GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects

Elodie Laine,*[1] Yasaman Karami,[1,2] and Alessandra Carbone *[1,3]

[1]Sorbonne Université, CNRS, IBPS, UMR 7238, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France

[2]Sorbonne Université, Institut du Calcul et de la Simulation, Sorbonne Université, Paris, France

[3]Institut Universitaire de France, Paris, France

***Corresponding authors:** E-mails: elodie.laine@upmc.fr; alessandra.carbone@lip6.fr.

Associate editor: Banu Ozkan

Abstract

The systematic and accurate description of protein mutational landscapes is a question of utmost importance in biology, bioengineering, and medicine. Recent progress has been achieved by leveraging on the increasing wealth of genomic data and by modeling intersite dependencies within biological sequences. However, state-of-the-art methods remain time consuming. Here, we present Global Epistatic Model for predicting Mutational Effects (GEMME) (www.lcqb.upmc.fr/GEMME), an original and fast method that predicts mutational outcomes by explicitly modeling the evolutionary history of natural sequences. This allows accounting for all positions in a sequence when estimating the effect of a given mutation. GEMME uses only a few biologically meaningful and interpretable parameters. Assessed against 50 high- and low-throughput mutational experiments, it overall performs similarly or better than existing methods. It accurately predicts the mutational landscapes of a wide range of protein families, including viral ones and, more generally, of much conserved families. Given an input alignment, it generates the full mutational landscape of a protein in a matter of minutes. It is freely available as a package and a webserver at www.lcqb.upmc.fr/GEMME/.

Key words: **protein, epistasis, mutational landscape, evolution, conservation, viral sequence, mutation.**

Introduction

Understanding which and how genetic variations affect proteins and their biological functions is a central question for bioengineering, medicine, and fundamental biology. In these fields, a fast and accurate assessment of the effects of every possible substitution at every position in a protein sequence (full single-site mutational landscape) or of combinations of mutations (pairs, triplets, etc.) would allow to reach some level of control over proteins, needed to improve the treatment of diseases, the design of new proteins, and the synthesis of molecular libraries. Deep mutational scans (Fowler and Fields 2014) or multiplexed assays for variant effects (Gasperini et al. 2016) have enabled the full description of the mutational landscapes of a few tens of proteins (see Riesselman et al. [2018] for a list of proteins and associated experiments). They have revealed that a protein contains a relatively small number of positions highly sensitive to mutations, where almost any substitution induces highly deleterious effects (McLaughlin et al. 2012; Firnberg et al. 2014). Although these methods represent major biotechnological advances, they remain resource intensive and are limited in their scalability. Moreover, the measured phenotype and the way it is measured vary substantially from one experiment to another, making it difficult to compare different measurements and/or proteins (Boucher et al. 2016). These limitations call for the development of efficient and accurate computational methods for high-throughput mutational scans.

Many computational methods predicting mutational effects exploit information coming from protein sequences observed in nature (Ng and Henikoff 2003; Capriotti et al. 2005; Cheng et al. 2005; Adzhubei et al. 2010; Dehouck et al. 2011; Sim et al. 2012; Ferguson et al. 2013; Mann et al. 2014; Hart and Ferguson 2015; Barton et al. 2016; Figliuzzi et al. 2016; Flynn et al. 2017; Hopf et al. 2017; Louie et al. 2018; Riesselman et al. 2018). They start from a multiple sequence alignment (MSA) and rely on the assumption that rarely occurring mutations induce deleterious effects. A straightforward way to estimate frequencies of occurrence is to treat each position in the alignment independently from the others. However, the amino acid residues comprising a protein are interdependent, and the effect of a mutation depends on the amino acids present at other positions, a phenomenon referred to as "epistasis" (Breen et al. 2012; McCandlish et al. 2016). By leveraging on the increasing wealth of genomic data, recent developments have enabled modeling interdependencies between positions and have significantly improved the accuracy of mutational effects predictions (Ferguson et al. 2013; Mann et al. 2014; Figliuzzi et al. 2016; Flynn et al. 2017; Hopf et al. 2017; Louie et al. 2018; Riesselman et al. 2018). Specifically, some statistical methods estimate couplings between pairs of positions (Ferguson et al. 2013;

Mann et al. 2014; Hart and Ferguson 2015; Barton et al. 2016; Figliuzzi et al. 2016; Flynn et al. 2017; Hopf et al. 2017; Louie et al. 2018). They are very accurate in identifying a few strong direct couplings responsible for the whole covariability observed in homologous sequences and corresponding to physical contacts in protein structures (Weigt et al. 2009; Stein et al. 2015; Louie et al. 2018). In the context of mutational outcome prediction, the ensemble of all pairwise couplings is used as a proxy to capture the influence of the whole sequence context on a particular position. One of the limitations of these methods is that the explicit calculation of higher-order couplings is computationally intractable. To circumvent this issue, a deep latent-variable model was proposed where the global sequence context is implicitly accounted for by coupling the observed positions to latent ("hidden") variables (Riesselman et al. 2018). The model is fully trained on each studied protein family to generate sequences likely to belong to the family. Deviations between outputs and inputs are then used as estimates of the mutational effects. The mutational landscapes of certain protein families are very well captured by this deep learning approach, but the results strongly depend on the variability of the input data. More generally, the statistical inference of a large body of parameters from a finite, and sometimes very low, sequence sampling is a challenging problem (Barton et al. 2014; Haldane and Levy 2019). It is particularly relevant in the case of viral proteins whose sequences are often highly conserved. Several technical advances employing regularization terms have improved the accuracy of interresidue coupling estimation when dealing with viral proteins (Ferguson et al. 2013; Mann et al. 2014; Hart and Ferguson 2015; Barton et al. 2016; Flynn et al. 2017; Louie et al. 2018). Moreover, the usage of very small position-specific amino acid alphabets has reduced the computational cost of the inference. These efforts have allowed achieving very good agreement between the fitness landscapes inferred from patient sequences and in vitro experiments for several proteins from HIV and HCV. Nevertheless, the available methods still remain computationally costly and have only been evaluated against low-throughput experimental data.

In this work, we present a fast, scalable, and simple algorithm to predict mutational effects by explicitly modeling interdependencies between all positions in a sequence. Key to our approach is the notion of an evolutionary history relating the sequences observed today in nature. We view each sequence as an evolutionary solution selected with respect to a mutation of interest. Our algorithm infers the evolutionary relationships relating natural sequences by quantifying their global similarities. These relationships, encoded in a tree, are used to determine the extent to which each position is *conserved* in evolution and to estimate the *evolutionary fit* required to accommodate mutations. Our measure of conservation is markedly different from measures that quantify frequencies of occurrence at single positions (columns) of an alignment. Indeed, we infer conservation levels by reconstructing phylogenetic trees from global similarities between sequences. This means that the conservation level of one position embeds the covariations between this position and all other positions in the sequence. For a position to be "conserved," we ask that the letter(s) appearing at that position is/are fully conserved in subtrees of ancient origin. Hence, two positions can have the same distribution of letters but different conservation levels. For example, this will happen if one position displays all occurrences of the most represented letter in a subtree of ancient origin, whereas the other displays them in several subtrees. This notion of conservation was inspired from that of evolutionary trace (Lichtarge et al. 1996; Mihalek et al. 2004). It is computed by the Joint Evolutionary Trees (JET) method (Engelen et al. 2009), and it proved to be useful in the identification of protein interfaces (Engelen et al. 2009; Laine and Carbone 2015; Ripoche et al. 2017). We use the computed conservation degrees to weight positions, the rationale being that changes occurring at more conserved positions likely have bigger impacts on the protein's function (Karami et al. 2018). Then, to be able to discriminate between different substitutions occurring at a given position, we combine two quantities. The first one is the relative frequency of occurrence of the mutation, relying on physicochemical similarities rather than amino acid identities. The second one is the minimum evolutionary fit required to accommodate the mutation. Namely, we estimate how far one has to go in the evolutionary tree to observe a natural sequence displaying the mutation. We compare our predictions with those of DeepSequence (Riesselman et al. 2018) and EVmutation (Hopf et al. 2017), which are currently the best state-of-the-art methods that have been evaluated on high-throughput experimental data. We show that our algorithm overall performs on-par with the nonlinear latent-variable model of DeepSequence and better than the pairwise epistatic model of EVmutation. The improvement over these two methods is particularly significant for much conserved protein families. We also show that our method's predictive performance is comparable with computational frameworks well suited to treat viral sequences (Flynn et al. 2017; Louie et al. 2018). Beyond prediction assessment, we provide a clear readout of the contribution of epistasis by looking at how sequences observed in nature are spread and evolutionary related. Our method is implemented as a fully automated tool, Global Epistatic Model for predicting Mutational Effects (GEMME), available as a downloadable package and as a webserver at www.lcqb.upmc.fr/GEMME/. We show that GEMME is faster than state-of-the-art methods by several orders of magnitude. Given an input alignment, the calculation of the full mutational landscape of a protein takes a few minutes. Hence, GEMME makes possible the systematic study of pairs or triplets of mutations appearing sequentially in time and associated with drug resistance, for example, in viruses. It could help in taking informed decisions regarding patients' treatment and public health by enabling real-time analysis of pathogenic sequence data (Neher and Bedford 2018).

Our results emphasize the usefulness of the information encoded in the way certain positions, because of their functional importance, are segregated along the topology of evolutionary trees. A functionally important position is expected to be associated with one or several subtrees of ancient origin and homogeneous with respect to that position (all sequences in the subtree display the same amino acid). This type of patterns is essentially what our measure of conservation captures. We demonstrate that this notion of evolutionary-informed conservation is more pertinent than conservation measures looking at individual positions independently and that it constitutes a valid alternative to the explicit estimation of pairwise couplings toward capturing coevolution signals. Hence, this work paves the way to new ideas and developments in sequence- and evolutionary-based mutational outcome prediction.

New Approaches

Sequences observed in nature have been selected for function through evolution. Hence, they can inform us on the constraints underlying evolutionary processes and help us estimate mutational effects. To assess the impact of a given mutation at a given position in a query sequence, we look at an ensemble of sequences homologous to that query (fig. 1*a*). These sequences can be organized in a tree based on their global similarities (fig. 1*b*). The topology of the tree reflects the evolutionary relationships between the sequences. Our main contribution is to extract conservation patterns in line with the topology of the tree and use them to determine the extent to which a mutation will be deleterious for the function of the query protein. For this, the first step of our approach consists in estimating the biological importance of each residue in the query by computing its evolutionary conservation (fig. 1*c*, first color strip). Highly conserved positions are likely important for the protein stability and/or function and thus likely sensitive to changes. For each position in the query, we look at the level in the tree where the amino acid at that position appeared and remained conserved thereafter. To illustrate this notion of conservation, we consider a toy example with two positions, namely *i* and *j* displaying S and G, respectively (fig. 1*a* and *b*). In the tree, one can observe that the amino acid G at position *j* was fixed much earlier in evolution compared with S at position *i* (fig. 1*b*, compare gray rectangles between the two panels). Hence, we will assign a much higher conservation degree to the former than to the latter. Since the tree is inferred from global similarities between entire sequences, the conservation degree of a given position accounts for the way all other positions have diverged along evolution. Here, we deal with a potentially large number of sequences, and the reconstruction of a unique tree relating all of them may lead to an unreliable topology. To cope with this issue, we sample the initial ensemble of sequences to extract representative subsets and reconstruct trees starting from these subsets (see Materials and Methods). Conservation degrees are averaged over all reconstructed trees to get statistically significant estimates. We then use them to compare mutations occurring at the same position and to compare different positions.

