RESEARCH ARTICLE

How Structural and Physicochemical Determinants Shape Sequence Constraints in a Functional Enzyme

Luciano A. Abriata[1]*, **Timothy Palzkill**[2], **Matteo Dal Peraro**[1]

1 Laboratory for Biomolecular Modeling, School of Life Sciences, and Swiss Institute of Bioinformatics, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, **2** Department of Pharmacology, Baylor College of Medicine, BCM-385, Houston, Texas, United States of America

* luciano.abriata@epfl.ch

Citation: Abriata LA, Palzkill T, Dal Peraro M (2015) How Structural and Physicochemical Determinants Shape Sequence Constraints in a Functional Enzyme. PLoS ONE 10(2): e0118684. doi:10.1371/journal.pone.0118684

Academic Editor: Vladimir N. Uversky, University of South Florida College of Medicine, UNITED STATES

Received: November 7, 2014

Accepted: January 21, 2015

Published: February 23, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: LAA was funded by an EMBO long-term fellowship during this work (http://www.embo.org); TP was funded by the National Institutes of Health grants AI32956 and AI106863 (http://www.nih.gov); MDP was funded by the Swiss National Science Foundation (http://www.snf.ch). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

The need for interfacing structural biology and biophysics to molecular evolution is being increasingly recognized. One part of the big problem is to understand how physics and chemistry shape the sequence space available to functional proteins, while satisfying the needs of biology. Here we present a quantitative, structure-based analysis of a high-resolution map describing the tolerance to all substitutions in all positions of a functional enzyme, namely a TEM lactamase previously studied through deep sequencing of mutants growing in competition experiments with selection against ampicillin. Substitutions are rarely observed within 7 Å of the active site, a stringency that is relaxed slowly and extends up to 15–20 Å, with buried residues being especially sensitive. Substitution patterns in over one third of the residues can be quantitatively modeled by monotonic dependencies on amino acid descriptors and predictions of changes in folding stability. Amino acid volume and steric hindrance shape constraints on the protein core; hydrophobicity and solubility shape constraints on hydrophobic clusters underneath the surface, and on salt bridges and polar networks at the protein surface together with charge and hydrogen bonding capacity. Amino acid solubility, flexibility and conformational descriptors also provide additional constraints at many locations. These findings provide fundamental insights into the chemistry underlying protein evolution and design, by quantitating links between sequence and different protein traits, illuminating subtle and unexpected sequence-trait relationships and pinpointing what traits are sacrificed upon gain-of-function mutation.

Introduction

The amino acid sequence of a protein defines its structure and function, but also a number of other entangled traits such as stability, water solubility/membrane integration, potential for interactions with other molecules, internal dynamics, hydration (referring to water molecules wetting the protein in its first solvation sphere and cavities), etc. These traits arise as a complex combination of the basic physicochemical properties of the constituent amino acids, in a way related to the ordered context of the protein's sequence. Given that all amino acids are chemically different from each other, all substitutions are expected to affect at least some of the traits associated to a particular protein sequence. If the effects produced by a substitution are sufficiently detrimental, the reduced function of the resulting protein will lead to decreased fitness and the mutation will likely not propagate in a population; on the contrary, if the effects are positive the mutation may increase in frequency and possibly be fixed (*i.e.* become permanent in the population). Thus, the physicochemical properties of the amino acids, and how they interact across a sequence, define not only the basic traits of a protein but also the potential of these traits to change and of the protein to evolve [1–3]. Whereas models of protein evolution account at most for protein structure at a coarse level [4–9], a fine, detailed understanding of how protein sequences evolve requires a full understanding of the effects of substitutions on all traits, to be incorporated into the evolutionary models [2,10–12]. This calls for a link between studies of genetic variation and structural biology [11].

From the side of structural biology, the problem fits into the goal of understanding the fundamental physical and chemical constraints that shape the sequence space available to functional proteins [11,13]. The most common approach towards capturing the relationships between protein sequence, structure, function and other traits has been to probe the effects of specific amino acid substitutions. Such studies, carried out on hundreds of proteins, have indeed provided most of our knowledge about protein physics and chemistry. However, they are inherently low-throughput and heterogeneous in terms of the recorded observables and the employed methods and materials. The new era of high-throughput deep sequencing technologies has allowed for much more complex experiments in which large libraries of mutants can be screened and then sequenced at once [14–17]. These technologies have been applied to full scans of a few proteins in studies about tolerance to substitution and distribution of fitness effects [18–22] and to monitor directed evolution experiments [23,24]. The produced data could provide at least part of the information required to connect genetic variability to structural biology [25–27] and to guide the development of better protein evolutionary models [28].

Based on deep sequencing technologies, a pioneering work by Deng et al. reported what is probably the highest-resolution map of an enzyme's tolerance to amino acid substitutions across its entire sequence [18]. That work quantified the likelihood of substituting each residue of TEM-1 β-lactamase by each of the other 19 amino acids while retaining enzymatic activity in the cell against ampicillin. As discussed in the original paper, the experiment revealed that TEM-1 is quite tolerant of amino acid substitutions, except in the region that surrounds the active site and other positions scattered throughout the protein. The low tolerance to substitutions around the active site was interpreted by Deng et al. on the basis of catalytic and/or substrate-binding roles for these residues, whereas tolerance farther away from the active site was attributed to general roles in conferring stability and solubility. As we show here, the high resolution of this dataset allows the retrieval of richer and finer quantitative information about the physical and chemical basis for why different amino acids are more or less favored at each position of the protein, at least in the context of ampicillin hydrolysis by TEM-1. We have analyzed the substitution patterns determined for each position of TEM-1 on the basis of its three-dimensional structure and considering which physicochemical properties of the amino acids could dictate the observed distributions. Expanding on the observation that the active site can barely tolerate substitutions, we found that this stringency is still strong within 5–10 Å of the active site, with upper limits that relax only slowly as the distance increases and extending up to 15–20 Å away from it. We also found that the substitution patterns observed for over one third of the 263 residues of mature TEM-1 is quantitatively explained by monotonic dependencies on simple amino acid descriptors and through predictions of stability changes. Most of the explained patterns reflect effects on stability of the protein core, stability of hydrophobic clusters beneath the surface, stability of polar networks and salt bridges on the protein surface, amino acid solubility and backbone flexibility. Other descriptors point at very specific features related to backbone conformations or even substrate binding and catalysis themselves. Notably, most of the wild type residues are already (nearly) optimal, corresponding to (nearly) optimal values of the relevant descriptors. Therefore, mutations will tend to be detrimental regarding the relevant descriptor at each site, posing important consequences on the physicochemical aspects of protein evolution and protein design.

Results and Discussion

The information produced by Deng et al. is available in the form of $\Delta\Delta G_{i,j}{}^{stat}$, which measures the frequency of each amino acid *j* at each position *i* of TEM-1's sequence relative to the frequency of the wild type residue, and k_i^* which quantifies the effective number of amino acids tolerated at each position *i*. The k^* parameter runs from 1 for a position fully restricted to one amino acid to 20 for a position in which any amino acid is equally likely. In turn, $\Delta\Delta G^{stat}$ is by definition zero for the wild type amino acid at a given position, negative for those substitutions that appear at a higher frequency than the wild type amino acid and positive for those at lower frequency than wild type.

