Identification of drug resistance mutations in HIV from constraints on natural evolution

Thomas C. Butler,[1,2] John P. Barton,[1,2,3] Mehran Kardar,[1,*] and Arup K. Chakraborty[1,2,3,4,†]

[1]*Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

[2]*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA*

[3]*Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02139, USA*

[4]*Departments of Chemistry and Biological Engineering, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

(Received 7 August 2015; revised manuscript received 4 December 2015; published 19 February 2016)

Human immunodeficiency virus (HIV) evolves with extraordinary rapidity. However, its evolution is constrained by interactions between mutations in its fitness landscape. Here we show that an Ising model describing these interactions, inferred from sequence data obtained prior to the use of antiretroviral drugs, can be used to identify clinically significant sites of resistance mutations. Successful predictions of the resistance sites indicate progress in the development of successful models of real viral evolution at the single residue level and suggest that our approach may be applied to help design new therapies that are less prone to failure even where resistance data are not yet available.

DOI: 10.1103/PhysRevE.93.022412

I. INTRODUCTION

Under selective pressure from suboptimal antiretroviral treatment regimens, HIV has been observed to evolve drug resistance within weeks of treatment initiation [1]. While modern combination therapies have greatly reduced the rate of evolution of drug resistance, resistant strains are found in greater than 14% of newly infected patients in the United States [2,3]. The rapid evolution of resistance is congruent with the overall observation that HIV evolution is remarkably fast, with studies indicating that in the absence of treatment a single patient's HIV infection will explore every possible point mutation many times daily [4–6]. However, empirical studies of viral sequence data indicate that HIV evolution is structured and exhibits reproducible patterns [1,7].

The existence of significant correlations in the evolution of HIV suggests that sequence data can be used to parametrize statistical mechanical models of HIV evolution that predict important features of its evolution, including the evolution of drug resistance. Previous researchers have used a variety of approaches to predict HIV fitness and aspects of its evolution using viral sequence data on its own [7,8] and with additional phenotypic properties such as drug resistance and replicative capacity [9]. Others have addressed the problem of predicting the sites of drug resistance mutations by detecting sites under positive selection during treatment [10], supervised learning [11], and structural modeling of protein-drug interactions [11,12].

Here we use HIV sequence data from 757 unique patients, obtained prior to the widespread clinical use of protease inhibitors, to parametrize a spin representation of the standard Eigen model of quasispecies evolution [13–15] (see Appendix A for details). This data was obtained from the Los Alamos National Laboratory HIV sequence database (www.hiv.lanl.gov). We then use the inferred model to predict sets of sites in HIV protease where joint mutations are unlikely to significantly impair viral fitness. We hypothesize that such sites are more likely to be sites of clinically relevant drug resistance mutations because resistance mutations that severely impair viral replication are unlikely to be selected. The exclusion of sequence data obtained after the clinical use of antiretroviral drugs limits the influence of selection for drug resistance, which may be present even in sequences obtained from drug-naive individuals (for example due to transmitted drug resistance), and focuses instead on intrinsic fitness constraints. Thus, our successful identification of major drug resistance sites (defined in [16]) here suggests that our techniques could be applied to predict HIV evolution in response to new treatment regimens or vaccine candidates.

We note, however, that this identification of resistance sites is "indirect" in the sense that only information about fitness is used. Thus, these predictions are not specific to a particular drug, and would be most useful in cases when resistance information is unknown. In order to make highly accurate, drug-specific predictions of resistance, additional information would be required to narrow down the list of potential drug resistance sites identified based on fitness constraints to the ones that are most relevant for a particular case.

II. FITNESS AND PREVALENCE LANDSCAPES FROM THE EIGENMODEL

We begin by inferring an estimate of the probability distribution of mutations in the viral protease from sequence data. Protease amino acid sequences are first translated into a binary form, with the amino acid at each site i encoded by $s_i \in \{0,1\}$, where 0 (1) denotes a wild type (mutant) amino acid at that site. Full sequences are thus represented as vectors $s = (s_1, s_2, \ldots, s_L)$, with $L = 99$ for protease. We assume that the joint distribution of mutations is adequately captured by the moments $\langle s_i s_j \rangle$ and find the maximum entropy distribution consistent with the observed moments (note that because $s_i^2 = s_i$, $\langle s_i \rangle = \langle s_i^2 \rangle$ all first moments are included) [8,17]. The

*kardar@mit.edu
†arupc@mit.edu

resulting probability distribution takes the form

$$P(s) = Z^{-1}\exp[-E(s)],$$
$$E(s) = \sum_{i<j}^{L} J_{ij}s_i s_j + \sum_{i=1}^{L} h_i s_i, \quad (1)$$

where Z is the partition function. The parameters $\{J_{ij}\}$, $\{h_i\}$ must be chosen such that the distribution $P(s)$ reproduces the observed moments $\langle s_i s_j\rangle$. Here the $\{J_{ij}\}$ can be thought of as capturing direct interactions between sites, disentangled from the network of correlations that include indirect effects mediated through intermediate sites [18–20]. Similar maximum entropy approaches have been fruitfully applied to analyze patterns of neural activity and predict contact residues in protein families [19–22]. The description of the selective cluster expansion algorithm used to infer $E(s)$ is given in [18,23]. Although only the pair correlations are constrained in Eq. (1), the inferred Ising model accurately predicts higher order correlations as well.

The form of the probability distribution gives rise to the notion of a "prevalence landscape" that expresses the relative frequencies of protease sequences. Previous work has shown that the inferred prevalences of sequences from HIV Gag proteins correlate with their replicative capacities, another proxy for fitness [8,24], in line with the intuition that fitter strains should be more prevalent. However, prevalence is affected by many factors other than fitness, including epidemiological dynamics, recombination, and demographic noise, which complicate this association [25–27].

