An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs

Jesse D. Bloom

Division of Basic Sciences and Computational Biology Program,

Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

E-mail: jbloom@fhcrc.org.

Submitted as an Article (Methods classification) to Mol Biol Evol

Abstract

Phylogenetic analyses of molecular data require a quantitative model for how sequences evolve. Traditionally, the details of the site-specific selection that governs sequence evolution are unknown, and so most phylogenetic models treat this selection crudely with a variety of free parameters designed to represent general features of mutation and selection. However, recent advances in high-throughput experiments have made it possible to quantify the effects of all single mutations on gene function. I have previously shown that such high-throughput experiments can be combined with knowledge of underlying mutation rates to create a parameter-free evolutionary model that describes the phylogeny of influenza nucleoprotein far better than existing models. Here I extend this work by showing that published experimental data on TEM-1 beta-lactamase (Firnberg et al., 2014) can be combined with a few mutation rate parameters to create an evolutionary model that describes beta-lactamase phylogenies much better than existing models. This experimentally informed evolutionary model is superior even for homologs that are substantially diverged (about 35% divergence at the protein level) from the TEM-1 parent that was the subject of the experimental study. These results suggest that experimental measurements can inform phylogenetic evolutionary models that are applicable to homologs that span a substantial range of sequence divergence.

Introduction

Most approaches for the phylogenetic analysis of gene sequences require a quantitative evolutionary model specifying the rate at which each site substitutes from one identity to another. In maximum-likelihood and Bayesian approaches, the evolutionary model is used to calculate the likelihood of the observed sequences given the phylogenetic tree (Felsenstein, 1981; Huelsenbeck et al., 2001). In distance-based approaches, the evolutionary model is used to calculate the distances between pairs of sequences (Saitou and Nei, 1987; Hasegawa et al., 1985). For all these approaches, inaccurate evolutionary models can lead to errors in inferred phylogenetic properties, including incorrect estimates of evolutionary distances (Halpern and Bruno, 1998) and incorrect tree topologies (Felsenstein, 1978; Huelsenbeck and Hillis, 1993).

Unfortunately, existing phylogenetic evolutionary models are extreme simplifications of the actual process of mutation and selection that shapes sequence evolution (Thorne et al., 2007). At least two major unrealistic assumptions afflict these models. First, in order to make phylogenetic algorithms computationally tractable, it is generally assumed that each site within a gene evolves independently. Second, most evolutionary models compound the first assumption of independence among sites with the second unrealistic assumption that all sites evolve identically – a severely flawed assumption since there is overwhelming evidence that proteins have strong preferences for certain amino acids at specific sites (Ashenberg et al., 2013; Halpern and Bruno, 1998). It is the second of these unrealistic assumptions that is remedied by the experimentally informed evolutionary model described here.

A major reason that most phylogenetic evolutionary models assume that sites evolve identically is because there has traditionally been insufficient information to do better. In the absence of *a priori* knowledge about selection on individual sites, the parameters of an evolutionary model must be inferred from the same sequences that are being analyzed phylogenetically. For instance, typical codon-level models infer parameters describing the equilibrium frequencies of different codons, the relative rates of transition and transversion mutations, the relative rates of nonsynonymous and synonymous mutations, and in many cases the shapes of distributions that allow some of these rates to be drawn from several categories (Goldman and Yang, 1994; Muse and Gaut, 1994; Yang et al., 2000; Kosiol et al., 2007). There are generally sufficient data to infer these parameters once for a single general model that applies to all sites in a gene – but inferring them separately for each site leads to a proliferation of free parameters that can overfit the sequence data (Posada and Buckley, 2004). Some studies have attempted to predict site-specific substitution rates or clas-

sify sites based on knowledge of the protein structure (Thorne et al., 1996; Goldman et al., 1998; Scherrer et al., 2012; Rodrigue et al., 2009; Kleinman et al., 2010) – however, such approaches are limited by the fact that the relationship between protein structure and site-specific selection is complex, and cannot be reliably predicted even by state-of-the-art molecular modeling (Potapov et al., 2009). An alternative approach is to treat the site-specific substitution probabilities as free parameters that are fit to the sequence data (Lartillot and Philippe, 2004; Le et al., 2008; Wu et al., 2013; Wang et al., 2008) – however, in order to restrain the proliferation of such parameters to a manageable level, these approaches must unrealistically constrain sites to fall in a small number of different substitution-model classes. Therefore, purely computational approaches have proven insufficient for creating evolutionary models that accurately represent the highly idiosyncratic site-specific selection that shapes sequence evolution.

However, this problem is beginning to be transformed by a new type of high-throughput experiment: deep mutational scanning (Fowler et al., 2010; Araya and Fowler, 2011). In deep mutational scanning, a gene is randomly mutagenized and subjected to functional selection in the laboratory, and then deep sequencing is used to quantify the relative frequencies of mutations before and after selection. In cases where the laboratory selection is sufficiently representative of the gene's real biological function, these experiments provide information that can be used to approximate the site-specific natural selection on mutations. To date, deep mutational scanning has been used to quantify the impact of most nucleotide or codon mutations to several proteins or protein domains (Fowler et al., 2010; Roscoe et al., 2013; Starita et al., 2013; Melamed et al., 2013; Traxlmayr et al., 2012; McLaughlin Jr et al., 2012; Firnberg et al., 2014; Bloom, 2014). For a few of these studies, the experimental coverage of possible mutations is sufficiently comprehensive to define site-specific amino-acid preferences for all positions in a gene.

I have previously shown that such experimentally determined site-specific amino-acid preferences can be combined with measurements of mutation rates to create a parameter-free evolutionary model that describes the phylogeny of influenza nucleoprotein far better than existing models that contain numerous free parameters (Bloom, 2014). Here I extend that work by showing that it is also possible to create an experimentally informed evolutionary model for another protein. I do this using deep mutational scanning data published by Firnberg et al. (2014) that quantifies the effects of nearly all amino-acid mutations on TEM-1 beta-lactamase. In this case, no measurements of mutation rates are available, so I construct an evolutionary model that is informed by the experimentally measured site-specific amino-acid preferences but also contains a few free parameters representing the mutation rates. I show that this evolutionary model greatly improves

the phylogenetic fit to both TEM and SHV beta-lactamases, the latter of which are substantially diverged (about 35% divergence at the protein level) from the TEM-1 parent that was the subject of the deep mutational scanning by Firnberg et al. (2014). These results generalize previous work on experimentally determined evolutionary models, and suggest that site-specific amino-acid preferences are sufficiently conserved during evolution to be applicable to gene homologs that span a substantial range of sequence divergence.

Results

Evolutionary model with known amino-acid preferences and unknown mutation rates

Summary of evolutionary model

I have previously described a codon-level phylogenetic evolutionary model for influenza nucleoprotein for which both the site-specific amino-acid preferences and the nucleotide mutation rates (assumed to be identical across sites) were determined experimentally (Bloom, 2014). The current work examines a protein for which the site-specific amino-acid preferences have been measured experimentally, but for which the nucleotide mutation rates are unknown. It is therefore necessary to extend the evolutionary model to treat the nucleotide mutation rates as unknown free parameters. Here I describe this extension.

In the model used here, the rate $P_{r,xy}$ of substitution from codon x to some other codon y at site r is

$$P_{r,xy} = Q_{xy} \times F_{r,xy}, \qquad \text{(Equation 1)}$$

where Q_{xy} denotes the rate of mutation from x to y, and $F_{r,xy}$ gives the probability that a mutation from x to y fixes if it occurs. This equation assumes that mutation rates are uniform across sites, and that the selection on mutations is site-specific but site-independent (i.e. the fixation probability at one site is not influenced by mutations at other sites).

