bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Measuring Selection Across HIV Gag: Combining Physico-Chemistry and Population Genetics

Elizabeth Johnson [1,2] and Michael A. Gilchrist* [2,3]

1 Microbiology, University of Tennessee, Knoxville, TN, United States

2 National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville, TN, United States

3 Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN, United States

Friday 14th September, 2018

* Corresponding author: mikeg@utk.edu

Abstract

We present physico-chemical based model grounded in population genetics. Our model predicts the stationary probability of observing an amino acid residue at a given site. Its predictions are based on the physico-chemical properties of the inferred optimal residue at that site and the sensitivity of the protein's functionality to deviation from the physico-chemical optimum at that site. We contextualize our physico-chemical model by comparing our model fit and parameters it to the more general, but less biologically meaningful entropy based metric: site sensitivity or $1/E$. We show mathematically that our physico-chemical model is a more restricted form of the entropy model and how $1/E$ is proportional to the log-likelihood of a parameter-wise 'saturated' model. Next, we fit both our physico-chemical and entropy models to sequences for subtype C's Gag poly-protein in the LANL HIV database. Comparing our model's site sensitivity parameters G' to $1/E$ we find they are highly correlated. We also compare the ability of G', $1/E$, and other indirect measures of HIV fitness to empirical *in vitro* and *in vivo* measures. We find G' does a slightly better job predicting empirical fitness measures of *in vivo* viral escape time and *in vitro* spreading rates. While our predictive gain is modest, our model can be modified to test more complex or alternative biological hypotheses. More generally, because of its explicit biological formulation, our model can be easily extended to test for stabilizing vs. diversifying selection. We conjecture that our model could also be extended include epistasis in a more realistic manner than Ising models, while requiring many fewer parameters than Potts models.

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

HIV's protein sequence has been described as having a great deal of 'genetic plasticity' due to the fact that amino acid substitutions at many different sites appear to have little or no consistent effect on HIV fitness (Lemey et al., 2006; Salemi, 2013; Cuevas et al., 2015; Rihn et al., 2013). Such genetic plasticity allows HIV populations to evade or 'escape' the patient's immune response by simply evolving to a new location in epitope space. HIV's remarkable ability to escape the host's immune response has impeded efforts to create an effective vaccine (Goulder and Watkins, 2004; Autran et al., 2008; Johnston and Fauci, 2008).

In response, researchers have tried to identify amino acid sites under strong and consistent stabilizing selection (Ferguson et al., 2013; Liu et al., 2013; Barton et al., 2016a), believing them to be promising vaccine targets (Rolland et al., 2013). In order to identify these sites, researchers have attempted to estimate HIV's fitness landscape (Deforche et al., 2008; Seifert et al., 2015; Kouyos et al., 2012; Hinkley et al., 2011; Shekhar et al., 2013; Ferguson et al., 2013; Mann et al., 2014; Lorenzo-Redondo et al., 2014; Moradigaravand et al., 2014; Barton et al., 2016a). Since first introduced by Wright (1932), fitness landscapes have become a conceptual cornerstone within the field of evolutionary biology (e.g. Lande and Arnold, 1983; Lande, 1985, 1986; Kauffman and Levin, 1987; Charlesworth and Rouhani, 1988; Kauffman, 1993; Niklas, 1994; Gavrilets, 1997; Fontana, 2002; Berg and Lässig, 2003; Gavrilets, 2004; Berg et al., 2004; Wilke and Drummond, 2006; Gilchrist, 2007; Lässig, 2007; Calcott, 2008; Mustonen and Lässig, 2009; Draghi et al., 2010; Gilchrist et al., 2009; Wallace et al., 2013; Gilchrist et al., 2015). In the simplest scenarios where the evolutionary process has either reached stationarity or we have no *a priori* information, the probability of observing a sequence i is proportional to its evolutionary fitness W_i raised to the effective population size N_e, i.e. $p_i \propto W_i^{N_e}$ (Wright, 1969; Iwasa, 1988; Berg and Lässig, 2003; Sella and Hirsh, 2005; McCandlish et al., 2015).

While often unaware of its theoretical foundation, HIV researchers have extensively used this link between the fitness contribution of an amino acid residue at a particular site and its observed genotype frequency (e.g. Rihn et al., 2013; Liu et al., 2006). For example, researchers have adapted the Potts and Ising models from statistical mechanics to quantify direct and epistatic fitness effects between amino acid sites (Ferguson et al., 2013; Mann et al., 2014; Barton et al., 2016a). While the Ising model, which categorizes amino acid residues as optimal or non-optimal, has been criticized as being overly simplistic, it has been effectively employed to identify epistatic interactions between sites (Ferguson et al., 2013; Mann et al., 2014). Instead of using a simple optimal/non-optimal categorization, the Potts model categorizes each of the 20 amino acid separately. While this greater categorization makes the Potts model more biologically realistic, it also makes it hard to fit. Specifically, in order to describe epistatic effects between any two sites, the Potts model requires the estimation of hundreds of parameters. As a result, confidently fitting this model to multiple

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

sequence alignments is infeasible, even with large data sets (Barton et al., 2016a). Researchers have also used Shannon entropy E, specifically the inverse entropy of a site $1/E$ (i.e. its 'conservation'), as a measure of the strength of consistent stabilizing selection for the presumably optimal consensus amino acid residue (Dietrich and Skipper, 2012; Acevedo et al., 2014). In this approach, sites with little variation in amino acid usage have high $1/E$ values and are, in turn, inferred to be experiencing strong and consistent stabilizing selection. In contrast, sites with substantial variation have low $1/E$ values and are inferred to be either under weak or variable stabilizing selection. As we show later, the conservation of a site $1/E$ is proportional to the expected log-likelihood of observing a randomly chosen amino acid under a saturated, multinomial model parameterized with a given data set.

Given its definition, site conservation $1/E$ can is best viewed as a summary statistic quantifying the ruggosity of HIV's fitness landscape (Kauffman, 1993) rather than a direct measure of consistent stabilizing selection for an optimal amino acid. Another shortcoming of $1/E$ as a biologically meaningful metric is the fact that it treats all amino acid residues as being equally dissimilar from the one another. As a result, $1/E$ ignores the fact that amino acid residues have differing degrees of physico-chemical dissimilarity. This is undesirable given that the physico-chemical properties of amino acids clearly affect the probability one amino acid will substitute for another (Grantham, 1974; Wilke and Drummond, 2010). As a result, $1/E$ has the potential to miss sites where there is strong, consistent stabilizing selection for a set of physico-chemical properties, but where these properties can be reasonably satisfied by more than one amino acid residue. More generally, the fact that $1/E$ ignores the physico-chemical properties of amino acid residues suggests it is glossing over information embedded within the data.

Despite the fact that site conservation $1/E$ ignores the physico-chemical characteristics and is not actually a measure of consistent stabilizing selection against non-optimal amino acid residues, it is widely used. One contributing factor to $1/E$'s wide use is the fact that it can be easily calculated from sequence alignment data (c.f. the Potts model). A second contributing factor to $1/E$'s popularity is its utility. Even though $1/E$ is a coarse meteric, it has proven useful for identifying 'fragile' sites, i.e. those with low genetic plasticity.

In an effort to overcome these aforementioned shortcomings, we introduce a model that is more biologically detailed than $1/E$, but less parameter rich than the Potts model. First, instead of treating amino acids residues as equally dissimilar categories, our model uses a subset set of physico-chemical properties and weighting terms to describe the differences between residues. Second, it explicitly estimates the sensitivity of viral fitness to deviations from the physico-chemical of the optimal amino acid for a given site. While the site sensitivity parameters G' we estimate correlate well with site conservation $1/E$, G' does a slightly better job predicting HIV genotype fitness from both *in vivo* and *in vitro* studies. Unlike site conservation $1/E$, our model allows us to test biologically motivated hypotheses. For example, we find that our physico-chemical

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

weightings and the distributions of G' vary between the different protein regions of Gag.