Insight into the relation between fitness and prevalence can be obtained through Eigen's model of evolution [13]. This model assumes an infinite population of viruses and accounts for mutation and selection, but neglects many of the important effects described above. However, these simplifications allow for the relationship between fitness and prevalence to be studied using methods adopted from statistical physics [14,15]. Following Eigen's model, the prevalence can also be written as the outcome of evolutionary dynamics over a large number of generations T, represented as a series of coupled Ising spin systems [14],

$$e^{-E(s^T)} \propto \sum_{\{s^t\}_{t=1}^{T-1}} \exp\left\{\sum_{t=1}^{T-1}[K(2s^t-1)(2s^{t+1}-1) - F(s^t)]\right\},$$
$$F(s) = \sum_{i<j}^{L} J_{ij}^f s_i s_j + \sum_{i=1}^{L} h_i^f s_i, \quad (2)$$

where K is related to the per site per generation mutation rate μ by $K = \frac{1}{2}\ln(\frac{1-\mu}{\mu})$. Here $F(s)$ is minus the log fitness of sequence s and is referred to as the fitness landscape. The superscripts $t \in \{1,2\ldots,T\}$ on the sequence vectors refer to discrete generations. The superscript f on the parameters $\{h_i^f\}$ and $\{J_{ij}^f\}$ indicates that these parameters are taken from the fitness landscape in Eq. (2) [assumed to have the same functional form as Eq. (1)], rather than the prevalence landscape of Eq. (1). The evolutionary dynamics described here applies to evolution within a population of hosts. Equations describing within host evolution would require accounting for differing immune pressure between individuals [15], though protease is not comparatively immunogenic [28]. Given that tens of millions of humans have been infected with HIV over the course of the epidemic and that both the number of infected cells and their rate of turnover are high [29], the limit of large population size and long sampling times that we consider here is not unreasonable.

Ideally, one would like to invert Eq. (2) to solve for $F(s)$ in terms of $E(s)$, because $E(s)$ is inferred directly from data. In principle, this could be achieved by matching the distribution of sequences at the final generation T in the Ising representation of Eigen's model with the prevalence landscape, given by Eq. (1). This is a challenging problem in general; however, approximate results can be obtained by studying a two site system, which can be solved by straightforward transfer matrix methods. While network effects influence the inferred couplings between sites, this simple approximation provides useful intuition. Furthermore, network effects exert a weaker influence on the $\langle s_i\rangle$, as most of their variance is explained by the single site h_i in Eq. (1).

Solving the two site version of Eq. (2) shows that the h^f are difficult to reliably infer, because the mutation coupling K is large enough ($K \simeq -\ln(\mu)$ and $\mu \simeq \mathcal{O}(10^{-4})$, using the microscopic mutation rate [30]) that very small h^f lead to large h in the prevalence landscape (see Appendix B for further details). However, large values of J in the prevalence landscape have a simple interpretation in the fitness landscape as couplings between pairs of sites where mutating both sites leads to only a small change in fitness compared to wild type (Fig. 1). In this case the double mutant could become advantageous with only a small increase in the fitness of one

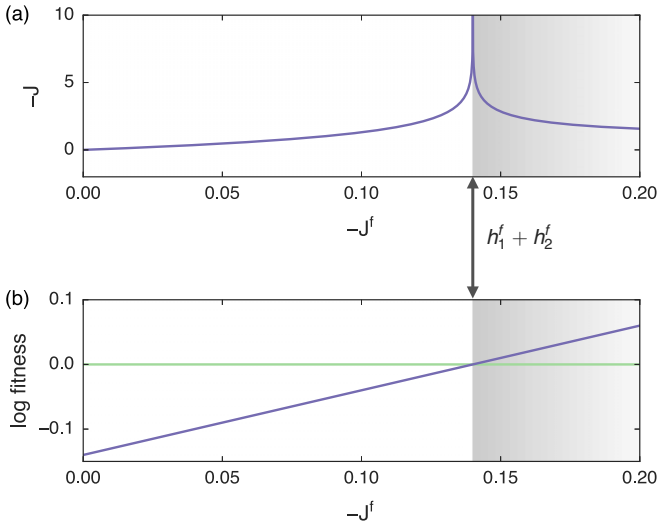

FIG. 1. The coupling $-J$ between a pair of sites increases sharply as the fitness of the double mutant approaches the fitness of the wild type sequence. (a) The peak in $-J$ occurs at the level crossing, where $-J^f = h_1^f + h_2^f$. If $-J^f$ becomes larger than $h_1^f + h_2^f$, so that the double mutant has higher fitness than the wild type sequence (shaded region), the corresponding coupling $-J$ in the prevalence landscape decreases. (b) Log fitness of the wild type strain and the double mutant strain as a function of J^f. The log fitnesses intersect at the point where $-J$ is maximized.