Fixation probabilities from amino-acid preferences

The fixation probability of a mutation from codon x to y is assumed to depend only on the encoded amino acids $\mathcal{A}(x)$ and $\mathcal{A}(y)$, as synonymous mutations are assumed to be selectively neutral. The fixation probabilities $F_{r,xy}$ are defined in terms of the experimentally measured amino-acid preferences at site r, where $\pi_{r,a}$ denotes the preference for amino-acid a at site r, and the preferences at each site sum to one ($1 = \sum_a \pi_{r,a}$). As in previous work (Bloom, 2014), I consider two different mathematical relationships between the amino-acid preferences and the fixation probabilities. The first relationship derives from considering the amino-acid preferences to be directly related to

selection coefficients, and is given by Halpern and Bruno (1998) as

$$F_{r,xy} = \begin{cases} 1 & \text{if } \pi_{r,\mathcal{A}(x)} = \pi_{r,\mathcal{A}(y)} \\ \frac{\ln\left(\frac{\pi_{r,\mathcal{A}(y)}}{\pi_{r,\mathcal{A}(x)}}\right)}{1-\frac{\pi_{r,\mathcal{A}(x)}}{\pi_{r,\mathcal{A}(y)}}} & \text{otherwise.} \end{cases} \quad \text{(Equation 2)}$$

The second relationship is based on considering the amino-acid preferences to reflect the fraction of genetic backgrounds that tolerate a specific mutation (Bloom et al., 2007), and is equivalent to the Metropolis et al. (1953) sampling criterion:

$$F_{r,xy} = \begin{cases} 1 & \text{if } \pi_{r,\mathcal{A}(y)} \geq \pi_{r,\mathcal{A}(x)} \\ \frac{\pi_{r,\mathcal{A}(y)}}{\pi_{\mathcal{A}(x)}} & \text{otherwise.} \end{cases} \quad \text{(Equation 3)}$$

Both of these relationships share the feature that mutations to higher-preference amino acids fix more frequently than mutations to lower-preference amino acids.

Mutation rates

The rate of mutation Q_{xy} from codon x to y is defined in terms of the underlying rates of nucleotide mutation. Let $R_{m\to n}$ denote the rate of mutation from nucleotide m to n. Then

$$Q_{xy} = \begin{cases} 0 & \text{if } x \text{ and } y \text{ differ by more than on nucleotide} \\ R_{m\to n} & \text{if } x \text{ differs from } y \text{ by a single-nucleotide change of } m \text{ to } n. \end{cases} \quad \text{(Equation 4)}$$

Assuming that the same mutation process operates on both the sequenced and complementary strands of the nucleic acid gives the constraint

$$R_{m\to n} = R_{m_c\to n_c} \quad \text{(Equation 5)}$$

where m_c denotes the complement of nucleotide m, since for example a mutation from A to G on one strand induces a mutation from T to C on the other strand. There are therefore six unique nucleotide mutation rates: $R_{A\to C} = R_{T\to G}$, $R_{A\to G} = R_{T\to C}$, $R_{A\to T} = R_{T\to A}$, $R_{C\to A} = R_{G\to T}$, $R_{C\to G} = R_{G\to C}$, and $R_{C\to T} = R_{G\to A}$. In principle, these mutation rates could be measured experimentally for the system of interest. Such experimental measurements were performed in my previous work on influenza nucleoprotein (Bloom, 2014), and the measured mutation rates

happened to be symmetric ($R_{m\to n} = R_{n\to m}$), which is sufficient to make the overall evolutionary model in Equation 1 reversible.

The protein studied here evolves in bacteria for which the mutation rates have not been measured experimentally. Therefore, these mutation rates must be treated as unknown free parameters. It turns out (see Methods) that the overall evolutionary model defined by Equation 1 is only reversible if the mutation rates are subject to the constraint

$$R_{C\to T} = \frac{R_{A\to G} \times R_{C\to A}}{R_{A\to C}}. \qquad \text{(Equation 6)}$$

This constraint lacks a clear biological motivation, and is assumed purely for the mathematical convenience that it makes the model reversible.

In the absence of independent information to calibrate absolute values for the branch lengths or mutation rates, one of the rates is confounded with the branch-length scaling and so can be assigned an arbitrary value > 0 without affecting the tree or its likelihood. Here the choice is made to assign

$$R_{A\to C} = 1 \qquad \text{(Equation 7)}$$

so that all other mutation rates are defined relative to this rate. With these constraints, there are now four independent mutation rates that must be treated as unknown free parameters:

$$\text{unknown mutation rate parameters} = \begin{cases} R_{A\to G} \\ R_{A\to T} \\ R_{C\to A} \\ R_{C\to G} \end{cases} \qquad \text{(Equation 8)}$$

In practice, these four mutation rate parameters will be estimated at their maximum likelihood values given the sequences and tree topology.

Equilibrium frequencies

Calculation of a phylogenetic likelihood requires assigning evolutionary equilibrium frequencies to the possible codons at the root node in addition to specifying the transition probabilities given by Equation 1. In many conventional phylogenetic models, these equilibrium frequencies are treated as free parameters that are estimated empirically from the sequence data. However, in reality the

equilibrium frequencies are the result of mutation and selection, and so can be calculated as the stationary state of the stochastic process defined by the evolutionary model. Specifically, it can be shown (see Methods) that for the evolutionary model in Equation 1, the equilibrium frequency $p_{r,x}$ of codon x at site r is

$$p_{r,x} = \frac{\pi_{r,\mathcal{A}(x)} \times q_x}{\sum_y \pi_{r,\mathcal{A}(y)} \times q_y} \qquad \text{(Equation 9)}$$

where q_x is given by

$$q_x = \frac{1}{8} \times \frac{\left(R_{A\to C} + R_{A\to G}\right)^{\mathcal{N}_{CG}(x)} \times \left(R_{C\to A} + R_{C\to T}\right)^{\left(3-\mathcal{N}_{CG}(x)\right)}}{\left(R_{A\to C} + R_{A\to G} + R_{C\to A} + R_{C\to T}\right)^3} \qquad \text{(Equation 10)}$$

where $\mathcal{N}_{CG}\left(x\right)$ is the number of C and G nucleotides in codon x. The equilibrium frequencies $p_{r,x}$ are therefore completely determined by knowledge of the experimentally determined amino-acid preferences $\pi_{r,a}$ and the four unknown mutation rate parameters in Equation 8.

Experimentally determined amino-acid preferences for beta-lactamase

The site-specific amino-acid preferences for beta-lactamase were determined using data from a previously published deep mutational scanning experiment performed by Firnberg et al. (2014). Specifically, Firnberg et al. (2014) created nearly all possible amino-acid mutants of TEM-1 beta-lactamase and then used antibiotic selection to enrich for functional variants at various antibiotic concentrations. Next, they used high-throughput sequencing to examine how the frequencies of mutations changed during this functional selection. They analyzed their data to estimate the impact of individual mutations on TEM-1 function, and had sufficient data to estimate the impact of 96% of the $297 \times 19 = 5,453$ possible amino-acid mutations.

Firnberg et al. (2014) report the impact of mutations in terms of what they refer to as the "fitness" effects. The analysis is not done in a true population-genetics framework, so the "fitness" values of Firnberg et al. (2014) may not correspond to fitnesses in the classical sense of the term – but these values certainly possess the basic feature of reflecting the effects of specific mutations on TEM-1 function.

Here I use the "fitness" values provided by Firnberg et al. (2014) to estimate the preferences for each of the 20 amino acids at each site in TEM-1. Specifically, let $w_{r,a}$ be the "fitness" value for mutation to amino-acid a at site r reported by Firnberg et al. (2014) in Data S2 of their paper.