Taken together, our physico-chemical model helps advance researchers ability to extract information from observational data and represents a biologically grounded framework which can be further extended and used to test clearly posed hypotheses. In its current form, our physico-chemical model only considers site independent effects and can be viewed as a constrained version of the more general, parameter saturated, entropy model underlying the calculation of $1/E$. Given the parallels between evolution and statistical mechanics (Sella and Hirsh, 2005), it should be possible to extend our physico-chemical model to include epistatic effects in a more realistic manner than the Ising models, which ignore physico-chemical properties, but in a substantially more efficient manner, in terms of the number of epistatic parameters and the ease of parameter estimation, than the physico-chemical informed Potts models.

Methods

We begin by presenting our physico-chemical model and, in the process, clearly define our site sensitivity parameter G'. Next, we review how site conservation $1/E$ is defined using Shannon entropy. We then clearly show the link between the physico-chemical model and the entropy model and their corresponding probabilities of observing each amino acid residue at a particular site. Finally, we describe the data and methods we used to parameterize the physico-chemical and entropy models and evaluate their ability to predict empirical measurements of HIV fitness. Definitions of all of our model parameters can be found in Table 1.

Modeling the HIV Fitness Landscape

In this study, we focus on two structurally similar models for describing a HIV fitness landscape: our physico-chemical based approach and Shannon entropy. Both models assume the fitness landscape is fixed and that each amino acid site affects viral fitness independent of the others. In the physico-chemical model the expected frequencies of the different amino acid residues are determined by their physico-chemical properties, the optimal residue for that site a_*, and the strength of consistent stabilizing selection G for a_*. These expected frequencies also depend on the physico-chemical weights $\vec{\theta}$, which are shared across a set of sites.

In contrast to our physico-chemical model, in the Shannon entropy model there are no shared parameters between sites, the optimal amino acid residue is assumed to be the most frequent one, and the frequency of the 19 other non-optimal amino acids residues are completely unconstrained and set to equal their observed frequencies for that site. Technically, the Shannon entropy model could be viewed as an unconstrained version of our physico-chemical model. In the Shannon entropy model there are 19 free parameters per site.

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

In contrast, in our physico-chemical model there are only 2, a_* and G, along with a minimum of 3 global parameters describing physico-chemical weights and a hierarchical distribution for G. We use AIC for model comparisons.

A Physico-Chemical Model

For a given site, we assume the fitness w of amino acid residue a_j declines exponentially as a product of a site specific sensitivity parameter G_i and the distance d, in physico-chemical space, of a_j from the optimal amino acid residue for that site a_*. That is,

$$w\left(a_j \middle| a_*, G_i, \vec{\theta}\right) \propto \exp\left[-G_i d\left(a_j, a_* \middle| \vec{\theta}\right)\right]. \tag{1}$$

We use the euclidean distance function

$$d\left(a_j, a_* \middle| \vec{\theta}\right) = \sqrt{\theta_c\left(\left(c_j - c_{a_*}\right)/\sigma_c\right)^2 + \theta_l\left(\left(l_j - l_{a_*}\right)/\sigma_l\right)^2 + \theta_v\left(\left(v_j - v_{a_*}\right)/\sigma_v\right)^2}, \tag{2}$$

the physico-chemical weighting terms $\vec{\theta} = (\theta_c, \theta_l, \theta_v)$ represent *a priori* unknown weights for the known amino acid residue physico-chemical properties: the ratio of carbon to non-carbon atoms or 'composition' c, polarity p, and molecular volume v. The physico-chemical weights $\vec{\theta}$ are are assumed to be shared across multiple sites and are estimated from the data. In order to account for the fact that physico-chemical properties are measured in different units and facilitate their interpretation of our weight terms $\vec{\theta}$, the differences in physico-chemical properties are scaled by the standard deviation of this property observed across the 20 canonical amino acids, i.e. σ_i. As a result of this scaling, the weighting term describe the sensitivity of protein function to the deviation in physico-chemical properties between amino acid relative to the total amount of variation possible. Our choice of physico-chemical properties follows Grantham (1974); other physico-chemical properties could be used instead (Sharma et al., 2013). Because G' is always multiplied by the distance function d in our physico-chemical model, there is an inherent lack of identifiability of these terms. To solve this problem and facilitate ease of interpretation, $\vec{\theta}$ was constrained so that the sum of its values equaled 1.

Following the example of other researchers in this field (Ferguson et al., 2013; Mann et al., 2014; Barton et al., 2016a) we treat HIV sequence data from different patients as independent samples from the evolutionary stationary distribution as described by Sella and Hirsh (2005). We define the composite parameter G'_i as the product of the site sensitivity and the effective population size N_e, i.e. $G'_i = G_i \times N_e$. For now we ignore the effects of mutation bias. In order to avoid issues that would result from the infinite MLE of G' at invariant sites, we employed hierarchical approach where G' is assumed to follow either a LogNormal or

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Gamma distribution. As a result, the parameters π_1 and π_2 represent the shape and scale parameters or the log scale mean and standard deviation for G', respectively. We allowed the G distributions and parameters to be shared or vary between each of the gag poly-peptide regions: the nucleo protein p6, nucleocapsid protein p7, matrix protein p17, and capsid protein p24. As a result, the probability p of observing amino acid residue j, at site i is,

$$p_j\left(a_*, G_i, \vec{\theta}\right) = \frac{\exp\left[-G_i' d\left(a_j, a_* \middle| \vec{\theta}\right)\right]}{\sum_k \exp\left[-G_i' d\left(a_k, a_* \middle| \vec{\theta}\right)\right]} f\left(G_i' | \pi_1, \pi_2\right) \tag{3}$$

Based on our model's assumptions and structure, G_i' represents a quantitative measure of the strength and efficacy of consistent stabilizing selection on a given site relative to genetic drift.

Given these assumptions, the probability of observing a set of amino acid residue counts $\vec{x} = \{x_1, x_2, \ldots, x_{20}\}$ at a given site i follows a multinomial distribution. That is,

$$\Pr\left(\vec{x} \middle| a_*, G_i', \vec{\theta}\right) = \binom{x_T}{x_1\ x_2\ \ldots x_{20}} \prod_{j=1}^{20} p_j\left(a_*, G_i', \vec{\theta}\right)^{x_j}$$

where $x_T = \sum_j x_j$ is the total number of observations made at site i and and a_* is the optimal amino acid for site i. Correspondingly, the Log-Likelihood L of the data $\vec{x}$ as a function of our physico-chemical the model parameters, a_*, G_i', and $\vec{\theta}$, is,

$$L\left(a_*, G_i', \vec{\theta} \middle| \vec{x}\right) = \sum_{j=1}^{20} x_j \ln\left[p_j\left(a_*, G_i', \vec{\theta}\right)\right] \tag{4}$$

By maximizing Eq.(4) with respect to our model parameters, we can identify the most likely parameter values and our confidence in them: the physico-chemical weights $\vec{\theta}$, the optimal amino acid residue at each site a_*, and the N_e scaled sensitivity of HIV fitness to deviations in physico-chemical space for each site G'. By linking fitness to the physico-chemical qualities of an amino acid residue, we effectively reduce the number of parameters in our multinomial model from 19 parameters per site to 2 parameters per site, G_i' and a_* plus, depending on how we partition the data, 2 or 8 shared physico-chemical weight parameters $\vec{\theta}$. As a result, our physico-chemical model is highly 'unsaturated'.

