Supporting Text

A Model Comparing Mutation and Recombination. Here we more rigorously derive Eqs. **2** and **3** from the main text, which quantify the probability with which mutants or chimeras with *m* substitutions retain function. Consider recombining two homologous parental proteins having *L* amino acid residues differing at *D* sites and a conserved structure (fold). We make three simplifying assumptions: *i*) the fraction of recombined proteins that retain function is an unbiased subset of those retaining fold, *ii*) the probability of retaining fold is determined by the independent probabilities that each residue is compatible with the parental structure and with all other residues, and *iii*) residues found in parental sequences are compatible with the structure and each other, while all other amino acids have an unknown average probability of incompatibility.

Under these assumptions, the probability that a protein containing residues $r_1 ... r_L$ retains the parental fold can be written

$$P_{\mathrm{f}}(r) = \prod_{i}^{L} \Pr(r_i \text{ compatible}) \prod_{j<k}^{L} \Pr(r_j, r_k \text{ compatible}) .$$

Although this probability cannot be practically computed for a particular protein because of the intricate details of the molecular interactions determining compatibility, we may estimate it on average over a large number of mutants or chimeras by examining the quantity $P_{\mathrm{f}}(m) = \langle P_{\mathrm{f}}(r) \rangle$, the average fraction of proteins with *m* substitutions that retain fold. Assumption 2 asserts independence, so

$$P_{\mathrm{f}}(m) = \langle P_{\mathrm{f}}(r) \rangle = \prod_{i}^{L} \langle \Pr(r_i \text{ compatible}) \rangle \prod_{j<k}^{L} \left\langle \Pr(r_j, r_k \text{ compatible}) \right\rangle ,$$

and according to Assumption 3 these average probabilities can be written in terms of an average residue–residue incompatibility p_{rr} and a residue–backbone incompatibility p_{rb},

$$\langle \Pr(r_i \text{ compatible}) \rangle = \begin{cases} 1 & \text{if } r_i \text{ is in a parental structure,} \\ p_{rb} < 1 & \text{otherwise;} \end{cases}$$
$$\left\langle \Pr(r_j, r_k \text{ compatible}) \right\rangle = \begin{cases} 1 & \text{if } r_j \text{ and } r_k \text{ are in a parental structure,} \\ p_{rr} < 1 & \text{otherwise.} \end{cases}$$

Our final assumption thus reduces determination of the probability of retaining fold to counting the number of possible residue–backbone and residue–residue incompatibilities resulting from *m* substitutions. In the case of random mutation, *m* substitutions create *m* possible residue–backbone incompatibilities and $m(L-(m+1)/2)$ residue–residue incompatibilities. Recombination, by contrast, does not create any residue–backbone incompatibilities, because residues from both parents have proven compatible with the conserved structure, but alters a possible $m(D-m)$ residue–residue compatibilities. As a result, we have

$$P_{\mathrm{f}}(m)_{\text{mutation}} = p_{rb}^{m} p_{rr}^{m(L-(m+1)/2)} \approx (p_{rb} p_{rr}^{L})^{m} \equiv \nu^{m} \qquad \textbf{[4]}$$

$$P_{\mathrm{f}}(m)_{\mathrm{recombination}} = p_{rr}^{m(D-m)} \equiv \rho^{\frac{m(D-m)}{D-1}} . \quad \textbf{[5]}$$

The definitions introduce the parameters ν and ρ to enable a direct comparison: the fraction of functional variants with a single substitution ($m = 1$) is ν for mutation and ρ for recombination. The approximation in Eq. **4** follows if $m << L$, which is generally true for random mutagenesis, and if p_{rr} is on average less than p_{rb}. We have now formulated $P_{\mathrm{f}}(m)$ in terms of two unknown parameters, which allow us to compare mutation and recombination in a simple way: ν (the neutrality) represents the average probability that a random residue substitution will preserve fold, and ρ (the recombinational tolerance) measures the average probability that a substitution coming from a homolog via recombination will preserve fold. ν <ρ indicates that substitutions created by recombination are more conservative than random substitutions, and ν >ρ the opposite. In all cases we expect ν <ρ because, as the intermediate expressions in Eqs. **4** and **5** show, $P_{\mathrm{f}}(m)_{\mathrm{recombination}}$ is strictly greater than $P_{\mathrm{f}}(m)_{\mathrm{mutation}}$. Moreover, Eqs. **4** and **5** indicate that ν and ρ should correlate through their mutual dependence on p_{rr}. As would be expected in this model, $P_{\mathrm{f}}(m)_{\mathrm{recombination}}$ is symmetric, such that it makes no difference which parent *m* is measured from.

Error Analysis and Fitting Procedure. Best-fit parameters and fit statistics were obtained using Mathematica's NonlinearRegress function with data weighted by inverse standard error on the dependent variable. Lactamase mutation data were fit to Eq. **1** and recombination data to Eq. **3**. For lactamase mutation data, standard error on the fraction functional was calculated using results from replicates, and standard error on the assessment of library average nucleotide mutation level $\langle m_{\mathrm{nt}} \rangle$ was calculated as described in (1). Standard errors for the lactamase recombination data were approximated under the assumption that each bin's fraction functional was generated by a binomial process with proportion equal to the minimum fraction functional. Lattice protein mutation data were fit to Eq. **2** and recombination data to Eq. **3**. We examined four values of *D* for each of ten lattice protein structures, and fits were performed independently on each of the four resulting 100-run sets of data. Standard errors were calculated over each 100-run set.

SCHEMA Disruption Calculations. In a previous study, we showed that the probability of retaining function, among lactamase chimeras exhibiting the same substitution level, depends on the number of residue-residue contacts broken (*E*), where a contact is defined as any two residues within 4.5 Å (2). Thus the particular choice of crossover sites for constructing a library of recombined sequences will affect the observed probability that function is conserved (P_{f}). This means that the P_{f} values for other PSE-4/TEM-1 libraries could differ from the values in Fig. 1. One baseline for the average effects of PSE-4 and TEM-1 recombination on lactamase function is the P_{f} for gene conversion events (e.g. double-crossover chimeras arising from the swapping of a single polypeptide element). To assess the effects of gene conversion on lactamase function, we calculated E and m for all possible PSE-4 and TEM-1 double-crossover chimeras ($N = 34{,}191$). At low sequence distances ($m < 20$), we found that the average disruption $\langle E \rangle$ of the double-crossover chimeras was similar to that calculated for chimeras in our unselected lactamase library (Fig. 5). At larger distances, however, double-crossover chimeras exhibited lower $\langle E \rangle$ than chimeras in our library. This finding suggests that double-crossover events are on average more conservative of function than estimated from analysis of our library. These differences arise because our lactamase library was constructed by using crossover sites that yield chimeras with even higher average disruption than in most randomly-selected, 13-crossover libraries.

Identified Functional Chimeras of TEM-1 and PSE-4. Table 2 lists the modular composition of functional chimeras isolated from the recombination library discussed in the main text. The polypeptide modules inherited from either PSE-4 (P) or TEM-1 (T) correspond to TEM-1 residues 1-39 (A), 40-57 (B), 58-67 (C), 68-

84 (D), 85-102 (E), 103-115 (F), 116-131 (G), 132-146 (H), 147-163 (I), 164-204 (J), 205-222 (K), 223-249 (L), 250-264 (M), 265-286 (N) and structurally related residues in PSE-4 identified using a structure-based alignment with Swiss-PDB Viewer (3). Substitution level (m) is the minimum number of mutations required to convert a chimera into PSE-4, excluding residues comprising the periplasmic secretory signal sequences.

Calculation of Neutrality ν from Error-Prone PCR Library Data. The fraction of functional clones in a mutant library generated by error-prone PCR can be modeled using experimental parameters and knowledge of protein neutrality (1). Multi-round error-prone PCR (see *Methods*) ensures that $\langle m_{\mathrm{nt}} \rangle$ is proportional to n_{cyc}, which in turn means that $P_{\mathrm{f}}(\langle m_{\mathrm{nt}} \rangle)$ will decline exponentially (1) with a slope related to ν, consistent with our data. In general, the observed $P_{\mathrm{f}}(\langle m_{\mathrm{nt}} \rangle)$ slope will be significantly higher than ν^{m} or even predictions which assume a Poisson distribution of mutations in the library, because error-prone PCR generates a mutation distribution of particularly high variance (1). The excess of sequences with fewer than average mutations inflate the fraction functional relative to the Poisson-based (smaller variance) expectation.

We calculated p_{ns} and p_{tr} from the sequencing data shown in Table 3. p_{ns} is the fraction of all mutations excluding deletions that were nonsynonymous = 0.677; p_{tr} is the fraction of all mutations that produced a deletion or a stop codon = 0.059. Our error-prone PCR protocol used 13 thermal cycles per round (n_{cyc} = number of rounds × 13), produced 9 DNA doublings per round for an efficiency $\lambda = 9/13 = 0.69$, and yielded the observed fractions functional at four values of $\langle m_{\mathrm{nt}} \rangle$ shown in Table 3.

To obtain a best-fit value for ν in a simple way, we made an auxiliary assumption that the number of thermal cycles n_{cyc} was proportional to the observed library average nucleotide mutation level $\langle m_{\mathrm{nt}} \rangle$, $n_{\mathrm{cyc}} = 13 \langle m_{\mathrm{nt}} \rangle / 8.37$, where 8.37 is the average number of nucleotide mutations introduced per round. Substituting this expression for n_{cyc} into Eq. **1** allowed us to express $P_{\mathrm{f}}(\langle m_{\mathrm{nt}} \rangle)$ as a function only of $\langle m_{\mathrm{nt}} \rangle$ and ν (the remaining values are constants). Using Mathematica's NonlinearRegress function on the five pairs of data for $P_{\mathrm{f}}(\langle m_{\mathrm{nt}} \rangle)$ [Table 3 and ($\langle m_{\mathrm{nt}} \rangle$=0, $P_{\mathrm{f}}(\langle m_{\mathrm{nt}} \rangle)$=1.05 ± 0.06] reported in the main text) with values weighted by the inverse standard error on $P_{\mathrm{f}}(\langle m_{\mathrm{nt}} \rangle)$ for each point, we obtained a best-fit value of ν = 0.54 ± 0.03 ($P < 0.0001$) (error is asymptotic standard error). To check that this result did not depend strongly on our auxiliary assumption, we then evaluated Eq. **1** for $P_{\mathrm{f}}(\langle m_{\mathrm{nt}} \rangle)$ using the actual number of thermal cycles at each round. The resulting data shown in Table 3 does not differ meaningfully from the predicted exponential line, and falls within a standard error of all but one datum.

1. Drummond, D. A., Iverson, B. L., Georgiou, G. & Arnold, F. H. (2005) arXiv: q-bio.QM/0411041.
2. Meyer, M. M., Silberg, J. J., Voigt, C. A., Endelman, J. B., Mayo, S. L., Wang, Z. G. & Arnold, F. H. (2003) *Protein Sci* **12,** 1686-93.
3. Guex, N. & Peitsch, M. C. (1997) *Electrophoresis* **18,** 2714-23.