I calculate the preference $\pi_{r,a}$ for a at site r as

$$\pi_{r,a} = \frac{w_{r,a}}{\sum_{a'} w_{r,a'}} \qquad \text{(Equation 11)}$$

where the sum over a' ranges over all 20 amino acids, the wild-type amino acid at site r is assigned a "fitness" of $w_{r,a} = 1$ in accordance with the normalization scheme used by Firnberg et al. (2014), and the $w_{r,a}$ values for the 4% of mutations for which no value is estimated by Firnberg et al. (2014) are set to the average $w_{r,a}$ of all non-wildtype amino acids at site r for which a $w_{r,a}$ value is provided.

The amino-acid preferences calculated in this manner are displayed graphically in Figure 1 along with information about residue secondary structure and solvent accessibility (see Supplementary file 1 for numerical data). As is extensively discussed by Firnberg et al. (2014) in their original description of the data, these preferences are qualitatively consistent with known information about highly constrained positions in TEM-1, and show the expected qualitative patterns of higher preferences for specific (particularly hydrophobic) amino acids at residues that are buried in the protein's folded structure. Here I focus on using these amino-acid preferences in a quantitative phylogenetic evolutionary model as described in the next section.

Experimentally determined amino-acid preferences improve phylogenetic fit

TEM and SHV beta-lactamase phylogenetic trees

To test if evolutionary models informed by the experimentally determined amino-acid preferences are superior to existing alternative models, I compared the fit of various models to beta-lactamase sequence phylogenies. Firnberg et al. (2014) performed their deep mutational scanning on TEM-1 beta-lactamase. There are a large number of TEM beta-lactamases with high sequence identity to TEM-1; the next closest group of lactamases is the SHV beta-lactamases (Bush et al., 1995), which on average have 62% nucleotide and 65% protein identity to TEM beta-lactamases. I assembled a collection of TEM and SHV beta-lactamases from the manually curated Lahey Clinic database (`http://www.lahey.org/Studies/`). These sequences were aligned to TEM-1, and highly similar sequences (sequences that differed by less than four nucleotides) were removed. The resulting alignment contained 85 beta-lactamase sequences (Supplementary file 2), of which 49 were TEM and 36 were SHV.

Maximum-likelihood phylogenetic trees of the TEM and SHV beta-lactamases were con-

structed using *codonPhyML* (Gil et al., 2013) with the codon substitution model of either Goldman and Yang (1994) or Kosiol et al. (2007). The resulting trees are displayed in Figure 2. The two different substitution models give extremely similar tree topologies. In both cases, the trees partition into two clades of closely related sequences, corresponding to the TEM and SHV beta-lactamases.

Experimentally informed models are superior for combined TEM and SHV phylogeny

To compare the evolutionary models, *HYPHY* (Pond et al., 2005) was used to optimize the branch lengths and free parameters of the evolutionary models to their maximum likelihood values on the fixed tree topologies in Figure 2. This analysis showed that the evolutionary models informed by the experimentally determined amino-acid preferences were clearly superior to commonly used alternative codon-substitution models.

Specifically, Table 1 and Table 2 show that the experimentally informed evolutionary models fit the combined TEM and SHV phylogeny with higher likelihoods than any of a variety of commonly used alternative models, regardless of whether the tree topology was estimated using the model of Goldman and Yang (1994) or Kosiol et al. (2007). This superiority is despite the fact that the alternative models (Goldman and Yang, 1994; Kosiol et al., 2007) contain many more free parameters. For instance, the most heavily parameterized alternative model contains 60 empirically estimated equilibrium frequency parameters plus an optimized parameter corresponding to the transition-transversion ratio, two optimized parameters corresponding to a gamma distribution of nonsynonymous-synonymous ratios across sites (Yang et al., 2000), and an optimized parameter corresponding to a distribution of substitution rates across sites (Yang, 1994). In contrast, the experimentally informed models only contain four free parameters (the mutation rates, Equation 8) – yet these experimentally informed models have substantially higher likelihoods. When AIC (Posada and Buckley, 2004) is used to penalize parameters, the superiority of the experimentally informed models is even more clear.

To confirm that the superiority of the experimentally informed models is due to the fact that the deep mutational scanning of Firnberg et al. (2014) captures information about the site-specific amino-acid preferences, I tested evolutionary models in which these preferences were randomized among sites (Table 1, Table 2). As expected, these randomized models were far worse than any of the alternatives, since the randomized preferences no longer correspond to the specific sites for which they were experimentally measured.

Experimentally informed models are superior for individual TEM and SHV phylogenies

The foregoing results show that experimentally informed models are superior for describing the combined TEM and SHV beta-lactamase phylogeny. Given that the amino-acid preferences were determined by experiments using a TEM-1 parent, it is worth asking whether these preferences accurately describe the evolution of both the TEM and SHV sequences, or whether they more accurately describe the TEM sequences (which are closely related to TEM-1, Figure 2) than the SHV sequences (which only have about 65% protein identity to TEM-1, Figure 2). This question is relevant because the extent to which site-specific amino-acid preferences are conserved during protein evolution remains unclear. For instance, while several experimental studies have suggested that such preferences are largely conserved among moderately diverged homologs (Ashenberg et al., 2013; Serrano et al., 1993), a simulation-based study has argued that preferences shift substantially during protein evolution (Pollock et al., 2012; Pollock and Goldstein, 2014). If the site-specific amino-acid preferences are largely conserved during the divergence of the TEM and SHV sequences, then the experimentally informed models should work well for both these groups – but if the preferences shift rapidly during evolution, then the experimentally informed models should be effective only for the closely related TEM sequences.

To test these competing possibilities, I repeated the analysis in the foregoing section separately for the TEM and SHV clades of the overall phylogenetic tree (the red versus blue clades in Figure 2). This analysis found that the experimentally informed evolutionary models were clearly superior to all alternative models for the SHV as well as the TEM clade (Table 3, Table 4, Table 5, Table 6). In fact, the extent of superiority of the experimentally informed model (as quantified by AIC) was greater for the SHV clade than the TEM clade, despite the fact that the former has fewer sequences. These results suggest that the applicability of the experimentally determined amino-acid preferences extends to beta-lactamase homologs that are substantially diverged from the TEM-1 parent that was the specific subject of the experiments of Firnberg et al. (2014).

Comparison of different methods for computing fixation probabilities

In the foregoing analyses, two different mathematical relationships were used to mathematically relate the experimentally determined amino-acid preferences to the substitution probabilities in the evolutionary models. One relationship (Equation 2) is based on a true population-genetics derivation by Halpern and Bruno (1998) under the assumption that the preferences are reflective of selection coefficients for amino acids at specific sites (as well as several other assumptions

that are unlikely to be strictly valid). The other relationship (Equation 3) is one that I suggested in previous work (Bloom, 2014) on the grounds that the amino-acid preferences might be best envisioned not as selection coefficients, but rather as measurements of the *fraction* of genetic backgrounds that tolerate a specific mutation, as would be implied by the evolutionary dynamics described in (Bloom et al., 2007). Although both relationships share the qualitative feature that mutations to higher-preference amino acids are favored over mutations to lower-preference ones, they differ in their quantitative details. In previous work on influenza nucleoprotein (Bloom, 2014), I reported that the relationship in Equation 3 outperformed the one in Equation 2 derived by Halpern and Bruno (1998).