The Entropy Model

While Shannon entropy is a measure of information of a message, it is also proportional to the log of the probability of the data $\vec{x}$ under a multinomial model where the probability of each category p_j is equal to its observed relative frequency x_j/x_T where, as before, x_T is the number of observations. Setting $p_j = x_j/x_T$

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

not only makes intuitive sense, it is also equal to the maximum likelihood estimate (MLE) of p_j under a multinomial model. Thus Shannon entropy of a given site i with a set of counts $\vec{x}$ can be written as

$$E\left(\vec{x}\right) = -\sum_{j=1}^{20} x_j/x_T \ln\left(x_j/x_T\right)$$

and is related to the probability of the data $\vec{x}$ at the maximum likelihood values of its parameters $\vec{p}$ under a multinomial model

$$\Pr\left(\vec{x}\middle|\vec{p} = \frac{\vec{x}}{x_T}\right) = \binom{x_T}{x_1\ x_2\ \ldots x_{20}} \exp\left[-x_T E\left(\vec{x}\right)\right] \tag{5}$$

Correspondingly, the maximum Log-Likelihood of the data $\vec{x}$ under the Shannon entropy model is,

$$L\left(\vec{p} = \frac{\vec{x}}{x_T}\middle|\vec{x}\right) = -x_T E\left(\vec{x}\right) = \sum_{j=1}^{20} x_j \ln\left(x_j/x_T\right) \tag{6}$$

The entropy of a site $E(\vec{x})$ can be interpreted in a number of different ways. One interpretation of $E(\vec{x})$ is as a diversity index for amino acid residues at a site or region (Jost, 2006). Another interpretation is, under certain conditions, $E(\vec{x})$ is proportional to the expected number of guesses one must make to infer the state of a site. Equivalently, $E(\vec{x})$ is also equal to the the mean contribution of a category to a site's log-likelihood.

Despite having many potential interpretations, because of its descriptive and 'many to one' nature, there is no clear way of linking the Shannon entropy of a site E and the strength of purifying selection for or against a particular amino acid. Finally, the inverse of E is used to describe the conservation of a site or region $1/E$ which has, in turn, been used as a heuristic measure of the strength of consistent stabilizing selection for the optimal amino acid residue at a site (Allen et al., 2005; Liu et al., 2012a). In addition, because the sum of p_i must equal one, from a likelihood perspective the entropy model has $20 - 1 = 19$ free parameters per site, making it a fully saturated model.

Data

Amino Acid Sequences

To parameterize the models we used the Gag poly-peptide of HIV subtype C MSA. We excluded linker regions and, as a result, analyzed 520 sites using 1058 curated sequences from the Los Alamos National Laboratory (LANL) database (Biophysics Group Los Alamos National Lab, 2016). This database contains filtered web alignments curated by the biophysics group at LANL. Details concerning the curation and filtering of the sequences in the alignment process can be found at `http://www.hiv.lanl.gov/`. Data was downloaded in

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

September 2016. Sequences were processed to obtain counts for amino acid residue j at site i for each HXB2 site.

Empirical Evaluation of Model Predictions

In order to test whether our site sensitivity parameter G' is more biologically informative than the standard conservation metric $1/E$, we evaluated its ability to predict empirical observations of *in vivo* and *in vitro* viral fitness. To test our physico-chemical model's ability to predict *in vivo* behavior of HIV, we used viral escape times from a patient's CD8 T-cell response as estimated by Barton et al. (2016a) using the patient data from Liu et al. (2012b). In that study, escape times of viral epitopes from the patient's immune response were estimated using sequential sequencing data from reactive epitopes sites in multiple HIV-1 subtype B and C infected patients over 3 years. These researchers found that that the mean conservation value $\overline{1/E}$ for an epitope is positively correlated with escape time. Epitopes were defined as 8-11 amino acid long regions of the proteome recognized by patient specific T cell immune response. Escape time t was defined as the number of days between detection of the T-cell response and the time viral variants bearing that respective reactive epitope fell below 50%. Because of limited sampling, the exact date at which a patient's reactive epitope response fell below the 50% threshold was not actually observed. Instead, Barton et al. (2016a) used a non-linear model to estimate the escape time. In order minimize the uncertainty in these estimates of escape time, we excluded data from patients who had already met the criteria before a patient's first sample or whose escape date estimated by extrapolation after a patient's final sample. In addition, because we fitted our model to HIV subtype C data, we only used epitopes whose consensus sequence was the same between subtypes B and C. As a result, we were left with only 14 estimates of escape times. Because these estimated escape times are defined at the epitope level, we compared escape time to the mean site sensitivity G' or $\bar{G}'$. Similarly, we also compared estimated escape times to the mean log conservation for each epitope $\overline{1/E}$.

To test our physico-chemical model's ability to predict *in vitro* data, we used viral replication rate in cell culture as a measure of HIV fitness. Rihn et al. (2013) estimated the replicative fitness of 31 HIV subtype B viruses bearing mutated capsid CA amino acid residues via spreading replication assay on human MT4 T-cell lines and peripheral blood monocytes. The CA mutants in this study were generated by creating a mutagenized CA library using a low fidelity PCR approach and then inserting the mutated CA sequences in replication competent proviral clones. Fitness was reported as % of the wild-type replication.

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Model Fitting and Inference

The model was fitted to the data using Mathematica 11 (Wolfram Research Inc., 2017). We used the built in, optimization function NMaximize[ ] with the "NelderMead" option to optimize all of the parameters other than the site specific G'. Within each evaluation of NMaximize[ ], we used FindMaximum[ ] to identify the optimal G'_i for each site. Using this approach, the model was independently fitted to each region from 200 different initial sets of parameter values. The first 100 fittings used initial values generated by fitting the model to only variable sites and a uniform hierarchical distribution for G'. For each of these fittings the estimated parameter values were multiplied by uniform random numbers to further vary their starting values. For the second 100 fittings, the initial values were generated using different random seeds for each run of NMaximize[ ].

We calculated the Fisher Information Matrix for the $\vec{\theta}$ for each region and used it to calculate the 95% Confidence Intervals (CI) for each of its parameter following Bolker (2008, p.197-200). We used these CI to quantify our uncertainty in our estimates and to compare parameters across regions. If the CI for the same parameter in two regions do not overlap, we concluded that the $p-$value for the hypothesis the regions shared the same parameter values was < 0.05.

In order to evaluate how well our physico-chemical model's site sensitivity G' and the Shannon entropy model's site conservation $1/E$ predicted available *in vivo* and *in vitro* fitness measures. Neither our G' or $1/E$ estimates nor our empirical *in vivo* and *in vitro* fitness estimates appear normally distributed; therefore, we used Kendall's non-parametric rank correlation τ to measure the association between our consistent stabilizing selection metrics and the empirical data. Significance tests for our τ estimates were done using Mathematica's KendallTauTest[]. Test statistics were generated by bootstrapping the data 10,000 times with replacement via the "Permutation" option.

Results

Briefly, we find that our population genetics and physico-chemical based site sensitivity terms G' are well correlated with the more commonly used entropy based conservation metric $1/E$. In terms of fitting the data, the entropy model, a saturated parameter model, has an astronomically better AIC score than our best fitting version of our physico-chemical model where Grantham weights and the distribution of site sensitivities vary by region (ΔAIC = 230,007). This poor performance of our model is discouraging, but is not surprising given the fact that the entropy model is a purely descriptive model and fits one parameter for every unconstrained data bin. As a result, the entropy model, by definition, produces the largest L value

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

possible given the data. This flexibility comes the cost, in terms of AIC, of a large number of estimated parameters: 19/site $\times$ 462 sites = 8778 parameters. This cost, however, is mitigated by the fact that we have more than a 1000 observations for each site. In contrast to its poorer AIC value, we find that our physico-chemical model does a better job predicting empirical *in vivo* and *in vitro* measurements of HIV fitness, if only slightly so. Further, our parameter estimates can be used to test more refined hypotheses such as whether the properties of natural selection, as described by our physico-chemical weights $\vec{\theta}$ and G' values, varies between protein regions. Thus, while our physico-chemical model appears to capture important biological information embedded in the data, there is clearly much room for future improvement.