Residues within 10 Å of the active site or with less than 20% of their surface exposed to the solvent experience strong constraints on amino acid identity

We first tested quantitatively the observation by Deng et al. that TEM-1 is generally tolerant of substitutions except in the surroundings of the active site and in distant positions that presumably confer solubility and stability. A plot of k^* against distance to the active site for all residues (Fig. 1A, where the distance of a given residue to the active site is defined as the distance from its Cα to the average coordinates of Ser70's, Lys73's and Glu166's Cα atoms) reveals that within a radius of 7 Å, each wild type residue tolerates substitutions by other 2–4 amino acids at most ($k^* < 5$). Even within 10 Å, all but two positions have $k^* < 6$, the two outliers being still below 10. The 7 Å sphere includes 9 residues into the "very sensitive" region surrounding the active site, whereas 22 additional residues lie between 7 and 10 Å. Beyond the 10 Å sphere the maximum observed k^* values rise slowly, crossing the limit of 15 for distances longer than 12.5 Å (24 additional residues). Still, within ∼20 Å of the active site (additional 110 residues) there is a higher density of points at low k^* values, and only at longer distances is the distribution of k^* values smoother. In summary, tight constraints shape the protein's sequence within 10 Å of its active site, and they are only slowly relaxed beyond that limit, thus extending to large numbers of residues. This "relaxation" region is approximated in Fig. 1A with a black line that arbitrarily indicates an upper limit of tolerance to substitutions. The slope of this line is ∼1.1 $Å^{-1}$, suggesting that roughly one additional amino acid can substitute for the wild type amino acid for every Å progressing away from the active site, at least for this dataset specific to TEM-1 hydrolyzing ampicillin. The long-reaching effect of the active site agrees with the finding, in several systems including lactamases, that interactions from the second shell and beyond very often fine-tune protein function [29–33].

The plot in Fig. 1B, which depicts k^* versus the fraction of exposed surface area for each residue, shows as a main trend a higher density of points in the region with very low exposed surface and very low k^*: 62.8% of the residues have less than 20% of their surface areas exposed, 75.9% of which have $k^* < 5$. This highlights strict amino acid constraints in most of the protein core, as pointed out by Deng et al. for ampicillin hydrolysis by TEM-1 but consistent with several other studies suggesting this is a rather general effect [3,34,35]. For residues with more

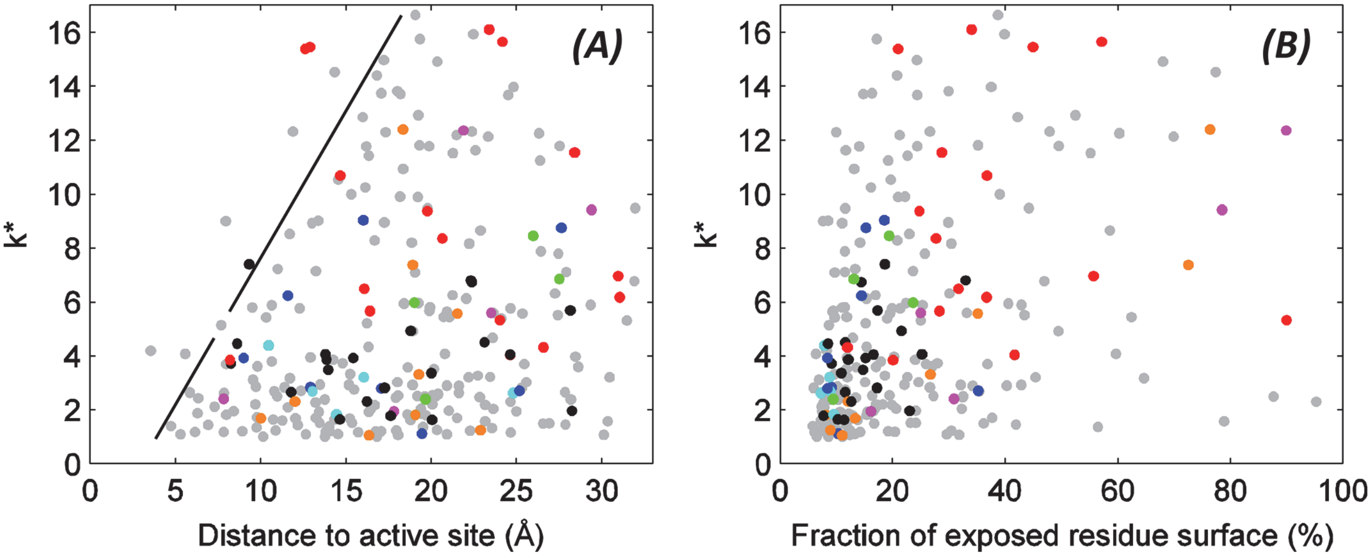

Fig 1. Effective number of substitutions versus distance to the active site and solvent exposure. Effective number of amino acids that appear at each position (*k**) plotted versus the distance from its Cα atom to the active site (A) and against the fraction of amino acid surface exposed to the solvent (B). The position of the active site was defined as the average position of the Ser70, Lys73 and Glu166 Cα atoms; surface exposure was computed with the POPS webserver [48,49] based on the X-ray structure deposited under PDB entry 1XPB [50]. Dot colors correspond to correlations against the most often matched properties: blue for *volume*, cyan for *volume/(P(helix)+P(sheet))*, green for *steric hindrance*, dark green for *steric hindrance / P(sheet)*, red for *hydrophobicity*, magenta for *log(solubility) x hydrophobicity*, orange for *flexibility x hydrophobicity*. Black dots represent correlations with FoldX predictions. Gray dots correspond to the rest of the residues (with other or no detected correlations). The line drawn in panel A arbitrarily indicates how the maximum possible *k** increases with distance to the active site.

doi:10.1371/journal.pone.0118684.g001

than 17% of their surface areas exposed, the smooth distribution of k^* values indicates variable tolerance to substitutions.

Notably, the maximum k^* of 16–17 implies that for any site in the protein, there are always at least 3 or 4 amino acids that are very strongly disallowed. As discussed by Deng et al., tryptophan and proline are the two least tolerated amino acids, bearing the largest $\Delta\Delta G^{stat}$ averages across the sequence and being favored only when they are the native amino acid. This metric can be expanded by computing for each amino acid the average $\Delta\Delta G^{stat}$ across sites where they are nonnative, which yields Trp, Pro, Gly, Phe, Cys, Glu and Arg as the least tolerated amino acids with values of 3.39, 3.23, 3.04, 3.02, 2.96, 2.92 and 2.85. (The average over sites where each amino acid is nonnative differs by > 0.2 from the average over the full sequence for Ala, Glu, Gly and Leu, and by 0.1–0.2 for Asp, Ile, Lys, Pro, Arg, Ser, Thr and Val, altering the relative order of amino acids in the list of least tolerated ones. For example, Gly ranks 5th across the full sequence but 3rd across nonnative sites, and Tyr ranks 7th across the full sequence but 9th across nonnative sites.). This trend argues against substitutions by these residues when designing mutations (for comparison, the other amino acids span $\Delta\Delta G^{stat}$ from 2.29 for Thr to 2.74 for Lys across sites where they are nonnative). Quoting Deng et al., the negative impact of substitutions by Trp may stem from steric clashes of its large side chain, whereas the negative impact of Pro could arise from effects on backbone conformation especially disruption of helical structures. Likewise, Phe can be hard to accommodate due to its hydrophobicity (notably, the more polar Tyr is well tolerated, so a size factor is in principle less important for these two amino acids); Gly can introduce flexibility in the backbone and/or produce void spaces that destabilize the structure; whereas cysteine can foster covalent dimerization or be prone to oxidation introducing large negative charges and possibly formation of reactive free radicals.