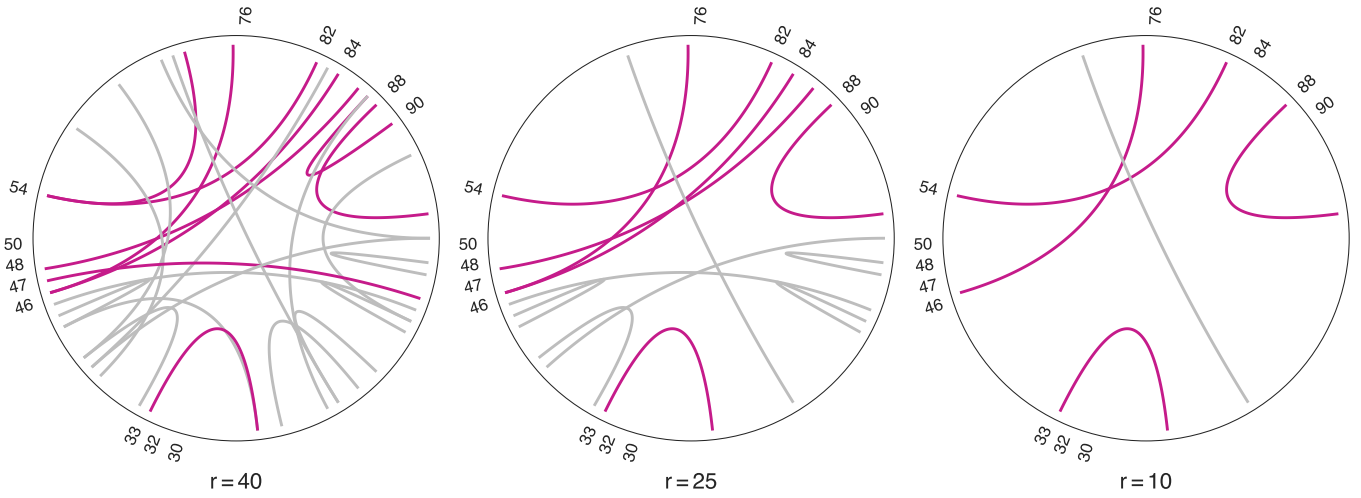

FIG. 2. Stronger couplings are more likely to link sites of major resistance mutations. Here we show the network of interactions between the top r ranked sites, from $r = 40$ (left) to $r = 10$ (right). Only the strongest couplings, those meeting or exceeding the largest coupling for the lowest ranked site, are displayed. Interactions linking at least one major resistance site are darkly shaded; links between nonresistance sites are lightly shaded.

of the mutations, as might occur when drugs are added to the environment, for example. Mathematically, this occurs as $-J^f$ approaches $h_1^f + h_2^f$. We refer to the point in parameter space where the coupling between sites allows the double mutant strain to have equal fitness to the wild type as a level crossing.

III. RESISTANCE MUTATIONS IN PROTEASE

To go from the interpretation of large values of $-J$ in the prevalence landscape as indicators of nearby level crossings to predictions of resistance mutations requires elucidating a relationship between level crossings and resistance mutations. A rigorous argument relating resistance mutations to the fitness landscape would require detailed knowledge of the drug, its binding sites, the structure of the target protein, and other details. However, generically we expect that when the environment in which HIV replicates changes due to the initiation of drug therapy, HIV must mutate in ways that abrogate drug binding, while at the same time preserving protein function. Large couplings $-J$ connect sites that are likely to be able to comutate with limited costs to fitness, even if the associated individual mutations are costly. Such sets of sites are therefore more likely to be associated with resistance. Here our assumption is that resistance cannot be achieved through selectively neutral mutations at single sites, in which case drug treatment would likely be ineffective. Indeed, this appears to be the typical case for HIV protease, where resistance mutations are usually deleterious [31].

To predict the sites of resistance mutations based on the above considerations, we consider the strongest couplings $-J_{ij}$ associated with each site i. Using the largest coupling values we then assign each site a rank $r \in \{1, \ldots, 99\}$ from strongest to weakest. We predict that the sites with the strongest interactions (i.e., the highest ranked sites) are most likely to be associated with drug resistance. Focusing on the highest ranked sites and the strong couplings between them can be seen as a process of pruning weaker interactions from the network. Three pruned versions of the network of mutational interactions in HIV protease are shown in Fig. 2. However, note that without any drug-specific information, we cannot specify which sites in a strongly coupled pair should be associated with resistance.

This model can be cast in the form of a classification rule by predicting sites ranked at or above some threshold rank r to be sites of drug resistance mutations and sites of lower rank to be unassociated with resistance. To test the model's performance, we take the set of resistance sites to be those classified as sites of major resistance mutations by the Stanford HIV drug resistance database (sites 30, 32, 33, 46, 47, 48, 50, 54, 76, 82, 84, 88, and 90) [16]. As higher ranked sites are selected, the proportion of sites that are associated with resistance should increase. This can be measured using positive prediction value (PPV) and negative prediction value (NPV), defined as

$$P(\text{true} = \text{resistance}|\text{predicted} = \text{resistance}),$$

$$P(\text{true} = \text{non-resistance}|\text{predicted} = \text{nonresistance}).$$

These are shown in Fig. 3 compared to benchmarks for a random classifier and demonstrate that the performance of the classification rule is substantially better than chance for higher ranked sites. Examination of the true positive rate (TPR) and false positive rate (FPR),

$$P(\text{predicted} = \text{resistance}|\text{true} = \text{resistance}),$$

$$P(\text{predicted} = \text{resistance}|\text{true} = \text{nonresistance}),$$

shown in Fig. 3, confirm that TPR $>$ FPR, indicating performance better than chance. We also note that the fraction of the strongest interactions which link at least one major drug resistance site is extremely high, as can be seen in Fig. 2 (further details in Appendix D).