Regional Variation in Model Parameters The variation and uncertainty in our estimates of $\vec{\theta}$ can be found in Table 2 and Figure S1. Although allowing the physico-chemical weights $\vec{\theta}$ and G' distributions to vary between poly-peptide regions required the addition of just 12 additional parameters and vastly improved the ability of our physico-chemical model to fit the sequence data by 1412 log-likelihood units ($\chi^2_{12} = 2824, p < 10^{-300}$). Surprisingly, despite this vast improvement in model fit, the differences in $\vec{\theta}$ were actually quite small. Further, given the sensitivity of our L function to $\vec{\theta}$, it is perhaps a bit surprising that the choice of a LogNormal or Gamma distribution for G' had no discernible effect on our estimates of a region's $\vec{\theta}$. This sensitivity is also reflected by fact that our CI for these parameters are on the order of $< 1\%$. These results clearly indicate that the effects of amino acid substitutions vary between protein regions. Consequently, all of the results we discuss below come from the model fit where $\vec{\theta}$ varies between proteins.

In terms of describing the distribution of site sensitivities G' across a given poly-peptide region, the LogNormal distribution performed substantially better than the Gamma distribution with the p6 nucleo protein and p24 matrix protein (ΔAIC = 7.2 and 15.6, respectively). The LogNormal distribution mean parameter μ was indistinguishable between these two regions, but the standard deviation parameter σ was significantly lower in p6 than p24. In contrast, the Gamma distribution performed substantially better than the LogNormal distribution for the p7 nucleocapsid protein and p17 capsid protein (ΔAIC = 9.4 and 24.8, respectively). The Gamma distribution shape parameter α was indistinguishable between these two regions, but the rate parameter β was significantly lower in p7 than p17. Taken together, these results indicate that the distribution of G' values varies between protein regions. This is in spite of the fact that the first two central moments of p6, p7, and p17 are statistically indistinguishable (Fig. 1 and Table S1).

Site Sensitivity G' vs. Site Conservation $1/E$ Kendall's rank correlation τ between our model's site sensitivity parameter G' and the entropy model's site conservation $1/E$ metric indicate that they are well correlated with one another ($\tau = 0.598 - 0.720$ with $p < 10^{-10}$ for all regions, Fig 2).

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Predicting Empirical HIV Fitness Data As expected, both site conservation $1/E$ and site sensitivity G' were positively correlated with the *in vivo* measure of fitness: escape time ($\tau = 0.291$ and 0.475, respectively; Fig 3 and Table S2). However, only the correlation between G' and escape time was significant ($p = 0.016$ vs. 0.141 for $1/E$).

Both G' and $1/E$ were also positively correlated with changes to the capsid protein on *in vitro* replication fitness ($\tau = 0.238$ and 0.211, respectively ; Fig. 4). However, as with the escape time predictions, only the G''s τ was significant ($p = 0.047$ vs. 0.080 for $1/E$). Similarly, G' and $1/E$ were also positively correlated with changes to the capsid protein on *in vitro* spreading fitness ($\tau = 0.231$ and 0.157), but neither predictor's τ was significantly greater than 0 ($p = 0.055$ and 0.192, respectively).

Discussion

Although initially introduced as a metaphor for describing how evolution works, fitness landscapes have proven to be useful tools in theoretical and empirical research (for example, Kauffman, 1993; Wilke and Drummond, 2006; Gilchrist, 2007; Gilchrist et al., 2009; Wallace et al., 2013; Gilchrist et al., 2015, and citations below). Ideally, HIV fitness landscapes can help researchers predict when and where a virus will escape immune control (Barton et al., 2016a) and develop effective vaccines (Ferguson et al., 2013; Shekhar et al., 2013). In a few cases HIV fitness landscapes have been estimated directly from experimental data (Hinkley et al., 2011; Kouyos et al., 2012; Mann et al., 2014; Rihn et al., 2013), but in most studies the landscape is inferred from the ever growing libraries of genotype frequency data (Barton et al., 2016a; Zanini et al., 2016; Mann et al., 2014; Ferguson et al., 2013; Deforche et al., 2008; Seifert et al., 2015).

In this study we utilize fundamental findings from the field of population genetics to infer HIV Gag's poly-peptide physico-chemical fitness landscape using count data from the LANL HIV database. In contrast to explicitly mapping out HIV's fitness landscape, other researchers have used site conservation $1/E$, where E is the Shannon entropy of a site, as a proxy for the strength of consistent stabilizing selection (Rihn et al., 2013; Ferrari et al., 2011; Liu et al., 2012b). While this interpretation may seem intuitive, $1/E$ is actually a summary statistic rather than a measure of the selection coefficient between an optimal amino acid and its alternatives. Further, entropy based metrics such as $1/E$ result from fitting saturated models, where the number of parameters is on par with the number of data categories. As a result, it is perhaps not surprising that the AIC for the entropy model fit is still substantially better than its value for our highly unsaturated physico-chemical model (ΔAIC = 230,007). We note that because, $1/E = \infty$ at invariant sites, $1/E$ for an invariant site will change dramatically if new data includes even a single alternative amino acid at that site. In contrast, because we assume G' comes from a probability distribution whose parameters are estimated

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

simultaneously with our G' values, the metric will not change as dramatically if an alternative amino acid is eventually observed at that site.

While saturated models have the advantage of being maximally flexible in terms of fitting data, this flexibility comes at the cost of being minimally informative about the processes generating the data. For example, our physico-chemical model's site sensitivity G' did a better job than the entropy model's site conservation $1/E$ in predicting *in vivo* and *in vitro* HIV fitness measurements (Figs 3 and 4). These results suggest that our physico-chemical model is more efficient at extracting biological meaningful information from the sequence data we used to fit our models.

In addition to extracting more meaning from the data, because our physico-chemical model is derived from biological principles, it can be used to evaluate biologically based hypotheses. For example, we find that the impact of our three different physico-chemical traits $\vec{\theta}$ on fitness varies substantially between between protein regions (Table 2). These results indicates the effects of substituting one amino acid residue for another varies with its broader genetic context. This is, perhaps, unsurprising given that previous researchers have found that interior regions of the protein are more sensitive to changes in polarity (Nakai et al., 1988). Nevertheless, our ability to detect these slight regional differences in the character of the consistent stabilizing selection illustrates the statistical power of our physico-chemical modeling approach.

In addition, site sensitivities G' values differ between protein regions in terms of the distribution that best describe the estimated values (LogNormal vs. Gamma). The fact that the G' values of different protein regions follow different distributions suggests that the contribution of each site to protein function varies between sites. For example, the LogNormal distribution suggests the G' values are the result many positive random independent variables acting in a multiplicative manner. In contrast, the Gamma distribution suggests that protein function depends on fluxing through different conformational states with exponential like waiting times for each state. Even amongst the p6 and p24 regions where site sensitivities G' are best described by a LogNormal distribution, they appear to come from distributions with different μ parameters (Table 2). For example, the average G' values for region p24 were greater than p7 and p17, which is consistent with previous findings (Rihn et al., 2013; Martinez-Picado et al., 2006) (Figure 1). Despite our ability to differentiate between different hypothesized distributions for G', our estimates of physico-chemical weightings $\vec{\theta}$ are actually insensitive to these distributional differences (Table 2). One plausible hypothesis for these peptide region specific differences in G' distributions could be their differences in secondary structure. This idea could be tested by determining whether grouping sites by secondary structure (or some other feature such as distance from surface or active site) provides a better fit than grouping sites by protein region as we do here.

