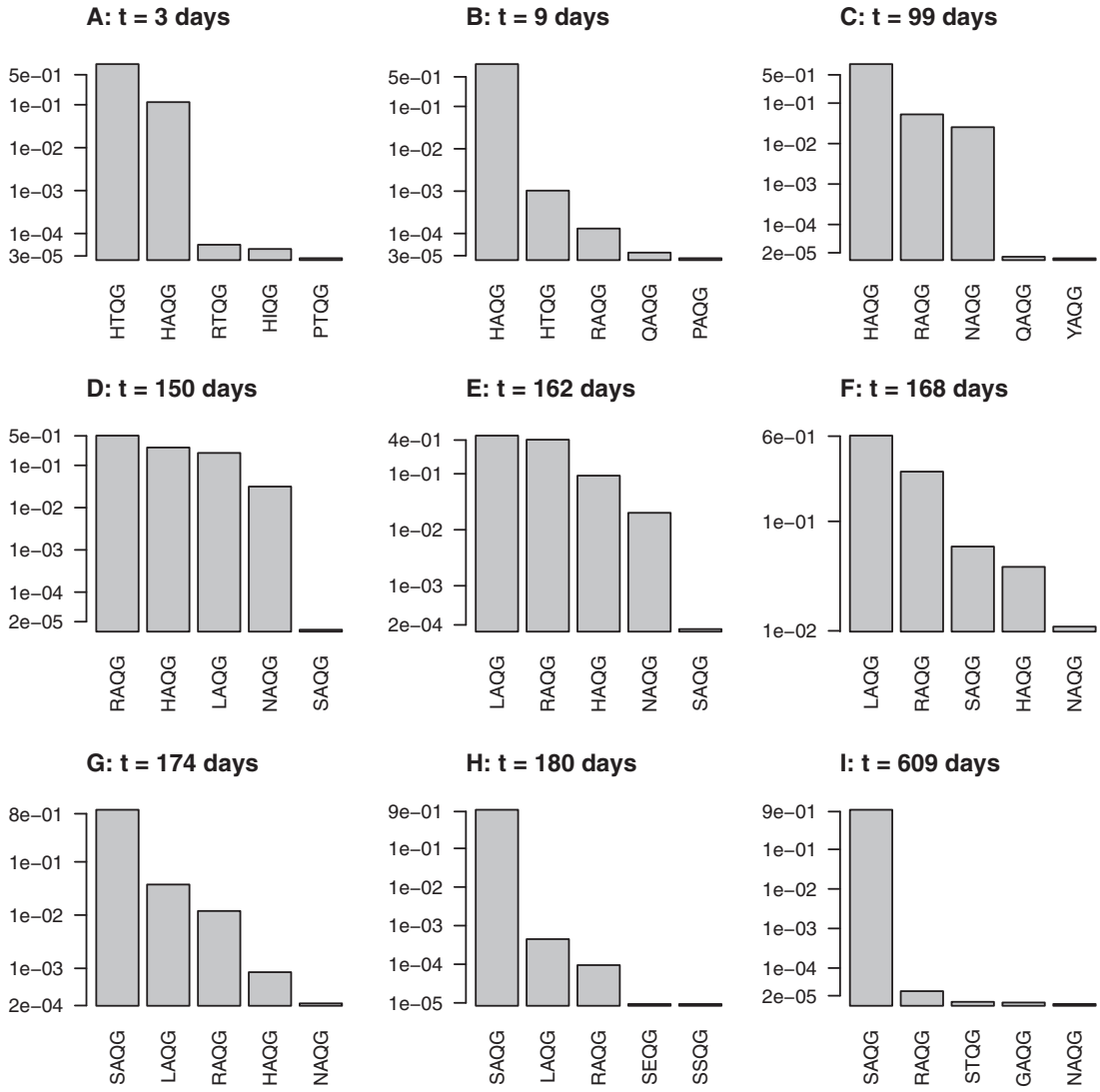

FIG 9 Time evolution of quasispecies composition during the H5-to-H3 adaptation of the model influenza virus segment, simulated under the empirically inferred fitness landscape of Fig. 8B. Panels A to I show the sequential snapshots at each time instance (immediately after a passage cycle), listing five genotypes with the highest frequencies (shown in logarithmic scale). The parameter values were as follows: $V_0 = 1 \times 10^4$, $U_0 = 1 \times 10^6$, $a = 1.0 \times 10^{-3}$ day^{-1}, $b = 1.0$ day^{-1}, $\mu = 1 \times 10^{-5}$, and $\tau = 3$ days.