To examine these results using classical statistical significance testing, we used the hypergeometric distribution to compute p values for the null model of randomly selecting the number of sites at or above each rank threshold and obtaining at least as many resistance mutations as found using the ranking classifier (see Appendix C). The predictions

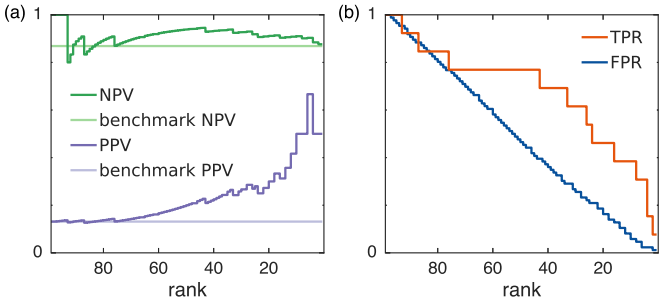

FIG. 3. Top-ranked sites, based on the maximum strength of their couplings, are far more likely to be sites of major drug resistance mutations than would be expected by chance. (a) Negative prediction value (NPV, upper curves) and positive prediction value (PPV, lower curves) for the classifier compared to the benchmark of random guessing as functions of rank. Collections of the highest ranked sites are clearly associated with improved PPV. (b) False positive rate (FPR) and true positive rate (TPR) as functions of rank. TPR $>$ FPR indicates performance better than chance.

have p values < 0.05 for essentially all rank thresholds from $r = 3$–50, which comports with the argument that strongly coupled sites are more likely to be sites of resistance mutations and supports the significance of the predictions of resistance among higher ranked sites. The lack of significance for the highest ranked pair is a consequence of the very small number of sites. Further tests also show a significant difference between the rank of resistance sites versus nonresistance sites (Mann-Whitney $U = 343$, $p = 0.0255$), another way to test the utility of interactions in predicting resistance sites. We also tested related classification rules constructed using direct information [19] and correlation matrices, with no improvement in performance. All methods based on pairwise interactions outperformed methods ranking sites according to their mutability, an intuitive result given that most resistance mutations in protease are deleterious (see Appendix D for details).

IV. DRUG COMBINATIONS AND BIOPHYSICAL INFORMATION

As virological failure occurs in patients undergoing treatment with protease inhibitors, new protease inhibitor drugs are administered [2]. To further assess the validity of our predictions, we used the model to infer pairs of protease inhibitors where multidrug resistance should be unlikely to evolve. We reasoned that resistance should be less likely if (1) a pair of drugs share few resistance mutations in common and if (2) fitness constraints make it difficult for the virus to tolerate mutations conferring resistance for both drugs simultaneously. The first condition can easily be checked by simply counting the number of common resistance mutations for each pair of drugs (see [16]). Information about the second condition can be obtained through the inferred couplings in our model. In the same way that large negative values of J indicate sites that can readily mutate together, positive values of J indicate sites where double mutations are suppressed. Thus, the interactions between the resistance mutations that are common to both drugs should be as positive as possible. We found three combinations (atazanavir-indinavir, atazanavir-fosamprenavir, and darunavir-nelfanavir) that are optimal for both of these criteria in the Pareto sense: Improvement in one criterion necessitates a reduction in the other criterion. Two of these, along with both near-optimal pairs (atazanavir-darunavir and atazanavir-lopinavir), incorporate atazanavir, consistent with clinical knowledge that the resistance profile of atazanavir appears distinct from other protease inhibitors [32].

The network of large interactions also captures important biophysical information. As a first example, the third strongest coupling is between sites 82 and 54. Site 82 is frequently the first resistance mutation site observed after the initiation of protease inhibitor treatment and is usually followed by mutation at site 54 [1]. Some couplings may also be associated with stabilizing mutations, which compensate for loss of fitness due to a destabilizing mutation. A recent biophysical study examined the melting temperatures of HIV protease with a major resistance mutation at site 84 [31]. The study showed that, on its own, the major resistance mutation reduced the stability of HIV protease considerably. When the mutation at site 84 is accompanied by one of a set of three known accessory mutations at sites 10, 63, and 71, stability is restored, or even enhanced. Couplings between sites 10 and 84, and sites 63 and 84, are strong, in the top 7% of all couplings (though weaker than the couplings shown in Fig. 2, which are within the top 1%). The coupling between sites 71 and 84 is slightly weaker, but still in the top 13% of all couplings. This suggests that links between destabilizing mutations and those that improve protein stability may be captured by the network of interactions inferred from sequence data.

V. CONCLUSIONS

Our results show that from sequence information alone, much of the evolutionary response of HIV protease to inhibitors can be reproduced. While in the case of protease inhibitors, the answer was known, the successful retrodictions indicate that our understanding of HIV evolution is becoming predictive at the level of individual residue sites. We anticipate that the methods developed above will contribute to the development of predictive theories of viral evolution and to the development of new treatments, such as integrase inhibitors [33], where resistance is not nearly as well characterized as in protease.

ACKNOWLEDGMENTS

We thank Daniel Kuritzkes, Martin Hirsch, Andrew Ferguson, Dariusz Murakowski, and Hanrong Chen for helpful discussions. This research was funded by the Ragon Institute of MGH, MIT, and Harvard and National Science Foundation under Grants No. PHY11-25915 and No. DMR-12-06323.

T.C.B. and J.P.B. contributed equally to this work.

APPENDIX A: SEQUENCE DATA AND CORRELATIONS

We downloaded a multiple sequence alignment (MSA) for the HIV-1 clade B protease protein from the Los Alamos National Laboratory HIV database (http://www.hiv.lanl.gov). Sequences labeled by the database as “problematic” were