To compare different mutations at a given position, we estimate the amount of changes required to accommodate a mutation over the entire sequence. This amount can be viewed as a proxy to quantify the "evolutionary fit" associated with the mutation of interest. We compute it by looking at how far natural sequences displaying the mutation are from the query sequence in the evolutionary tree. Our working hypothesis is that the more distant these sequences, the more deleterious the mutation. For the sake of simplicity, let us consider two mutations at position *i*, namely S-to-T and S-to-A, which we want to compare, and see how they are associated with changes at another position *j* (fig. 1*a*). Although the S-to-T mutation is sometimes associated with the wild-type G at *j* (sequences in blue), the S-to-A mutation is systematically accompanied by a mutation at *j* (namely G-to-V, sequences in green). Since position *j* is much more conserved than *i*, this will result in sequences bearing S-to-A at *i* being much further away in the tree, with respect to the query, than sequences bearing S-to-T at *i* (fig. 1*b*, compare the locations of the green and blue sequences). Intuitively, this observation suggests that it will be more difficult for the query to accommodate A compared with T at position *i*, and hence that S-to-A will be more deleterious than S-to-T. We can easily generalize this reasoning over two positions to the whole sequence. For this, we define an evolutionary distance between the query *q* and some sequence *s* which explicitly accounts for the conservation degrees of all variable positions between *q* and *s* (see Materials and Methods and fig. 1*c*). For each studied mutation, we look for the closest sequence to *q* displaying that mutation, and we use its evolutionary distance to estimate the minimal evolutionary fit associated with the mutation. We combine evolutionary fits with site-independent frequencies calculated using a reduced amino acid alphabet to get more precise estimates (see Materials and Methods).

Then, to be able to compare mutations occurring at different positions, we rely on the hypothesis that more conserved positions will be more sensitive to any mutation than less conserved positions. To implement this idea, we reweight the predicted mutational effects by the evolutionary conservation degrees (fig. 1*c*, compare the two matrices). As a result, highly deleterious mutations will be mainly found at highly conserved positions (fig. 1*c*, second matrix, dark squares are mainly localized at conserved positions, highlighted by arrows). In our toy example, although the evolutionary distance computed for the G-to-V mutation at *j* is lower than that computed for the S-to-A mutation at *i* (fig. 1*b*, compare the locations of the starred orange sequence on the right and the starred green sequence on the left), the former will be predicted as more deleterious than the latter because position *j* is much more conserved than position *i*.

GEMME's predictive model globally accounts for epistasis by explicitly looking at the whole sequence context when assessing the effect of a particular mutation. It is applicable to single-site mutations and also to combinations of mutations (see Materials and Methods).

Results

We assessed GEMME's predictive power against experimental measures collected from 41 high-throughput mutational scans of 33 proteins and 1 protein complex, representing

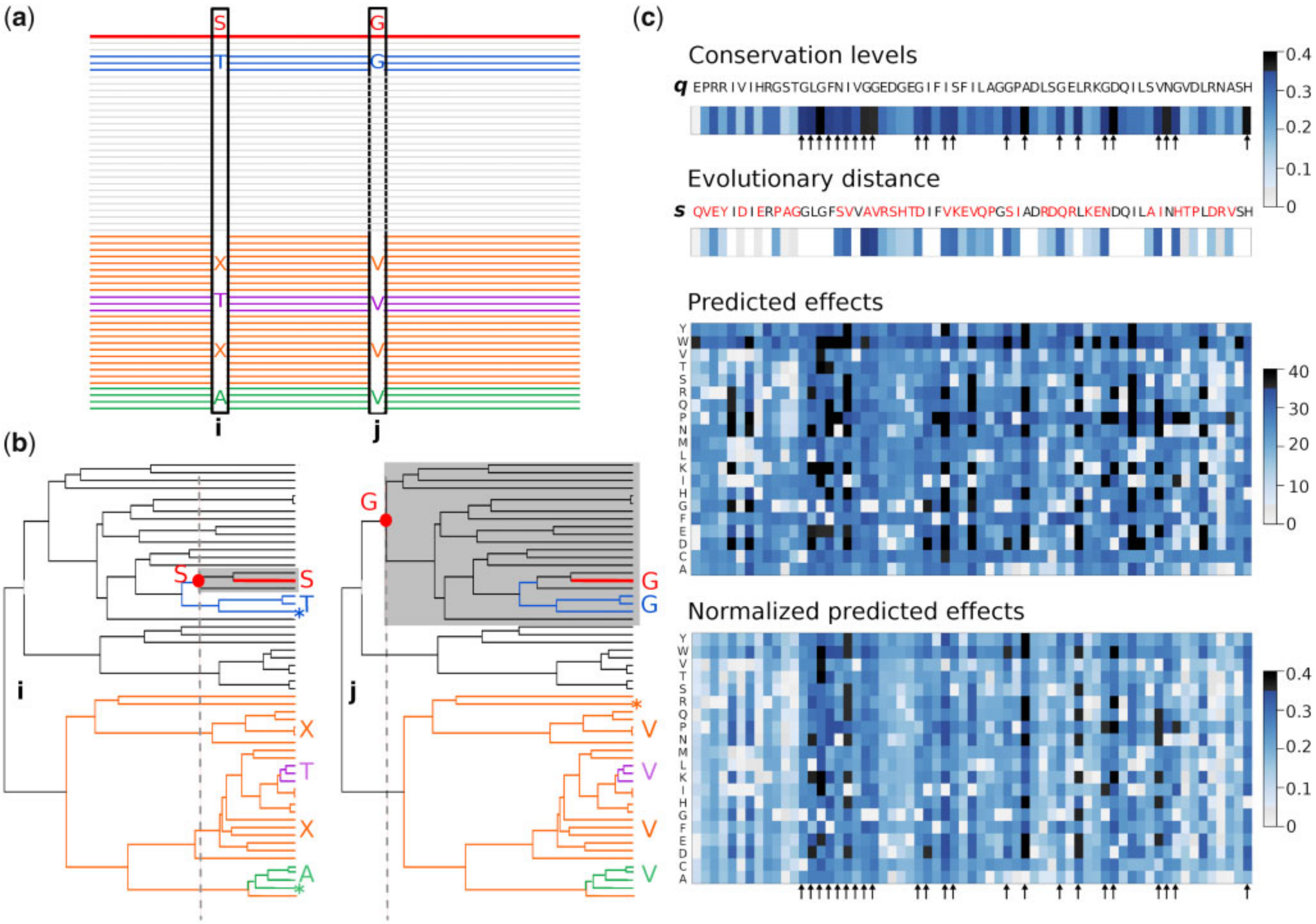
(a)
S
G
T
G
X
V
T
V
X
V
A
V
i
j
(b)
S
S
T
G
G
G
X
T
X
A
V
V
V
V
i
j
(c)
Conservation levels
q EPRRIVIHRGSTGLGFNIVGGEDGEGIFISFILAGGPADLSGELRKGDQILSVNGVDLRNASH
0.4
0.3
0.2
0.1
0
Evolutionary distance
s QVEYIDIERPAGGLGFSVVAVRSHTDIFVKEVQPGSIADRDQRLKENDQILAINHTPLDRVSH
Predicted effects
Y W V T S R Q P N M L K I H G F E D C A
40
30
20
10
0
Normalized predicted effects
Y W V T S R Q P N M L K I H G F E D C A
0.4
0.3
0.2
0.1
0

FIG. 1. Principle of the method. (*a*) Ensemble of sequences related to a query sequence, on top and in red. The query displays a serine (S) at position *i* and a glycine (G) at position *j*. Some sequences are colored according to the amino acids they display at the two positions: T-G in blue, T-V in purple, X-V in orange (X stands for any amino acid, except for T and A), and A-V in green. (*b*) Tree representing the evolutionary relationships between the related sequences. The color code is the same as in (*a*). Information concerning positions *i* and *j* is reported on the left and on the right, respectively. The red dots and dotted gray lines indicate the levels where S and G appeared at positions *i* and *j* and remained conserved thereafter. The associated subtrees are highlighted by gray rectangles. The stars indicate the closest sequences to the query displaying the S-to-T mutation at *i* (left, in blue), the S-to-A mutation at *i* (left, in green) and the G-to-V mutation at *j* (right, in orange). (*c*) Workflow of the method applied on the third PDZ domain of PSD95 (DLG4). The color strip on top gives conservation levels computed for the query sequence *q*. Positions highlighted by arrows are highly conserved. A homologous sequence *s* is displayed below, with its mutations highlighted in red. The second color strip indicates the squared conservation levels for the positions of the mutations. The two matrices give the predicted effects and NPEs, respectively, for all possible substitutions at all positions in *q*.

657,840 mutations (supplementary table S1, Supplementary Material online). Among them, 38 scans deal with single-site mutations. One scan describes the complete 2-point mutational landscape of a 96-residue long protein, and 2 scans contain multiple mutations. The largest scan reports measures for 496,137 mutants, where the number of variable positions between each mutant and the wild-type query ranges between 1 and 26. The average Spearman rank correlation computed between all predicted and experimental values is $\bar{\rho} = 0.53 \pm 0.13$. The best agreement is obtained for the bacterial β-lactamase, with a correlation of 0.74 (fig. 2*a*). There is a only a weak correlation between the proportion of disruptive mutations in the experiment and GEMME's predictive performance (fig. 2*b*). Compared with the state-of-the-art methods DeepSequence and EVmutation, GEMME performs equally well or better (fig. 2*a*). Namely, its overall performance are similar to those of DeepSequence (Riesselman et al. 2018) ($\Delta\rho_{\text{GEMME}-\text{DEEP}} \geq 0$ in 19/41 scans, with an average of 0.02 ± 0.12 and a median of −0.01) and significantly better than those of EVmutation (Hopf et al. 2017) ($\Delta\rho_{\text{GEMME}-\text{EVmutation}} \geq 0$ in 32/41 scans, with an average of 0.03 ± 0.05 and a median of 0.03). Importantly, GEMME achieves much higher correlations for the five viral sequences of the data set (fig. 2*a*, on the right, and fig. 2*b*, orange dots), up to $\Delta\rho = 0.5$ compared with DeepSequence and $\Delta\rho = 0.1$ compared with EVmutation. The input alignments for these proteins display a very low degree of diversity, with >60% of sequences sharing >60% of identity with the query sequence (fig. 3*a*, underlined proteins). More generally, the lower the diversity of the input alignment, the higher the improvement of GEMME over the two other methods (fig. 3*a*).

We also assessed GEMME's performance against 128 experimental measures coming from 9 low-throughput mutational studies of two HIV proteins, namely the envelope

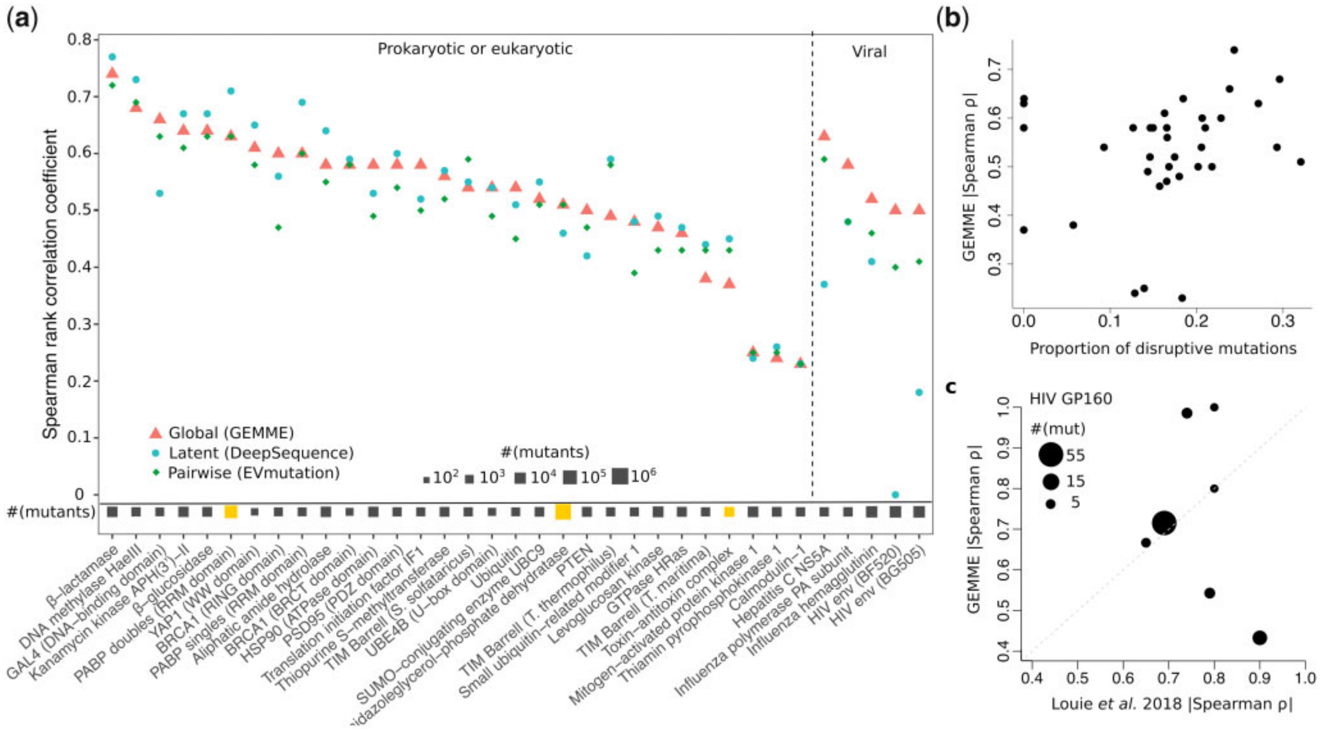

Fig. 2. Comparison of predictive performances between GEMME and state-of-the-art methods. (*a*) Spearman rank correlation coefficients ρ between predicted and experimental measures for 35 high-throughput experiments corresponding to 34 proteins (see Materials and Methods). Scans comprising multiple mutations are highlighted by a gold rectangle. (*b*) Spearman rank correlation coefficients in function of the proportion of disruptive mutations in the experiment. Disruptive mutations were defined as those displaying an experimental measure below $\mu - \sigma$, with μ the mean and σ the standard deviation. (*c*) Spearman rank correlation coefficients ρ between predicted and experimental measures for seven low-throughput experiments performed on gp160 from HIV. The size of each point indicates the number of mutants.

protein gp160 and the protease (supplementary fig. S1, Supplementary Material online). The input alignments for these proteins are composed of sequences coming from patients (see Materials and Methods) and sharing >60% sequence identity with the query. The considered mutants contain between 1 and 46 mutations. GEMME's Spearman rank correlations range from 0.43 to 1, with a weighted averaged value of 0.70. These results are similar to those obtained by two coevolution-based computational frameworks (Flynn et al. 2017; Louie et al. 2018) recently applied to viral sequences (compare supplementary fig. S1*a*–*g*, Supplementary Material online, with supplementary fig. S1 in Louie et al. [2018] and supplementary fig. S1*h* and *i*, Supplementary Material online, with fig. 3A in Flynn et al. [2017]). In particular, we obtain a weighted correlation of 0.70 against seven experiments performed on gp160, whereas Louie et al. (2018) reported a value of 0.74. The difference in performance between the two methods varies significantly from one experiment to another (fig. 2*c*). This may be explained by the relatively small number of measures (a few tens at most) coming from each experiment.

Epistasis Helps Discriminate between Equally Frequent Mutations

To compare different mutations occurring at the same position, GEMME's model combines two contributions, namely the minimal evolutionary fit required to accommodate each mutation and the relative frequency of occurrence of the mutation (see Materials and Methods). The first term accounts for intersite dependencies and can thus be qualified as "epistatic," whereas the second term is computed in a site-independent manner. If a mutation is rare and appears far away in the evolutionary tree, then both terms will be high and the mutation will be predicted as deleterious. On the contrary, if a mutation is frequent and found in sequences very close to the query, both terms will be small and hence, the predicted impact of the mutation will be small. The two terms will disagree in case of a rare mutation appearing in a sequence very similar to the query, or in case of a frequent mutation appearing only in highly divergent sequences. By default, GEMME puts a higher weight on the epistatic term (see Materials and Methods).

We systematically assessed the predictive power of each contribution taken separately (see Materials and Methods, fig. 4*a*, and supplementary table S1, Supplementary Material online). Overall, the predictions issued by the epistatic contribution were found in better agreement with the experimental measures than those from the independent contribution (average $\Delta\rho = 0.04 \pm 0.09$, median $\Delta\rho = 0.02$). Yet, in about one third of the cases (11/35), it is better to rely on site-independent frequencies rather than evolutionary distances (fig. 4*a*, on the right). Hence, the predictive power of the information coming from epistasis varies substantially from one protein to another. Such variability may arise from some intrinsic properties of the studied proteins, the

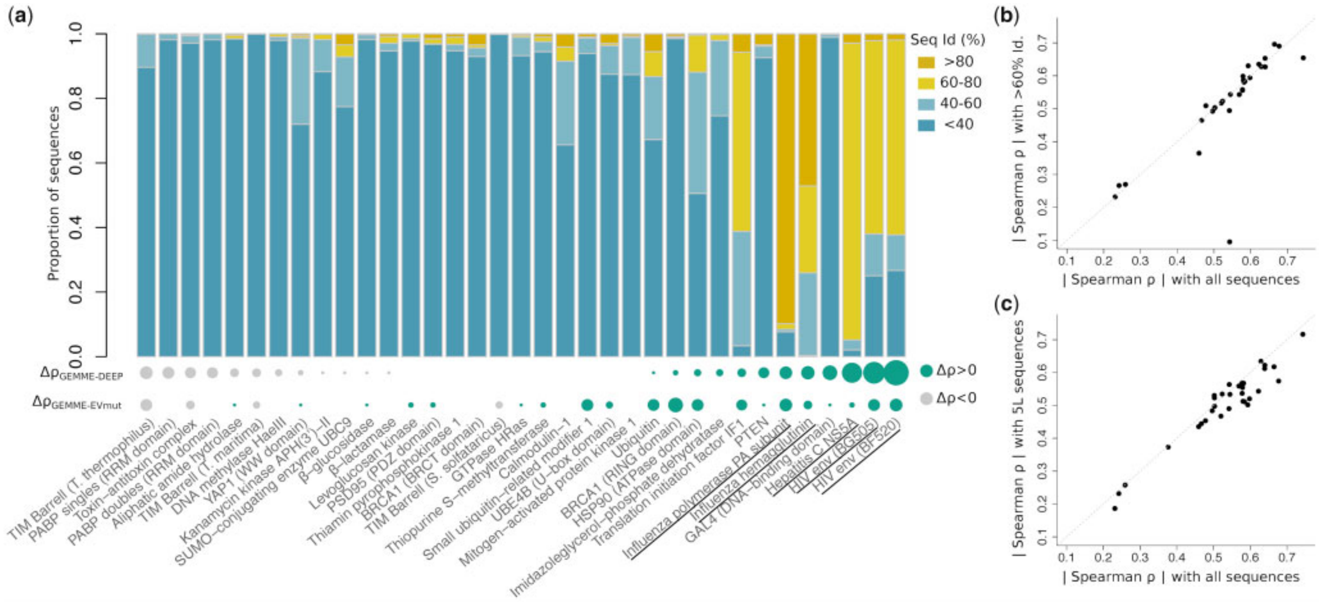

Fig. 3. Influence of the input alignment on the predictive performance. (*a*) Sequence identities of the input alignments. The percentages of sequences sharing <40%, between 40% and 60%, between 60% and 80%, and >80% with the query sequence are reported. Viral sequences are underlined. The dots at the bottom indicate the gain of performance of GEMME with respect to DeepSequence ($\Delta\rho_{\text{GEMME-DEEP}}$) and EVmutation ($\Delta\rho_{\text{GEMME-EVmut}}$). The size is proportional to the absolute value of $\Delta\rho$ and the color depends on the sign. $\Delta\rho$ is positively correlated with the proportion of sequences sharing >60% identity with the query (Pearson correlation coefficient $R = 0.71$ for $\Delta\rho_{\text{GEMME-DEEP}}$ and $R = 0.48$ for $\Delta\rho_{\text{GEMME-EVmut}}$). (*b*) Comparison of Spearman rank correlations obtained by GEMME when the conservations levels are computed from the whole ensemble of input sequences (*x* axis) or with only sequences displaying >60% identity with the query (*y* axis). (*c*) Comparison of Spearman rank correlations obtained by GEMME starting from the full input alignment (*x* axis) or from a subset of the alignment containing 5*L* sequences (chosen randomly from the full alignment), where *L* is the length (number of residues) of the query (*y* axis).

experimental setup (e.g., measured phenotypes) and/or the properties of the input alignment. The bacterial DNA methyltransferase *Hae*III and the nonstructural protein 5A (NS5A) from Hepatitis C virus provide two archetypal examples that help understand the influence of the input alignment. Both proteins display much contrasted performances between the two contributions, but although the epistatic term is the best one for the bacterial methyltransferase (fig. 4*a*, on the left), the most accurate predictions for the viral NS5A are issued from the independent term (fig. 4*a*, on the right). In the first case, a lot of mutations are rather frequent in the input alignment (fig. 4*b*, on the left, in blue). More precisely, half of the mutations are at least 22% as frequent as the wild-type amino acid. By contrast, in the second case, half of the mutations are very rare (<0.03% as frequent as the wild-type amino acid) or simply not found in the alignment (fig. 4*b*, on the left, in red). More generally, as the mutation average frequency of occurrence increases, so does the gain of the epistatic contribution over the independent one (fig. 4*b*, in the middle, Pearson correlation coefficient $R = 0.61$). The shape of the mutation frequency distribution is also a good indicator (fig. 4*b*, on the right, $R = -0.60$ with the distribution's skewness). The best case scenario for the epistatic contribution is when a large body of mutations are rather frequent, and few mutations are rarer than average and widely spread (see the long left tail of the distribution in blue in fig. 4*b*, on the left). Hence, when dealing with equally frequent mutations, looking at the overall divergence of the sequences where they appear, as quantified by our evolutionary fit, helps improving the discrimination between them.

Evolutionary Conservation Is a Valid Proxy for Position-Specific Sensitivity to Mutations

Besides providing full protein mutational landscapes, mutational scans can be used to determine which positions in the protein are particularly sensitive to mutations. Such positions typically represent a small portion of the protein and can be viewed as its "weak" spots. The sensitivity of a position can be estimated by averaging its mutational outcomes over the 19 possible substitutions. In GEMME's predictions, this average is strongly correlated to the position's degree of conservation (supplementary fig. S2, Supplementary Material online). This is expected as GEMME reweights positions according to their evolutionary conservation to compare mutations occurring at different positions (see Materials and Methods). This results in highly conserved positions displaying overall higher predicted mutational effects than lowly conserved positions. We found that the conservation degrees alone provide estimates of position-specific sensitivities to mutations that are only slightly less accurate than the averages computed from GEMME's full predicted matrices (fig. 5*a*). This indicates that our conservation measure is already a good indicator of the extent to which a position will be sensitive to mutations. Importantly, this holds true even when the variability of the available sequence data is low. This means that we are able to

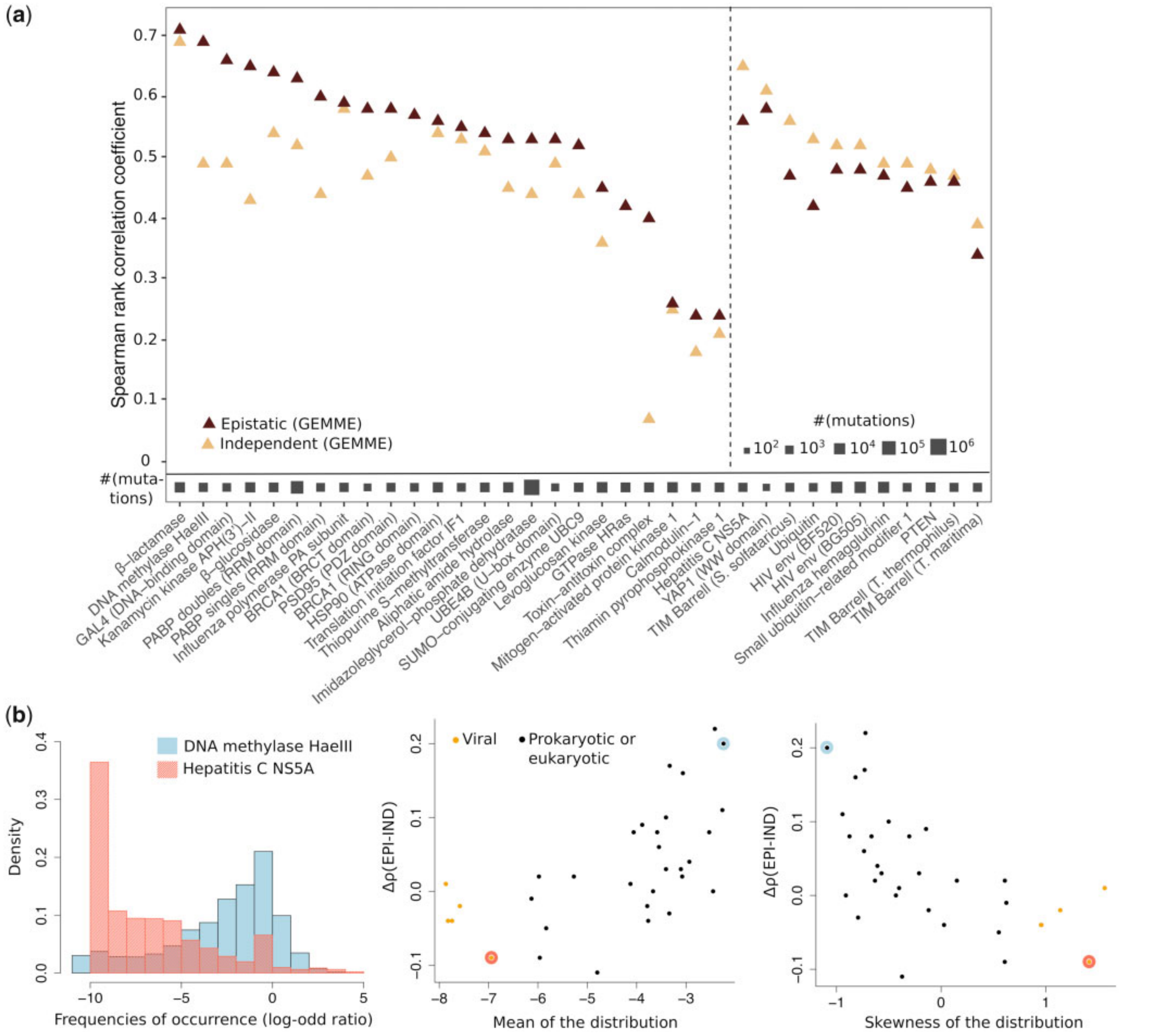

Fig. 4. Comparison of predictive performances between the epistatic and independent contributions of GEMME's model. (*a*) On the *x* axis, proteins are divided in two groups according to the contribution yielding the highest correlation with experimental data (epistatic contribution on the left, independent one on the right). (*b*) Left panel. Examples of distributions for the mutations' site-independent relative frequencies of occurrence. For each mutation, the reported value is the log-odd ratio between the number of sequences displaying the mutation over the number of sequences displaying the wild-type amino acid (see Materials and Methods). Middle and right panels. Difference in Spearman ρ coefficient in function of the mean (in the middle) and the skewness (on the right) of the log-odd ratio distribution. The skewness reflects the asymmetry of the distribution (positive skewness indicates a left tail, whereas negative skewness indicates a right tail). The dots corresponding to the proteins taken as examples on the left panel are encircled.

capture meaningful signals in contexts where the content of information is poor. Moreover, our conservation measure compares well with the predictions issued by DeepSequence and EVmutation (fig. 5*a*). In several cases, it better reflects experimentally determined mutational sensitivities than the averages computed from these predictions (fig. 5*a*, points highlighted in blue). Finally, calculating conservation levels only from sequences close to the query (>60% identity) instead of the full ensemble of input sequences does not significantly affect the quality of the predictions in most cases (fig. 3*b*).

GEMME's Results Are Robust and Its Model Is Transferable to Other Proteins

To assess the transferability of GEMME's model to other systems, we systematically evaluated the influence of its two main parameters and of the input alignment's depth on the quality of the predictions. The first parameter is the

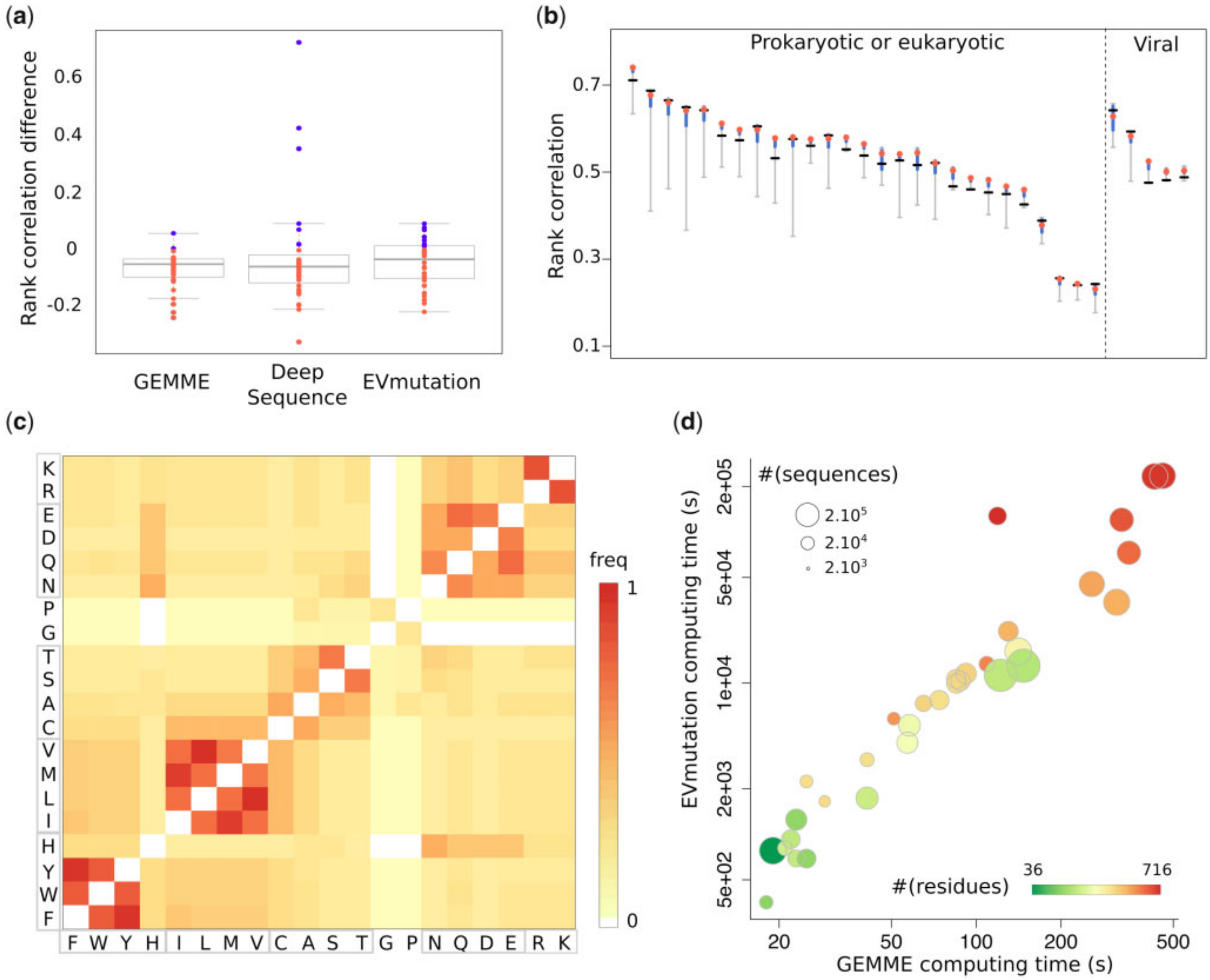

FIG. 5. Analysis of GEMME's parameters and computing time. (*a*) Differences in position rank correlation between the evolutionary conservation degrees computed by GEMME and the mutational effects predicted by GEMME, DeepSequence, and EVmutation. Each point stands for a scan. Positive and negative differences are highlighted in blue and red, respectively. (*b*) Ranges of rank correlation obtained when varying the relative importance of GEMME's independent and epistatic contributions. Each vertical segment corresponds to a deep mutational scan (same order as in fig. 2). The red dot indicates the correlation obtained with GEMME's default model, where the epistatic term is assigned a weight of 0.6 (and 0.4 for the independent term, see Materials and Methods). The black dash indicates the best performance achieved by the independent or epistatic contribution alone. The blue thick segments highlight the range of values obtained when varying the epistatic term's weight between 0.5 and 0.8 (see also supplementary figs. S3 and S4, Supplementary Material online). (*c*) Amino acid grouping preferences observed in GEMME's best performing models (parameters optimized for each scan). The color code goes from white (grouping never observed) to red (grouping observed for all scans). The amino acids are ordered so as to highlight the reduced alphabet used by default in GEMME. (*d*) Computing times of EVmutation and GEMME (in seconds, with logarithmic scales).