and sequence parameters are the key inputs to model specification. In this work, we first adopted a random sequence model where the genomic segment under consideration ($L_a = 1$ to 3 aa; $L = 3$ to 9 nt) can exhibit fitness values assumed to be random variables around a mean fitness, which decays exponentially with distance from an MF sequence (see Materials and Methods). This choice reflects both the typical distribution of mutant fitness over deleterious, neutral, and beneficial ranges and the presence of one or multiple peaks in the fitness landscape. This assumption has also been previously validated using published experimental characterizations of HIV-1 protease fitness (28, 48).

The conclusions drawn from this part of our study pertain to general features one expects to see in observations of adaptation processes common to many different viral pathogens: if we perform passages in host environments to which the viruses have already been well adapted, no significant new drift in population structure occurs (Fig. 2). In reality, however, perfect adaptation would be rare and one almost always expects drifts, e.g., toward a more fit genotype 2 aa away from the starting WT (Fig. 3). The speed of this adaptation process is characterized by the time (or the number of passage rounds) required for the population to first discover the MF sequence and then become dominated by it. The jumping rate is the mean inverse time of this discovery/growth process.

The multimodal distribution of jumping rates shown in Fig. 4 reveals an important signature of the topology of fitness landscapes (49) constrained by the genetic code: if we visualize the network of amino acids, where edges denote single substitutions (see Fig. S5 in reference 28 and Fig. 10), an evolutionary path would frequently get stuck in genotypes with amino acid sequences with significantly higher fitness than the WT but without a direct access to the MF. We thus expect that there is always a nonzero probability for a certain evolutionary adaptation pathway to never reach a high-fitness "target" genotype even when it is

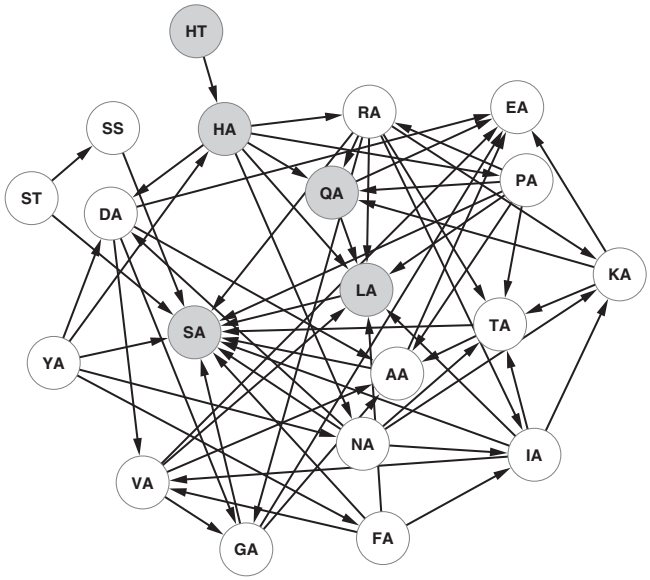

FIG 10 Influenza virus genotype network corresponding to the fitness landscape in Fig. 8B. Nodes represent the genotypes (first two amino acids) and edges connect pairs for which there exists at least one single-nucleotide mutation separating the two nodes. The arrowheads on the edges reflect the direction of fitness increases (Fig. 8B). The shaded nodes are the dominant genotypes shown in Fig. 11. The length of a given path does not necessarily equal the number of nucleotide mutations: QA → LA → SA requires multiple mutations within the Leu-coding codons.