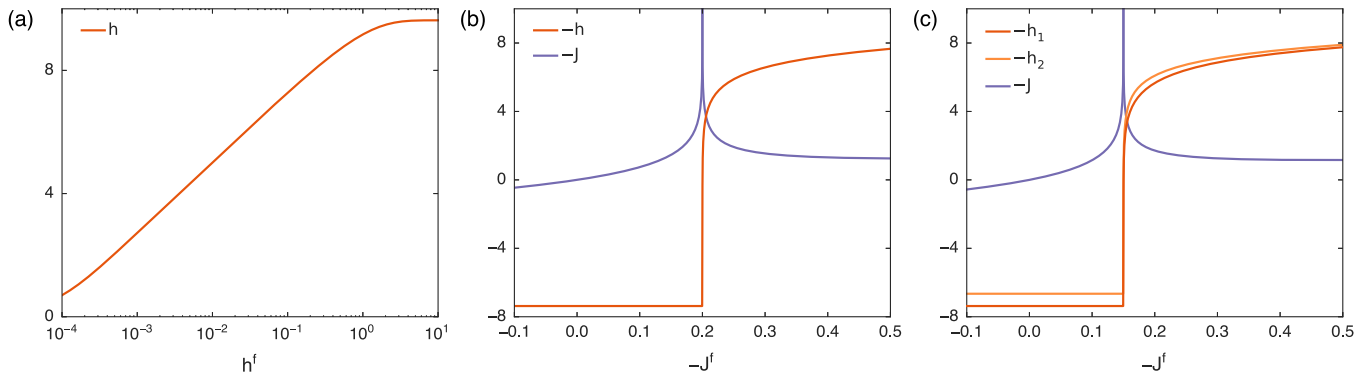

FIG. 4. (a) Plot of h versus h^f showing the sensitivity of h. The inferred field h approaches a μ-dependent cutoff as $h^f \to \infty$. (b) Plot of inferred h and J as a function of J_f for a two site approximation assuming $h_1^f = h_2^f = 0.1$. The inferred value of h is insensitive to J_f until the "level crossing" point $h_1^f + h_2^f + J^f = 0$ is reached. (c) Similar results are obtained when h_1^f and h_2^f are not identical. Here $h_1^f = 0.1$ and $h_2^f = 0.05$.

excluded. To minimize evolved drug resistance [3,34], we only selected sequences obtained in the year 1996 or earlier, and we removed sequences from trial studies of protease inhibitors, as described in the main text, yielding a total of 6701 sequences from 757 unique patients. After downloading, the MSA data was processed to remove insertions relative to the HXB2 reference sequence [35]. Ambiguous amino acids were then imputed with simple mean imputation.

The binarized MSA data consists of sequences from B patients, which we label $k = 1, \ldots, B$. Let us call the number of sequences from the kth patient as B_k, and let us write the ath sequence from patient k as $s^{(k,a)} = \{s_1^{(k,a)}, \ldots, s_{99}^{(k,a)}\}$, with the single site variables $s_i \in \{0,1\}$. To obtain a representative sample of the population, we averaged over multiple sequences from the same patient, so the one- and two-point correlations we obtain from the data are then

$$p_i = \frac{1}{B}\sum_{k=1}^{B}\left[\frac{1}{B_k}\sum_{a=1}^{B_k} s_i^{(k,a)}\right],$$
$$p_{ij} = \frac{1}{B}\sum_{k=1}^{B}\left[\frac{1}{B_k}\sum_{a=1}^{B_k} s_i^{(k,a)} s_j^{(k,a)}\right]. \tag{A1}$$

The one-point correlations p_i measure the frequency of mutations at each position i, and the two-point correlations p_{ij} measure the frequency of pairs of mutations occurring simultaneously at two positions i, j.

APPENDIX B: RELATIONSHIP BETWEEN h_f, J_f AND h, J

In general, it is difficult to show the precise relationship between the underlying h_f, J_f parameters and the corresponding inferred h, J. However, some insight can be obtained in simple cases.

First, let us consider a single site approximation in the Eigen model in Ising form [14]. Here the formula for $\exp[-E(s^T)]$ is as in (2), but with $F(s)$ given by

$$F(s) = \sum_{i=0}^{L} h_i^f s_i. \tag{B1}$$

We can solve for each site by decomposing the sum in (2) into a product of transfer matrices,

$$M = \begin{pmatrix} \exp\left(K - h^f\right) & \exp(-K) \\ \exp\left(-K - h^f\right) & \exp(K) \end{pmatrix}. \tag{B2}$$

In the limit of many generations, we can rewrite (2) as

$$\exp[-E(s^T)] \propto \lim_{T\to\infty} M^T v^0, \tag{B3}$$

where v^0 is a vector with the proportion of the population initially in the wild type and mutant states. This implies that we can obtain all of the information about the asymptotic state by looking at the eigenvector associated with the largest eigenvalue of M. Solving for the corresponding field yields

$$h = -\ln\left[\left(\frac{1-e^{h_f}}{2}\right)e^{2K} + \sqrt{e^{h_f} + \left(\frac{1-e^{h_f}}{2}\right)^2 e^{4K}}\right]. \tag{B4}$$

We find then that h is highly sensitive to small changes in h_f for small h_f [see Fig. 4(a); note that $K \simeq 4$ for amino acid mutations in HIV]. Precisely inferring h^f from h is thus a difficult problem in practice. However, it is likely that these issues are moderated at population sizes that are finite. Expressions for h, J inferred through a two-site approximation are unwieldy, but can easily be evaluated numerically [Figs. 4(b) and 4(c)].