Finding out the amino acid descriptors that best explain the observed $\Delta\Delta G^{stat}$ distributions

We then attempted to rationalize the observed $\Delta\Delta G^{stat}$ distributions on the basis of the physicochemical properties of the amino acids. A set of 28 descriptors was assembled that describes different properties of the amino acids with low correlation to each other (S1 Table), based on previous works that developed minimal complete sets of amino acid properties [36,37] and combinations of them. The descriptors were evaluated for those that could explain through linear correlations the distribution of $\Delta\Delta G^{stat}$ values observed at each position of TEM-1 for ampicillin hydrolysis. Dependencies were selected when they produced correlations better than $r = 0.67$ ($r^2 = 0.45$) and RMSD < 1 between experimental and linearly back-predicted $\Delta\Delta G^{stat}$ values, as a proxy for monotonicity. The exact procedure and list of descriptors are provided in Methods.

A total of 105 correlations were obtained between $\Delta\Delta G^{stat}$ values and amino acid descriptors, 73 of which are best correlations (*i.e.* they map to 73 different residues) and 32 are second-best correlations (*i.e.* they satisfy the r and RMSD criteria but slightly less well than another, best correlation). Of the 28 descriptors, 20 could fit at least one site. *Hydrophobicity*, *Volume* and *log(solubility) x Hydrophobicity* were those most often selected by the procedure, accounting for almost one third of the total (S1 Fig.). Nine additional descriptors were selected more than five times: number of O atoms in the side chain, number of N atoms in the side chain, *Volume / log(solubility)*, *P(helix) + P(sheet)* (standing for alpha helical and beta sheet propensities, respectively), *Volume / P(helix)*, *Flexibility* (sensitive to backbone flexibility), *Hydrophobicity x Flexibility*, *Volume / (P(helix) + P(sheet))* and *Steric hindrance*. Examples of correlations between $\Delta\Delta G^{stat}$ and the corresponding descriptors are shown in Fig. 2 (A-R) and discussed in the next subsection. All the correlations are mapped to the sequence in S2 Fig. and can be explored interactively in the spreadsheet provided in the supporting material (S2 Table).

Given the important role of stability constraints in shaping proteins, we further scanned the $\Delta\Delta G^{stat}$ distributions against the changes in folding stability predicted by the program FoldX [38] for each of the 19 mutants at each position of TEM-1 ($\Delta\Delta G^{FoldX}$ values). Like other programs of its kind, FoldX performs poorly at quantitatively reproducing experimental changes in stability but is considered to correctly reflect trends [34,35,39,40], which could potentially fit our goal by helping to explain the $\Delta\Delta G^{stat}$ distributions of residues under complex constraints. The search for correlations between $\Delta\Delta G^{stat}$ and $\Delta\Delta G^{FoldX}$ values results in 25 residue positions for which $r > 0.67$ (S3 Table). All these residues form portions of hydrophobic clusters inside the protein core, indicating that FoldX is especially good at identifying them (*i.e.* black points in Fig. 1B). Only five of them overlap with those to which descriptors were mapped above, hence we can add 20 new explained distributions to our list to reach a total of 93 residue sites for which descriptors accurately model experimental results. Examples of selected dependencies between $\Delta\Delta G^{stat}$ and $\Delta\Delta G^{FoldX}$ are given in Fig. 2S and 2T and discussed below; all of them are available as an interactive spreadsheet in S4 Table.

A third of the amino acid substitutions tolerated by TEM-1 can be quantitatively modeled by the amino acid descriptors

The protocol employed to search for relationships between $\Delta\Delta G^{stat}$ distributions and amino acid descriptors assumes dependencies on only one property, or two at most in the case of composite variables, and that the involved relationships are close to linear. Despite its simplicity, though, it allowed us to quantitatively explain, together with FoldX predictions, over 35% of the observed substitution patterns. This means that the identities of the amino acids in approximately one third of the TEM-1 sequence are shaped mainly by one or two fundamental