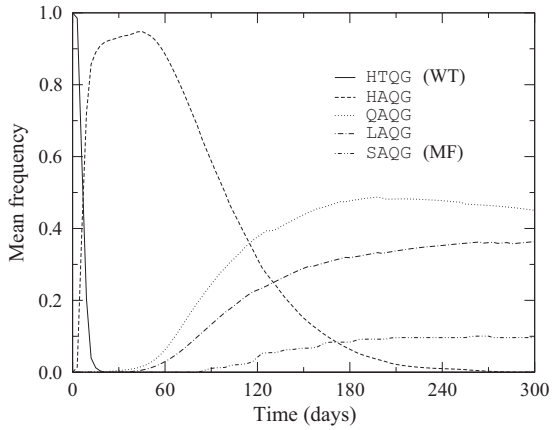

FIG 11 Mean frequencies of the top five dominant genotypes as functions of time from the influenza virus adaptation simulations. Up to 10^3 realizations under the condition of Fig. 9 were averaged. Other genotypes not shown have mean frequencies less than 1%.

close by in sequence space. This feature can be interpreted as a consequence of the "ruggedness" of the fitness landscape, leading to frequent trapping of evolutionary dynamics of adapting organisms (50).

The jumping rate J depends sensitively on the distance from WT to MF sequences (Fig. 5); for μ of $\sim 10^{-5}$, L_a values from 1 to 3 aa yield J values decreasing from 10^{-1} day^{-1} to 10^{-2} day^{-1}, with the latter corresponding to jumping times of ~100 days, which is beyond the upper limit of typical passage durations. Recent experimental demonstrations of species jump-capable mutations induced by serial passages are consistent with this observation: Sheahan et al. used serial passages in cell culture (12) to probe the adaptations presumed to have been responsible for the SARS-CoV outbreak in 2004, with the species jump from bats to humans via palm civets as host species. They engineered a SARS-CoV epidemic strain in which the S protein receptor-binding domain was replaced by its counterpart in the civet strain (the WT) and passaged it in human airway epithelial cells. Only after they manually introduced to the WT a key mutation previously known to significantly affect receptor binding (51) did serial passages up to 22 days produce strains with significantly enhanced growth and replication in human cells. Sequencing of these adapted strains revealed two additional mutations in the receptor-binding domain that arose during passages (12). This observation agrees with our conclusion regarding the likelihood of species jump-capable mutations spontaneously arising during serial passages: with physically reasonable parameter sets, more fit genotypes present in the sequence space requiring mutations up to a d of 2 aa can easily be discovered within time periods of ~20 days (jumping rate $J = 5 \times 10^{-2}$ day^{-1} in Fig. 5), while those further away will be much less likely to appear.

To examine more specific instances of such adaptation processes with direct relevance to experimental investigations, we specialized the model to the respiratory infection adaptation of H5N1 under serial passages and simulated the recent gain-of-function experiments by Herfst et al. (6) using our stochastic simulation algorithm. In this experiment, passaging WT H5N1 in ferrets did not yield adaptation, but after three mutations previously known to play important roles in influenza virus infection were introduced, 10 passage rounds produced strains capable of respiratory infection between ferrets. Analyses of sequences revealed two key mutations that arose during passages. The large number of sequences currently available for influenza allowed us to determine fitness landscapes using a computational fitness inference procedure. By choosing different sequence sets for this inference (H5 and H3 in Fig. 8A and B, respectively), we obtained landscapes specific to different host environments to which a viral population adapts. The four amino acid sites within HA chosen for investigation (His103, Thr156, Gln222, and Gly224) are those previously shown to exhibit adaptive mutations enabling respiratory-route infection in experiments (6). In the experiments, with two mutations (Q222L/G224S) preintroduced, serial passaging produced adapted strains with two additional mutations (H103Y/T156A). The first two mutations did not feature prominently in the H5 or H3 fitness landscapes (Fig. 8), reflecting the engineered nature of these mutations that compensate for the inefficiency of H5 binding to mammalian receptors (46). Our choice of taking the WT and MF genotypes in our simulation as HTQG and SAQG, respectively, each the (near) highest fitness sequences in H5 and H3 landscapes, respectively, mimics the situation where an H5 strain undergoes adaptation with neither the compensatory mutations Q222L/G224S nor the help of collective changes to other sites.