In contrast, for the beta-lactamase sequences studied here, the relationship of Halpern and Bruno (1998) outperforms the one that I suggested in my previous work (Table 1, Table 2, Table 3, Table 4, Table 5, Table 6). The reason for and relevance of these discordant results remains unclear. There are almost certainly differences in the evolutionary conditions (population size, etc) for influenza nucleoprotein and beta-lactamase that influence the relationship between selection coefficients and fixation probabilities. In addition, there are substantial differences between the experiments of Firnberg et al. (2014) on beta-lactamase and my previous work on nucleoprotein – in particular, Firnberg et al. (2014) examine the effects of single mutations to the parental gene, whereas my previous work examined the average effects of individual mutations in variants that often contained multiple mutations. Finally, the experimental measurements are imperfect – in my previous work, the preferences determined by independent biological replicates of the experiments only had a Pearson correlation coefficient of 0.79; Firnberg et al. (2014) do not provide data on the consistency of their measurements across biological replicates, but it seems safe to assume that their experiments are also imperfect. Therefore, further work is probably needed to determine if any meaning can be ascribed to the differences in fit for Equation 2 versus Equation 3, as well as to identify the optimal mathematical relationship for connecting experimentally measured amino-acid preferences to substitution probabilities in evolutionary models. However, both the past and current work strongly suggest that using any reasonable mathematical relationship to inform evolutionary models with experimentally determined amino-acid preferences is sufficient to lead to dramatic improvements in phylogenetic fit.

Discussion

I have shown that an evolutionary model informed by experimentally determined site-specific amino-acid preferences fits beta-lactamase phylogenies better than a variety of existing models that do not utilize experimental information. When considered in combination with prior work demonstrating that an experimentally determined evolutionary model dramatically improves phylogenetic fit for influenza nucleoprotein (Bloom, 2014), these results suggest that experimentally informed models are generally superior for phylogenetic analyses of protein-coding genes. The explanation for this superiority is obvious: proteins have strong preferences for certain amino acids at specific sites (Ashenberg et al., 2013; Halpern and Bruno, 1998), but existing evolutionary models treat all sites identically (or at best partition them into a few crude categories). The use of experimental data on site-specific amino-acid preferences improves evolutionary models by informing them about the complex selection that shapes actual sequence evolution. Use of this information also makes evolutionary models more interpretable by replacing heuristic free parameters with experimentally measurable quantities that can be directly related to the underlying processes of mutation and selection – an important step if one hopes to connect such models to population genetics in a meaningful way (Thorne et al., 2007; Halpern and Bruno, 1998).

The major drawback of experimentally informed models is their more limited scope. Most existing codon-based evolutionary models can be applied to any gene (Goldman and Yang, 1994; Muse and Gaut, 1994; Kosiol et al., 2007) – but experimentally informed models require experimental data for the gene in question. However, this requirement may not be as crippling as it initially appears. The first experimentally determined evolutionary model for influenza nucleoprotein required direct measurement of both the site-specific amino-acid preferences and the underlying mutation rates (Bloom, 2014). However, the model presented here only requires measurement of the amino-acid preferences, as the mutation rates are treated as free parameters. Rapid advances in the experimental technique of deep mutational scanning are making such data available for an increasing number of proteins (Fowler et al., 2010; Roscoe et al., 2013; Starita et al., 2013; Melamed et al., 2013; Traxlmayr et al., 2012; McLaughlin Jr et al., 2012; Firnberg et al., 2014; Bloom, 2014).

In this respect, it is encouraging that the site-specific amino-acid preferences determined experimentally for TEM-1 improve phylogenetic fit to substantially diverged (35% protein sequence divergence) SHV beta-lactamases as well as highly similar TEM beta-lactamases. As discussed in the Introduction, there are two major limitations to most existing evolutionary models: they

treat sites identically, and they treat sites independently. Experimentally informed evolutionary models of the type described here have the potential to completely remedy the first limitation, as experiments define site-specific selection with increasing precision. However, such models still treat sites independently – and this limitation will never be completely overcome by experiments, since the unforgiving math of combinatorics means that no experiment can examine all arbitrary combinations of mutations (for example, TEM-1 has only 5453 single amino-acid mutants, but it has 14815801 double mutants, 26742520805 triple mutants, and over 10^{13} quadruple mutants). The utility of experimentally informed evolutionary models therefore depends on the extent to which site-specific amino-acid preferences measured for one protein can be extrapolated to other homologs – in other words, are sites sufficiently independent that the preferences at a given position remain similar after mutations at other positions? This question remains a topic of active debate, with experimental studies suggesting that site-specific preferences are largely conserved among closely and moderately related homologs (Ashenberg et al., 2013; Serrano et al., 1993), but some computational studies emphasizing substantial shifts in preferences during evolution (Pollock et al., 2012; Pollock and Goldstein, 2014). The fact that the TEM-1 experimental data informs a model that accurately describes the substantially diverged SHV homologs suggests reasonable conservation of site-specific amino-acid preferences among beta-lactamase homologs.

This apparent conservation of site-specific amino-acid preferences suggests that the phylogenetic utility of experimentally informed evolutionary models may extend well beyond the immediate proteins that were experimentally characterized. This type of experimental generalization would have precedent: only a tiny fraction of proteins have been crystallized, but because structure is largely conserved during protein evolution, it is frequently possible to use a structure determined for one protein to draw insights about a range of related homologs (Lesk and Chothia, 1980; Sander and Schneider, 1991). It seems plausible that the conservation of site-specific amino-acid preferences could similarly enable deep mutational scanning (Fowler et al., 2010; Araya and Fowler, 2011) to provide the experimental data to inform evolutionary models of sufficient scope to improve the accuracy and interpretability of phylogenetic analyses for a substantial number of proteins of interest.

Methods

Availability of computer code and data

The phylogenetic analyses were performed using the software package *phyloExpCM* (**phylo**genetic analyses with **exp**erimental **c**odon **m**odels, https://github.com/jbloom/phyloExpCM), which primarily serves as an interface to run *HYPHY* (Pond et al., 2005). Input data, computer code, and a description sufficient to enable replication of all analyses reported in this paper are available via http://jbloom.github.io/phyloExpCM/example_2014Analysis_lactamase.html.

Equilibrium frequencies and reversibility

Here I show that the evolutionary model defined by Equation 1 is reversible (satisfies detailed balance), and has $p_{r,x}$ defined by Equation 9 as its stationary state.

First, note that the fixation probabilities $F_{r,xy}$ defined by both Equation 2 and Equation 3 satisfy reversibility with respect to the amino-acid preferences $\pi_{r,\mathcal{A}(y)}$ – namely that

$$\pi_{r,\mathcal{A}(x)} \times F_{r,xy} = \pi_{r,\mathcal{A}(y)} \times F_{r,yx}, \qquad \text{(Equation 12)}$$

as can be verified by direct substitution.

Next, observe that with the constraints in Equation 5 and Equation 6, the mutation rates Q_{xy} are reversible with respect to the q_x values defined by Equation 10 – namely that

$$q_x \times Q_{xy} = q_y \times Q_{yx}. \qquad \text{(Equation 13)}$$

Here I show this for the specific case where $x = AAT$ and $y = CAT$; other cases can be verified

similarly. In this specific case,

$$\begin{aligned} q_x \times Q_{xy} &= \frac{1}{8}\left(R_{C\to A}+R_{C\to T}\right)^3 \times R_{A\to C} \\ &= \frac{1}{8}\left(R_{C\to A}+R_{C\to T}\right)^3 \times \frac{R_{A\to G}\times R_{C\to A}}{R_{C\to T}} \\ &= \frac{1}{8}\left(R_{C\to A}+R_{C\to T}\right)^2 \times \left(\frac{R_{A\to G}\times \left(R_{C\to A}\right)^2}{R_{C\to T}} + R_{A\to G}\times R_{C\to A}\right) \\ &= \frac{1}{8}\left(R_{C\to A}+R_{C\to T}\right)^2 \times \left(\frac{R_{A\to G}\times R_{C\to A}}{R_{C\to T}} + R_{A\to G}\right) \times R_{C\to A} \\ &= \frac{1}{8}\left(R_{C\to A}+R_{C\to T}\right)^2 \times \left(R_{A\to C}+R_{A\to G}\right) \times R_{C\to A} \\ &= q_y \times Q_{yx}, \end{aligned} \qquad \text{(Equation 14)}$$

which verifies Equation 13.