To compute the solution for the Eigen model in the two site approximation, the following transfer matrix was used:

$$M = \begin{pmatrix} \exp\left(2K - h_1^f - h_2^f - J\right) & \exp\left(-h_1^f\right) & \exp\left(-h_2^f\right) & \exp(-2K) \\ \exp\left(-h_1^f - h_2^f - J\right) & \exp\left(2K - h_1^f\right) & \exp\left(-2K - h_2^f\right) & 1 \\ \exp\left(-h_1^f - h_2^f - J\right) & \exp\left(-2K - h_1^f\right) & \exp\left(2K - h_2^f\right) & 1 \\ \exp\left(-2K - h_1^f - h_2^f - J\right) & \exp\left(-h_1^f\right) & \exp\left(-h_2^f\right) & \exp(2K) \end{pmatrix}. \tag{B5}$$

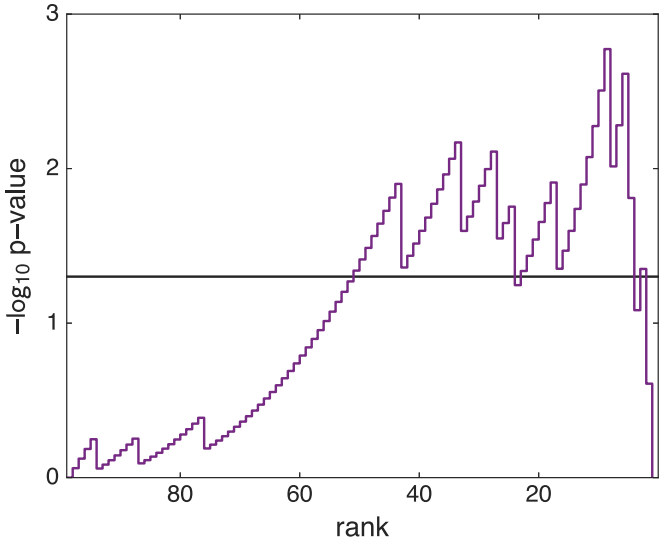

FIG. 5. Minus log p values (base 10) as a function of rank. The solid line indicates the standard significance threshold 0.05.

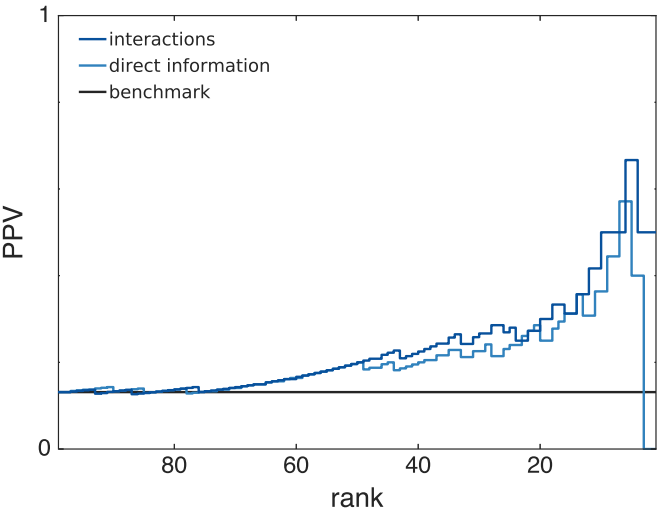

FIG. 7. Comparison of direct information approach and the direct interaction approach from the paper to classifying drug resistance mutations using PPV.

The normalized elements of the eigenvector associated with the largest eigenvalue give the fraction of the population in each state and are trivially algebraically related to the parameters of the prevalence landscape.

APPENDIX C: STATISTICAL SIGNIFICANCE OF RESISTANCE MUTATION DETECTION

As a further check of the significance of the results, we computed p values for the null hypothesis that predicted resistance sites were drawn randomly. This results in a p value that is a function of number of predicted resistance sites. If there are r sites randomly drawn out of a total of $N = 99$ sites, and m of the sites drawn are resistance sites (out of $M = 13$), the p value is given by

$$p = \sum_{k=m}^{M} \frac{\binom{M}{k}\binom{N-M}{r-k}}{\binom{N}{r}}. \tag{C1}$$

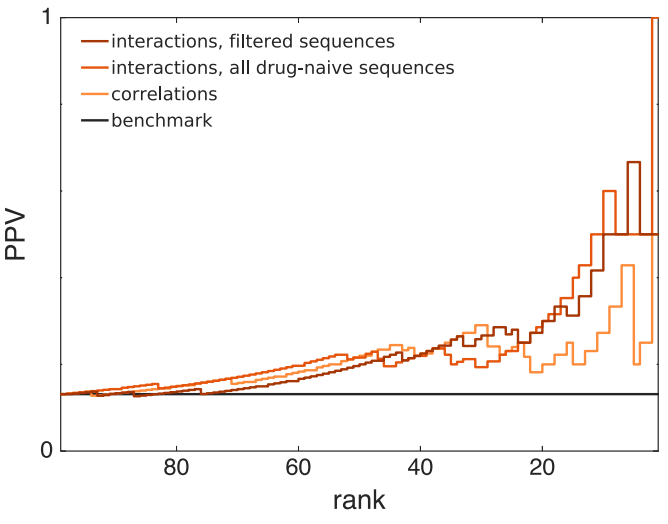

FIG. 6. Comparison of classification results for PPV.

The p values are plotted in Fig. 5 as a function of rank r. As noted in the main text, $p < 0.05$ for almost all ranks between 3 and 50, supporting the significance of the results, as the classification rule is not expected to perform well for weakly coupled sites (low ranks).

