Fitness-based prediction of sites of drug resistance mutations using sequence data from HIV protease inhibitor-naïve patients

Thomas C Butler [*] [†], **John P Barton** [*] [†] [‡], **Mehran Kardar** [*], **and Arup K Chakraborty** [*] [†] [‡] [§]

[*]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139,[†]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139,[‡]Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology, and Harvard University, Boston, MA 02129, and [§]Departments of Chemistry and Biological Engineering; Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139

Submitted to Proceedings of the National Academy of Sciences of the United States of America

HIV evolves with extraordinary rapidity. However, its evolution is constrained by its fitness landscape and the associated pattern of epistatic interactions between mutations. Computational models of HIV that exploit these constraints to anticipate its evolutionary response to new environments could aid in searching for therapies that provide improved protection against resistance. Here we infer a statistical model describing patterns of mutations in HIV protease sequence data obtained prior to widespread clinical use of protease inhibitors. Guided by a simple model of evolutionary dynamics, we use the inferred statistical model to identify sets of mutations likely to be able to co-occur with low fitness costs, which we hypothesize to be likely sites of drug resistance mutations. The resulting predictions of the sites of HIV protease resistance mutations perform significantly better than chance, despite the exclusion of resistance data from the model. Successful predictions in the absence of labeled resistance data suggest that our approach may be applied to help design new therapies that are less prone to failure even where resistance data is not yet available, as well as indicating progress in the development of a model predicting features of HIV evolution at the single residue level.

statistical mechanics | HIV | drug resistance | machine learning

Abbreviations: HIV, human immunodeficiency virus

Under selective pressure from sub-optimal anti-retroviral treatment regimens, evolved drug resistance leading to HIV virological failure has been observed to occur 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 HIV patients in the United States [2, 3]. The rapid evolution of resistance is an instance of 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, 5, 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 infer predictive models of HIV evolution. Previous researchers have used a variety of approaches to attempt to predict HIV fitness and aspects of its evolution using viral sequence data on its own [7, 8], and also sequences labeled according to phenotypic properties such as drug resistance and replicative capacity [9]. Other researchers have addressed the related problem of effectively predicting the sites of resistance mutations by detecting sites under positive selection during treatment [10], supervised learning [11], and structural modelling [11, 12]. Despite the success of these latter approaches in predicting resistance sites, each requires treatment or structural data, limiting their usefulness in contexts (such as the design or introduction of a new combination therapy) where such data are not available.

In the present work, we use sequence data obtained prior to the widespread clinical use of protease inhibitors, combined with mathematical modeling, to predict sets of sites in HIV protease where coordinated mutations are unlikely to significantly impair viral fitness. We predict that such sites are more likely to be sites of clinically relevant drug resistance mutations because mutations that confer drug resistance but destroy or severely impair viral replication are unlikely to be selected. Excluding data obtained after the clinical use of anti-retroviral drugs allows us to explore how the natural evolution of HIV predicts its evolution under drug pressure. Our successful prediction of major drug resistance sites (defined in [13]) using natural evolution data suggests that the techniques of this paper can be applied to predict aspects of HIV evolution in response to new treatment regimens and vaccine candidates. Such predictions will enhance clinical outcomes by improving treatment rollout and vaccine design. As an illustration, we consider pairs of protease inhibitors that can be used in successive treatments that are predicted to inhibit the evolution of resistance to both drugs. We emphasize that the purpose of the present work is both to develop a clinically useful technique for anticipating HIV evolution, and also to advance scientific understanding of viral evolution by creating a model of HIV evolution that is mechanistically interpretable and predictive at the single residue level.

Significance

Anticipating the mutational paths most likely to lead to drug resistance could help inform drug design and treatment strategies against pathogens such as HIV. Here we combine statistical analysis of sequence data and a model of viral evolution to identify sets of mutations likely to have low fitness costs, and demonstrate that this information can be used to predict sites of drug resistance mutations in HIV protease. Importantly, these predictions are based on evolutionary considerations alone, using sequence data obtained prior to the clinical use of protease inhibitor drugs. Thus, our approach could be used to anticipate sites of drug resistance mutations even when detailed resistance data is not available.