Next, I show that $p_{r,x}$ defined by Equation 9 defines the evolutionary equilibrium frequencies. Define

$$P_{r,xx} = -\sum_{y\neq x} P_{r,xy}. \qquad \text{(Equation 15)}$$

With this definition, the matrix $\mathbf{I}+\mathbf{P_r}$ is a stochastic matrix, where $\mathbf{I}$ is the identity matrix and $\mathbf{P_r}=[P_{r,xy}]$. For plausible values for the mutation rates and amino-acid preferences, $\mathbf{I}+\mathbf{P_r}$ will also be irreducible and acyclic. Therefore, according to the Perron-Frobenius theorems, it has a unique (within a scaling constant) principal eigenvector $\mathbf{p_r}=[p_{r,x}]$ with an eigenvalue of one that represents the equilibrium frequencies. In other words, $\mathbf{p_r}=\mathbf{p_r}\left(\mathbf{I}+\mathbf{P_r}\right)$. It can be verified that Equation 9 defines such an eigenvector by writing this eigenvector equation in element-wise form. Immediately below, I verify this for the case where $F_{r,xy}$ is defined by Equation 3 and

$\pi_{r,\mathcal{A}(y)} > \pi_{r,\mathcal{A}(x)}$; other cases can be verified similarly. For this specific case:

$$\begin{aligned}
p_{r,x} &= p_{r,x} + \sum_y p_{r,y} P_{r,yx} \\
&= p_{r,x} + (\sum_{y \neq x} p_{r,y} P_{r,yx}) + p_{r,x} P_{r,xx} \\
&= p_{r,x} + (\sum_{y \neq x} p_{r,y} P_{r,yx}) - p_{r,x} \sum_{y \neq x} P_{r,xy} \\
&= p_{r,x} + \sum_{y \neq x} (p_{r,y} P_{r,yx} - p_{r,x} P_{r,xy}) \\
&= p_{r,x} + \sum_{y \neq x} (p_{r,y} Q_{yx} F_{r,yx} - p_{r,x} Q_{xy} F_{r,xy}) \\
&= p_{r,x} + \sum_{y \neq x} \left(\pi_{r,\mathcal{A}(y)} q_y Q_{yx} F_{r,yx} - \pi_{r,\mathcal{A}(x)} q_x Q_{xy} F_{r,xy}\right) \\
&= p_{r,x} + \sum_{y \neq x} \left(\pi_{r,\mathcal{A}(y)} q_y Q_{yx} F_{r,yx} - \pi_{r,\mathcal{A}(x)} q_x Q_{xy} F_{r,xy}\right) \\
&= p_{r,x} + \sum_{y \neq x} \left(\pi_{r,\mathcal{A}(y)} q_y Q_{yx} \frac{\pi_{r,\mathcal{A}(x)}}{\pi_{r,\mathcal{A}(y)}} - \pi_{r,\mathcal{A}(x)} q_x Q_{xy}\right) \\
&= p_{r,x} + \pi_{r,\mathcal{A}(x)} \sum_{y \neq x} (q_y Q_{yx} - q_x Q_{xy}) \\
&= p_{r,x} \qquad \text{(Equation 16)}
\end{aligned}$$

where the last line follows from Equation 13. This verifies that $p_{r,x}$ is the stationary state of the Markov process defined by $P_{r,xy}$.

To verify that the substitution model is reversible, it is necessary to show that $0 = p_{r,x} P_{r,xy} - p_{r,y} P_{r,yx}$. This is shown below for the case where $F_{r,xy}$ is defined by Equation 3 and $\pi_{r,\mathcal{A}(y)} > \pi_{r,\mathcal{A}(x)}$; other cases can be verified similarly. For this specific case,

$$\begin{aligned}
0 &= p_{r,x} P_{r,xy} - p_{r,y} P_{r,yx} \\
&= \pi_{r,\mathcal{A}(x)} \times q_x \times Q_{xy} \times F_{r,xy} - \pi_{r,\mathcal{A}(y)} \times q_y \times Q_{yx} \times F_{r,yx} \\
&= \pi_{r,\mathcal{A}(x)} \times q_x \times Q_{xy} \times F_{r,xy} - \pi_{r,\mathcal{A}(y)} \times q_y \times Q_{yx} \times F_{r,yx} \\
&= \pi_{r,\mathcal{A}(x)} \times q_x \times Q_{xy} - \pi_{r,\mathcal{A}(y)} \times q_y \times Q_{yx} \times \frac{\pi_{r,\mathcal{A}(x)}}{\pi_{r,\mathcal{A}(y)}} \\
&= q_x \times Q_{xy} - q_y \times Q_{yx} \\
&= 0 \qquad \text{(Equation 17)}
\end{aligned}$$

where the last line follows from Equation 13.

The fact that $P_{r,xy}$ defines a reversible Markov process with stationary state $p_{r,x}$ means that it is possible to define a symmetric matrix $\mathbf{S_r}$ such that

$$\mathbf{S_r}\,\mathbf{diag}\left(\ldots, p_{r,x}, \ldots\right) = \mathbf{P_r} \qquad \text{(Equation 18)}$$

where $\mathbf{diag}\left(\ldots, p_{r,x}, \ldots\right)$ is the diagonal matrix with $p_{r,x}$ along its diagonal. Noting $\mathbf{S_r} = \mathbf{P_r}\,\mathbf{diag}\left(\ldots, \frac{1}{p_{r,x}}, \ldots\right)$, we have

$$S_{r,xy} = \begin{cases} \frac{P_{r,xy}}{p_{r,y}} = 0 & \text{if } x \text{ and } y \text{ differ by more than one nucleotide mutation,} \\ \frac{P_{r,xy}}{p_{r,y}} = \left(\sum_z \pi_{r,\mathcal{A}(z)} \times q_z\right) \frac{Q_{xy}}{q_y} \frac{F_{r,xy}}{\pi_{r,\mathcal{A}(y)}} & \text{if } x \text{ and } y \text{ differ by one nucleotide,} \\ \frac{P_{r,xx}}{p_{r,x}} & \text{otherwise.} \end{cases} \qquad \text{(Equation 19)}$$

This matrix is symmetric since $S_{r,xy} = S_{r,yx}$ as can be verified from the fact that $\frac{Q_{xy}}{q_y} = \frac{Q_{yx}}{q_x}$ and $\frac{F_{r,xy}}{\pi_{r,\mathcal{A}(y)}} = \frac{F_{r,yx}}{\pi_{r,\mathcal{A}(x)}}$ as is guaranteed by Equation 12 and Equation 13.

Acknowledgments

Thanks to T. Bedford and J. Felsenstein for helpful discussions. This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (grant number R01 GM102198).

References

Ambler R, Coulson A, Frère JM, Ghuysen JM, Joris B, Forsman M, Levesque R, Tiraby G, Waley S. 1991. A standard numbering scheme for the class a beta-lactamases. *Biochemical Journal*. 276:269.

Araya CL, Fowler DM. 2011. Deep mutational scanning: assessing protein function on a massive scale. *Trends in biotechnology*. 29:435–442.

Ashenberg O, Gong LI, Bloom JD. 2013. Mutational effects on stability are largely conserved during protein evolution. *Proc. Natl. Acad. Sci. USA*. 110:21071–21076.

Bloom JD. 2014. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *bioRxiv*. http://biorxiv.org/content/early/2014/03/05/002899.