The evolutionary path described in Fig. 9 illustrates an example of how the adaptation process could unfold in the sequence space under the selective forces of realistic fitness landscapes. The series of changes to the highest-frequency genotypes within the quasispecies (HTQG → HAQG → RAQG → LAQG → SAQG) as well as their timings reflect both the features of the local landscape visited and the mutational constraints imposed by the genetic

code (Fig. 10). The first change (T156A) occurred almost immediately ($t = 3$ days), driven both by a large fitness increase (4.9 → 12.3) and by the proximity of Thr- and Ala-coding codons. The remaining part of the trajectory and the time evolution of mean frequencies of dominant genotypes (Fig. 11) show the typical dynamical patterns of adaptive evolution we found in these simulations: fairly long periods of drift in which the master sequence remains unchanged, while the group of minority genotypes keeps evolving (the quasispecies "explores" the immediate neighborhood of an established peak), the gradual growth in frequency of some minor variants, one of which takes over as the new master sequence, and the eventual discovery of the global MF. The last event occurred only in 10% of the simulation runs within ~300 days (Fig. 11), but if it did occur, the MF invariably became dominant. The series of amino acid changes at position 103 (H → R → L → S) and its timing in Fig. 9 are the results of the compromise between fitness gains (Fig. 8B) and the nucleotide substitution required: e.g., CAT → CGT → CTT → TCT, the last step requiring two substitutions, explaining the rare occurrence of SAQG as the master sequence within the timescale of ~100 days (Fig. 11).

The experimental MF genotypes for positions 103 and 156 are Tyr and Ala, respectively. The genotype YAQG had a relatively low fitness in our inferred landscape (ranked 19 in Fig. 8B), and its frequency remained low in the simulations. This discrepancy is primarily due to the inadequacy of the sequence data underlying the inferred landscape, underrepresenting rare compensatory mutations. In addition, it reflects both the constraints imposed by the fitness gradient (Fig. 8B) and the genetic code: the genotype YA is only accessible from the WT (HT) via HA but involves a fitness decrease (Fig. 10).

We may infer insights into a better understanding of evolutionary dynamics of viral pathogens in natural settings from the study of serial passages. Systematic investigation of the effects of factors influencing the relative ease of species jump, such as those shown in Fig. 6, can play important roles in such interpretations. For instance, the host cell number and bottleneck size roughly correspond to the size of susceptible host populations and the degree and severity of selection present during the natural spread of a viral infection, whereas the infection and death rates characterize the population dynamics within this host environment. In this viewpoint, the longitudinal patterns of population size and quasispecies structure shown in Fig. 3 can be regarded as idealizations of what typically happens in natural infectious cycles: a viral strain ventures into a new host environment, completing multiple cycles of rapid growth followed by stagnation due to host depletion and reemergence in a fresh host population, but with factors characterizing each round highly variable and unpredictable rather than uniform as in passages. Our study of evolutionary dynamics under passages suggests that adaptation mostly occurs in the early growth phase (Fig. 3), an adapting viral population may frequently get stuck without a direct route to a neighboring highly fit strain (Fig. 4), and the speed of adaptation is most sensitive to the distance (Fig. 5) and target/bottleneck sizes (Fig. 6). The specific instance of adaptation simulated under the empirically derived influenza A virus fitness landscape (Fig. 8) explicitly demonstrates that in reality, a quasispecies adaptation involves the stochastic evolution of both the master sequence and its clouds (Fig. 9 and 11), instead of a simple jump from the WT genotype to a global MF sequence.

We restricted our study to the evolutionary dynamics of small genomic segments on the order of a few amino acids. With the empirical inference of fitness landscapes from MSAs, however, it is conceivable to expand the scope to larger domains, possibly containing multiple proteins, with the direct coupling analysis taking into account correlated evolution of multiple amino acid sites. In addition, in modeling influenza virus evolution with large genomic segments, an extension taking into account reassortment events that play important roles in influenza virus adaptation (7, 26, 52) could provide more realistic descriptions.

ACKNOWLEDGMENTS

This work was supported by a competitive In-House Laboratory Independent Research Award by the U.S. Army Assistant Secretary of the Army for Acquisition, Logistics, and Technology, and by the U.S. Army Medical Research and Materiel Command (Ft. Detrick, Maryland).

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the U.S. Department of Defense.