relative importance given to the epistatic and independent contributions. We observed that our default model, where the epistatic contribution is given more weight, systematically achieves similar or better correlation with experiments than each contribution taken alone (fig. 5*b*, compare red dots and black dashes). In cases where the independent contribution performs better than the epistatic one, combining the two leads to only slightly lower performances. Moreover, varying the relative weights of the two terms around their default values has a very small impact on the quality of the predictions in most of the cases (fig. 5*b*, blue segments). The second degree of freedom is the choice of the amino acid alphabet. GEMME relies on similarities between amino acids rather than identities to compute the mutations' frequencies of occurrence. By default, we consider seven classes of amino acids, grouping together the aromatic ones (F, W, Y, H), the hydroxyl-containing ones with alanine (C, A, S, T), the aliphatic hydrophobic ones (I, L, M, V), the positively charged ones (K, R), and the polar and negatively charged ones (N, Q, D, E). Glycine and proline are in a separate class each. We tested 164 different alphabet reduction schemes (see Materials and Methods and supplementary table S2, Supplementary Material online) and found that the impact of the alphabet's choice on the predictive performance is limited (average correlation standard deviation of 0.01, supplementary figs. S3 and S4, Supplementary Material online). So is the correlation gain obtained by optimizing the parameters for each scan (average $\Delta\rho = 0.02 \pm 0.01$, and median $\Delta\rho = 0.02$, supplementary fig. S5*a*, Supplementary Material online). Moreover, the amino acid grouping preferences exhibited by the best performing models are in good agreement with the alphabet chosen by default (fig. 5*c*). In addition to the model's parameters, the input alignment's depth may also influence the quality of the predictions. The initial input alignments comprise between 10 and 2,000 times as many sequences as the length L of the query protein (supplementary table S3, Supplementary Material online) and display a wide range of variability degrees (fig. 3*a*). Reducing the number of input sequences to 5L results in a limited loss of correlation (fig. 3*c*, average $\Delta\rho = 0.03 \pm 0.03$, and median $\Delta\rho = 0.02$). Even with 0.5L sequences, the average correlation is of $\bar{\rho} = 0.46$ (compared with 0.53 with the full alignments, see supplementary fig. S6*a*, Supplementary Material online). Reducing the variability of the alignment, in addition to its size, further degrades performance (supplementary fig. S6*b*, Supplementary Material online). Overall, this analysis shows that our results are robust to parameter changes and alignment depth and that our choices lead to predictions whose quality is close to the best one can hope for within GEMME's framework. This is true overall and on most of the scans studied here, which makes us confident that our default model is directly transferable to other proteins.

GEMME Is Faster than State-of-the-Art Methods by Several Orders of Magnitude

To be applicable at large scale, computational scans should be fast. Given a query sequence of length L associated with an alignment of N_0 sequences, GEMME estimates $L \times 21 + N_0$ quantities to predict a full single-site mutational landscape (see Materials and Methods). It took <10 min, on a single-core processor, for GEMME to generate any of the complete single-site mutational landscapes considered here. The corresponding proteins are of various lengths, comprising between $L = 36$ and 716 residues, and are associated with up to several hundreds of thousands of homologous sequences (fig. 5*d*). By comparison, EVmutation requires several days of computation to deal with the biggest proteins of the data set. Overall, GEMME is faster than EVmutation by a factor ranging between 19 and 1,072 (supplementary table S3, Supplementary Material online). It should be stressed that EVmutation disregards some positions and some sequences from the input alignment, whereas GEMME does not. DeepSequence is expected to be even more computationally expensive. Training one deep latent model on the β-lactamase family (to estimate several millions of parameters) and generating the corresponding full mutational landscape required almost 7 h (24,175 s) on a powerful graphics card (see Materials and Methods). This computing time has to be multiplied by 5 to obtain results similar to those reported in Riesselman et al. (2018). GEMME took only about 1 min to treat the same protein on a single-core processor (fig. 5*d* and supplementary table S3, Supplementary Material online). Hence, we estimate DeepSequence to be several thousands times more computationally expensive than GEMME. The method described in Louie et al. (2018) required 2.5 days of CPU time on a 16-core node to estimate the parameters (about 4.4 million) for HIV gp160. The method described in Flynn et al. (2017) required 4 h of computation on two powerful graphics cards to estimate the parameters (about 40,000) for HIV protease. For comparison, GEMME's calculations on these proteins estimated took <10 and 3 min, respectively, on a single-core processor.

Discussion

We have presented GEMME, a computational method for performing mutational scans of protein sequences. It exclusively exploits protein sequence data available in public databases. It relies on a few biologically sound assumptions about the relationship between protein sequence and function. Its algorithm is straightforward and requires setting only two parameters. It uses the mathematical tree structure underlying the evolution of natural sequences and it explores it by using simple new concepts (smallest path between wild-type and mutated sequences). This is markedly different from what has been developed previously. State-of-the-art methods feature tens to hundreds of thousands of parameters, infer some of them using sophisticated machine learning techniques, others empirically, reweight the input sequences to correct for sampling bias, and do not explicitly model the evolutionary history relating these sequences. Despite its apparent simplicity, our method achieves similar or better performance than the state of the art, and it can deal with highly variable as well as highly conserved sequences. Importantly, although GEMME was designed to treat any protein family, its performance on viral proteins is similar to recent computational frameworks well suited to treat these proteins. It has the advantage of being faster than recently published methods by several orders of magnitude.

An important ingredient of GEMME is the inclusion of dependencies between the different positions in the sequence of interest. We are not the first ones to propose to account for "epistasis" in the prediction of mutational effects. What has been done before was to explicitly model couplings between pairs of positions, inferred from co-occurring patterns in the input sequence data, or to implicitly model higher-order dependencies by coupling each position to a "hidden" variable. Let us stress that the introduction of such hidden variables do no ease interpretability. We adopt an orthogonal approach by using the notion of evolutionary history relating the sequences observed today in nature. We infer such evolutionary history by quantifying global similarities between sequences, thus accounting for all positions and their interdependencies. Then we use the reconstructed evolutionary trees to identify functionally important positions, the rationale being that such positions should display a few amino acids that appeared and were fixed early in evolution.

Contrary to statistical inference-based methods, GEMME does not try to estimate a joint probability distribution. This means that we do not make any assumption on the space of all possible sequences. Instead, we directly exploit the information encoded in natural sequences and we do it in a query-centered way. Specifically, each tree constructed to compute conservation levels contains the query, and the evolutionary distances are computed with respect to the query. Hence, our predictions estimate deviations from the query, whereas predictions from statistical inference-based methods correspond to ratios of probabilities of belonging to the protein "family" represented by the input alignment. On the one hand, this constitutes a limitation of our method. For instance, in case of a mutant with a higher fitness than the wild type and evolutionary far away from it, GEMME will simply predict a strong mutational effect. On the other hand, this can be an advantage when the input alignment contains protein "subfamilies" performing different functions and displaying different functionally relevant patterns of conservation and coevolution (as is the case of the cryptochrome/photolyase family for instance).

We have implemented our method as a fully automated package and webserver, available at www.lcqb.upmc.fr/GEMME/. It can deal with single mutations as well as combinations of mutations. It has been carefully evaluated against experimental measurements from 9 low-throughput experiments and 41 high-throughput mutational scans comprising up to 496,137 data points. Moreover, we have provided an understanding of the contribution of "epistasis" to the discrimination of mutations, based on the analysis of the variability of the input sequence data, and have demonstrated that our evolutionary-informed measure of conservation is a good indicator of the extent to which a position is sensitive to mutations. Finally, we have systematically assessed the influence of changing GEMME's two parameters and the input alignment's depth on the quality of the predictions. Our results are consistent and suggest that GEMME's model is transferable to other systems than the few tens studied here.

This work proposes an original and efficient approach to the characterization of protein mutational landscapes. Working with a few parameters, versus many, and relying on a "simple" approach allow to get a comprehensive understanding of why a mutation will be predicted as deleterious or not. We believe ease of interpretability is important and can help foster conceptual breakthroughs. Indeed, it has been shown in various occasions in Riesselman et al. (2018) for instance that the success of machine learning methods largely depends on biologically meaningful priors used to impose some structure on the predictive model. Perspectives for this work include the systematic description of the effects of mutations on protein interaction networks, a field that will continue to expand in the coming years.

Materials and Methods

GEMME's Workflow

The GEMME method takes as input a MSA in FASTA format, with the query sequence appearing on top. First, evolutionary conservation levels are computed using JET (Engelen et al. 2009). Then, GEMME predicts the mutational landscape of the query sequence. By default, it estimates the mutational outcomes of all possible single mutations. Alternatively, the user can provide an ensemble of single or multiple mutations of interest.

Homologous Sequence Retrieval and Selection

The user can ask GEMME to compute conservation levels directly on the input MSA. Alternatively, GEMME will automatically launch a PSI-BLAST (Altschul et al. 1997) search to retrieve up to 5,000 sequences related to the query. Then, a number of selection criteria will be applied to filter the set of related sequences. By default, sequences redundant with the query (>98% identity) or too far (<20% identity), too small (<80% coverage), too gapped (>10% of the size of the alignment) or not significant enough (*e*-value $\geq 10^{-5}$) are removed. If the number of remaining sequences is too low (<100), the selection criteria are progressively relaxed as described in Engelen et al. (2009). All parameters are adjustable by the user.

Evolutionary Conservation Levels (T_{JET})

The calculation of evolutionary conservation levels relies on a Gibbs-like sampling of the filtered set of related sequences (Engelen et al. 2009). Sequences are classified into four groups, depending on the degree of identity they share with the query (20–39%, 40–59%, 60–79%, and 80–98%). Starting from an ensemble of N sequences, $\sqrt{N}$ sequences are randomly picked up from the four classes to construct a subset representing the diversity of the whole set. The sequences are then aligned and a distance tree is constructed from the alignment using the Neighbor-Joining algorithm (Studier and Keppler 1988). For each position in the query sequence, a *tree trace level* is computed: it corresponds to the level l in the tree where the amino acid at this position appeared and remained conserved thereafter (see Engelen et al. [2009] for a more precise definition). Let us recall that this definition of evolutionary trace is notably different from the measure defined by Lichtarge and coworkers to rank protein residues (Lichtarge et al. 1996; Mihalek et al. 2004).