APPENDIX D: RESULTS FOR ALTERNATIVE CLASSIFICATION PROCEDURES AND DRUG-NAIVE DATA

Here we show predictions of resistance sites using alternative classification rules and data. We first examine the predictions made with the same model, but including all sequences from drug-naive patients up until the present. The results are shown in Fig. 6, along with the calculation from the main text for comparison, and are not significantly

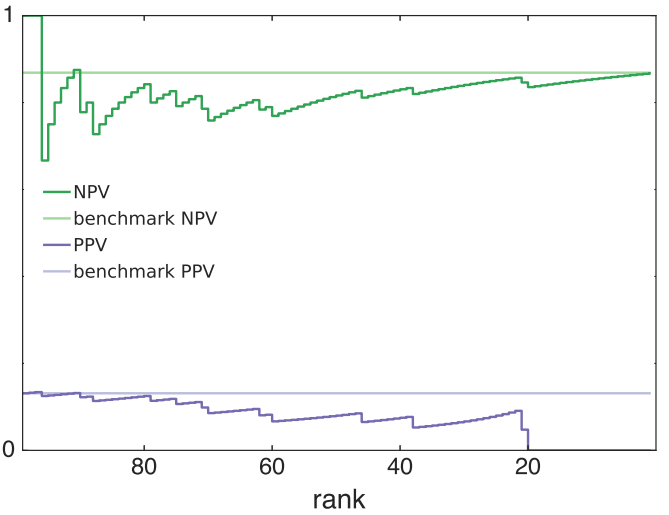

FIG. 8. Negative predictive value (NPV, upper curves) and positive predictive value (PPV, lower curves) for sites ranked according to the frequency of mutations at that site.

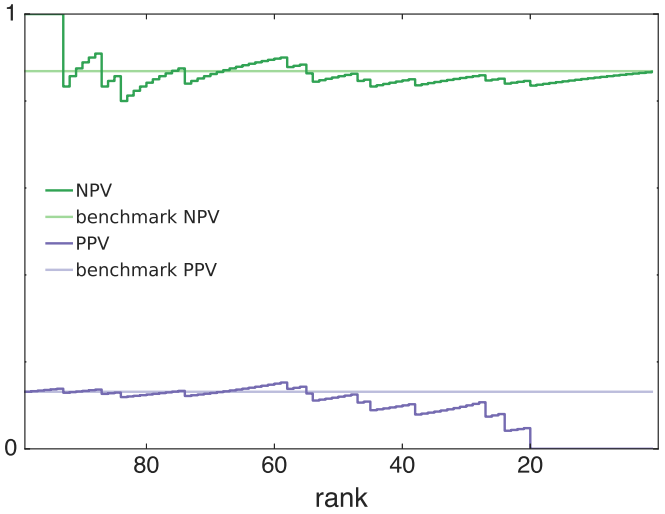

FIG. 9. Negative predictive value (NPV, upper curves) and positive predictive value (PPV, lower curves) for sites ranked according to the inferred field at that site.

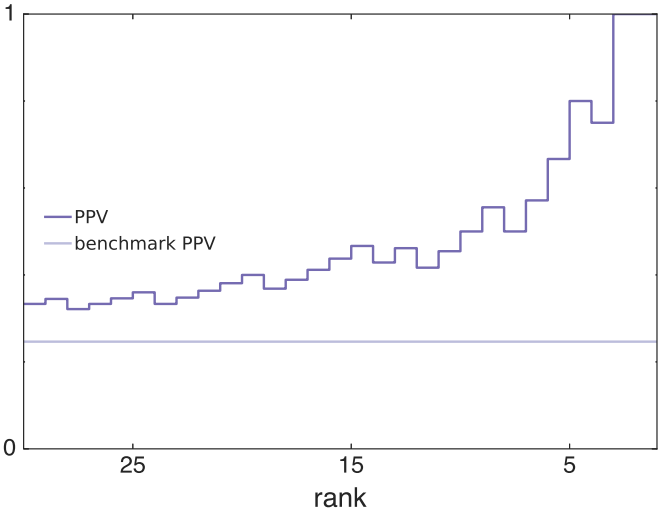

FIG. 10. Performance on the classification problem of identifying pairs of sites where one or more sites are associated with major drug resistance using the top 30 ranked couplings, measured by positive predictive value (PPV).

different. This is probably because transmitted protease inhibitor resistance is relatively rare [3,34]. However, the performance is slightly better at the extremely high threshold limit for the drug-naive sequence case, a possible signature of transmitted drug resistance.

Another very simple way to make predictions is to simply threshold the observed correlation matrix, defined by

$$C_{ij} = \frac{\langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle}{\sqrt{\langle s_i \rangle (1 - \langle s_i \rangle) \langle s_j \rangle (1 - \langle s_j \rangle)}}. \tag{D1}$$

In principle, all of the arguments developed in the main text apply to correlations as well. However, the presumed advantage of the direct interactions approach is that it disentangles indirect from direct interactions, which the correlation matrix does not. Predictions using the correlation matrix compared with the direct interactions approach (with all sequences from drug-naive patients, as well as the restricted sequence set used in the main text) are in Fig. 6. The direct interaction approach clearly performs better for the high ranked sites.

In protein contact prediction, a common measure of interactions is the direct information. Direct information is defined with respect to a two site model:

$$P(s_i, s_j) = Z^{-1} \exp\left(J_{ij} s_i s_j + \tilde{h}_i s_i + \tilde{h}_j s_i\right). \tag{D2}$$

The coupling J_{ij} is taken from the full solution of the inverse Ising problem with all sites, and the fields $\tilde{h}_i$ and $\tilde{h}_j$ are chosen to match the single site probabilities $P(s_i)$ and $P(s_j)$. The direct information between sites i and j is then constructed as

$$DI_{ij} = \sum_{s_i, s_j} P(s_i, s_j) \ln \left[\frac{P(s_i, s_j)}{P(s_i) P(s_j)} \right]. \tag{D3}$$

Thresholding the direct information matrix and following the usual procedure results in predictions of resistance sites. The results are shown in Fig. 7.

All methods based on interactions perform better than ranking sites according to their mutability, either directly by mutation frequency (Fig. 8) or by the inferred field h (Fig. 9). This is because mutations in protease that confer drug resistance in protease tend to be deleterious; thus, directly ranking sites according to the ease of single mutations leads to poor predictions of resistance.