Reserved for Publication Footnotes

Our analysis begins by inferring an estimate of the probability distribution of multiple mutations in the viral protease protein from large amounts of sequence data. The form of the probability distribution gives rise to a natural notion of a “prevalence landscape” analogous to a fitness landscape, that expresses the relative probabilities of protease sequences. The form of the prevalence landscape is given by a disordered Ising model from statistical physics [8, 14]. As in the problem of predicting protein contact residues, representing the prevalence landscape as a disordered Ising model has the advantage that the parameters in the model have a simple interpretation as the strength of direct interactions between mutations in the prevalence landscape [15, 16, 17]. However, to make predictions about evolution, it is of greater interest to infer fitness interactions between mutations than prevalence interactions.

Using the intuitive notion that highly fit strains should be more prevalent, previous work has shown that the inferred prevalences of sequences from HIV Gag proteins correlate with their replicative capacities, another proxy for fitness [8, 18]. However, the prevalence landscape is determined by many factors other than fitness, including epidemiological dynamics, recombination, and demographic noise, so fitness cannot be simply equated with prevalence [19, 20, 21].

While a complete understanding of how fitness and frequency are related is elusive, an insight into their relationship can be obtained by studying the relation between fitness and frequency in Eigen’s model of evolution [22]. 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 frequency to be studied for short proteins or simple fitness landscapes using methods adopted from statistical physics [23, 24]. This analysis allows for a more precise interpretation of the fitness implications of the direct interactions between mutations in the prevalence landscape, informing the prediction of the sites of resistance mutations. We note that while the results of the Eigen model provide a useful interpretation of the frequency landscape that motivates the procedure used to generate predictions, the actual predictions of resistance mutations do not rely on this interpretation.

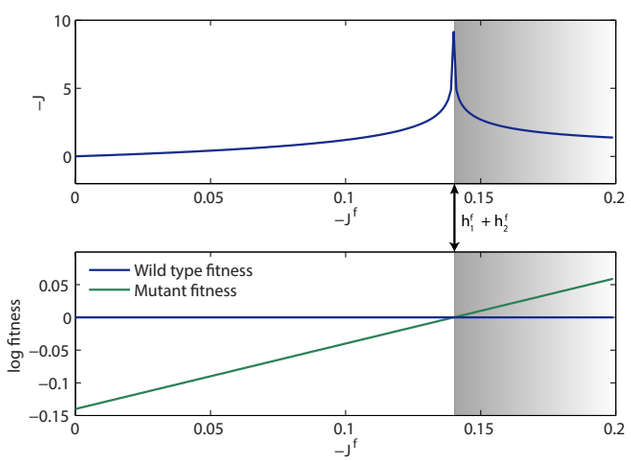

Fig. 1. The coupling J between a pair of sites where single mutations are deleterious increases sharply as the fitness of the double mutant approaches the fitness of the wild type sequence. Top panel: Coupling J in the prevalence landscape as a function of J^f from the fitness landscape. The peak occurs at the level crossing, where $-J^f = h_1^f + h_2^f$. Note 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. Bottom panel: 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. This shows that large negative inferred values of J in the prevalence landscape are associated with level crossings in the two site limit.

Results

Inferring the HIV prevalence landscape. The large size of the sequence space for HIV protease (naïvely 20^{99} possible sequences for protease, which is 99 amino acids long), and the comparatively small amount of available sequence data, preclude direct estimation of the probability of sequences with multiple mutations from their frequencies in the data. Instead we adopt the procedure of previous work on HIV Gag proteins [8] to construct a probability distribution consistent with the observed frequency of mutations at each site and pair of sites, which can be reliably estimated from existing sequence data. Protease amino acid sequences are first translated into a binary form by coding the wild type (consensus) amino acid at each site as 0, and a mutant as 1. The amino acid identity at each site i in a sequence is thus coded as a binary variable $s_i \in \{0, 1\}$, and full sequences are represented as vectors of binary variables $s = (s_1, s_2, .., s_{99})$. Representing sequences in this way disallows predictions about mutations to particular amino acids, but enables the accurate inference of the prevalence landscape by greatly reducing the dimensionality of the inference problem. Future work on HIV protease will relax the binary approximation.