This procedure is repeated $\sqrt{N}$ times and the *tree trace levels* are averaged over the $\sqrt{N}$ trees to get more statistically significant values, which we denote *relative trace significances*, or T_{JET}, and which are expressed as (Engelen et al. 2009)

$$T_{\text{JET}}(i) = \frac{1}{M_i}\sum_{t=1}^{M_i}\frac{L_t - l_i^t}{L_t}, \qquad (1)$$

where l_i^t is the *tree trace level* of residue r_i in tree t, L_t is the maximum level of t, and M_i is the number of trees where a nonnull *tree trace level* was computed for r_i. This procedure efficiently handles sequence sampling bias without requiring to explicitly reweight sequences and setting a sequence identity cutoff. T_{JET} values vary in the interval [0, 1] and represent averages over all trees of residues' evolutionary conservation levels.

To produce the results reported here, we used the most recent version of the JET method, namely JET^2 (Laine and Carbone 2015) (available at www.lcqb.upmc.fr/JET2). JET^2 uses MUSCLE (Edgar 2004) to align sequences. JET^2 was launched in its iterative mode: The procedure described above was repeated ten times and the maximum conservation value obtained over the ten runs was retained for each residue. Note that running only one iteration leads to very similar predictions (99.6 ± 0.4% Pearson correlation coefficient on average). Hence, on the webserver, the default number of iterations is set to 1.

Predicted Effects: Comparison of Mutations Occurring at the Same Position

To compare mutations occurring at the same position, GEMME combines two contributions. The first one is termed *epistatic* and corresponds to the minimal evolutionary fit required to accommodate the mutation of interest. The second one is termed *independent* and reflects the relative frequency of occurrence of the mutation. Hence, the predicted effect of a mutation X-to-Y at position i is expressed as

$$\mathrm{PE}(Y_i) = -\alpha \min\left\{1, \frac{\mathrm{PE}^{\mathrm{Epi}}(Y_i)}{\max_{Z,k}[\mathrm{PE}^{\mathrm{Epi}}(Z_k)]}\right\} \times \log\left[\frac{1}{\max_k(\sum_Z |S_{Z_k}|)}\right] + (1-\alpha)\mathrm{PE}^{\mathrm{Ind}}(Y_i), \tag{2}$$

where $\mathrm{PE}^{\mathrm{Epi}}(Y_i)$ and $\mathrm{PE}^{\mathrm{Ind}}(Y_i)$ are the values of the epistatic and independent contributions (defined below), respectively. The term $\min\left\{1, \frac{\mathrm{PE}^{\mathrm{Epi}}(Y_i)}{\max_{Z,k}[\mathrm{PE}^{\mathrm{Epi}}(Z_k)]}\right\}$ scales the value of the epistatic contribution between 0 and 1. The maximum value $\max_{Z,k}[\mathrm{PE}^{\mathrm{Epi}}(Z_k)]$ is determined over all positions k and all possible substituting amino acids Z for a given protein. The term $\log\left[\frac{1}{\max_k(\sum_Z S_{Z_k}|)}\right]$, where S_{Z_k} is the ensemble of sequences displaying Z at position k, gives the lowest possible log-ratio value. It corresponds to the extreme case where all nongapped sequences at position k display the wild-type amino acid. It is used as a multiplying factor here so as to be able to combine the predictions coming from the two different contributions.

In case of a mutation U not observed in the input alignment, we consider that sequences displaying that mutation are infinitely far from the query sequence and we use a pseudo-count to estimate its frequency of occurrence (see below for more detailed explanations). In practice, the predicted effect of that mutation will be expressed as

$$\mathrm{PE}(U_i) = -\alpha \log\left[\frac{1}{\max_k(\sum_Z |S_{Z_k}|)}\right] - (1-\alpha)\log\left(\frac{1}{|S_{X_i}|}\right), \tag{3}$$

where S_{Z_k} and S_{X_i} are the ensembles of sequences displaying Z at position k and X at position i. Hence, the effect predicted for a nonobserved mutation will solely depends on the position of the mutation.

The coefficient α determines the relative weight of each contribution. By default, α is set to 0.6, and sequence counts are calculated using a reduced representation of the amino acid alphabet composed of seven classes: FWYH, ILMV, CAST, G, P, NQDE, and RK (LW-I-7 in supplementary table S2, Supplementary Material online).

Epistatic Contribution ($\mathrm{PE}^{\mathrm{Epi}}$)

The evolutionary relationships between all the sequences in the input MSA can be represented by a tree, which is not explicitly computed here. The topology of that tree is implicitly reflected by the T_{JET} values, which were computed and averaged over many small trees. We illustrate this by considering two positions i and j at which the query sequence q displays S and G, respectively (fig. 1*a* and *b*). Position i is lowly conserved ($T_{\mathrm{JET}} = 0.2$), whereas position j is highly conserved ($T_{\mathrm{JET}} = 0.8$). This implies that q belongs to a smaller subtree of sequences displaying S at position i (fig. 1*b*, on the left, dark gray rectangle), and to a bigger subtree of sequences displaying G at position j (fig. 1*b*, on the right, light gray rectangle).

To estimate how close some sequence s is from the query sequence q, we define the evolutionary distance $D_{\mathrm{evol}}(q, s)$ as

$$D_{\mathrm{evol}}(q, s) = \sum_{k=1}^{n} T_{\mathrm{JET}}(k)^2 \times 1_{X_k^q \neq X_k^s}, \tag{4}$$

where n is the length of q, X_k^q is the amino acid of q at position k, and $1_{X_k^q \neq X_k^s}$ is the indicator function. Only positions where the amino acid in s is different from the amino acid in q (fig. 1*c*, in red) contribute to the sum, and the level of contribution depends on the level of conservation of the position (fig. 1*c*, second color strip).

To assess the effect of a mutation X-to-Y at position i in q, we select the subset S_{Y_i} of sequences displaying the mutation, and look for the sequence within S_{Y_i} being the closest to q. The resulting minimal evolutionary distance estimates how far from q one has to go in the tree to observe a sequence bearing Y at i. Hence, the predicted effect of mutation Y at position i, $PE^{Epi}(Y_i)$, is expressed as

$$\mathrm{PE}^{\mathrm{Epi}}(Y_i) = \min_{s \in S_{Y_i}}[D_{\mathrm{evol}}(q, s)]. \tag{5}$$

To avoid bias due to the presence of a peculiar sequence or of an alignment error in the MSA, we require that there exists at least one sequence different from the closest one and at a similar distance to the query. For this, we rank all evolutionary distances in ascending order and compute the difference between the first and second ones. If the difference is lower than an arbitrarily chosen cutoff of 5, then we keep the first one. Otherwise, we replace it by the second one. In case of a mutation U not observed in the alignment, the ensemble S_{U_i} is empty (i.e., no sequence displaying U at position i could be found) and we set $PE^{Epi}(U_i) = +\infty$.

This metric enables to directly compare and rank several substitutions at a given position. Given two amino acids Y and Z substituting X at position i, one can express the difference $\mathrm{PE}^{\mathrm{Epi}}(Y_i) - \mathrm{PE}^{\mathrm{Epi}}(Z_i)$ as

$$\Delta \mathrm{PE}^{\mathrm{Epi}}(Y_i, Z_i) = \min_{S_{Y_i}} \left[\sum_{k=1,k\neq i}^{n} T_{\mathrm{JET}}(k)^2 \times 1_{W_k^q \neq W_k^s} \right] - \min_{S_{Z_i}} \left[\sum_{k=1,k\neq i}^{n} T_{\mathrm{JET}}(k)^2 \times 1_{W_k^q \neq W_k^s} \right], \qquad (6)$$

where the sums are computed over all positions except i, as the contribution of i cancels out. Consequently, this difference quantifies how much the global sequence context of position i has to change to accommodate the X-to-Y substitution versus X-to-Z. The substitution displaying the highest change will be predicted as the most deleterious one at that position.

As an example, let us consider mutations S-to-T and S-to-A at position i in figure 1*a* and *b*. The S-to-T mutation induces a smaller minimal amount of changes than the S-to-A mutation, as the closest sequence displaying T at i (indicated by a blue star in fig. 1*b*, left panel) has a lower evolutionary distance to the query than the closest sequence displaying A at i (green star). The S-to-A substitution at i systematically implies a G-to-V mutation at another position j, whereas the S-to-T substitution does not.

Independent Contribution ($\mathrm{PE}^{\mathrm{Ind}}$)

This contribution focuses only on the position where the mutation occurs. Hence, the effect of a substitution X-to-Y at positions i will be estimated as

$$\mathrm{PE}^{\mathrm{Ind}}(Y_i) = -\log \left[\frac{\max(1, |S_{Y_i}|)}{|S_{X_i}|} \right], \qquad (7)$$

where $|S_{X_i}| \geq 1$ since at least the query sequence q displays X at position i. The value 1 in the numerator serves as a pseudo-count in case the mutation of interest U is not observed in the alignment and thus $|S_{U_i}| = 0$. According to this model, the fewer sequences displaying the mutation, the more deleterious the mutation, for a given position.

Normalized Predicted Effects: Comparison of Mutations Occurring at Different Positions

The matrix of predicted effects (fig. 1*c*) enables comparing the 19 possible substitutions at each position in the query sequences. To be able to compare substitutions at different positions, we proceed through a normalization step. The normalized predicted effect (NPE) of a mutation X-to-Y at position i is expressed as

$$\mathrm{NPE}(Y_i) = T_{\mathrm{JET}}(i) \times \mathrm{PE}(Y_i). \qquad (8)$$

If position i is highly conserved, then its $T_{\mathrm{JET}}(i)$ value will be close to one and the predicted effects at that position will remain unchanged. By contrast, if position i is lowly conserved, then its $T_{\mathrm{JET}}(i)$ value will be close to 0 and the predicted effects at that position will be largely reduced. Hence, the normalization will result in highly conserved positions being predicted as highly intolerant to mutations, whereas any substitution at a poorly conserved position will have a small effect. This is illustrated in figure 1c where one can see that the predictions for the highly conserved positions (columns highlighted by arrows in the NPE matrix) remain essentially unchanged upon normalization, whereas the other positions "whiten up."

In the toy example pictured in figure 1*a* and *b*, the effect predicted for mutation S-to-A at i is higher than that predicted for mutation G-to-V at j. And as one can observe, this is explained by the closest sequence displaying A at position i (indicated by a green star in fig. 1*b*, on the left) being further away than the closest sequence displaying V at position j (fig. 1*b*, on the right, orange star). However, since position i is much less conserved than position j, the normalization step will result in a large reduction of the effect predicted for S-to-A at i, which will thus end up as less deleterious than G-to-V at j.

We extended the global epistatic model to deal with combinations (pairs, triplets, etc.) of mutations. The NPE of a given combination of p mutations is expressed as

$$\mathrm{NPE}(Y_1, Y_2, \ldots, Y_p) = \sum_{j=1}^{p} \mathrm{NPE}(Y_j). \qquad (9)$$

Computational Complexity

Given an input alignment of N_0 sequences, the method first proceeds through a filtering step using various criteria to retain N sequences. N is typically in the order of a few hundreds or thousands. Then, $\sqrt{N}$ trees are built from $\sqrt{N}$ alignments of $\sqrt{N}$ sequences to compute L evolutionary trace values, L being the length of the query. To generate a full single-site mutational landscape, the independent contribution requires computing $L \times 20$ amino acid frequencies. The epistatic contribution requires computing all distances between the query and the N_0 sequences in the input alignment. Hence, in total $L \times 21 + N_0$ quantities need to be estimated. The independent and epistatic contributions are linearly combined and the computed predictions are finally multiplied by the evolutionary traces.

Parameters Setup

To determine the default value of α and the default reduced amino acid alphabet scheme, we systematically computed predictions for all values of α, ranging between 0 and 1 by increments of 0.1, and for 164 reduced alphabets (see below and supplementary table S2, Supplementary Material online). For each combination (α, alphabet), we computed its mean squared displacement from the best performing combination. Among the three combinations displaying the lowest mean squared displacements, we chose the combination with the lowest median squared displacement, namely $\alpha = 0.6$ and LW-I-7 as the alphabet scheme. To identify the model yielding the best performance, for each experimental scan, the coefficient α was varied between 0 and 1 by increments of 0.1, and the 164 amino acid alphabets were systematically tested.

Experimental Data Sets and Input Alignments

To assess GEMME's performance and compare it fairly with several state-of-the-art methods, we considered the same experimental data and the same input alignments as those reported in the corresponding studies. Specifically, the experimental measures determined by 41 deep mutational scans were taken from Riesselman et al. (2018) (see Araya et al. 2012; Deng et al. 2012; McLaughlin et al. 2012; Jacquier et al. 2013; Melamed et al. 2013; Roscoe et al. 2013; Starita et al. 2013; Melnikov et al. 2014; Qi et al. 2014; Roscoe and Bolon 2014; Aakre et al. 2015; Kitzman et al. 2015; Rockah-Shmuel et al. 2015; Romero et al. 2015; Stiffler et al. 2015; Wu et al. 2015; Doud and Bloom 2016; Mishra et al. 2016; Firnberg et al. 2016 for details about each experiment). The input alignments were also taken from Riesselman et al. (2018). The 41 scans were performed across 34 full proteins, protein domains or protein complexes. Among those, 38 scans comprise only single-site mutations and 3 scans comprise combinations of mutations. Two scans are associated with two different domains of the same protein (BRCA1), and one scan is associated with a protein complex (toxin–antitoxin complex). In the main text and figures, we report the performance obtained against one measured phenotype from one scan, for each protein. In case of multiple scans associated with the same protein, we focus on the most recent one. There is one exception, namely PABP, for which two scans were retained because one deals with single-site mutations and the other one with multiple mutations. The selected measured phenotype is the one yielding the best agreement with the predictions. All results for all measured phenotypes from all scans are reported in supplementary table S1, Supplementary Material online.

The experimental measures determined by seven low-throughput experiments performed on gp160 from HIV were taken from Louie et al. (2018) (see Anastassopoulou et al. 2007; Lobritz et al. 2007; Kassa et al. 2009; Troyer et al. 2009; da Silva et al. 2010; Liu et al. 2013 for details about each experiment). Each mutant within each experiment contain between 1 and 46 mutations with respect to the query wild-type sequence. The input sequence alignment (comprising 20,043 sequences coming from 1,918 patients, retrieved from https://www.hiv.lanl.gov/) was also taken from Louie et al. (2018). The experimental measures determined by two low-throughput experiments performed on the HIV protease were taken from Flynn et al. (2017) (see Chang and Torbett 2011; Henderson et al. 2012 for details about each experiment). The measured phenotypes are melting temperature (Chang and Torbett 2011) and replicative capacity (Henderson et al. 2012). For each phenotype, Flynn et al. (2017) mixed several experiments together and assessed their performance against these mixtures. Aggregating the measurements collated from different experiments is not obvious, even if the measured phenotype is the same, due to technical variations between the experiments (resulting in batch effects). Consequently, we decided to apply GEMME on the experiment comprising the largest number of measures, for each phenotype. Each mutant within each experiment contain one or two mutations with respect to the query wild-type sequence. The input alignment (comprising 5,610 drug-experienced sequences from 4,604 patients, retrieved from https://www.hiv.lanl.gov/) was also taken from Flynn et al. (2017).

Reduced Amino Acid Alphabets

A reduced alphabet is a clustering of amino acids based on their relative similarity. We tested 164 different alphabet schemes (Peterson et al. 2009), whose names and characteristics are reported in supplementary table S3, Supplementary Material online. They comprise between 2 and 19 letters. AB schemes were defined based on the ability of standard methods to correctly predict secondary structure from the simplified sequences (Andersen and Brunak 2004). CB schemes were produced by using the Miyazawa–Jernigan interaction matrix (Miyazawa and Jernigan 1996) and a distance-based clustering scheme (Cieplak et al. 2001). DSSP and GBMR schemes were designed to maximally preserve structural information (Solis and Rackovsky 2000). HSDM and SDM schemes were defined based on new substitution matrices derived from structural alignments of proteins with low-sequence identity (Prlic et al. 2000). LR is a 10-letter alphabet intended to increase the sensitivity of protein alignment searches (Landes and Risler 1994). LW-I and LW-NI schemes were designed to preserve information in global sequence alignments between a sequence and its reduced-alphabet version (Li et al. 2003). Notice that LW-I and LW-NI are identical at the level of 2, 3, and 15–19 letters, and that CB and LW are identical at the 2-letter level. LZ-MJ and LZ-BL were defined based on the identification of deviations of pair frequency counts from a random background, computed on the Miyazawa–Jernigan and BLOSUM 50 matrices (Liu et al. 2002). ML schemes are based on the BLOSUM 50 substitution matrix (Murphy et al. 2000). MM is a five-letter alphabet based on the Johnson–Overington matrix (Johnson and Overington 1993) which proved useful for aligning homologous sequences and assessing folds (Melo and Marti-Renom 2006). MS is a six-letter alphabet based upon intuition and a study of the effects of disulfide bonds on protein folding which suggested separating aliphatic hydrophobic and aromatic hydrophobic residues (Mirny and Shakhnovich 1999). TD schemes are based on intuitive physicochemical considerations (Thomas and Dill 1996). WW is a five-letter alphabet derived from the Miyazawa–Jernigan matrix by preserving maximal similarity between a reduced-alphabet version of the matrix and the full 20 × 20 matrix (Wang and Wang 1999).

Comparison of Performance

Predictions for DeepSequence and EVmutation were directly taken from Riesselman et al. (2018). To evaluate and compare performances, we used Spearman rank correlation coefficient ρ as the primary metric. This choice was also made in previous studies (Figliuzzi et al. 2016; Hopf et al. 2017; Riesselman et al. 2018) and is justified by the fact that we do not expect a linear relationship between predicted and experimental values. For each high-throughput experiment, correlations were computed on the set of mutations for which both experimental measures and predictions from DeepSequence and EVmutation were available. For the low-throughput experiments, both individual and weighted average correlations were computed. Weighted averages of individual Spearman correlations are commonly used in meta-analysis (Field 2001) and allow accounting for inconsistencies between experiments. The weighted average correlation $\bar{\rho}$ of N experiments is expressed as

$$\bar{\rho} = \frac{\sum_{i=1}^{N} n_i \rho_i}{\sum_{i=1}^{N} n_i}, \tag{10}$$

where ρ_i is the correlation computed for the ith experiment and n_i is the number of measures in the ith experiment.

Comparison of Computing Times

To measure GEMME's computing time, we ran the tool starting from the input MSA and the result of the PSI-BLAST search. Hence, for each protein, we measured the elapsed (wall clock) time required to compute the conservation levels and the predictions from all models (default, independent and epistatic). The computation of the conservation levels was realized by running one iteration of JET^2. The resulting matrices of predicted mutational effects were very similar (99.6 ± 0.4% Pearson correlation coefficient on average) to those obtained when running ten iterations of JET^2 and retaining the maximum conservation level over the ten runs. EVmutation's code was downloaded from https://github.com/debbiemarkslab/EVmutation and https://github.com/debbiemarkslab/plmc. For each protein, we measured the elapsed (wall clock) time required to compute the pairwise couplings and the predictions from both the independent and epistatic models. The same input MSAs, taken from Riesselman et al. (2018), were given to GEMME and EVmutation. In these alignments, some positions are flagged because they are highly gapped. GEMME considered all positions whereas EVmutation disregarded the flagged ones. EVmutation also disregarded highly gapped sequences. The numbers of positions and sequences considered by each tool are reported in supplementary table S3, Supplementary Material online. Calculations were realized on a single-core processor Intel Xeon E5-2630 v4 at 2.20 GHz. DeepSequence's code was downloaded from https://github.com/debbiemarkslab/DeepSequence. We ran the tool to train one model on the input MSA for the β-lactamase and compute the predictions. The calculation was realized on a NVIDIA TITAN Xp graphics card. We tried and ran the tool on a single-core processor but it produced an error.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was partially undertaken in the framework of the LabEx CALSIMLAB, supported by the public grant ANR-11-LABX-0037-01 constituting a part of the "Investissements d'Avenir" program (reference : ANR-11-IDEX-0004-02; YK); funds from the Institut Universitaire de France (AC). We thank Tristan Bitard-Feildel for providing computing time for DeepSequence, Raymond Louie for providing us with an MSA of HIV gp160 and some experimental fitness data, and Anne Lopes and Hugues Richard for careful reading of the manuscript.

References

Aakre CD, Herrou J, Phung TN, Perchuk BS, Crosson S, Laub MT. 2015. Evolving new protein–protein interaction specificity through promiscuous intermediates. *Cell* 163(3):594–606.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods*. 7(4):248–249.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*. 25(17):3389–3402.

Anastassopoulou CG, Marozsan AJ, Matet A, Snyder AD, Arts EJ, Kuhmann SE, Moore JP. 2007. Escape of HIV-1 from a small molecule CCR5 inhibitor is not associated with a fitness loss. *PLoS Pathog*. 3(6):e79.

Andersen CAF, Brunak S. 2004. Representation of protein-sequence information by amino acid subalphabets. *AI Mag*. 25(1):97–104.

Araya CL, Fowler DM, Chen W, Muniez I, Kelly JW, Fields S. 2012. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc Natl Acad Sci U S A*. 109(42):16858–16863.

Barton JP, Cocco S, De Leonardis E, Monasson R. 2014. Large pseudocounts and L2-norm penalties are necessary for the mean-field inference of Ising and Potts models. *Phys Rev E Stat Nonlin Soft Matter Phys*. 90(1):012132.