We note also that many of the largest couplings link sites where just one site is classified as a major site of drug resistance. Based on the methods presented here, we have no way to distinguish which site or sites in a strongly linked pair should be associated with drug resistance. One alternate approach, then, would be to rank the couplings in order of their strength and attempt to predict how often either one or both coupled sites are sites of major drug resistance. Performance on this classification problem is also substantially better than random for the largest couplings, as shown in Fig. 10.

[1] A. Molla, M. Korneyeva, Q. Gao, S. Vasavanonda, P. J. Schipper, H.-M. Mo, M. Markowitz, T. Chernyavskiy, P. Niu, N. Lyons *et al.*, Nat. Med. **2**, 760 (1996).

[2] P. A. Volberding and S. G. Deeks, Lancet **376**, 49 (2010).

[3] W. H. Wheeler, R. A. Ziebell, H. Zabina, D. Pieniazek, J. Prejean, U. R. Bodnar, K. C. Mahle, W. Heneine, J. A. Johnson, H. I. Hall *et al.*, AIDS **24**, 1203 (2010).

[4] A. Rambaut, D. Posada, K. A. Crandall, and E. C. Holmes, Nat. Rev. Genet. **5**, 52 (2004).

[5] J. M. Coffin, Science **267**, 483 (1995).

[6] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho, Science **271**, 1582 (1996).

[7] V. Dahirel, K. Shekhar, F. Pereyra, T. Miura, M. Artyomov, S. Talsania, T. M. Allen, M. Altfeld, M. Carrington, D. J. Irvine *et al.*, Proc. Natl. Acad. Sci. USA **108**, 11530 (2011).

[8] A. L. Ferguson, J. K. Mann, S. Omarjee, T. Ndungu, B. D. Walker, and A. K. Chakraborty, Immunity **38**, 606 (2013).
[9] T. Hinkley, J. Martins, C. Chappey, M. Haddad, E. Stawiski, J. M. Whitcomb, C. J. Petropoulos, and S. Bonhoeffer, Nat. Genet. **43**, 487 (2011).
[10] L. Chen, A. Perlina, and C. J. Lee, J. Virol. **78**, 3722 (2004).
[11] Z. W. Cao, L. Y. Han, C. J. Zheng, Z. L. Ji, X. Chen, H. H. Lin, and Y. Z. Chen, Drug Discovery Today **10**, 521 (2005).
[12] N. Beerenwinkel, B. Schmidt, H. Walter, R. Kaiser, T. Lengauer, D. Hoffmann, K. Korn, and J. Selbig, Proc. Natl. Acad. Sci. USA **99**, 8271 (2002).
[13] M. Eigen, Naturwissenschaften **58**, 465 (1971).
[14] I. Leuthäusser, J. Chem. Phys. **84**, 1884 (1986).
[15] K. Shekhar, C. F. Ruberman, A. L. Ferguson, J. P. Barton, M. Kardar, and A. K. Chakraborty, Phys. Rev. E **88**, 062705 (2013).
[16] S.-Y. Rhee, M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer, Nucleic Acids Res. **31**, 298 (2003).
[17] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[18] S. Cocco and R. Monasson, Phys. Rev. Lett. **106**, 090601 (2011).
[19] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc. Natl. Acad. Sci. USA **108**, E1293 (2011).
[20] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature (London) **440**, 1007 (2006).
[21] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. USA **106**, 67 (2009).
[22] D. S. Marks, T. A. Hopf, and C. Sander, Nat. Biotechnol. **30**, 1072 (2012).
[23] J. Barton and S. Cocco, J. Stat. Mech.: Theory Exp. (2013) P03002.
[24] J. K. Mann, J. P. Barton, A. L. Ferguson, S. Omarjee, B. D. Walker, A. K. Chakraborty, and T. Ndung'u, PLoS Comput. Biol. **10**, e1003776 (2014).
[25] B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes, Science **303**, 327 (2004).
[26] P. Lemey, A. Rambaut, and O. G. Pybus, AIDS Rev. **8**, 125 (2006).
[27] R. A. Neher and T. Leitner, PLoS Comput. Biol. **6**, e1000660 (2010).
[28] I. Bartha, J. M. Carlson, C. J. Brumme, P. J. McLaren, Z. L. Brumme, M. John, D. W. Haas, J. Martinez-Picado, J. Dalmau, C. López-Galíndez *et al.*, eLife **2**, e01123 (2013).
[29] R. J. De Boer, R. M. Ribeiro, and A. S. Perelson, PLoS Comput. Biol. **6**, e1000906 (2010).
[30] R. Sanjuan, M. R. Nebot, N. Chirico, L. M. Mansky, and R. Belshaw, J. Virol. **84**, 9733 (2010).
[31] M. W. Chang and B. E. Torbett, J. Mol. Biol. **410**, 756 (2011).
[32] R. Colonno, R. Rose, C. McLaren, A. Thiry, N. Parkin, and J. Friborg, J. Infect. Dis. **189**, 1802 (2004).
[33] Y. Pommier, A. A. Johnson, and C. Marchand, Nat. Rev. Drug Discovery **4**, 236 (2005).
[34] R. K. Gupta, M. R. Jordan, B. J. Sultan, A. Hill, D. H. Davis, J. Gregson, A. W. Sawyer, R. L. Hamers, N. Ndembi, D. Pillay *et al.*, Lancet **380**, 1250 (2012).
[35] T. Leitner, B. Korber, M. Daniels, C. Calef, and B. Foley, in *HIV-1 Subtype and Circulating Recombinant form (CRF) Reference Sequences*, edited by T. Leitner, B. Foley, B. Hahn, P. Marx, F. McCutchan, J. Mellors, S. Wolinsky, and B. Korber (New Mexico: Los Alamos National Laboratory, 2005), pp. 41–48.