We proceed by assuming that the joint distribution of mutations is mostly specified by the moments $\langle s_i s_j \rangle$ and finding the least structured (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) [25, 8]. The resulting probability distribution takes the form

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

where Z is a normalization constant (the partition function), s is a sequence vector with elements $\{s_i\}$, and $\{J_{ij}\}$, $\{h_i\}$ parametrize the distribution. The parameters $\{J_{ij}\}$, $\{h_i\}$ are chosen such that the distribution $P(s)$ reproduces the observed moments $\langle s_i s_j \rangle$. Inferring the maximum entropy estimate of the joint distribution of discrete random variables in this manner has also been fruitfully applied to the analysis of neuronal network states and protein contact prediction [26, 15, 17]. The description of the algorithm used to infer $E(s)$ is given in the Supporting Information and in [27, 28]. The assumption that the distribution is mostly specified by its first two moments is justified by noting that the inferred probability distribution in Eq. 1 predicts the observed higher order moments well (Supporting Information).

For the purpose of the present work, the primary advantage of inferring the probability distribution Eq. 1 is to distinguish direct interactions between mutations from correlations. While correlations between two sites may be due to interactions with a common intermediate site, accurate estimates of the parameters $\{J_{ij}\}$, $\{h_i\}$ in Eq. 1 capture direct interactions, disentangled from indirect effects mediated through intermediate sites [15, 28, 26].

The probability distribution Eq. 1 also allows estimation of the relative prevalences of different sequences. In line with the expectation that fitness plays a substantial role in determining

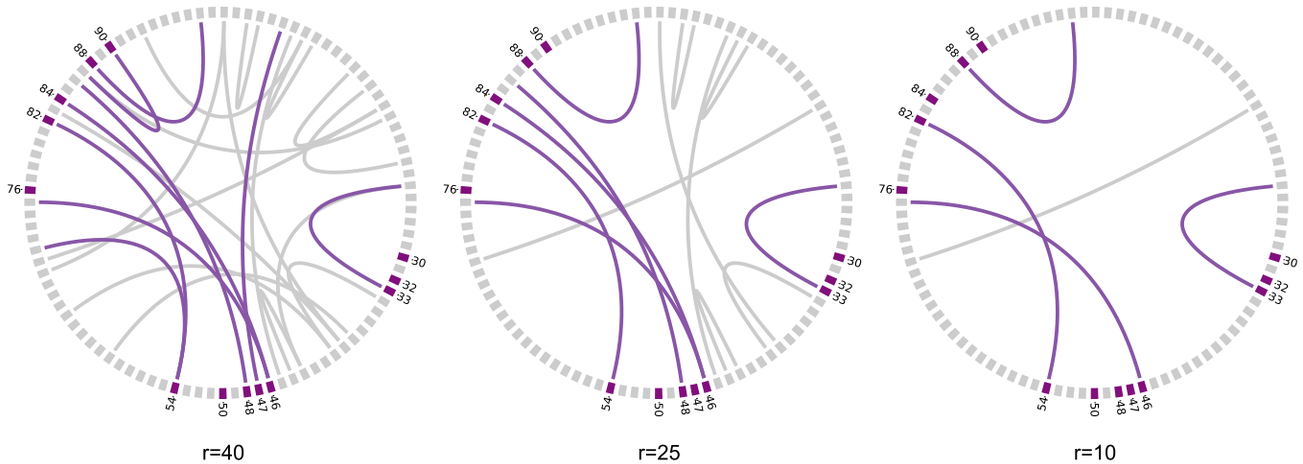

Fig. 2. Stronger couplings are more likely to link sites of major resistance mutations. The network of interactions between the top r ranked sites, from $r = 40$ (left) to $r = 10$ (right), are plotted with Circos [30]. Only the strongest couplings, those meeting or exceeding the largest coupling for the lowest ranked site, are displayed. Major resistance sites and interactions linking them are colored, while other sites and links between non-resistance sites are grey.

the frequency of a viral strain, previous work on Gag proteins shows that the prevalence in Eq. 1 is a reasonable predictor of fitness [8], even in comparison with regression methods using supervised learning and labeled data [9]. These results provide evidence that the prevalence landscape approach captures meaningful information about HIV fitness. However, as noted above, it is also the case that the dynamics of mutations, recombination events, and other factors contribute to frequency, and so Eq. 1 does not translate directly into fitness and should not be expected to correlate with it perfectly.

The sequences used to infer Eq. 1 were taken from the Los Alamos HIV database (www.hiv.lanl.gov). As transmitted drug resistance to protease inhibitors is observed in contemporary HIV populations [29, 3], we excluded all sequences after 1996, or with the phrase “protease inhibitor” in their meta data. These filters were designed to remove as much evolved drug resistance from the sequence data as possible so that the landscape in Eq. 1 reflects the intrinsic fitness of HIV protease, uninfluenced by drug pressure. We also carried out the same analysis on all available drug naïve sequences from the Los Alamos data base with slight changes in our results that indicate population level selective pressure on HIV protease to evolve resistance (see Supporting Information).