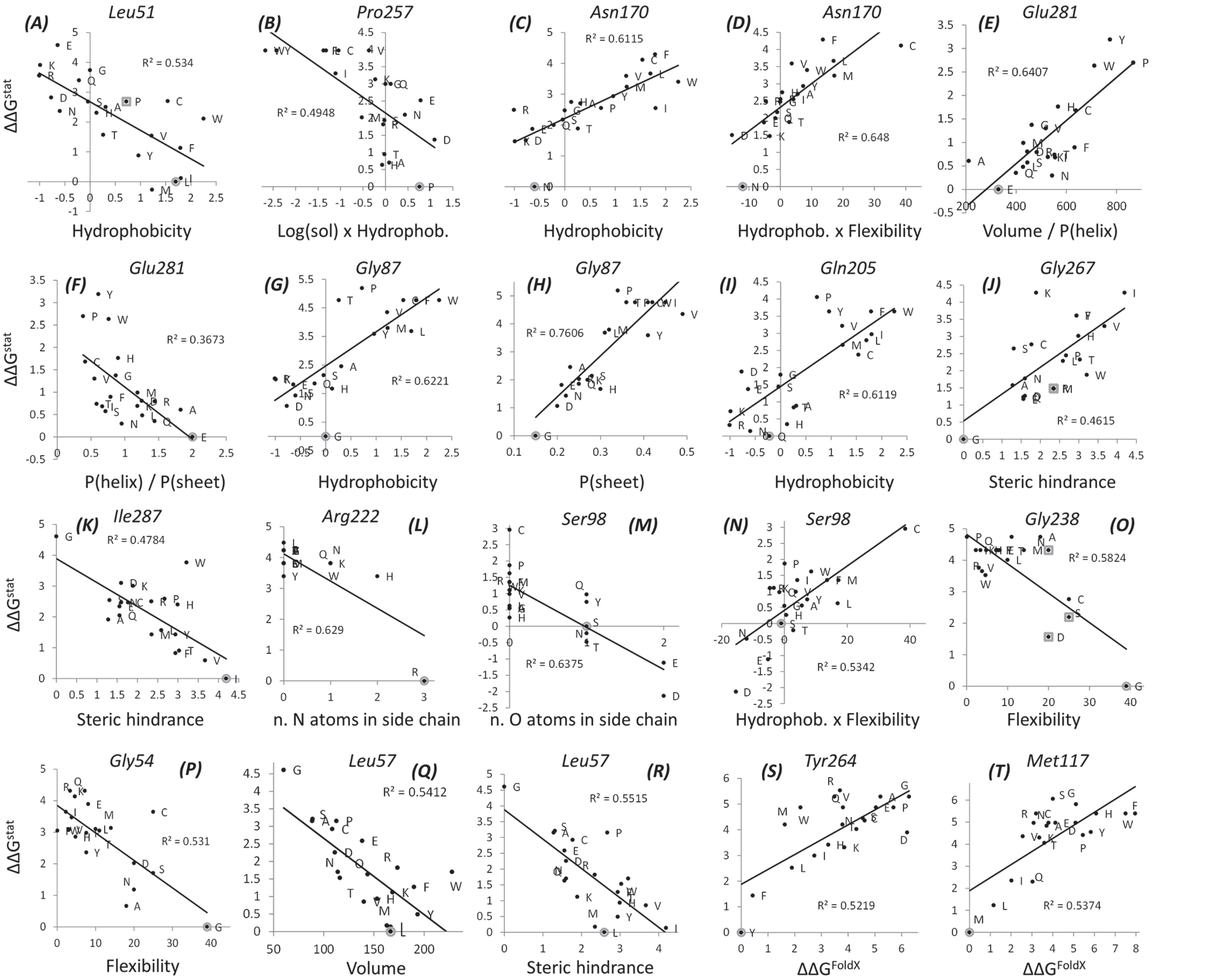

Fig 2. Examples of correlations detected between ΔΔG^{stat} values and amino acid descriptors (a-r) or ΔΔG^{FoldX} (s-t). The gray circles point at the wild type amino acid, gray squares point at substitutions that have been observed in natural TEM variants. Lines correspond to best linear fits.

doi:10.1371/journal.pone.0118684.g002

properties in a monotonic fashion and by well-captured effects on stability, whereas other variables have much less weight in defining the most likely residues at those positions.

Besides the general conclusions outlined in the next subsection, the correlation plots can be interpreted in detail. For example, Leu51, whose side chain forms part of a small hydrophobic cluster with the methyl group of Thr195 and the Pro257 side chain directly beneath the protein surface (Fig. 3A), is sensitive to hydrophobicity favoring more hydrophobic residues (Fig. 2A). Position 257 (Fig. 2B) responds to *log(solubility) x Hydrophobicity*, which means that amino acids simultaneously hydrophobic (to pack with Leu51) and soluble (possibly because polar side chains at this position could snorkel out of the protein) are favored. No property was selected for Thr195, due to the small dispersion of ΔΔG^{stat} values at that position (0.78 against average and median deviations of 1.14 and 1.15 across the full sequence) and its high tolerance to

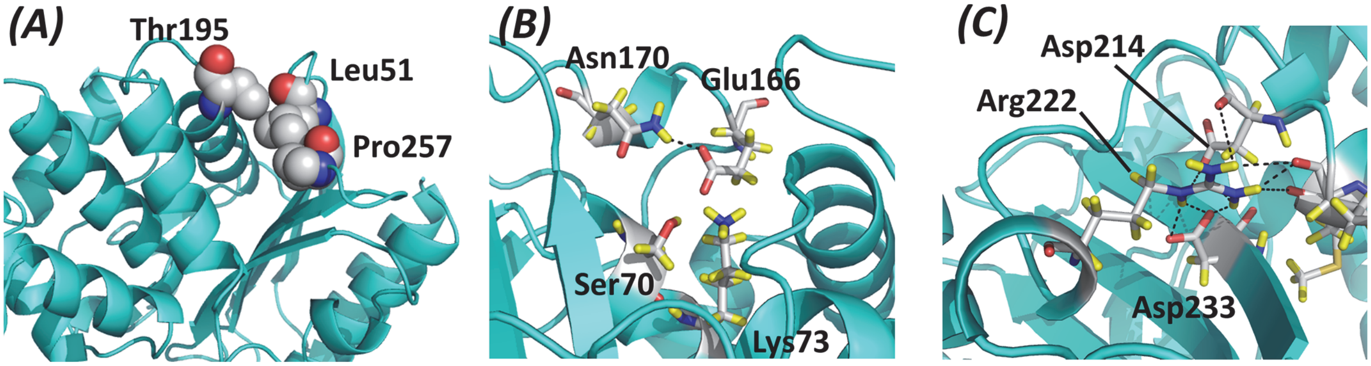

Fig 3. Structural representation of the (a) Leu51 (b) Asn170 (c) Arg222 regions in the TEM-1 β-lactamase structure. Pictures rendered from PDB ID 1XPB [50] using the program PyMOL [51]. Atom colors are red for oxygen, blue for nitrogen, gray for carbon and yellow for hydrogen.

doi:10.1371/journal.pone.0118684.g003

substitution ($k^* = 15.9$). These are examples of constraints that act to stabilize superficial hydrophobic clusters, as opposed to those clusters truly at the protein core which are better described by dependencies on amino acid volume, steric hindrance and FoldX predictions (see next subsection).

Another example is the solvent-exposed Gln205 (Fig. 2I) which exhibits a dependence on hydrophobicity contrary to that of Leu51. Gln205 is selected for its hydrophilicity, probably to aid in solubilizing the protein. Notably in this case, like in a few others, the $\Delta\Delta G^{stat}$ values distribute in a cluster of high probability ($\Delta\Delta G^{stat} \approx 0$ to 1.75) and another of low probability ($\Delta\Delta G^{stat} \approx 2.25$ to 4). Accordingly, hydrophobicity acts as a discrete classifier (being < 0.5 for the tolerated amino acids and > 0.5 for amino acids with low probability) rather than as a variable for continuous modelling.

Asn170, where hydrophilic residues are also favored, provides a more interesting example (Fig. 2C). This residue is on the protein surface but just 20.1% of its surface is exposed, thus it is not clear whether it contributes to solubility. However, Asn170 forms a hydrogen bond to Glu166, an active site residue, and could thus play an auxiliary catalytic role modulating the pKa and/or orientation of its carboxylate group (Fig. 3B). This is expected to be an important constraint, and indeed, the native amino acid is an outlier itself in the plot against hydrophobicity, being much more preferred than any other amino acid. This offset is observed in many correlations, suggesting that the wild type residues are largely preferred due to very specific reasons at some locations, although they can be substituted under certain constraints. In this example, Asn170 might be replaced by other residues that will not form exactly the same hydrogen bond with Glu166 but will at least preserve the polarity of the region. Notice that the composite variable *Hydrophobicity* x *Flexibility* explains slightly better the $\Delta\Delta G^{stat}$ distribution for Asn170, which could account for a secondary need for rigidity to better position the side chains for hydrogen bond formation.

The next two cases are examples of unsuspected relationships that this analysis helped unveil. First, Glu281 correlates positively with *Volume/P(helix)* (Fig. 2E) and negatively with *P(helix)/P(sheet)* (Fig. 2F). Other properties, such as hydrophobicity or solubility, do not correlate well ($r^2 = 0.1879$ and 0.040, respectively) although they would be hypothesized to be important given that this residue's side chain is exposed. Instead, the detected dependences reflect helical constraints on the backbone conformation, which must be preferably achieved with small amino acids. In the second case, for Gly87, the two descriptors *Hydrophobicity* and *P(sheet)* explain equally well the observed distributions (Fig. 2G and 2H). This residue is located in a tight turn on the protein surface and is 42% exposed, suggesting that the negative correlation with hydrophobicity reflects its role in conferring solubility. Notice that glycine is an important outlier in the plot against *Hydrophobicity*, probably due to the important conformational constraint revealed by its correlation against *P(sheet)*. The finding of two important dependencies might point at the requirement of a malleable (*i.e.* capable of adopting several defined conformations) and polar amino acid at this position rather than simply a "flexible" one, because flexibility does not correlate as well ($r^2 = 0.24$). Notice that although glycine residues are usually attributed a role in conferring flexibility (which is indeed observed for most glycines located in loops, see next subsection), our analysis suggests that this particular glycine would fulfill other roles.