While our physico-chemical model improves our ability to predict empirical HIV fitness data, there is still

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

a substantial amount of noise in our predictions. This variation likely has a number of different sources. In terms of the data we are trying to predict, the *in vitro* spreading fitness measures suffer from the unnatural qualities of cell culture. Similarly, the *in vivo* epitope escape fitness measures likely includes substantial effects due to biological variation between patients, is based on a limited number of sample time points, and limited sequencing depth to determine genotype frequencies.

In terms of model shortcomings, there are many. For example, our choice of physico-chemical properties to include in our distance function was based solely on Grantham's classic work (Grantham, 1974). Fortunately, testing the power of other physico-chemical properties is straightforward with our model. Further, in its current form, our physico-chemical model ignores the effect of mutation bias. Mutation bias, i.e. the fact that mutation rates between residues differ from one another, can also contribute to the probability of observing a particular codon and, in turn, amino acid. Although the effects of mutation bias are likely to be unimportant at sites with large site sensitivities G', mutation bias can dominate the evolutionary outcome of sites under weak selection.

Moving on to shortcomings that are more challenging to over come, both the physico-chemical and entropy models assume statistical independence between patient samples. These assumptions also shared by the more complex Ising and Potts models and could be addressed by extending our approach to a phylogenetic framework (Beaulieu et al., In Review). Both the physico-chemical and entropy models also assume a single, invariant fitness peak centered around a site's optimal amino acid residue a_*. While it should be possible to extend our physico-chemical to allow for more than one peak in the amino acid residue landscape by modifying our distance function, we suspect our ability to reject the simpler hypothesis of a single peak would be weak, especially if they occurred in similar points in physico-chemical space. Allowing a_* to explicitly switch over time, would be even more challenging than allowing for multiple fitness peaks. We expect detecting such switches would be very difficult without large, high quality and high resolution datasets.

Perhaps most glaringly shortcoming is the fact that our physico-chemical ignores epistatic effects. Epistatic interactions are likely ubiquitous and have been well documented in the virus literature (Koek et al., 2012; Brockman et al., 2007, 2010). Fortunately, we could extend our physico-chemical model to include the effects of epistatic interactions between sites in a similar manner to the Ising and Potts model. In the same way that we use a physico-chemical distance function and site sensitivity G' to generate the predicted frequency of the 20 canonical amino acid residues, we could define a more complex distance function that would allow us to describe epistatic effects between sites and predict the probability of the 400 different possible site pairs of amino acid residues using a relatively small number of parameters. The end result would ideally be a more realistic model than the Ising model, but one that requires many fewer parameters, is more biologically informative, and is easier to fit than the Potts model.

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

In conclusion, we argue that one promising way to improve our ability to extract biologically meaningful information from sequence databases is to use well defined and biologically grounded models. Here we show how our physico-chemical model can improve our ability predict HIV fitness. Because our model is grounded in the field of population genetics, it is inherently well suited to describe evolutionary data. Because our model is well defined biologically, it provides a clear framework to test specific hypotheses about HIV protein sequence evolution and can serve as the basis for more complex, but parameter limited, models.

Acknowledgements

M.A.G received financial support from NSF grants MCB-1546402 (Primary Investigator: A. VonArnim), MCB-1120370 (Primary Investigator: M.A.G), and DEB-1355033 (Primary Investigator: Brian O'Meara) as well as the Department of Ecology and Evolutionary Biology at the University of Tennessee and the National Institute for Mathematical and Biological Synthesis (NIMBioS, NSF:DBI-1300426 with additional support from the University of Tennessee). E.G.H. received financial support from the Departments of Mathematics and Microbiology at the University of Tennessee and fellowship from NIMBioS. We also gratefully acknowledge assistance from Dr. Vitaly V. Ganusov.

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

Acevedo, A., L. Brodsky, and R. Andino. 2014. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature 505:686–690.

Allen, T. M., M. Altfeld, S. C. Geer, E. T. Kalife, C. Moore, et al. 2005. Selective escape from CD8+ T-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution. Journal of Virology 79:13239–13249.

Autran, B., R. L. Murphy, D. Costagliola, R. Tubiana, B. Clotet, et al. 2008. Greater viral rebound and reduced time to resume antiretroviral therapy after therapeutic immunization with the ALVAC-HIV vaccine (vCP1452). AIDS (London, England) 22:1313–1322.

Barton, J. P., N. Goonetilleke, T. C. Butler, B. D. Walker, A. J. McMichael, et al. 2016a. Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. Nature Communications 7:11660.

Barton, J. P., N. Goonetilleke, T. C. Butler, B. D. Walker, A. J. McMichael, et al. 2016b. Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. Nature Communications 7:11660.

Beaulieu, J., B. O'Meara, C. Landerer, J. J. Chai, and M. A. Gilchrist. In Review. Population genetics based phylogenetics under stabilizing selection for an optimal amino acid sequence: A nested modeling approach. Systematic Biology .

Berg, J. and M. Lässig. 2003. Stochastic evolution and transcription factor binding sites. Biophysics 48:S36–S44.

Berg, J., S. Willmann, and M. Lassig. 2004. Adaptive evolution of transcription factor binding sites. BMC Evolutionary Biology 4:1–12.

Biophysics Group Los Alamos National Lab. 2016. Sequence Alignments.

Bolker, B. 2008. Ecological Models and Data in R. Princeton University Press, Princeton, NJ.

Brockman, M. A., Z. L. Brumme, C. J. Brumme, T. Miura, J. Sela, et al. 2010. Early Selection in Gag by Protective HLA Alleles Contributes to Reduced HIV-1 Replication Capacity That May Be Largely Compensated for in Chronic Infection. Journal of Virology 84:11937–11949.

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Brockman, M. A., A. Schneidewind, M. Lahaie, A. Schmidt, T. Miura, et al. 2007. Escape and Compensation from Early HLA-B57-Mediated Cytotoxic T-Lymphocyte Pressure on Human Immunodeficiency Virus Type 1 Gag Alter Capsid Interactions with Cyclophilin A. Journal of Virology 81:12608–12618.

Calcott, B. 2008. Assessing the fitness landscape revolution. Biology & Philosophy 23:639–657.

Charlesworth, B. and S. Rouhani. 1988. The probability of peak shifts in a founder population .2. an additive polygenic trait. Evolution 42:1129–1145.

Cuevas, J. M., R. Geller, R. Garijo, J. Lpez-Aldeguer, and R. Sanjun. 2015. Extremely High Mutation Rate of HIV-1 In Vivo. PLoS Biology 13.

Davey, N. E., V. P. Satagopam, S. Santiago-Mozos, C. Villacorta-Martin, T. A. M. Bharat, et al. 2014. The hiv mutation browser: A resource for human immunodeficiency virus mutagenesis and polymorphism data. Plos Computational Biology 10:e100395.

Deforche, K., R. Camacho, K. V. Laethem, P. Lemey, A. Rambaut, et al. 2008. Estimation of an in vivo fitness landscape experienced by HIV-1 under drug selective pressure useful for prediction of drug resistance evolution during treatment. Bioinformatics 24:34–41.

Dietrich, M. R. and R. Skipper. 2012. A Shifting Terrain: A Brief History of the Adaptive Landscape. The adaptive landscape in evolutionary biology :3–15.

Draghi, J. A., T. L. Parsons, G. P. Wagner, and J. B. Plotkin. 2010. Mutational robustness can facilitate adaptation. Nature 463:353–355.

Ferguson, A. L., J. K. Mann, S. Omarjee, T. Ndungu, B. D. Walker, et al. 2013. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. Immunity 38:606–617.

Ferrari, G., B. Korber, N. Goonetilleke, M. K. P. Liu, E. L. Turnbull, et al. 2011. Relationship between Functional Profile of HIV-1 Specific CD8 T Cells and Epitope Variability with the Selection of Escape Mutants in Acute HIV-1 Infection. PLOS Pathog 7:e1001273.