Extracting compensatory pair predictions from the prevalence landscape. To relate the prevalence landscape in Eq. 1 to fitness, we note that prevalence can also be written as the outcome of evolutionary dynamics in the Eigen model, represented as an Ising model from statistical physics [23]

$$\exp(-E(s^T)) \propto \sum_{\{s^t\}_{t=1}^{T-1}} \exp\left[\sum_{t=1}^{T-1} K(2s^t - 1)(2s^{t+1} - 1) - F(s^t))\right]$$

$$F(s) = \sum_{i<j}^{L} J_{ij}^f s_i s_j + \sum_{i=1}^{L} h_i^f s_i, \qquad [2]$$

where K is related to the per site per generation mutation rate μ by $K = \frac{1}{2}\log(\frac{1-\mu}{\mu})$, and $F(s)$ is minus the log fitness of sequence s and will be referred to as the fitness landscape. Here the superscripts $t \in \{1, 2 \ldots, T\}$ on the sequence vectors refer to discrete generations in Eigen's model of evolution. The superscript f on the parameters $\{h_i^f\}$ and $\{J_{ij}^f\}$ indicates that the 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 [24], though protease is not comparatively immunogenic [31].

Ideally, one would like to invert Eq. 2 to solve for the function $F(s)$ in terms of $E(s)$, because $E(s)$ is inferred directly from data. Because this is a very high dimensional and difficult problem, we proceed instead by considering a two site approximation, where only pairs of sites are coupled. While network effects influence the inferred couplings between sites, this simple approximation provides useful conceptual insight. Furthermore, most of the variance in the on-diagonal terms $\langle s_i \rangle$ is explained by the single site h_i in Eq. 1, indicating that network effects exert a weaker influence on the $\langle s_i \rangle$ (Supporting Information).

Solving the two site limit of Eq. 2 shows that the $\{h_i^f\}$ are difficult to reliably infer, because the mutation coupling K is large enough ($K \simeq -\log(\mu)$ and $\mu \simeq \mathcal{O}(10^{-4})$) that very small h_i^f lead to large h_i in the prevalence landscape (see Supporting Information). However, large values of J_{ij} 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 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 shall 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.

It is important to note that large values of $-J$ in prevalence are, within the limits of the two site approximation, only related to the actual numerical value of the underlying J^f through a non-linear transformation that is too sensitive to be reliably inverted with real data at present, and is not justified due to the simplicity of the model (see Fig. 1). The major contribution of the model is to support an interpreta-

tion of large values of J_{ij} as associated with pairs of sites that can co-mutate with low fitness cost. As will be seen below, this interpretation is useful for understanding the connection between the inferred prevalence landscape and resistance.

A distinction should also be made between the concept of level crossings introduced here and the classical notion of intragene epistasis. Epistasis can be defined in this context as $e = -F(1,1) + F(0,1) + F(1,0) = -J^f$, or the departure for two sites from multiplicative combination of the fitness of mutations [32, 33]. A level crossing is more stringently defined as a place in parameter space where $J^f - h_1^f - h_2^f = 0$ so that the wild type and the double mutant are equally fit. With the parameters $\{J^f, h_1^f, h_2^f\}$ tuned near a level crossing, the fitness of the double mutant is near that of the wild type, even if the fitness penalty for the associated single mutants is large. Epistasis is known to affect the evolutionary dynamics of HIV-1 [9], but the relationship between the level crossings in this work and previous empirical studies of epistasis in HIV [34, 35] will require further research to be fully understood.

Predicting the sites of resistance mutations. 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, the following heuristic argument suggests a way to proceed with the limited information we have assumed in the present study. When the environment HIV is replicating in changes due to the initiation of drug therapy, HIV must mutate in ways that alter its sequence in order to abrogate drug binding, while at the same time preserving protein function. Large couplings J_{ij} connect sites that are likely to be able to co-mutate with very 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.

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.

The model above can be cast in the form of a standard classification rule from supervised learning (as is typically obtained using labeled data) 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. While this allows standard performance measures for classification to be applied to the present model, it is important to emphasize that no labeled data was used to construct the classification rule.