Bloom JD, Raval A, Wilke CO. 2007. Thermodynamics of neutral protein evolution. *Genetics*. 175:255–266.

Bush K, Jacoby GA, Medeiros AA. 1995. A functional classification scheme for beta-lactamases and its correlation with molecular structure. *Antimicrobial agents and chemotherapy*. 39:1211.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. Weblogo: a sequence logo generator. *Genome research*. 14:1188–1190.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*. 27:401–410.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17:368–376.

Firnberg E, Labonte JW, Gray JJ, Ostermeier M. 2014. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol. Biol. Evol.* p. msu081.

Fonzé E, Charlier P, To'Th Y, Vermeire M, Raquet X, Dubus A, Frere JM. 1995. TEM1-lactamase structure solved by molecular replacement and refined structure of the S235A mutant. *Acta Crystallographica Section D: Biological Crystallography*. 51:682–694.

Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, Fields S. 2010. High-resolution mapping of protein sequence-function relationships. *Nat. Methods*. 7:741–746.

Gil M, Zanetti MS, Zoller S, Anisimova M. 2013. Codonphyml: Fast maximum likelihood phylogeny estimation under codon substitution models. *Mol. Biol. Evol.* 30:1270–1280.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*. 149:445–458.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution probabilities for protein-coding DNA sequences. *Mol. Biol. Evol.* 11:725–736.

Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.* 15:910–917.

Hasegawa M, Kishino H, Yano Ta. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of molecular evolution*. 22:160–174.

Huelsenbeck JP, Hillis DM. 1993. Success of phylogenetic methods in the four-taxon case. *Systematic Biology*. 42:247–264.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*. 294:2310–2314.

Kleinman CL, Rodrigue N, Lartillot N, Philippe H. 2010. Statistical potentials for improved structurally constrained evolutionary models. *Mol. Biol. Evol.* 27:1546–1560.

Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. *Mol. Biol. Evol.* 24:1464–1479.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* 21:1095–1109.

Le SQ, Lartillot N, Gascuel O. 2008. Phylogenetic mixture models for proteins. *Phil. Trans. R. Soc. B*. 363:3965–3976.

Lesk AM, Chothia C. 1980. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *J. Mol. Biol.* 136:225–270.

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

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*. 19:1537–1551.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*. 21:1087–1092.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*. 11:715–724.

Pollock DD, Goldstein RA. 2014. Strong evidence for protein epistasis, weak evidence against it. *Proc. Natl. Acad. Sci. USA*. p. 201401112.

Pollock DD, Thiltgen G, Goldstein RA. 2012. Amino acid coevolution induces an evolutionary stokes shift. *Proc. Natl. Acad. Sci. USA*. 109:E1352–E1359.

Pond SK, Delport W, Muse SV, Scheffler K. 2010. Correcting the bias of empirical frequency parameter estimators in codon models. *PLoS One*. 5:e11230.

Pond SL, Frost SD, Muse SV. 2005. Hyphy: hypothesis testing using phylogenies. *Bioinformatics*. 21:676–679.

Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*. 53:793–808.

Potapov V, Cohen M, Schreiber G. 2009. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Prot. Eng. Des. Sel.* 22:553–560.

Rodrigue N, Kleinman CL, Philippe H, Lartillot N. 2009. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Mol. Biol. Evol.* 26:1663–1676.

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.* .

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406–425.

Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Bioinformatics*. 9:56–68.

Scherrer MP, Meyer AG, Wilke CO. 2012. Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evol. Bio.* 12:179.

Serrano L, Day AG, Fersht AR. 1993. Step-wise mutation of barnase to binase: a procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. *J. Mol. Biol.* 233:305–312.

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. USA*. 110:E1263–E1272.

Thorne JL, Choi SC, Yu J, Higgs PG, Kishino H. 2007. Population genetics without intraspecific data. *Mol. Biol. Evol.* 24:1667–1677.

Thorne JL, Goldman N, Jones DT. 1996. Combining protein evolution and secondary structure. *Mol. Biol. Evol.* 13:666–673.

Tien M, Meyer AG, Spielman SJ, Wilke CO. 2013. Maximum allowed solvent accessibilites of residues in proteins. *PLoS One*. 8:e80635.

Traxlmayr MW, Hasenhindl C, Hackl M, Stadlmayr G, Rybka JD, Borth N, Grillari J, Rüker F, Obinger C. 2012. Construction of a stability landscape of the ch3 domain of human igg1 by combining directed evolution with high throughput sequencing. *J. Mol. Biol.* .

Wang HC, Li K, Susko E, Roger AJ. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC evolutionary biology*. 8:331.

Wu CH, Suchard MA, Drummond AJ. 2013. Bayesian selection of nucleotide substitution models and their site assignments. *Mol. Biol. Evol.* 30:669–688.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39:306–314.

Yang Z, Nielsen R, Goldman N, Pedersen AMK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*. 155:431–449.

model	ΔAIC	log likelihood	parameters (optimized + empirical)
experimentally informed (Equation 2)	0.0	-4044.8	4 (4 + 0)
experimentally informed (Equation 3)	39.2	-4064.5	4 (4 + 0)
GY94, gamma ω, gamma rates	346.0	-4208.8	13 (4 + 9)
KOSI07, gamma ω, gamma rates	364.3	-4167.0	64 (4 + 60)
GY94, gamma ω, single rate	414.5	-4244.1	12 (3 + 9)
KOSI07, gamma ω, single rate	420.6	-4196.1	63 (3 + 60)
GY94, one ω, gamma rates	482.5	-4278.1	12 (3 + 9)
KOSI07, one ω, gamma rates	504.9	-4238.3	63 (3 + 60)
KOSI07, one ω, one rate	586.4	-4280.1	62 (2 + 60)
GY94, one ω, one rate	609.7	-4342.7	11 (2 + 9)
randomized (Equation 3)	1218.4	-4654.1	4 (4 + 0)
randomized (Equation 2)	1428.0	-4758.9	4 (4 + 0)

Table 1: Experimentally informed evolutionary models fit the combined TEM and SHV beta-lactamase sequence phylogeny (Figure 2A) much better than evolutionary models that do not utilize experimental data. Show are the difference in AIC relative to the best-fitting model (smaller ΔAIC indicates better fit), the log likelihood, and the number of free parameters. For each model, the branch lengths and model parameters were optimized on a fixed tree topology (Figure 2A) estimated with the model of Goldman and Yang (1994). The experimentally informed models use amino-acid preferences derived from the data of Firnberg et al. (2014) plus the four mutation rate parameters (Equation 8). For the randomized models, the experimentally measured amino-acid preferences are randomized among sites – these models are far worse since the preference are no longer assigned to the correct positions. GY94 denotes the model of Goldman and Yang (1994) with 9 equilibrium frequency parameters calculated using the CF3x4 method (Pond et al., 2010). KOSI07 denotes the model of Kosiol et al. (2007) with 60 equilibrium frequency parameters calculated using the F methods. All variants of GY94 and KOSI07 have a single transition-transversion ratio (κ) estimated by maximum likelihood. Different model variants either have a single nonsynonymous-synonymous ratio (ω) or values drawn from four discrete gamma-distributed categories (Yang et al., 2000), and either a single rate or rates drawn from four discrete gamma-distributed categories (Yang, 1994). The data and source code used to generate these data are provided via `http://jbloom.github.io/phyloExpCM/example_2014Analysis_lactamase.html`.

model	ΔAIC	log likelihood	parameters (optimized + empirical)
experimentally informed (Equation 2)	0.0	-4045.1	4 (4 + 0)
experimentally informed (Equation 3)	37.3	-4063.7	4 (4 + 0)
GY94, gamma ω, gamma rates	349.6	-4210.9	13 (4 + 9)
KOSI07, gamma ω, gamma rates	353.8	-4162.0	64 (4 + 60)
KOSI07, gamma ω, single rate	406.8	-4189.5	63 (3 + 60)
GY94, gamma ω, single rate	416.1	-4245.1	12 (3 + 9)
GY94, one ω, gamma rates	479.5	-4276.9	12 (3 + 9)
KOSI07, one ω, gamma rates	481.5	-4226.9	63 (3 + 60)
KOSI07, one ω, one rate	560.0	-4267.1	62 (2 + 60)
GY94, one ω, one rate	603.6	-4339.9	11 (2 + 9)
randomized (Equation 3)	1216.8	-4653.5	4 (4 + 0)
randomized (Equation 2)	1425.7	-4758.0	4 (4 + 0)

Table 2: Experimentally informed evolutionary models also provide a superior phylogenetic fit when the tree topology is estimated using the model of Kosiol et al. (2007) rather than that of Goldman and Yang (1994). This table differs from Table 1 in that the phylogenetic fit is to all TEM and SHV sequences using the tree topology in Figure 2B rather than that in Figure 2A.

model	ΔAIC	log likelihood	parameters (optimized + empirical)
experimentally informed (Equation 2)	0.0	-2386.8	4 (4 + 0)
experimentally informed (Equation 3)	60.5	-2417.1	4 (4 + 0)
GY94, gamma ω, gamma rates	229.1	-2492.4	13 (4 + 9)
GY94, one ω, gamma rates	294.4	-2526.1	12 (3 + 9)
GY94, gamma ω, single rate	295.3	-2526.5	12 (3 + 9)
KOSI07, gamma ω, gamma rates	303.8	-2478.7	64 (4 + 60)
KOSI07, gamma ω, single rate	371.8	-2513.8	63 (3 + 60)
KOSI07, one ω, gamma rates	388.9	-2522.3	63 (3 + 60)
GY94, one ω, one rate	460.2	-2609.9	11 (2 + 9)
KOSI07, one ω, one rate	533.6	-2595.7	62 (2 + 60)
randomized (Equation 3)	953.5	-2863.6	4 (4 + 0)
randomized (Equation 2)	984.7	-2879.2	4 (4 + 0)

Table 3: Experimentally informed evolutionary models also provide a superior phylogenetic fit when the analysis is limited only to TEM beta-lactamase sequences. This table differs from Table 1 in that the phylogenetic fit is only to the TEM sequences (the portion of the tree shown in red in Figure 2A.)

model	ΔAIC	log likelihood	parameters (optimized + empirical)
experimentally informed (Equation 2)	0.0	-1782.7	4 (4 + 0)
experimentally informed (Equation 3)	10.3	-1787.8	4 (4 + 0)
KOSI07, gamma ω, gamma rates	382.7	-1914.0	64 (4 + 60)
KOSI07, one ω, gamma rates	393.4	-1920.4	63 (3 + 60)
GY94, gamma ω, gamma rates	399.6	-1973.4	13 (4 + 9)
GY94, one ω, gamma rates	407.8	-1978.5	12 (3 + 9)
KOSI07, gamma ω, single rate	449.5	-1948.4	63 (3 + 60)
KOSI07, one ω, one rate	467.4	-1958.3	62 (2 + 60)
GY94, gamma ω, single rate	475.3	-2012.3	12 (3 + 9)
GY94, one ω, one rate	496.4	-2023.8	11 (2 + 9)
randomized (Equation 3)	940.8	-2253.1	4 (4 + 0)
randomized (Equation 2)	965.0	-2265.2	4 (4 + 0)

Table 4: Experimentally informed evolutionary models also provide a superior phylogenetic fit when the analysis is limited only to SHV beta-lactamase sequences. This table differs from Table 1 in that the phylogenetic fit is only to the SHV sequences (the portion of the tree shown in blue in Figure 2A.)

model	ΔAIC	log likelihood	parameters (optimized + empirical)
experimentally informed (Equation 2)	0.0	-2392.2	4 (4 + 0)
experimentally informed (Equation 3)	58.2	-2421.4	4 (4 + 0)
GY94, gamma ω, gamma rates	232.2	-2499.3	13 (4 + 9)
GY94, one ω, gamma rates	292.2	-2530.4	12 (3 + 9)
GY94, gamma ω, single rate	298.7	-2533.6	12 (3 + 9)
KOSI07, gamma ω, gamma rates	300.2	-2482.4	64 (4 + 60)
KOSI07, gamma ω, single rate	368.0	-2517.3	63 (3 + 60)
KOSI07, one ω, gamma rates	377.0	-2521.7	63 (3 + 60)
GY94, one ω, one rate	462.5	-2616.5	11 (2 + 9)
KOSI07, one ω, one rate	525.0	-2596.7	62 (2 + 60)
randomized (Equation 3)	955.3	-2869.9	4 (4 + 0)
randomized (Equation 2)	989.2	-2886.8	4 (4 + 0)

Table 5: Experimentally informed evolutionary models also provide a superior phylogenetic fit to the TEM beta-lactamases when the tree topology is estimated using the model of Kosiol et al. (2007) rather than that of Goldman and Yang (1994). This table differs from Table 3 in that the phylogenetic fit is to the TEM sequences using the red portion of tree topology in Figure 2B rather than the red portion of the tree topology in Figure 2A.

model	ΔAIC	log likelihood	parameters (optimized + empirical)
experimentally informed (Equation 2)	0.0	-1778.5	4 (4 + 0)
experimentally informed (Equation 3)	10.2	-1783.7	4 (4 + 0)
KOSI07, gamma ω, gamma rates	382.2	-1909.6	64 (4 + 60)
KOSI07, one ω, gamma rates	387.3	-1913.2	63 (3 + 60)
GY94, gamma ω, gamma rates	392.9	-1966.0	13 (4 + 9)
GY94, one ω, gamma rates	397.1	-1969.1	12 (3 + 9)
KOSI07, gamma ω, single rate	443.2	-1941.2	63 (3 + 60)
KOSI07, one ω, one rate	458.2	-1949.6	62 (2 + 60)
GY94, gamma ω, single rate	463.6	-2002.4	12 (3 + 9)
GY94, one ω, one rate	481.9	-2012.5	11 (2 + 9)
randomized (Equation 3)	936.6	-2246.8	4 (4 + 0)
randomized (Equation 2)	959.5	-2258.3	4 (4 + 0)

Table 6: Experimentally informed evolutionary models also provide a superior phylogenetic fit to the SHV beta-lactamases when the tree topology is estimated using the model of Kosiol et al. (2007) rather than that of Goldman and Yang (1994). This table differs from Table 4 in that the phylogenetic fit is to the SHV sequences using the blue portion of tree topology in Figure 2B rather than the blue portion of the tree topology in Figure 2A.

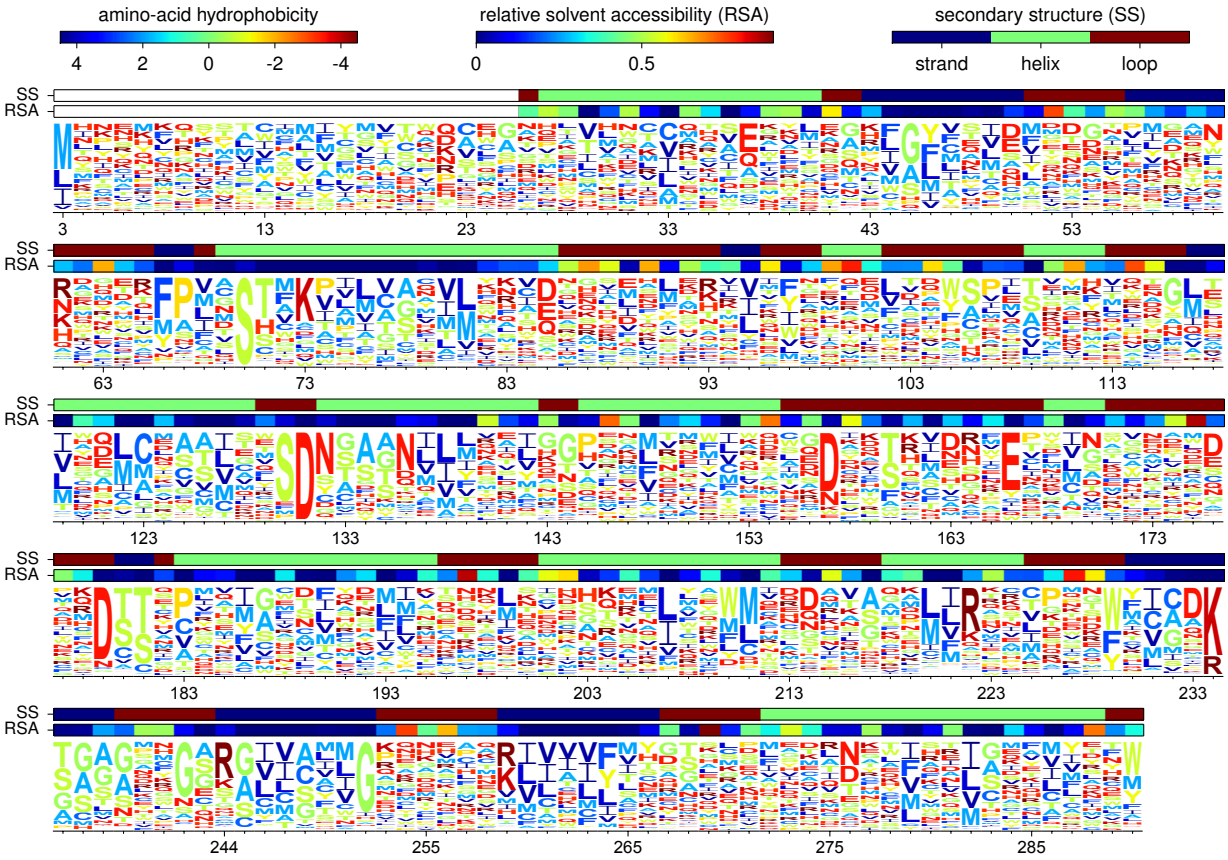

Figure 1: The amino-acid preferences for TEM-1 beta-lactamase, calculated from the data of Firnberg et al. (2014). The heights of letters are proportional to the preference for that amino acid at that position in the protein. Residues are numbered using the scheme of Ambler et al. (1991). Letters are colored according to the hydrophobicity of the amino acid. Bars above the letters indicate the secondary structure and relative solvent accessibility as calculated from the crystal structure in PDB entry 1XPB (Fonzé et al., 1995), with maximum solvent accessibilities taken from Tien et al. (2013). The figure was generated using *WebLogo* (Crooks et al., 2004) integrated into the *mapmuts* software package (Bloom, 2014). The data and source code used to create this plot are provided via `http://jbloom.github.io/phyloExpCM/example_2014Analysis_lactamase.html`.

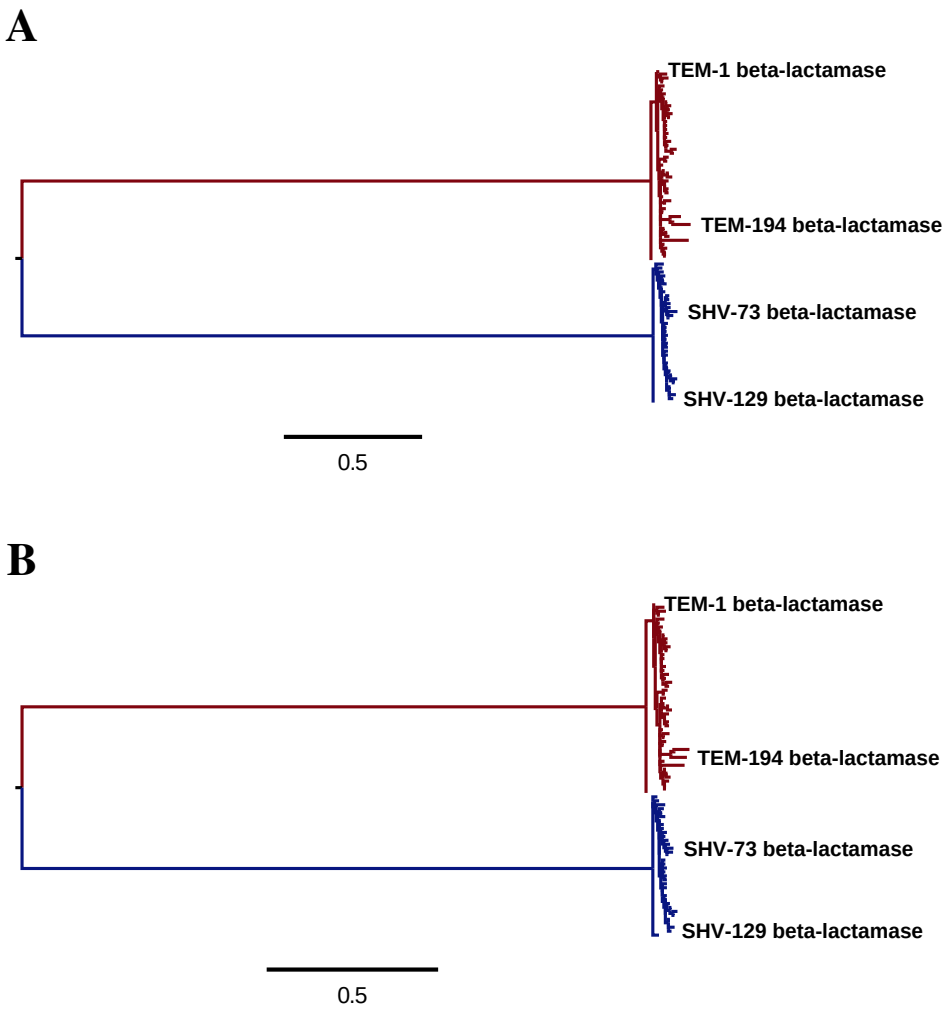

Figure 2: Phylogenetic trees of TEM (red) and SHV (blue) beta-lactamases inferred using *codonPhyML* (Gil et al., 2013) with the codon substitution model of **(A)** Goldman and Yang (1994) or **(B)** Kosiol et al. (2007). The inferred trees are very similar for both models. The TEM and SHV sequences each cluster into closely related clades, with extensive divergence between these two clades. The average pairwise divergence between the TEM and SHV clades is 38% at the nucleotide level and 35% at the protein level. For both models, a single transition-transversion ratio (κ) and four discrete gamma-distributed nonsynonymous-synonymous ratios (ω) were estimated by maximum likelihood. The equilibrium codon frequencies were determined empirically using the CF3x4 method (Pond et al., 2010) for the model of Goldman and Yang (1994), or the F method for the model of Kosiol et al. (2007) The data and source code used to create these trees are provided via `http://jbloom.github.io/phyloExpCM/example_2014Analysis_lactamase.html`.

Supplementary file 1: This text file contains the amino-acid preferences displayed graphically in Figure 1. In this file, the amino acids are numbered sequentially starting at one with the N-terminal methionine, rather than using numbering scheme of Ambler et al. (1991) that is employed in Figure 1.

Supplementary file 2: This FASTA file contains the alignment of TEM and SHV beta-lactamase sequences used to create the phylogenetic trees in Figure 2.