Barton JP, Goonetilleke N, Butler TC, Walker BD, McMichael AJ, Chakraborty AK. 2016. Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. *Nat Commun*. 7:11660.

Boucher JI, Bolon DN, Tawfik DS. 2016. Quantifying and understanding the fitness effects of protein mutations: laboratory versus nature. *Protein Sci*. 25(7):1219–1226.

Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. 2012. Epistasis as the primary factor in molecular evolution. *Nature* 490(7421):535–538.

Capriotti E, Fariselli P, Casadio R. 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res*. 33(Web Server issue):W306–W310.

Chang MW, Torbett BE. 2011. Accessory mutations maintain stability in drug-resistant HIV-1 protease. *J Mol Biol*. 410(4):756–760.

Cheng J, Randall A, Baldi P. 2005. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* 62(4):1125–1132.

Cieplak M, Holter NS, Maritan A, Banavar JR. 2001. Amino acid classes and the protein folding problem. *J Chem Phys*. 114(3):1420–1423.

da Silva J, Coetzer M, Nedellec R, Pastore C, Mosier DE. 2010. Fitness epistasis and constraints on adaptation in a human immunodeficiency virus type 1 protein region. *Genetics* 185(1):293–303.

Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. 2011. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics* 12(1):151.

Deng Z, Huang W, Bakkalbasi E, Brown NG, Adamski CJ, Rice K, Muzny D, Gibbs RA, Palzkill T. 2012. Deep sequencing of systematic combinatorial libraries reveals beta-lactamase sequence constraints at high resolution. *J Mol Biol*. 424(3-4):150–167.

Doud MB, Bloom JD. 2016. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses* 8(6):155.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32(5):1792–1797.

Engelen S, Trojan LA, Sacquin-Mora S, Lavery R, Carbone A. 2009. Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. *PLoS Comput Biol*. 5(1):e1000267.

Ferguson AL, Mann JK, Omarjee S, Ndung'u T, Walker BD, Chakraborty AK. 2013. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* 38(3):606–617.

Field AP. 2001. Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. *Psychol Methods*. 6(2):161–180.

Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. 2016. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol Biol Evol*. 33(1):268–280.

Firnberg E, Labonte JW, Gray JJ, Ostermeier M. 2014. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol Biol Evol*. 31(6):1581–1592.

Firnberg E, Labonte JW, Gray JJ, Ostermeier M. 2016. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol Biol Evol*. 33(5):1378.

Flynn WF, Haldane A, Torbett BE, Levy RM. 2017. Inference of epistatic effects leading to entrenchment and drug resistance in HIV-1 protease. *Mol Biol Evol*. 34(6):1291–1306.

Fowler DM, Fields S. 2014. Deep mutational scanning: a new style of protein science. *Nat Methods*. 11(8):801–807.

Gasperini M, Starita L, Shendure J. 2016. The power of multiplexed functional analysis of genetic variants. *Nat Protoc*. 11(10):1782.

Haldane A, Levy RM. 2019. Influence of multiple-sequence-alignment depth on Potts statistical models of protein covariation. *Phys Rev E* 99(3-1):032405.

Hart GR, Ferguson AL. 2015. Empirical fitness models for hepatitis C virus immunogen design. *Phys Biol*. 12(6):066006.

Henderson GJ, Lee SK, Irlbeck DM, Harris J, Kline M, Pollom E, Parkin N, Swanstrom R. 2012. Interplay between single resistance-associated mutations in the HIV-1 protease and viral infectivity, protease activity, and inhibitor sensitivity. *Antimicrob Agents Chemother*. 56(2):623–633.

Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, Marks DS. 2017. Mutation effects predicted from sequence co-variation. *Nat Biotechnol*. 35(2):128.

Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E, Glodt J, Bercot B, Petit E, Poulain J, Barnaud G, et al. 2013. Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc Natl Acad Sci U S A*. 110(32):13067–13072.

Johnson MS, Overington JP. 1993. A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J Mol Biol*. 233(4):716–738.

Karami Y, Bitard-Feildel T, Laine E, Carbone A. 2018. "Infostery" analysis of short molecular dynamics simulations identifies highly sensitive residues and predicts deleterious mutations. *Sci Rep*. 8(1):16126.

Kassa A, Finzi A, Pancera M, Courter JR, Smith AB, Sodroski J. 2009. Identification of a human immunodeficiency virus type 1 envelope glycoprotein variant resistant to cold inactivation. *J Virol*. 83(9):4476–4488.

Kitzman JO, Starita LM, Lo RS, Fields S, Shendure J. 2015. Massively parallel single-amino-acid mutagenesis. *Nat Methods*. 12(3):203–206.

Laine E, Carbone A. 2015. Local geometry and evolutionary conservation of protein surfaces reveal the multiple recognition patches in protein–protein interactions. *PLoS Comput Biol*. 11(12):e1004580.

Landes C, Risler JL. 1994. Fast databank searching with a reduced amino-acid alphabet. *Comput Appl Biosci*. 10(4):453–454.

Li T, Fan K, Wang J, Wang W. 2003. Reduction of protein sequence complexity by residue grouping. *Protein Eng*. 16(5):323–330.

Lichtarge O, Bourne HR, Cohen FE. 1996. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*. 257(2):342–358.

Liu X, Liu D, Qi J, Zheng W-M. 2002. Simplified amino acid alphabets based on deviation of conditional probability from random background. *Phys Rev E* 66(2):021906.

Liu Y, Holte S, Rao U, McClure J, Konopa P, Swain JV, Lanxon-Cookson E, Kim M, Chen L, Mullins JI. 2013. A sensitive real-time PCR based assay to estimate the impact of amino acid substitutions on the competitive replication fitness of human immunodeficiency virus type 1 in cell culture. *J Virol Methods*. 189(1):157–166.

Lobritz MA, Marozsan AJ, Troyer RM, Arts EJ. 2007. Natural variation in the V3 crown of human immunodeficiency virus type 1 affects replicative fitness and entry inhibitor sensitivity. *J Virol*. 81(15):8258–8269.

Louie RHY, Kaczorowski KJ, Barton JP, Chakraborty AK, McKay MR. 2018. Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proc Natl Acad Sci U S A*. 115(4):E564–E573.

Mann JK, Barton JP, Ferguson AL, Omarjee S, Walker BD, Chakraborty A, Ndung'u T. 2014. The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. *PLoS Comput Biol*. 10(8):e1003776.

McCandlish DM, Shah P, Plotkin JB. 2016. Epistasis and the dynamics of reversion in molecular evolution. *Genetics* 203(3):1335–1351.

McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. 2012. The spatial architecture of protein function and adaptation. *Nature* 491(7422):138–142.

Melamed D, Young DL, Gamble CE, Miller CR, Fields S. 2013. Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-binding protein. *RNA* 19(11):1537–1551.

Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS. 2014. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res*. 42(14):e112.

Melo F, Marti-Renom MA. 2006. Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets. *Proteins* 63(4):986–995.

Mihalek I, Res I, Lichtarge O. 2004. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol*. 336(5):1265–1282.

Mirny LA, Shakhnovich EI. 1999. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol*. 291(1):177–196.

Mishra P, Flynn JM, Starr TN, Bolon DNA. 2016. Systematic mutant analyses elucidate general and client-specific aspects of Hsp90 function. *Cell Rep*. 15(3):588–598.

Miyazawa S, Jernigan RL. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol*. 256(3):623–644.

Murphy LR, Wallqvist A, Levy RM. 2000. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng*. 13(3):149–152.

Neher RA, Bedford T. 2018. Real-time analysis and visualization of pathogen sequence data. *J Clin Microbiol*. 56(11):e00480-18.

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 31(13):3812–3814.

Peterson EL, Kondev J, Theriot JA, Phillips R. 2009. Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. *Bioinformatics* 25(11):1356–1362.

Prlic A, Domingues FS, Sippl MJ. 2000. Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng*. 13(8):545–550.

Qi H, Olson CA, Wu NC, Ke R, Loverdo C, Chu V, Truong S, Remenyi R, Chen Z, Du Y, et al. 2014. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis C viral fitness and drug sensitivity. *PLoS Pathog*. 10(4):e1004064.

Riesselman AJ, Ingraham JB, Marks DS. 2018. Deep generative models of genetic variation capture the effects of mutations. *Nat Methods*. 15(10):816–822.

Ripoche H, Laine E, Ceres N, Carbone A. 2017. JET2 Viewer: a database of predicted multiple, possibly overlapping, protein-protein interaction sites for PDB structures. *Nucleic Acids Res*. 45(D1):D236–D242.

Rockah-Shmuel L, Toth-Petroczy A, Tawfik DS. 2015. Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. *PLoS Comput Biol*. 11(8):e1004421.

Romero PA, Tran TM, Abate AR. 2015. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proc Natl Acad Sci U S A*. 112(23):7159–7164.

Roscoe BP, Bolon DN. 2014. Systematic exploration of ubiquitin sequence, E1 activation efficiency, and experimental fitness in yeast. *J Mol Biol*. 426(15):2854–2870.

Roscoe BP, Thayer KM, Zeldovich KB, Fushman D, Bolon DN. 2013. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J Mol Biol*. 425(8):1363–1377.

Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*. 40(Web Server issue): W452–W457.

Solis AD, Rackovsky S. 2000. Optimized representations and maximal information in proteins. *Proteins* 38(2):149–164.

Starita LM, Pruneda JN, Lo RS, Fowler DM, Kim HJ, Hiatt JB, Shendure J, Brzovic PS, Fields S, Klevit RE. 2013. Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc Natl Acad Sci U S A*. 110(14):E1263–E1272.

Stein RR, Marks DS, Sander C. 2015. Inferring pairwise interactions from biological data using maximum-entropy probability models. *PLoS Comput Biol*. 11(7):e1004182.

Stiffler MA, Hekstra DR, Ranganathan R. 2015. Evolvability as a function of purifying selection in TEM-1 beta-lactamase. *Cell* 160(5):882–892.

Studier JA, Keppler KJ. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol*. 5(6):729–731.

Thomas PD, Dill KA. 1996. An iterative method for extracting energy-like quantities from protein structures. *Proc Natl Acad Sci U S A*. 93(21):11628–11633.

Troyer RM, McNevin J, Liu Y, Zhang SC, Krizan RW, Abraha A, Tebit DM, Zhao H, Avila S, Lobritz MA, et al. 2009. Variable fitness impact of HIV-1 escape mutations to cytotoxic T lymphocyte (CTL) response. *PLoS Pathog*. 5(4):e1000365.

Wang J, Wang W. 1999. A computational approach to simplifying the protein folding alphabet. *Nat Struct Biol*. 6(11):1033–1038.

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. 2009. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci U S A*. 106(1):67–72.

Wu NC, Olson CA, Du Y, Le S, Tran K, Remenyi R, Gong D, Al-Mawsawi LQ, Qi H, Wu TT, et al. 2015. Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. *PLoS Genet*. 11(7):e1005310.