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) [13]. The suitable statistics for evaluating a classifier's performance are dependent on the algorithm chosen and on the problem to be solved [36]. For the method described here, we note that not every resistance mutation is necessarily associated with a nearby level crossing, and also that it is not necessary that large couplings should be associated with presently identified resistance (although it is likely that targeting such sites would lead to unusually rapid evolution of resistance). However, as higher ranked sites are selected, the proportion of sites that are associated with resistance should increase. To measure this, a suitable statistic is the positive prediction value (PPV), which is defined as

$$P(\text{true} = \text{resistance}|\text{predicted} = \text{resistance}). \qquad [3]$$

Similarly the negative prediction value (NPV) is defined as

$$P(\text{true} = \text{non-resistance}|\text{predicted} = \text{non-resistance}). \qquad [4]$$

These are shown in Fig. 3 compared to random benchmarks, and demonstrate that the performance of the classification rule is substantially better than chance as higher ranked sites are selected. Further understanding of the performance can be had by investigating standard measures of classification performance, including the true positive rate (TPR),

$$P(\text{predicted} = \text{resistance}|\text{true} = \text{resistance}), \qquad [5]$$

and false positive rate (FPR),

$$P(\text{predicted} = \text{resistance}|\text{true} = \text{non-resistance}), \qquad [6]$$

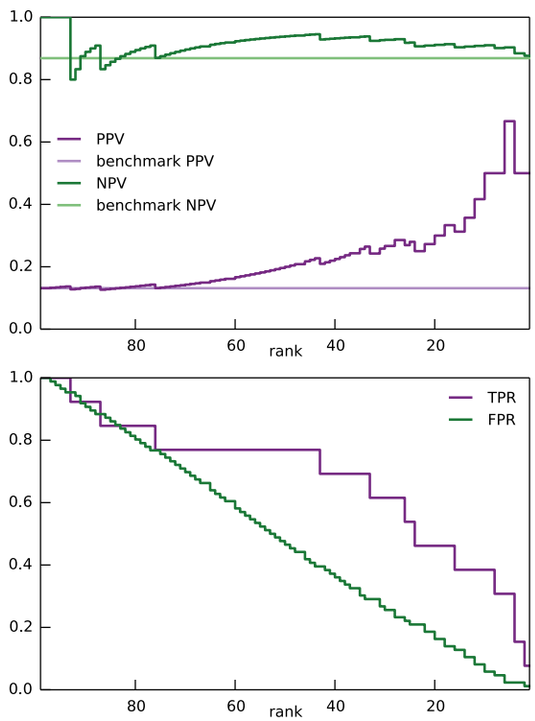

Fig. 3. Top-ranked sites, based on the maximum strength of their couplings, are far more likely to be sites of major drug resistance mutations than would be expected by chance. The top panel shows the positive prediction value (PPV) and negative prediction value (NPV) for the classifier defined in the text compared to the benchmark of random guessing, as a function of rank. Collections of the highest ranked sites are clearly associated with improved PPV. The bottom panel shows the false positive rate (FPR) and the true positive rate (TPR) as functions of rank. TPR > FPR indicates performance better than chance.

as a function of rank. The results in Fig. 3 confirm that the true positive rate is larger than the false positive rate, indicating performance better than chance. Beyond these performance measures, we 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 Supporting Information).

To examine these results using classical statistical significance testing, we used the hyper-geometric 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. For the results above, the predictions 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 the 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. Details of the significance testing are in the Supporting Information. We also tested related classification rules constructed using direct information [15] and correlation matrices, with no improvement in performance (see Supporting Information).

In contrast to standard supervised learning where complex models can lead to over fitting and poor generalization [37], the results above cannot be due to over fitting. This is because the model is fully blind to the sites of resistance mutations, and therefore cannot fit to them at all. The exclusion of over fitting as an explanation for these results increases confidence that the techniques above can be fruitfully applied where no labeled data is available. Future work will examine other proteins that rapidly evolve resistance to drugs, as well as secondary resistance mutations in HIV protease to increase the sample size and better assess the performance of the approach developed above.

Protease Inhibitor pair therapies. 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 that are optimized to protect patients from evolving overlapping resistance.

To protect against resistance, a pair of drugs should have as many non-overlapping resistance mutations as possible. Additionally, we would like to identify pairs of drugs whose associated resistance mutations are difficult to make simultaneously due to fitness constraints. In the same way that large positive values of $-J$ indicate sites that can readily mutate together, negative values of $-J$ indicate sites where double mutations are suppressed. Thus, the interactions between the resistance mutations that are not common to both drugs should be as negative as possible. We found three combinations (atazanivir-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 [38].