REFERENCES

1. **Domingo E, Holland JJ.** 1997. RNA virus mutations and fitness for survival. Annu. Rev. Microbiol. **51:**151–178. http://dx.doi.org/10.1146/annurev.micro.51.1.151.
2. **Holmes EC.** 2009. The evolution and emergence of RNA viruses. Oxford University Press, Oxford, United Kingdom.
3. **Peng X, Chan EY, Li Y, Diamond DL, Korth MJ, Katze MG.** 2009. Virus-host interactions: from systems biology to translational research. Curr. Opin. Microbiol. **12:**432–438. http://dx.doi.org/10.1016/j.mib.2009.06.003.
4. **Law GL, Korth MJ, Benecke AG, Katze MG.** 2013. Systems virology: host-directed approaches to viral pathogenesis and drug targeting. Nat. Rev. Microbiol. **11:**455–466. http://dx.doi.org/10.1038/nrmicro3036.
5. **Pepin KM, Lass S, Pulliam JRC, Read AF, Lloyd-Smith JO.** 2010. Identifying genetic markers of adaptation for surveillance of viral host jumps. Nat. Rev. Microbiol. **8:**802–813. http://dx.doi.org/10.1038/nrmicro2440.
6. **Herfst S, Schrauwen EJ, Linster M, Chutinimitkul S, de Wit E, Munster VJ, Sorrell EM, Bestebroer TM, Burke DF, Smith DJ, Rimmelzwaan GF, Osterhaus AD, Fouchier RA.** 2012. Airborne transmission of influenza A/H5N1 virus between ferrets. Science **336:**1534–1541. http://dx.doi.org/10.1126/science.1213362.
7. **Holmes EC, Ghedin E, Miller N, Taylor J, Bao Y, St George K, Grenfell BT, Salzberg SL, Fraser CM, Lipman DJ, Taubenberger JK.** 2005. Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. PLoS Biol. **3:**e300. http://dx.doi.org/10.1371/journal.pbio.0030300.
8. **Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, Fouchier RA.** 2004. Mapping the antigenic and genetic evolution of influenza virus. Science **305:**371–376. http://dx.doi.org/10.1126/science.1097211.
9. **Josset L, Belser JA, Pantin-Jackwood MJ, Chang JH, Chang ST, Belisle SE, Tumpey TM, Katze MG.** 2012. Implication of inflammatory macrophages, nuclear receptors, and interferon regulatory factors in increased virulence of pandemic 2009 H1N1 influenza A virus after host adaptation. J. Virol. **86:**7192–7206. http://dx.doi.org/10.1128/JVI.00563-12.
10. **Bolles M, Donaldson E, Baric R.** 2011. SARS-CoV and emergent coronaviruses: viral determinants of interspecies transmission. Curr. Opin. Virol. **1:**624–634. http://dx.doi.org/10.1016/j.coviro.2011.10.012.
11. **Graham RL, Baric RS.** 2010. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. J. Virol. **84:**3134–3146. http://dx.doi.org/10.1128/JVI.01394-09.
12. **Sheahan T, Rockx B, Donaldson E, Sims A, Pickles R, Corti D, Baric R.** 2008. Mechanisms of zoonotic severe acute respiratory syndrome coronavirus host range expansion in human airway epithelium. J. Virol. **82:**2274–2285. http://dx.doi.org/10.1128/JVI.02041-07.
13. **Wu K, Peng G, Wilken M, Geraghty RJ, Li F.** 2012. Mechanisms of host receptor adaptation by severe acute respiratory syndrome coronavirus. J. Biol. Chem. **287:**8904–8911. http://dx.doi.org/10.1074/jbc.M111.325803.
14. **Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF,**