Arg222 and Ser98 provide examples where the atomic composition of the side chain describes the observed distributions better than any physicochemical property. The guanidinum group of Arg222 forms multiple hydrogen bonds and salt bridges with the carboxylate groups of Asp214 and Asp233 and with three backbone oxygen atoms, which effectively closes a loop located at the protein surface (Fig. 3C). Thus, arginine (3 nitrogen atoms on the side chain) is by far the preferred residue, followed by histidine (2 N atoms) and then the other amino acids (Fig. 2L). Similarly, Ser98 is roughly as likely as Thr or Asn, with the three of them having one oxygen atom in the side chain, whereas Asp and Glu (2 oxygens) are more favored than any of those three, and amino acids with no oxygens in their side chains are the least favored (Fig. 2M). Ser98 is located in a small loop closed through extensive hydrogen bonds to its alcohol group, where carboxylate groups could accommodate more interactions.

Major constraints focus on stability of the protein core and of superficial hydrophobic clusters, followed by constraints on salt bridges and polar interactions at the surface, flexibility and solubility

A more general interpretation of the results follows. For this, the descriptors most often selected are color-coded into the dots in Fig. 1 and mapped on TEM-1's structure in Fig. 4. Different descriptors explain the $\Delta\Delta G^{stat}$ distributions (or in other words, shape the sequence constraints) at different locations of the protein. Correlations with $\Delta\Delta G^{FoldX}$ reflect the importance of hydrophobic clusters at the core (Fig. 4A and black dots in Fig. 1B) suggesting in turn that FoldX might perform particularly well for predictions of destabilization induced upon mutation deep inside protein cores. On top, correlations with *Volume/(P(helix)+P(sheet))*, *Volume*, *Steric hindrance* and *Steric hindrance/P(sheet)* mirror the need for small volumes and special residue packing at many locations inside the protein (Fig. 4B, D, E, H and Fig. 1B in cyan, blue, green, dark green, respectively). Finally, correlations against numbers of O or N atoms, *number of hydrogen bonds* or *Isoelectric point* map to residues involved in salt bridges and hydrogen bonds on the polar surface, like in the examples presented for Arg222 and Ser98.

Dependencies on *Hydrophobicity*, *Hydrophobicity x Flexibility* and *log(solubility) x Hydrophobicity* map to superficial sites of variable solvent exposure (Fig. 4C, F, I and Fig. 1B in red, magenta and orange, respectively). These residues would play roles in conferring solubility (cases where more polar or soluble amino acids are favored) and in compacting superficial hydrophobic clusters (more hydrophobic or less soluble amino acids). Six positions where flexibility was selected by the procedure (Fig. 4G) suggest an important contribution of dynamics at these locations, five of them actually mapping to loops. Moreover, three map to loops around the active site in the broad region that pivots the hinge motion linked to active site opening and closing as computed by normal mode analysis (S3 Fig.).

As a final note we would like to posit the idea that despite the analyzed data is specific for TEM-1 hydrolyzing ampicillin and thus most interpretations are valid under that setting, a

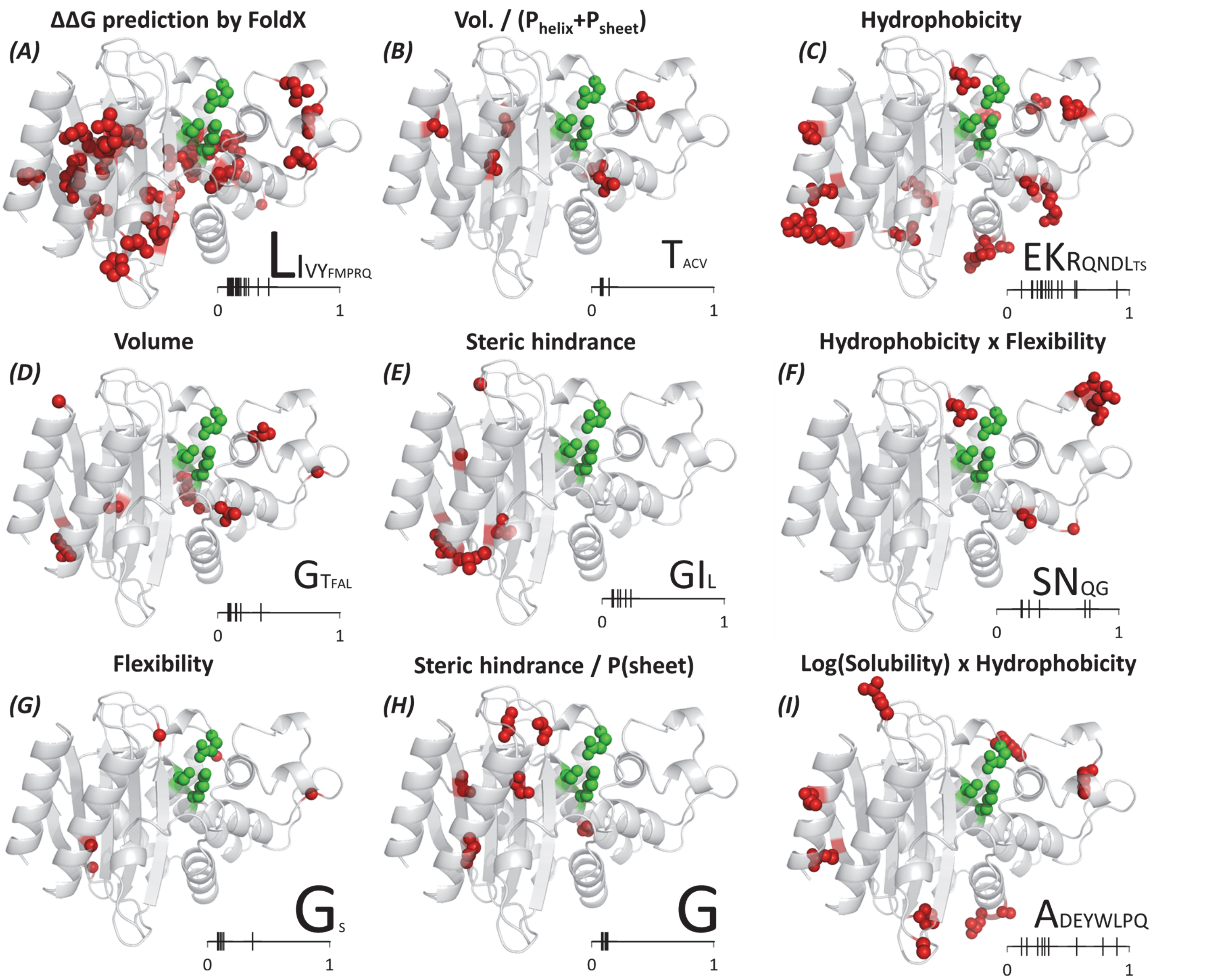

Fig 4. Structure mapping of the residues whose substitution patterns can be explained by the nine most common descriptors. The mapped amino acids are shown as red spheres, and residues Ser70, Lys73 and Glu166 as green spheres. All residue representations lack the main chain nitrogen, carbonyl carbon and oxygen atoms for clarity. The letters on the bottom right of each panel indicate the wild type amino acids most often found at the indicated locations, with the font size being roughly proportional to the relative number of occurrences of the amino acid. The small bar on the bottom right of each panel measures the fractional solvent exposure of the wild type residues to which the descriptor was mapped.

doi:10.1371/journal.pone.0118684.g004

significant fraction of the fits could reflect generalizable trends, especially those concerning protein stability and solubility.