Biophysical interpretation of large couplings. In order to make good predictions about HIV evolution and adaptation, features of our model should relate to physical and biological properties of HIV protease. We find that the network of large interactions does indeed capture 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 [39]. The study showed that on its own, the major resistance mutation reduced the stability of HIV protease considerably, as measured by melting temperature. 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 (Supporting Information). 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.

The analysis described here is very closely related to that used to predict physical contacts in protein structures from sequence data in widely divergent organisms [16, 15, 17]. In the contact prediction problem, the strongest couplings $\{J_{ij}\}$ (typically assessed using direct information, see Supporting Information) are taken to indicate contacts. We also observe some connection between large couplings and contacts: out of the top 40 couplings, we find 9 are contacts in the protease dimer, and 6 of these correspond to pairs of sites directly adjacent to one another. Approximately 10% of all pairs of sites in protease are in contact, so contact pairs are enriched among the strongest couplings, but the correlation is much weaker than previous results for protein families. However, the effective capture of links between destabilizing and stabilizing pairs of mutations discussed above suggests that the model describes additional structurally and functionally important relationships between sites other than contacts. One possible reason for this discrepancy in interpretations versus the protein contact literature is that the sequences in the present study are taken only from HIV, rather than from a variety of evolutionarily diverged organisms as in the contact prediction cases.

Discussion

We have shown that by analysing sequences of HIV protease the sites of major resistance mutations can be statistically identified. While the predictions were only tested against previously known resistance sites, that the model was completely blind to resistance data provides confidence that therapeutically useful predictions based on the techniques presented above may prove accurate.

It is interesting to ask how well the techniques developed here will generalize. We expect that the answer depends on the protein studied. It seems possible that some proteins have fewer sites connected by large couplings, or have lower typical fitness costs for individual mutations. Drugs targeting such proteins may not lead to resistance that is predictable using the methods developed here. One possible example of such a protein may be HIV reverse transcriptase. For most protease inhibitors currently in use, resistance typically requires > 2 mutations on average, but in reverse transcriptase, resistance to most inhibitors evolves with only 1-2 mutations [40]. However, it is also known that there are compensatory inter-

actions between resistance mutations in reverse transcriptase [41], which provides evidence that the present approach will generalize well. It is possible that analysis of the energy functions of other proteins that are drug targets, including the strength of the couplings J_{ij} between different sites, can provide some insight into the emergence of resistance mutations.

It is also worth noting that neither false positive or false negative resistance mutation predictions necessarily indicate error in the prevalence landscape, or in the inference from strong couplings. This is because the association of resistance mutations with level crossings is simply a heuristic that should not be expected to apply to all resistance mutations, or to all level crossings. It is likely that therapeutically targeting sites associated with inferred level crossings that are not presently associated with resistance will lead to resistance at those sites.

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 scientific 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 new treatments, such as integrase inhibitors [42], where resistance is not nearly as well characterized as in protease.

Acknowledgements

We thank Daniel Kuritzkes, Martin Hirsch, Andrew Furguson and Dariusz Murakowski for helpful discussions. This research was funded by the Ragon Institute of MGH, MIT, & Harvard, and NSF under Grant No. PHY11-25915.