Fontana, W. 2002. Modelling 'evo-devo' with RNA. Bioessays 24:1164–1177.

Gavrilets, S. 1997. Evolution and speciation on holey adaptive landscapes. Trends In Ecology & Evolution 12:307–312.

Gavrilets, S. 2004. Fitness Landscapes and the Origin of Species. No. 41 in Monographs in Population Biology, Princeton University Press, Princeton, NJ.

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Gilchrist, M., P. Shah, and R. Zaretzki. 2009. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. Genetics 183:1493–1505.

Gilchrist, M. A. 2007. Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. Molecular Biology and Evolution 24:2362–2373.

Gilchrist, M. A., W.-C. Chen, P. Shah, C. L. Landerer, and R. Zaretzki. 2015. Estimating gene expression and codon-specific translational efficiencies, mutation biases, and selection coefficients from genomic data alone. Genome Biology and Evolution 7:1559–1579.

Goulder, P. J. R. and D. I. Watkins. 2004. HIV and SIV CTL escape: implications for vaccine design. Nature Reviews Immunology 4:630–640.

Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862–864.

Hinkley, T., J. Martins, C. Chappey, M. Haddad, E. Stawiski, et al. 2011. A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. Nature Genetics 43:487–489.

Iwasa, Y. 1988. Free fitness that always increases in evolution. Journal of Theoretical Biology 135:265–281.

Johnston, M. I. and A. S. Fauci. 2008. An HIV Vaccine Challenges and Prospects. New England Journal of Medicine 359:888–890.

Jost, L. 2006. Entropy and diversity. Oikos 113:363–375.

Kauffman, S. and S. Levin. 1987. Towards a general-theory of adaptive walks on rugged landscapes. Journal of Theoretical Biology 128:11–45.

Kauffman, S. A. 1993. Origins of Order. Oxford University Press.

Kouyos, R. D., G. E. Leventhal, T. Hinkley, M. Haddad, J. M. Whitcomb, et al. 2012. Exploring the Complexity of the HIV-1 Fitness Landscape. PLoS Genet 8:e1002551.

Koek, M., S. Henke, K. G. akov, G. B. Jacobs, A. Schuch, et al. 2012. Mutations in HIV-1 gag and pol Compensate for the Loss of Viral Fitness Caused by a Highly Mutated Protease. Antimicrobial Agents and Chemotherapy 56:4320–4330.

Lande, R. 1985. Expected time for random genetic drift of a population between stable phenotypic states. Proceedings Of The National Academy Of Sciences Of The United States Of America 82:7641–7645.

Lande, R. 1986. The dynamics of peak shifts and the pattern of morphological evolution. Paleobiology 12:343–354.

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Lande, R. and S. J. Arnold. 1983. The measurement of selection on correlated characters. Evolution 37:1210–1226.

Lässig, M. 2007. From biophysics to evolutionary genetics: statistical aspects of gene regulation. BMC Bioinformatics 8 Suppl 6:S7.

Lemey, P., A. Rambaut, and O. G. Pybus. 2006. HIV evolutionary dynamics within and among hosts. AIDS reviews 8:125–140.

Liu, M. K., N. Hawkins, A. J. Ritchie, V. V. Ganusov, V. Whale, et al. 2013. Vertical t cell immunodominance and epitope entropy determine hiv-1 escape. The Journal of Clinical Investigation 123:380–393.

Liu, M. K., N. Hawkins, A. J. Ritchie, V. V. Ganusov, V. Whale, et al. 2012a. Vertical t cell immunodominance and epitope entropy determine hiv-1 escape. Journal of Clinical Investigation .

Liu, M. K., N. Hawkins, A. J. Ritchie, V. V. Ganusov, V. Whale, et al. 2012b. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. Journal of Clinical Investigation .

Liu, Y., J. McNevin, J. Cao, H. Zhao, I. Genowati, et al. 2006. Selection on the Human Immunodeficiency Virus Type 1 Proteome following Primary Infection. Journal of Virology 80:9519–9529.

Lorenzo-Redondo, R., S. Delgado, F. Morn, and C. Lopez-Galindez. 2014. Realistic Three Dimensional Fitness Landscapes Generated by Self Organizing Maps for the Analysis of Experimental HIV-1 Evolution. PLoS ONE 9:e88579.

Mann, J. K., J. P. Barton, A. L. Ferguson, S. Omarjee, B. D. Walker, et al. 2014. The Fitness Landscape of HIV-1 Gag: Advanced Modeling Approaches and Validation of Model Predictions by In Vitro Testing. PLoS Comput Biol 10:e1003776.

Martinez-Picado, J., J. G. Prado, E. E. Fry, K. Pfafferott, A. Leslie, et al. 2006. Fitness cost of escape mutations in p24 Gag in association with control of human immunodeficiency virus type 1. Journal of virology 80:3617–3623.

McCandlish, D. M., C. L. Epstein, and J. B. Plotkin. 2015. Formal properties of the probability of fixation: Identities, inequalities and approximations. Theoretical Population Biology 99:98–113.

Moradigaravand, D., R. Kouyos, T. Hinkley, M. Haddad, C. J. Petropoulos, et al. 2014. Recombination Accelerates Adaptation on a Large-Scale Empirical Fitness Landscape in HIV-1. PLoS Genet 10:e1004439.

Mustonen, V. and M. Lässig. 2009. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. Trends Genet 25:111–119.

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Nakai, K., A. Kidera, and M. Kanehisa. 1988. Cluster-analysis of amino-acid indexes for prediction of protein-structure and function. Protein Engineering 2:93–100.

Niklas, K. J. 1994. Morphological evolution through complex domains of fitness. Proceedings of the National Academy of Sciences of the United States of America 91:6772–6779.

Rihn, S. J., S. J. Wilson, N. J. Loman, M. Alim, S. E. Bakker, et al. 2013. Extreme genetic fragility of the HIV-1 capsid. PLoS Pathogens :e1003461.

Rolland, M., S. Manocheewa, J. V. Swain, E. C. Lanxon-Cookson, M. Kim, et al. 2013. HIV-1 Conserved-Element Vaccines: Relationship between Sequence Conservation and Replicative Capacity. Journal of Virology 87:5461–5467.

Salemi, M. 2013. The Intra-Host Evolutionary and Population Dynamics of Human Immunodeficiency Virus Type 1: A Phylogenetic Perspective. Infectious Disease Reports 5.

Seifert, D., F. D. Giallonardo, K. J. Metzner, H. F. Gnthard, and N. Beerenwinkel. 2015. A Framework for Inferring Fitness Landscapes of Patient-Derived Viruses Using Quasispecies Theory. Genetics 199:191–203.

Sella, G. and A. E. Hirsh. 2005. The application of statistical physics to evolutionary biology. Proceedings of the National Academy of Sciences of the United States of America 102:9541–9546.

Sharma, A., K. K. Paliwal, A. Dehzangi, J. Lyons, S. Imoto, et al. 2013. A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition. BMC Bioinformatics 14:233.

Shekhar, K., C. F. Ruberman, A. L. Ferguson, J. P. Barton, M. Kardar, et al. 2013. Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics 88:062705.

Wallace, E. W. J., E. M. Airoldi, and D. A. Drummond. 2013. Estimating selection on synonymous codon usage from noisy experimental data. Molecular Biology And Evolution 30:1438–1453.

Wilke, C. O. and D. A. Drummond. 2006. Population genetics of translational robustness. Genetics 173:473–481.

Wilke, C. O. and D. A. Drummond. 2010. Signatures of protein biophysics in coding sequence evolution. Current Opinion In Structural Biology 20:385–389.

Wolfram Research Inc. 2017. Mathematica 11.

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Wright, S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *In* D. F. Jones, ed., Proceedings of the Sixth International Congress on Genetics, vol. 1. Austin, TX.

Wright, S. 1969. Evolution and the genetics of populations. Vol. 2. The theory of gene frequencies., vol. 2. University of Chicago Press.

Zanini, F., V. Puller, J. Brodin, J. Albert, and R. Neher. 2016. In-vivo mutation rates and fitness landscape of HIV-1. bioRxiv :045039.

Zar, J. 1999. Biostatistical Analysis. Prentice Hall, Upper Saddle River, NJ, 4th ed.

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Tables

491

Table 1. Symbols used in this work and their verbal definitions.

Symbol	Definition
w_{ij}	Evolutionary fitness of amino acid residue j at site i.
N_e	Effective population size.
a_*	The optimal amino acid residue for site i.
G_i	Sensitivity of fitness to deviation from a_* at site i.
G'_i	Product of G_i and N_e.
$d(a_j, a_*)$	physico-chemical distance between residue j and a_*.
$\theta_c, \theta_l, \theta_v$	Weighting terms for physico-chemical properties composition c, polarity p, and volume v.
π_1, π_2	Parameters for G' $\sim$LogNormal$(\mu = \pi_1, \sigma = \pi_2)$ or Gamma$(\alpha = \pi_1, \beta = \pi_2)$.
p_j	Probability of observing amino acid residue j at a given site.
x_j	Count of amino acid residue j plus 1 at a given site.
$\vec{x}$	Vector of amino acid residue counts at a given site.
$E(\vec{x})$	Shannon entropy for a set of residue counts $\vec{x}$ at a given site.
$1/E$	Measure of conservation of site as inverse of Shannon entropy.
L	The log-likelihood function in equations (4) and (6).

Table 2. Differences in AIC values ΔAIC and MLE $\pm$95%CI of physico-chemical model parameters for each poly-peptide region under Log-Normal and Gamma distributions of site sensitivities G'. Note that because both distributions have two parameters, their ΔAIC values are simply twice their differences in L. The parameters θ_c and θ_l, represent physico-chemical model weights for residue composition and polarity, respectively. By definition, $\theta_v = 1 - (\theta_c + \theta_l)$ and is not presented. For $G' \sim$ LogNormal$(\mu = \pi_1, \sigma = \pi_2)$ and for $G' \sim$ Gamma$(\alpha = \pi_1, \beta = \pi_2)$ where β is the rate parameter. 95%CI calculated using hessian of LLik surface at MLE (Bolker, 2008, p.197-200).

Region	Sites	Distribution	ΔAIC	θ_c	θ_l	π_1	π_2
p6	44	LogNormal	0	0.280 ± 0.006	0.495 ± 0.006	1.973 ± 0.100	0.803 ± 0.071
		Gamma	7.176	0.279 ± 0.006	0.495 ± 0.006	1.607 ± 0.261	6.288 ± 1.196
p24	231	LogNormal	0	0.408 ± 0.004	0.411 ± 0.003	2.669 ± 0.044	0.798 ± 0.032
		Gamma	15.648	0.408 ± 0.004	0.412 ± 0.003	1.804 ± 0.131	10.802 ± 0.920
p7	55	Gamma	0	0.729 ± 0.001	0.176 ± 0.001	2.036 ± 0.300	7.880 ± 1.329
		LogNormal	9.362	0.729 ± 0.001	0.177 ± 0.005	2.511 ± 0.095	0.829 ± 0.066
p17	132	Gamma	0	0.356 ± 0.004	0.374 ± 0.004	0.666 ± 0.004	18.032 ± 0.010
		LogNormal	24.760	0.354 ± 0.004	0.375 ± 0.004	2.097 ± 0.070	0.973 ± 0.052

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figures

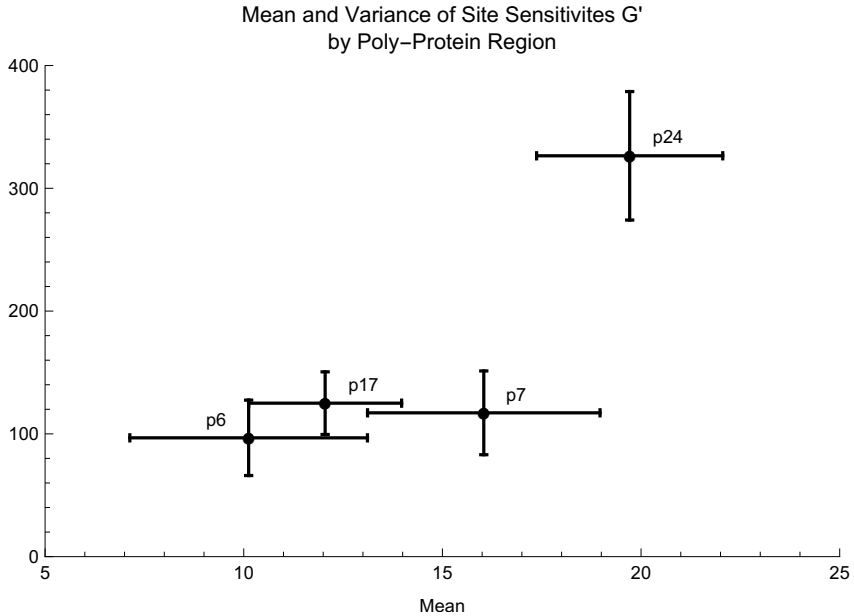

Fig 1. Comparison of estimates of mean and variance of G' for poly-peptide regions. Error bars represent parameter 95% CI. CI estimated using the t distribution and the likelihood ratio test squares after Zar (1999, p. 99 and 111). Note that even thought the central moments of p6, p7, and p17 are statistically indistinguishable, the model parameters describing their distribution are distinguishable.

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

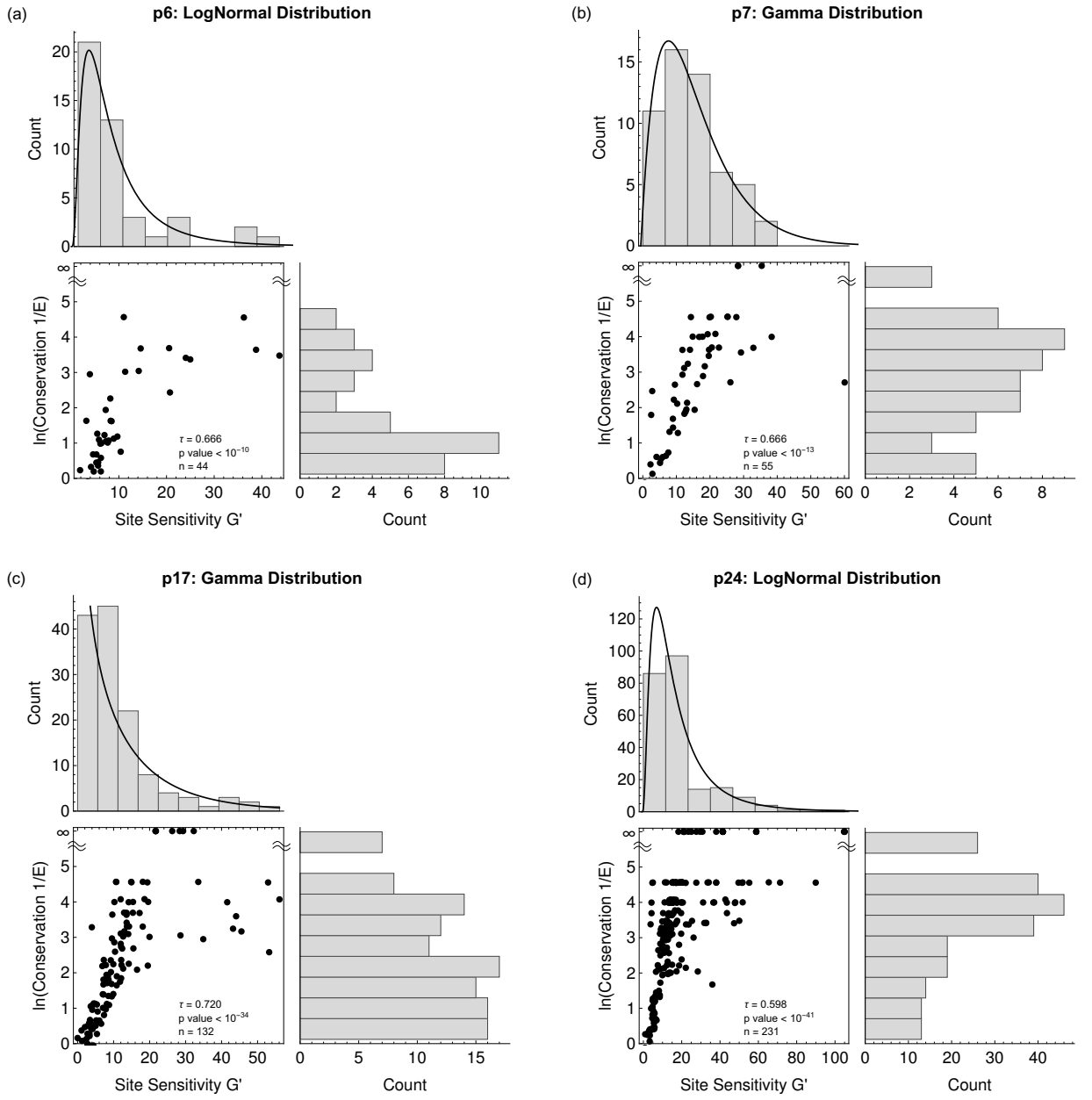

Fig 2. Comparison of site specific physico-chemical sensitivity G' and site conservation $1/E$ for each Gag poly-peptide region using Kendall τ correlation statistic. Invariant sites, by definition, have a $1/E$ value of ∞. In contrast, because our model assumes G' is pulled from either a LogNormal or Gamma probability distribution whose parameters are estimated simultaneously with G', invariant sites have finite G' values. G' values, their distribution, and the parameters used are the best fitting model for each region (Table 2). Regions: (a) nucleo protein p6 with $G' \sim$ Gamma, (b) nucleocapsid protein p7 $G' \sim$ LogNormal, (c) matrix protein p17 $G' \sim$ LogNormal, and (d) capsid protein p24 $G' \sim$ Gamma. Note, n is the number of sites fitted for each region.

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

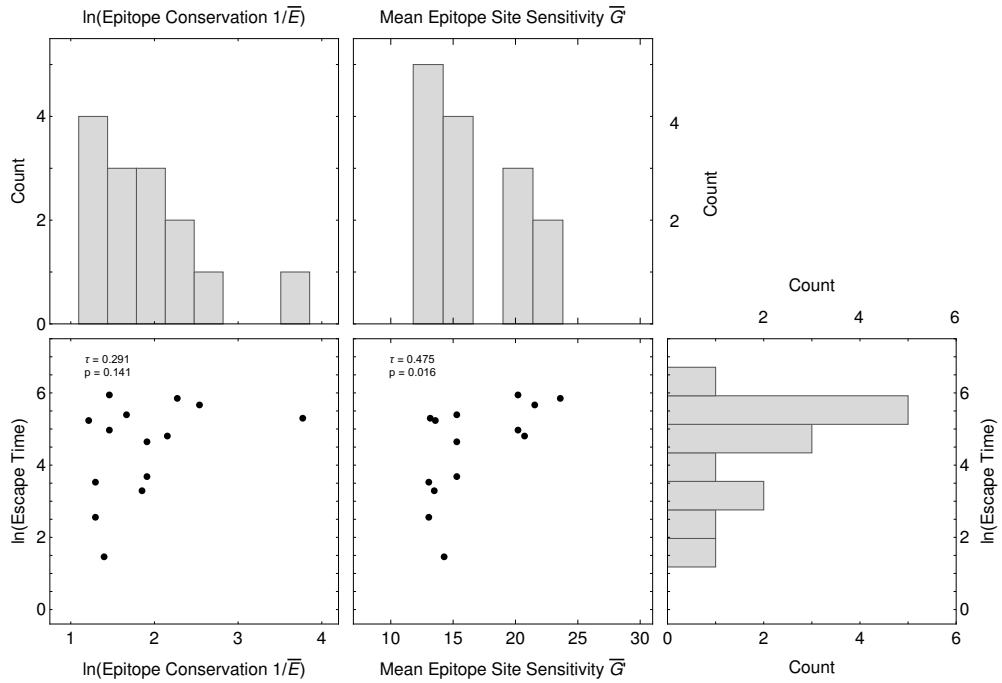

Fig 3. Kendall rank correlation τ of mean log conservation $\ln(\overline{1/E})$ and mean site sensitivity $\overline{G'}$ with ln(escape time) of 14 epitopes of the Gag poly-peptide whose immune escape was observed (original data from Liu et al. (2013) results from Barton et al. (2016b) used here). Correlation p values estimated by bootstrapping data 10,000 times.

bioRxiv preprint doi: https://doi.org/10.1101/204297; this version posted September 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

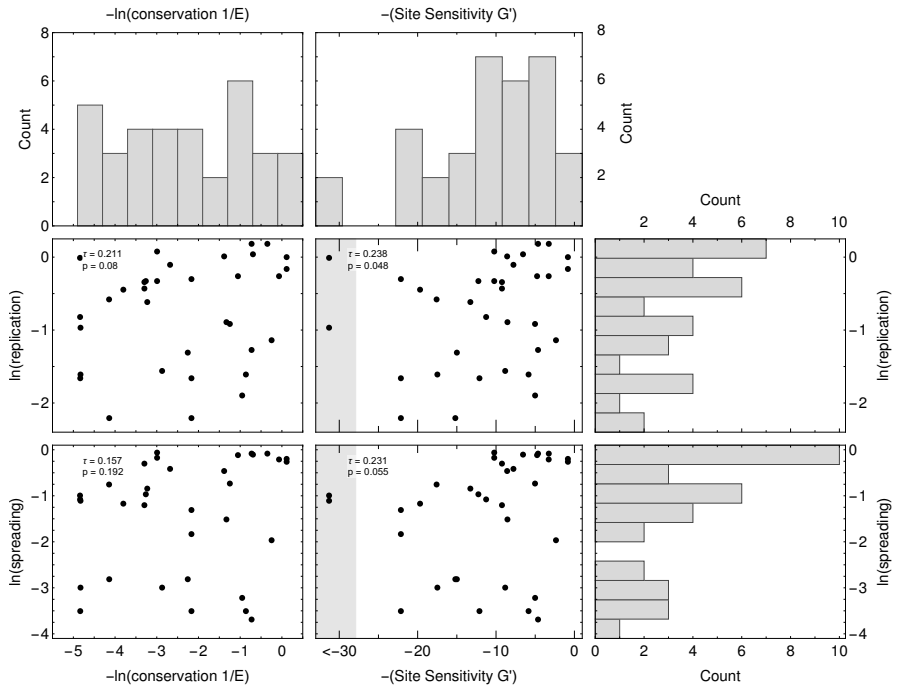

Fig 4. Metrics vs Spreading fitness Correlation of conservation and sensitivity of residues with spreading fitness of 31 viral strains bearing mutations in the corresponding residues A) Correlation of the assayed spreading fitness of the mutated virus in days with the correspondingly position's entropy B) Correlation of the assayed spreading fitness of the mutated virus in days with the corresponding position's physico-chemical sensitivity.