The unexplained distributions

We have thought of three main possible explanations for why no descriptors were retrieved for the $\Delta\Delta G^{stat}$ distributions of the remaining residues. First, as in the case of Thr195 described above, many positions have too low or too high k^* and too little dispersion of the $\Delta\Delta G^{stat}$ values, which simply means that there is no trend to search for, or that a trend can be blurred by the uncertainties in $\Delta\Delta G^{stat}$. Indeed, low standard deviations in the $\Delta\Delta G^{stat}$ values are associated with either very low or very high k^* values (S4 Fig.), implying large stringency or large tolerance to substitutions at those positions, respectively. We estimate that $\sim$ 10% of the protein sites are under such situation.

A second important difficulty in finding descriptors to account for $\Delta\Delta G^{stat}$ distributions might arise from multiple constraints that affect several properties simultaneously, and/or that follow non-monotonic dependencies on the descriptor(s) such that intermediate (rather than minimum or maximum) values are optimal. In principle, fitting to nonlinear functions and combinations of descriptors could unveil these patterns, but it would be difficult to assess the statistical significance of different fits. By focusing only on monotonic dependencies, we have concentrated on strong effects. Therefore, our analysis was limited only to the monotonic dependencies presented above, together with the good FoldX predictions. In a test protocol fitting unexplained $\Delta\Delta G^{stat}$ distributions to quadratic equations, which could roughly point at optimal values of a descriptor in which $\Delta\Delta G^{stat}$ is maximized or minimized, it was found that an additional 10–15% of distributions could be explained (examples of such fits are shown in S5 Fig.).

Finally, $\Delta\Delta G^{stat}$ distributions may be complicated by effects that do not mirror the inherent physics and chemistry of the protein, as discussed in literature [11,19] For example, codon substitutions could affect transcription or translation leading to premature transcription stops or ribosome stalling. Furthermore, effects on protein stability could be impacted by contributions from differential sensitivity to proteases. From what is left after the approximations given above, these undefined effects would apply to roughly 40–45% of the TEM-1 sequence positions.

Concluding Remarks

The analysis presented here highlights the fundamental factors that shape an enzyme's tolerance to substitutions, which ultimately dictate its accessible sequence space. These relationships are tightly linked to the principles of protein engineering and protein evolution. In particular, these analyses detect which protein trait(s) are expected to be sacrificed upon mutation either in an evolutionary setting or during rational design, beyond stability as usually recognized. For example, substitution of Gly238 by Ser, as observed in many extended-spectrum TEM variants, produces a decrease in flexibility, which is the most important property at that position (Fig. 2, panel O). If the detriment is too strong, the substitution might need to be buffered or compensated by additional substitutions. Regarding protein design, the analysis also suggests quantitative guidelines on where not to incorporate substitutions, which residues should generally be avoided when designing mutations, and which properties must be considered when designing substitutions based on the location of the residue in the protein structure. Our analysis does not allow us to unambiguously distinguish contributions to general traits (such as stability or solubility) from those to specific traits (such as effects on substrate binding or on the catalytic mechanism), although many observations could be rather general considering what is known about protein physical chemistry. In this regard, it would be very interesting and informative to analyze a similar dataset acquired under selection against a different β-lactam.

β-lactamases have been a particularly interesting target of structural biology studies, stemming from their clinical relevance and from a large experience accumulated on them, with tens to hundreds of characterized variants [41–46]. But as shown here and in the original work by Deng et al., competition experiments taking advantage of deep sequencing technologies can provide much larger amounts of information, and under homogeneous backgrounds and laboratory conditions. We expect that analyses of the kind introduced here will allow extraction of the fine structure information available in future datasets. Notably, no results like those reported here could be obtained from an alignment of natural TEM sequences, probably because of a lack of variability and/or confounding effects from epistatic interactions, or from other deep sequencing datasets of lower resolution [19,20]. The specific problem of epistatic interactions cannot be addressed with the available data, but deep sequencing technologies applied to large libraries of pairwise substitutions could shed light on how it builds up from non-additive effects of mutations. This is hard to test from low-throughput experiments or from natural variability but is an essential element of protein evolution and engineering. Our study demonstrates the power of deep sequencing experiments, particularly when large numbers of sequence reads are obtained.

Methods

The dataset analyzed here corresponds exactly to that in ref. [18], which expands on a study at lower resolution as feasible two decades ago [47]. Specifically, the dataset was built by performing deep sequencing on a plasmid library with nearly full randomization of two to four contiguous codons, selected against 1 mg/ml ampicillin in *E. coli*. Such substitution scheme resulted in relatively smooth variations of the determined tolerance to substitutions at each single residue, thus enabling the detection of trends against amino acid descriptors (in contrast, very few correlations were retrieved from another study based on exclusively single substitutions).

For the analysis of $\Delta\Delta G^{stat}$ distributions in terms of amino acid properties, 28 descriptors were compiled in three sets from values available in references [36,37]. The first set contains five simple, uncorrelated properties ($r^2 < 0.3$): *Volume*, *log(Solubility)*, *Hydrophobicity*, *Isoelectric point*, and *Helix propensity* (*P(helix)*). The second set contains another 17 descriptors, including simple ones that are correlated with those in the first set or with each other by $0.3 < r^2 < 0.61$, the products and ratios of descriptors from the first set ("composite variables"), and discrete descriptors reflecting the atomic composition of the side chain (number of O atoms, number of N atoms) and the maximum number of possible hydrogen bonds. The third set contains six additional simple and composite properties correlated by $0.61 < r^2 < 0.72$ to those in the first two sets. The inclusion of composite variables aims at accounting for $\Delta\Delta G^{stat}$ distributions of sites that are similarly sensitive to two different properties. All the 28 descriptors, their sources, values for each amino acid and correlations to each other are given in Supporting S1 Table.

The search for descriptors that could explain $\Delta\Delta G^{stat}$ distributions at each site was implemented as MS Excel and Matlab scripts. The procedure follows three incremental stages. First, we searched for correlations with the five basic properties (*i.e.* from the simplest set described above). Next, we searched for correlations with variables of the second set only for sites that did not correlate sufficiently well with any property from the first set. The same was then done with properties from the third set. The aim of this protocol was to minimize the inclusion of the more complex descriptors when correlations to simpler variables could explain the $\Delta\Delta G^{stat}$ distributions, yet allow them to be incorporated when no simpler descriptor was sufficient. In all cases, a descriptor was selected if it produced correlations better than 0.67 ($r^2 > 0.45$) and $RMSD < 1$ between experimental and linearly back-predicted $\Delta\Delta G^{stat}$ values. The best and second-best descriptors selected for each position are listed in S2 Table. Notice that in some cases the correlation plots suggest dependences other than linear, but it is hard to assess with statistical significance the true underlying model.