Cummins LB, Arthur LO, Peeters M, Shaw GM, Sharp PM, Hahn BH. 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature **397:**436–441. http://dx.doi.org/10.1038/17130.
15. **Ebert D.** 1998. Experimental evolution of parasites. Science **282:**1432–1435. http://dx.doi.org/10.1126/science.282.5393.1432.
16. **Webster DP, Farrar J, Rowland-Jones S.** 2009. Progress towards a dengue vaccine. Lancet Infect. Dis. **9:**678–687. http://dx.doi.org/10.1016/S1473-3099(09)70254-3.
17. **Mardis ER.** 2008. Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. **9:**387–402. http://dx.doi.org/10.1146/annurev.genom.9.081307.164359.
18. **Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD.** 1996. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science **271:**1582–1586. http://dx.doi.org/10.1126/science.271.5255.1582.
19. **Nowak MA, Bangham CR.** 1996. Population dynamics of immune responses to persistent viruses. Science **272:**74–79. http://dx.doi.org/10.1126/science.272.5258.74.
20. **Baccam P, Beauchemin C, Macken CA, Hayden FG, Perelson AS.** 2006. Kinetics of influenza A virus infection in humans. J. Virol. **80:**7590–7599. http://dx.doi.org/10.1128/JVI.01623-05.
21. **Lee HY, Perelson AS, Park SC, Leitner T.** 2008. Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression. PLoS Comput. Biol. **4:**e1000240. http://dx.doi.org/10.1371/journal.pcbi.1000240.
22. **Lim KI, Lang T, Lam V, Yin J.** 2006. Model-based design of growth-attenuated viruses. PLoS Comput. Biol. **2:**e116. http://dx.doi.org/10.1371/journal.pcbi.0020116.
23. **Lim KI, Yin J.** 2009. Computational fitness landscape for all gene-order permutations of an RNA virus. PLoS Comput. Biol. **5:**e1000283. http://dx.doi.org/10.1371/journal.pcbi.1000283.
24. **Lázaro E, Escarmis C, Domingo E, Manrubia SC.** 2002. Modeling viral genome fitness evolution associated with serial bottleneck events: evidence of stationary states of fitness. J. Virol. **76:**8675–8681. http://dx.doi.org/10.1128/JVI.76.17.8675-8681.2002.
25. **Lázaro E, Escarmis C, Perez-Mercader J, Manrubia SC, Domingo E.** 2003. Resistance of virus to extinction on bottleneck passages: study of a decaying and fluctuating pattern of fitness loss. Proc. Natl. Acad. Sci. U. S. A. **100:**10830–10835. http://dx.doi.org/10.1073/pnas.1332668100.
26. **Russell CA, Fonville JM, Brown AE, Burke DF, Smith DL, James SL, Herfst S, van Boheemen S, Linster M, Schrauwen EJ, Katzelnick L, Mosterín A, Kuiken T, Maher E, Neumann G, Osterhaus AD, Kawaoka Y, Fouchier RA, Smith DJ.** 2012. The potential for respiratory droplet-transmissible A/H5N1 influenza virus to evolve in a mammalian host. Science **336:**1541–1547. http://dx.doi.org/10.1126/science.1222526.
27. **Imai M, Watanabe T, Hatta M, Das SC, Ozawa M, Shinya K, Zhong G, Hanson A, Katsura H, Watanabe S, Li C, Kawakami E, Yamada S, Kiso M, Suzuki Y, Maher EA, Neumann G, Kawaoka Y.** 2012. Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. Nature **486:**420–428.
28. **Woo HJ, Reifman J.** 2012. A quantitative quasispecies theory-based model of virus escape mutation under immune selection. Proc. Natl. Acad. Sci. U. S. A. **109:**12980–12985. http://dx.doi.org/10.1073/pnas.1117201109.
29. **Eigen M.** 1971. Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften **58:**465–523. http://dx.doi.org/10.1007/BF00623322.
30. **Swetina J, Schuster P.** 1982. Self-replication with errors: a model for polynucleotide replication. Biophys. Chem. **16:**329–345. http://dx.doi.org/10.1016/0301-4622(82)87037-3.
31. **Gillespie DT.** 1977. Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. **81:**2340–2361. http://dx.doi.org/10.1021/j100540a008.
32. **Weigt M, White RA, Szurmant H, Hoch JA, Hwa T.** 2009. Identification of direct residue contacts in protein-protein interaction by message passing. Proc. Natl. Acad. Sci. U. S. A. **106:**67–72. http://dx.doi.org/10.1073/pnas.0805923106.
33. **Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M.** 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. U. S. A. **108:**E1293–E1301. http://dx.doi.org/10.1073/pnas.1111471108.
34. **Ferguson AL, Mann JK, Omarjee S, Ndung'u T, Walker BD, Chakraborty AK.** 2013. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. Immunity **38:**606–617. http://dx.doi.org/10.1016/j.immuni.2012.11.022.
35. **Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D.** 2008. The influenza virus resource at the National Center for Biotechnology Information. J. Virol. **82:**596–601. http://dx.doi.org/10.1128/JVI.02005-07.
36. **Jones DT, Taylor WR, Thornton JM.** 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8:**275–282.
37. **Pawelek KA, Huynh GT, Quinlivan M, Cullinane A, Rong L, Perelson AS.** 2012. Modeling within-host dynamics of influenza virus infection including immune responses. PLoS Comput. Biol. **8:**e1002588. http://dx.doi.org/10.1371/journal.pcbi.1002588.
38. **Murphy BR, Rennels MB, Douglas RG, Jr, Betts RF, Couch RB, Cate TR, Jr, Chanock RM, Kendal AP, Maassab HF, Suwanagool S, Sotman SB, Cisneros LA, Anthony WC, Nalin DR, Levine MM.** 1980. Evaluation of influenza A/Hong Kong/123/77 (H1N1) *ts*-1A2 and cold-adapted recombinant viruses in seronegative adult volunteers. Infect. Immun. **29:**348–355.
39. **Nobusawa E, Sato K.** 2006. Comparison of the mutation rates of human influenza A and B viruses. J. Virol. **80:**3675–3678. http://dx.doi.org/10.1128/JVI.80.7.3675-3678.2006.
40. **Gago S, Elena SF, Flores R, Sanjuan R.** 2009. Extremely high mutation rate of a hammerhead viroid. Science **323:**1308. http://dx.doi.org/10.1126/science.1169202.
41. **Yang Z.** 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39:**306–314. http://dx.doi.org/10.1007/BF00160154.
42. **Kampmann ML, Fordyce SL, Avila-Arcos MC, Rasmussen M, Willerslev E, Nielsen LP, Gilbert MT.** 2011. A simple method for the parallel deep sequencing of full influenza A genomes. J. Virol. Methods **178:**243–248. http://dx.doi.org/10.1016/j.jviromet.2011.09.001.
43. **Barnard DL.** 2009. Animal models for the study of influenza pathogenesis and therapy. Antiviral Res. **82:**A110–A122. http://dx.doi.org/10.1016/j.antiviral.2008.12.014.
44. **Bouvier NM, Lowen AC.** 2010. Animal models for influenza virus pathogenesis and transmission. Viruses **2:**1530–1563. http://dx.doi.org/10.3390/v20801530.
45. **Chutinimitkul S, van Riel D, Munster VJ, van den Brand JM, Rimmelzwaan GF, Kuiken T, Osterhaus AD, Fouchier RA, de Wit E.** 2010. *In vitro* assessment of attachment pattern and replication efficiency of H5N1 influenza A viruses with altered receptor specificity. J. Virol. **84:**6825–6833. http://dx.doi.org/10.1128/JVI.02737-09.
46. **Zhang W, Shi Y, Lu X, Shu Y, Qi J, Gao GF.** 2013. An airborne transmissible avian influenza H5 hemagglutinin seen at the atomic level. Science **340:**1463–1467. http://dx.doi.org/10.1126/science.1236787.
47. **Mansky LM, Temin HM.** 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J. Virol. **69:**5087–5094.
48. **Fernàndez G, Clotet B, Martinez MA.** 2007. Fitness landscape of human immunodeficiency virus type 1 protease quasispecies. J. Virol. **81:**2485–2496. http://dx.doi.org/10.1128/JVI.01594-06.
49. **Aguirre J, Buldu JM, Stich M, Manrubia SC.** 2011. Topological structure of the space of phenotypes: the case of RNA neutral networks. PLoS One **6:**e26324. http://dx.doi.org/10.1371/journal.pone.0026324.
50. **Kauffman SA.** 1993. The origins of order: self-organization and selection in evolution. Oxford University Press, Oxford, United Kingdom.
51. **Li F, Li W, Farzan M, Harrison SC.** 2005. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science **309:**1864–1868. http://dx.doi.org/10.1126/science.1116480.
52. **Ince WL, Gueye-Mbaye A, Bennink JR, Yewdell JW.** 2013. Reassortment complements spontaneous mutation in influenza A virus NP and M1 genes to accelerate adaptation to a new host. J. Virol. **87:**4330–4338. http://dx.doi.org/10.1128/JVI.02749-12.