1. Molla A, et al. (1996) Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. Nature Medicine 2:760–766.
2. Volberding PA, Deeks SG (2010) Antiretroviral therapy and management of HIV infection. The Lancet 376:49–62.
3. Wheeler WH, et al. (2010) Prevalence of transmitted drug resistance associated mutations and HIV-1 subtypes in new HIV-1 diagnoses, US-2006. AIDS 24:1203–1212.
4. Rambaut A, Posada D, Crandall KA, Holmes EC (2004) The causes and consequences of HIV evolution. Nature Reviews Genetics 5:52–61.
5. Coffin JM (1995) HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science 267:483–483.
6. 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.
7. Dahirel V, et al. (2011) Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. Proceedings of the National Academy of Sciences 108:11530–11535.
8. Ferguson AL, et al. (2013) Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. Immunity 38:606–617.
9. Hinkley T, et al. (2011) A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. Nature Genetics 43:487–489.
10. Chen L, Perlina A, Lee CJ (2004) Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. Journal of Virology 78:3722–3732.
11. Cao ZW, et al. (2005) Computer prediction of drug resistance mutations in proteins. Drug Discovery Today 10:521–529.
12. Beerenwinkel N, et al. (2002) Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. Proceedings of the National Academy of Sciences 99:8271–8276.
13. Rhee SY, et al. (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Research 31:298–303.
14. Binder K, Young AP (1986) Spin glasses: Experimental facts, theoretical concepts, and open questions. Reviews of Modern Physics 58:801.
15. Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences 108:E1293–E1301.
16. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein–protein interaction by message passing. Proceedings of the National Academy of Sciences 106:67–72.
17. Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. Nature Biotechnology 30:1072–1080.
18. Mann JK, et al. (2014) The fitness landscape of HIV-1 Gag: Advanced modeling approaches and validation of model predictions by in vitro testing. PLoS Computational Biology In Press.
19. Grenfell BT, et al. (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303:327–332.
20. Lemey P, Rambaut A, Pybus OG (2006) HIV evolutionary dynamics within and among hosts. AIDS Reviews 8:125–140.
21. Neher RA, Leitner T (2010) Recombination rate and selection strength in HIV intra-patient evolution. PLoS Computational Biology 6:e1000660.
22. Eigen M (1971) Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften 58:465–523.
23. Leuthäusser I (1986) An exact correspondence between Eigens evolution model and a two-dimensional Ising system. The Journal of Chemical Physics 84:1884.
24. Shekhar K, et al. (2013) Spin models inferred from patient data faithfully describe HIV fitness landscapes and enable rational vaccine design. arXiv preprint arXiv:1306.2029.
25. Jaynes ET (1957) Information theory and statistical mechanics. Physical Review 106:620.
26. Schneidman E, Berry MJ, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. Nature 440:1007–1012.
27. Cocco S, Monasson R (2011) Adaptive cluster expansion for inferring Boltzmann machines with noisy data. Physical Review Letters 106:090601.
28. Barton J, Cocco S (2013) Ising models for neural activity inferred via selective cluster expansion: structural and coding properties. Journal of Statistical Mechanics: Theory and Experiment 2013:P03002.
29. Gupta RK, et al. (2012) Global trends in antiretroviral resistance in treatment-naive individuals with HIV after rollout of antiretroviral treatment in resource-limited settings: a global collaborative study and meta-regression analysis. The Lancet 380:1250–1258.
30. Krzywinski M, et al. (2009) Circos: an information aesthetic for comparative genomics. Genome Research 19:1639–1645.
31. Bartha I, et al. (2013) A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. eLife 2:e01123.
32. Kouyos RD, Silander OK, Bonhoeffer S (2007) Epistasis between deleterious mutations and the evolution of recombination. Trends in Ecology & Evolution 22:308–315.
33. Phillips PC (2008) Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics 9:855–867.
34. Parera M, Perez-Alvarez N, Clotet B, Martínez MA (2009) Epistasis among deleterious mutations in the HIV-1 protease. Journal of Molecular Biology 392:243–250.
35. Bonhoeffer S, Chappey C, Parkin NT, Whitcomb JM, Petropoulos CJ (2004) Evidence for positive epistasis in HIV-1. Science 306:1547–1550.
36. Hand DJ (2012) Assessing the performance of classification methods. International Statistical Review 80:400–414.
37. Hastie T, Tibshirani R, Friedman JJH (2001) The elements of statistical learning (Springer New York) Vol. 1.
38. Colonno R, et al. (2004) Identification of i50l as the signature atazanavir (atv)-resistance mutation in treatment-naive hiv-1-infected patients receiving atv-containing regimens. Journal of Infectious Diseases 189:1802–1810.
39. Chang MW, Torbett BE (2011) Accessory mutations maintain stability in drug-resistant HIV-1 protease. Journal of Molecular Biology 410:756–760.
40. Tang MW, Shafer RW (2012) HIV-1 antiretroviral resistance. Drugs 72:e1–e25.
41. Hu Z, Kuritzkes DR (2011) Interaction of reverse transcriptase (RT) mutations conferring resistance to lamivudine and etravirine: effects on fitness and RT activity of human immunodeficiency virus type 1. Journal of Virology 85:11309–11314.
42. Pommier Y, Johnson AA, Marchand C (2005) Integrase inhibitors to treat HIV/AIDS. Nature Reviews Drug Discovery 4:236–248.