Scanning the first set of five variables led to the substitution patterns for 26 positions correlating with a variable: 16 with hydrophobicity, eight with volume and two with isoelectric point (no positions correlated with *P(helix)* or *log(solubility)*). Extending the analysis to the second set of descriptors, correlations were found for a total of 98 positions; and including also the third set yielded a total of 105 correlations, carried on to the analysis presented under Results and Discussion. All the correlations can be explored interactively in the spreadsheet provided in the supporting material (S2 Table).

Supporting Information

S1 Fig. Distribution of picked descriptors. Distribution of picked descriptors. across the 105 correlations with amino acid properties (FoldX predictions not included).
(DOCX)

S2 Fig. Sequence distribution of descriptors. Sequence distribution of descriptors picked to explain the observed $\Delta\Delta G^{stat}$ values at each site and of sites where $\Delta\Delta G^{stat}$ correlates with $\Delta\Delta G^{FoldX}$, together with plots showing the distance to the active site and the fraction of exposed area for the wild type residue (both extracted from PDB ID 1XPB). When two descriptors were found, the second-best is shown on top.
(DOCX)

S3 Fig. Open and closed conformations of TEM-1. Conformations as retrieved from normal mode analysis on the 1XPB structure (second mode computed with ProDy with a 15 Å cutoff). Spheres map to the Cα atoms of all glycines; green spheres map to residues for which flexibility was selected as the main descriptor explaining $\Delta\Delta G^{stat}$. Picture rendered in VMD.
(DOCX)

S4 Fig. Plot of the standard deviation in $\Delta\Delta G^{stat}$ versus k^*. Plot of the standard deviation in $\Delta\Delta G^{stat}$ versus k^* for all the residues in mature TEM. Green dots correspond to residues whose distribution was explained by the descriptors; red dots are the unexplained residues.
(DOCX)

S5 Fig. Examples of fits to quadratic dependencies. Four examples of fits to quadratic dependencies on the descriptors, which satisfy $r^2 > 0.45$ and RMSD < 1 between experimental and back-predicted $\Delta\Delta G^{stat}$. Notice that these plots do not necessarily imply a strict quadratic dependence on the variable. For example, the plot for Lys73 seems to point out that only very polar residues are allowed, while the plot for Gly143 could point out that only amino acids with very low beta sheet propensity are allowed. On the other hand, the plots for Ser82 and Thr266 seem to truly reflect an optimal balance between two main amino acid properties.
(DOCX)

S1 Table. Amino acid descriptors employed in this work. The 28 amino acid descriptors compiled here, which were scanned to find correlations with $\Delta\Delta G^{stat}$ distributions. The five simplest properties are shown in dark green background, the second set in orange, the third one in pink. Correlations between all descriptors are given.
(XLSX)

S2 Table. Interactively scanning descriptors for correlations with $\Delta\Delta G^{stat}$ data. Interactive spreadsheet in which the 28 descriptors can be scanned for the best and second-best correlations. Clicking "up" and "down" moves through the protein sequence. Only cases for which $r^2 > 0.45$ and RMSD < 1 are displayed. This file is compatible with Microsoft Excel version 2010 and up (active content must be enabled), but not with Excel for Mac OS.
(XLSM)

S3 Table. List of descriptors selected by the scanning procedure. List of TEM-1 residues showing the best descriptor accounting for each $\Delta\Delta G^{stat}$ distribution (labeled "1"), the second-best descriptor ("2") and good correlations with $\Delta\Delta G^{FoldX}$.
(XLSX)

S4 Table. Interactive testing of $\Delta\Delta G^{stat}$ fits against $\Delta\Delta G^{FoldX}$ predictions. Interactive spreadsheet with which the $\Delta\Delta G^{stat}$ values can be compared against $\Delta\Delta G^{FoldX}$ predictions. Clicking "up" and "down" moves through the protein sequence. This file is compatible with Microsoft Excel version 2010 and up (active content must be enabled), but not with Excel for Mac OS. (XLSM)

Author Contributions

Conceived and designed the experiments: LAA MDP. Performed the experiments: LAA. Analyzed the data: LAA TP MDP. Contributed reagents/materials/analysis tools: LAA MDP. Wrote the paper: LAA TP MDP.

References

1. Aharoni A, Gaidukov L, Khersonsky O, McQ Gould S, Roodveldt C, et al. (2005) The "evolvability" of promiscuous protein functions. Nat Genet 37, 73–76 PMID: 15568024
2. Thorne JL (2007) Protein evolution constraints and model-based techniques to study them. Curr Opin Struct Biol 17, 337–341 PMID: 17572082
3. Tokuriki N, Tawfik DS (2009) Stability effect of mutations and protein evolvability. Curr Opin Struct Biol 19, 596–604 doi: 10.1016/j.sbi.2009.08.003 PMID: 19765975
4. Fornasari MS, Parisi G, Echave J (2007) Quaternary structure constraints on evolutionary sequence divergence. Mol Biol Evol 24, 349–351 PMID: 17124181
5. Meyer AG, Wilke CO (2013) Integrating sequence variation and protein structure to identify sites under selection. Mol Biol Evol 30, 36–44 doi: 10.1093/molbev/mss217 PMID: 22977116
6. Parisi G, Echave J (2005) Generality of the structurally constrained protein evolution model: assessment on representatives of the four main fold classes. Gene 345, 45–53 PMID: 15716088
7. Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL (2003) Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 20, 1692–1704 PMID: 12885968
8. Rodrigue N, Lartillot N, Bryant D, Philippe H (2005) Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene 347, 207–217 PMID: 15733531
9. Thorne JL, Choi SC, Yu J, Higgs PG, Kishino H (2007) Population genetics without intraspecific data. Mol Biol Evol 24, 1667–1677 PMID: 17470435
10. Goldman N, Thorne JL, Jones DT (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149, 445–458 PMID: 9584116
11. Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, et al. (2012) The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. 21, 769–785 doi: 10.1002/pro.2071 PMID: 22528593
12. Yu J, Thorne JL (2006) Testing for spatial clustering of amino acid replacements within protein tertiary structure. J Mol Evol 62, 682–692 PMID: 16752209
13. Kaltenbach M, Tokuriki N (2014) Dynamics and constraints of enzyme evolution. J Exp Zool B Mol Dev Evol doi: 10.1002/jez.b.22562
14. Araya CL, Fowler DM (2011) Deep mutational scanning: assessing protein function on a massive scale. Trends Biotechnol. 29 435–442 doi: 10.1016/j.tibtech.2011.04.003 PMID: 21561674
15. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, et al. (2010) High-resolution mapping of protein sequence-function relationships. Nat Methods 7, 741–746 doi: 10.1038/nmeth.1492 PMID: 20711194
16. Hecht M, Bromberg Y, Rost B (2013) News from the protein mutability landscape. J Mol Biol 425, 3937–3948 doi: 10.1016/j.jmb.2013.07.028 PMID: 23896297
17. Hietpas R, Roscoe B, Jiang L, Bolon DN (2012) Fitness analyses of all possible point mutations for regions of genes in yeast. Nat Protoc 7, 1382–1396 doi: 10.1038/nprot.2012.069 PMID: 22722372
18. Deng Z, Huang W, Bakkalbasi E, Brown NG, Adamski CJ, et al. (2012) Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. J Mol Biol 424, 150–167 doi: 10.1016/j.jmb.2012.09.014 PMID: 23017428
19. Firnberg E, Labonte JW, Gray JJ, Ostermeier M (2014) A comprehensive, high-resolution map of a gene's fitness landscape. Mol Biol Evol 31, 1581–1592 doi: 10.1093/molbev/msu081 PMID: 24567513

20. Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E, et al. (2013) Capturing the mutational landscape of the beta-lactamase TEM-1. Proc Natl Acad Sci USA 110, 13067–13072 doi: 10.1073/pnas.1215206110 PMID: 23878237

21. McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R (2012) The spatial architecture of protein function and adaptation. Nature 491, 138–142 doi: 10.1038/nature11500 PMID: 23041932

22. 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, 1363–1377 doi: 10.1016/j.jmb.2013.01.032 PMID: 23376099

23. Starita LM, Pruneda JM, Lo RS, Fowler DM, Kim HJ, et al. (2013) Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. Proc Natl Acad Sci USA 110, E1263–1272 doi: 10.1073/pnas.1303309110 PMID: 23509263

24. Traxlmayr MW, Hasenhindl C, Hackl M, Stadlmayr G, Rybka JD, et al. (2012) Construction of a stability landscape of the CH3 domain of human IgG1 by combining directed evolution with high throughput sequencing. J Mol Biol 423, 397–412 doi: 10.1016/j.jmb.2012.07.017 PMID: 22846908

25. Kolkman JA, Stemmer WP (2001) Directed evolution of proteins by exon shuffling. Nat Biotechnol 19, 423–428 PMID: 11329010

26. Arnold FH, Wintrode PL, Miyazaki K, Gershenson A (2001) How enzymes adapt: lessons from directed evolution. Trends Biochem Sci 26, 100–106 PMID: 11166567

27. Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol 10, 866–876 doi: 10.1038/nrm2805 PMID: 19935669

28. Bloom JD (2014) An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. Mol Biol Evol doi: 10.1093/molbev/msu220

29. Drawz SM, Bethel CR, Hujer KM, Hurless KN, Distler AM, et al. (2009) The role of a second-shell residue in modifying substrate and inhibitor interactions in the SHV beta-lactamase: a study of ambler position Asn276. Biochemistry 48, 4557–4566 doi: 10.1021/bi9003292 PMID: 19351161

30. González LJ, Moreno DM, Bonomo RA, Vila AJ (2014) Host-specific enzyme-substrate interactions in SPM-1 metallo-β-lactamase are modulated by second sphere residues. PLoS Pathog 10, e1003817 doi: 10.1371/journal.ppat.1003817 PMID: 24391494

31. Marshall NM, Garner DK, Wilson TD, Gao YG, Robinson H, et al. (2009) Rationally tuning the reduction potential of a single cupredoxin beyond the natural range. Nature 462, 113–116 doi: 10.1038/nature08551 PMID: 19890331

32. Morgada MN, Abriata LA, Zitare U, Alvarez-Paggi D, Murgida DH, et al. (2014) Control of the electronic ground state on an electron-transfer copper site by second-sphere perturbations. Angew Chem Int Ed Engl 53, 6188–6192 doi: 10.1002/anie.201402083 PMID: 24777732

33. Tomatis PE, Rasia RM, Segovia L, Vila AJ (2005) Mimicking natural evolution in metallo-beta-lactamases through second-shell ligand mutations. Proc Natl Acad Sci U S A 102, 13761–6 PMID: 16172409

34. Huang F, Nau WM (2003) A conformational flexibility scale for amino acids in peptides. Angew Chem Int. Ed Engl 42, 2269–2272 PMID: 12772159

35. Meiler J, Müller M, Zeidler A, Schmäschke F (2001) Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. J Mol Model 7, 360–369

36. Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320, 369–387 PMID: 12079393

37. Khan S, Vihinen M (2010) Performance of protein stability predictors. Hum Mutat 31, 675–684 doi: 10.1002/humu.21242 PMID: 20232415

38. Potapov V, Cohen M, Schreiber G (2009) Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Eng Des Sel 22, 553–560 doi: 10.1093/protein/gzp030 PMID: 19561092

39. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS (2007) The stability effects of protein mutations appear to be universally distributed. J Mol Biol 369, 1318–1332 PMID: 17482644

40. Tokuriki N., Stricher F, Serrano L, Tawfik DS (2008) How protein stability and new functions trade off. PLoS Comput Biol 4, e1000002 doi: 10.1371/journal.pcbi.1000002 PMID: 18463696

41. Abriata LA, Salverda MLM, Tomatis PE (2012) Sequence-function-stability relationships in proteins from datasets of functionally annotated variants: the case of TEM β-lactamases. FEBS Lett 586, 3330–3335 doi: 10.1016/j.febslet.2012.07.010 PMID: 22850115

42. Galán J-C, González-Candelas F, Rolain J-M, Cantón R (2013) Antibiotics as selectors and accelerators of diversity in the mechanisms of resistance: from the resistome to genetic plasticity in the β-lactamases world. Front Microbiol 4, 9 doi: 10.3389/fmicb.2013.00009 PMID: 23404545

43. Salverda MLM, De Visser JAGM, Barlow M (2010) Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. FEMS Microbiol Rev 34, 1015–1036 doi: 10.1111/j.1574-6976.2010.00222.x PMID: 20412308

44. Thai QK, Bös F, Pleiss J (2009) The Lactamase Engineering Database: a critical survey of TEM sequences in public databases. BMC Genomics 10, 390 doi: 10.1186/1471-2164-10-390 PMID: 19698099

45. Widmann M, Pleiss J, Oelschlaeger P (2012) Systematic analysis of metallo-β-lactamases using an automated database. Antimicrob Agents Chemother 56, 3481–3491 doi: 10.1128/AAC.00255-12 PMID: 22547615

46. Novais A, Comas I, Baquero F, Cantón R, Coque TM, et al. (2010) Evolutionary trajectories of beta-lactamase CTX-M-1 cluster enzymes: predicting antibiotic resistance. PLoS Pathog 6, e1000735 doi: 10.1371/journal.ppat.1000735 PMID: 20107608

47. Huang W, Petrosino J, Hirsch M, Shenkin PS, Palzkill T (1996) Amino acid sequence determinants of beta-lactamase structure and activity. J Mol Biol 258, 688–703 PMID: 8637002

48. Cavallo L, Kleinjung J, Fraternali F (2003) POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. Nucleic acids research 31, 3364–3366 PMID: 12824328

49. Fornili A, Autore F, Chakroun N, Martinez P, Fraternali F (2012) Protein-water interactions in MD simulations: POPS/POPSCOMP solvent accessibility analysis, solvation forces and hydration sites. Methods Mol Biol 819, 375–392 doi: 10.1007/978-1-61779-465-0_23 PMID: 22183548

50. Fonzé E, Charlier P, To'th Y, Vermeire M, Raquet X, et al. (1995) TEM1 beta-lactamase structure solved by molecular replacement and refined structure of the S235A mutant. Acta Crystallogr D Biol Crystallogr 51, 682–694 PMID: 15299797

51. DeLano WL (2002) The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA.