Protein Engineering, Design & Selection vol. 19 no. 12 pp. 563–570, 2006
Published online November 6, 2006 doi:10.1093/protein/gzl045

Structure-guided SCHEMA recombination of distantly related β-lactamases

Michelle M.Meyer[1], Lisa Hochrein[2] and Frances H.Arnold[12,3]

[1]Biochemistry and Molecular Biophysics, California Institute of Technology, Mail Code 210-21 and [2]Division of Chemistry and Chemical Engineering, California Institute of Technology, Mail Code 210-41, Pasadena, CA 91125, USA

[3]To whom correspondence should be addressed.
E-mail: frances@cheme.caltech.edu

We constructed a library of β-lactamases by recombining three naturally occurring homologs (TEM-1, PSE-4, SED-1) that share 34–42% sequence identity. Most chimeras created by recombining such distantly related proteins are unfolded due to unfavorable side-chain interactions that destabilize the folded structure. To enhance the fraction of properly folded chimeras, we designed the library using SCHEMA, a structure-guided approach to choosing the least disruptive crossover locations. Recombination at seven selected crossover positions generated 6561 chimeric sequences that differ from their closest parent at an average of 66 positions. Of 553 unique characterized chimeras, 111 (20%) retained β-lactamase activity; the library contains hundreds more novel β-lactamases. The functional chimeras share as little as 70% sequence identity with any known sequence and are characterized by low SCHEMA disruption *(E)* compared to the average nonfunctional chimera. Furthermore, many nonfunctional chimeras with low *E* are readily rescued by low error-rate random mutagenesis or by the introduction of a known stabilizing mutation (TEM-1 M182T). These results show that structure-guided recombination effectively generates a family of diverse, folded proteins even when the parents exhibit only 34% sequence identity. Furthermore, the fraction of sequences that encode folded and functional proteins can be enhanced by utilizing previously stabilized parental sequences.

Keywords: chimera/directed evolution/mutational robustness/protein design

Introduction

Directed evolution has proven to be an effective technique for engineering proteins with desired properties. Because the probability of a protein retaining its fold and function decreases exponentially with the number of random substitutions introduced (Bloom *et al*., 2005), only a few mutations are made in each generation in order to maintain a reasonable fraction of functional proteins for screening (Voigt *et al*., 2001). Creating libraries with higher levels of mutation while maintaining structure and function requires identifying mutations that are less likely to disrupt the structure (Lutz and Patrick, 2004). One strategy to accomplish this is homologous recombination: mutations introduced by recombination are less deleterious than random mutations because they are compatible with the backbone structure (Drummond *et al*., 2005). Random recombination of highly similar proteins often generates libraries with a high fraction of functional sequences (Ness *et al*., 1999). However, as more distantly related proteins are recombined, the fraction of chimeric proteins that fold correctly decreases significantly (Ostermeier *et al*., 1999; Sieber *et al*., 2001; Ostermeier, 2003).

Computational methods that rely on sequence and structure information have been developed to predict which chimeras are likely to function (Voigt *et al*., 2002; Moore and Maranas, 2003; Saraf and Maranas, 2003; Saraf *et al*., 2004). We have developed the SCHEMA energy function to aid in designing libraries of protein chimeras. SCHEMA uses structural information to identify interacting amino acid residue pairs; interactions that are broken upon recombination then count toward a disruption score, or *E* (Voigt *et al*., 2002). We have shown that β-lactamase (Meyer *et al*., 2003) and cytochrome P450 heme domain (Otey *et al*., 2006) chimeras with lower *E* are more likely to retain fold and function. In a SCHEMA-designed library of cytochromes P450 sharing ∼63% sequence identity, 47% of the chimeras correctly bound the heme cofactor, indicating a folded structure (Otey *et al*., 2006). Of the folded chimeras, at least 72% were catalytically active. Thus SCHEMA enables us to produce a synthetic family of >2300 diverse, catalytically active, P450s.

The ∼63% sequence identity shared by the cytochromes P450 in the Otey *et al*. (2006) study is still high compared to that of many known homolog pairs. For proteins of approximately the same length, recombining more distantly related homologs generates greater sequence diversity and a higher mutation level in the chimeras. More mutation generally leads to more disruption, but the nature of the disruption can also change as the proteins diverge. For example, proteins tend to accumulate more mutations in core positions as they diverge; disruption of core interactions may be more destabilizing on average than disruption of surface interactions. To examine these effects, we tested SCHEMA recombination of proteins sharing only 34–42% sequence identity. The three β-lactamases [PSE-4, TEM-1 and SED-1 (Jelsch *et al*., 1993; Lim *et al*., 2001; Petrella *et al*., 2001)] recombined in this study are much closer to the 'twilight zone' (20–35% identity) where sequence identity can no longer be used as a surrogate for homology (Doolittle, 1986; Rost, 1999).

Recombination, while less disruptive than random mutation, nonetheless introduces disruptive mutations. Many 'global suppressor' substitutions have been identified that can increase a protein's tolerance to random mutation by increasing stability (Shortle and Lin, 1985; Poteete *et al*., 1997; Bloom *et al*., 2005). However, it is unknown whether such mutations can increase a protein's tolerance to the multiple disruptions introduced by recombination. We therefore tested the extent to

which the nonfunctional chimeras from the library could be 'rescued' by random mutagenesis.

Materials and methods

SCHEMA calculations

The SCHEMA disruption (E) for a chimera was calculated according to

$$E = \sum_i \sum_{j>i} C_{ij}\Delta_{ij}, \quad (1)$$

where C_{ij} =1 if any side-chain heavy atoms or main-chain carbons in amino acid residues i and j are within 4.5 Å (Voigt *et al*., 2002). The β-lactamase parent sequences are assumed to fold into (approximately) the same 3-dimensional structures [described by the contact map (C_{ij})]. The structures of TEM-1 (Maveyraud *et al*., 1998) and PSE-4 (Lim *et al*., 2001) have only 0.98 Å RMSD over the backbone atoms. No structure is available for SED-1. The structure of PSE-4 (1G68) was used to calculate C_{ij}; using a TEM-1 structure (1BT5) causes only slight changes. The Δ_{ij} is calculated from the sequence alignment of the parent proteins: Δ_{ij} = 0 if amino acids i and j in the chimera are found at the same positions in any parental protein sequence, otherwise the interaction is broken and Δ_{ij} = 1. The sequences of TEM-1, SED-1 and PSE-4 were aligned using clustalW (Chenna *et al*., 2003). This alignment shows no differences from a structural alignment between TEM-1 (1BT5) and PSE-4 (1G68) generated in Swiss-PDB Viewer (Guex and Peitsch, 1997). Python scripts for calculating E are available on the Arnold lab website http://www.che.caltech.edu/groups/fha/.

Library design

The RASPP (Recombination as a Shortest Path Problem) algorithm (Endelman *et al*., 2004) was used to identify the combinatorial libraries with the lowest average SCHEMA disruption <E> at many levels of diversity. RASPP was run iteratively with a minimum block length L of 5–33 amino acids and a <m> bin size of 1. Python scripts to perform RASPP can be found at the Arnold lab website http://www.che.caltech.edu/groups/fha/.

Library construction

The parental genes PSE-4, TEM-1 and SED-1 were described previously (Jelsch *et al*., 1993; Lim *et al*., 2001; Petrella *et al*., 2001). The SCHEMA library was constructed following the method of Hiraga and Arnold (2003) and Meyer *et al*. (2006), using the type IIb restriction endonuclease *BsaX*1. All restriction endonucleases and other enzymes for molecular biology were purchased from New England Biolabs, and oligonucleotides from Operon. Due to the small size of block 2 (24 nt), the parental gene fragments were added to the ligation reaction as annealed and phosphorylated oligonucleotides. The library of full-length chimeras was ligated into pProTet E.333 (Clontech) and transformed into *Escherichai coli* XL-1 Blue (Stratagene) where the protein is constitutively expressed. Additional details of library construction can be found in the Supplementary data available at *PEDS* online.

Sequence analysis

The sequences of 1100 chimeras were determined using high-throughput probe hybridization as described previously (Meinhold *et al*., 2003) and detailed in the supplementary information. From these sequences, 811 complete sequences were obtained, of which 553 were unique.

Functional screen

To screen for chimera function, deep-well 96-well plates containing 500 μl of LB medium with 35 μg/ml chloramphenicol were inoculated from previously grown cultures in 384-well plates and incubated with shaking for 18 h at 37°C, 80% humidity. Approximately 2 μl aliquots of each culture were transferred to duplicate LB agar plates containing varying concentrations of ampicillin (0, 5, 10, 25, 50, 100, 250, 500, 1000, 2000 μg/ml) using a 96-well stamp and were allowed to grow at 37°C. After 18 h the plates were observed for growth. Colonies growing at concentrations of ampicillin 25 μg/ml or greater were considered to express functional chimeras. XL-1 Blue cells containing pProTet E.333 with no β-lactamase insert survive to 5 μg/ml ampicillin in this assay. The concentration of ampicillin necessary to prevent growth was recorded as the MIC (minimal inhibitory concentration). Chimeras that grew on the 2000 μg/ml plates are recorded as 2000+.

Random mutagenesis

DNA for nonfunctional chimeras was sequenced prior to mutagenesis to confirm that no point mutations were present. Error-prone PCR was performed on each chimera in the following 100 μl reaction: 3 ng template, 1 μM forward and reverse primers matched to the parental sequences of blocks 1 and 8, 7 mM $MgCl_2$, 75 μM $MnCl_2$, 200 μM dATP and dGTP, 50 μM dTTP and dCTP, 1× PCR buffer without $MgCl_2$ and 5 U of *Taq* polymerase (Applied Biosystems). Reactions were heated to 95°C for 5 min and 14 cycles of 30 s at 95°C, 30 s at 55°C and 1 min at 72°C were completed. PCR products were digested with *Kpn*I and *Pst*I, cloned into pProTet E.333 (Clontech) cut with the same enzymes and transformed into *E.coli* XL-1 Blue (Stratagene).

Transformed *E.coli* XL-1 Blue were plated onto selective medium (35 μg/ml chloramphenicol and 10 μg/ml ampicillin) to identify sequences conferring resistance to ampicillin. Untransformed XL-1 is resistant to <5 μg/ml of ampicillin. To estimate the number of independent clones in the selected sample, an aliquot was plated on nonselective medium (35 μg/ml chloramphenicol). Colonies present on selective plates after 18 h of growth at 37°C were picked and the DNA extracted. The DNA was sequenced to identify mutations and retransformed into *E.coli*. For all functional sequences reported, the survival rate of retransformed *E.coli* on selective medium was high, indicating plasmid conferred resistance. A minimum of ~20 000 colonies were examined for each chimera. If no colonies were present on selective plates, 10 colonies were picked from nonselective plates to determine the insert incorporation frequency. Typically five of these colonies were sequenced to verify successful random mutagenesis (error rate was 1.85 ± 0.8 nt changes per gene).

Site-directed mutagenesis

The TEM-1 M182T mutation was introduced into 29 selected chimeras using QuikChange Mutagenesis (Stratagene) with the following primer and its reverse complement: 5′-CGT GAC ACC ACG ACC CCT GTA GCA ATG G. The altered codon is underlined. Mutagenesis reactions were transformed into *E.coli* XL-1 Blue (Stratagene) and equal aliquots were plated

onto the selective and nonselective media described above. Colonies growing on selective medium (35 μg/ml chloramphenicol and 10 μg/ml ampicillin) after 18 h at 37°C were picked and the DNA extracted for sequencing. For chimeras for which no colonies appeared on selective plates, two colonies were picked from the nonselective plates and the DNA was sequenced to verify that the mutation was properly incorporated. All sequences were retransformed into *E.coli* to verify that the plasmid conferred resistance.

Results

Library design using SCHEMA

We generated a library of chimeric β-lactamases by recombining fragments of the genes for PSE-4, TEM-1 and SED-1. These proteins are ~265 amino acids in length and share between 34 and 42% sequence identity. We chose to construct a combinatorial library with eight blocks (seven recombination sites), giving $3^8 = 6561$ possible chimeras. To ensure that a significant fraction of the chimeras fold, we used the optimization algorithm RASPP (Recombination as a Shortest Path Problem) (Endelman *et al*., 2004) to choose recombination sites that minimize the library average SCHEMA energy (<*E*>). Because minimizing sequence changes also minimizes <*E*>, RASPP is performed with minimum length constraints (*L*) on the sequence fragments between the recombination sites to ensure a diverse population of chimeras. RASPP was iterated using different *L* to identify optimal libraries over a range of average mutation levels with respect to the closest parent sequence, <*m*>. Optimal libraries identified by RASPP are shown in Figure 1 for a wide range of <*m*>.

β-Lactamases are often considered single-domain proteins (Jones *et al*., 1997) but are divided into two subdomains by some structural classification schemes (Murzin *et al*., 1995). The subdomains consist of an α+β sandwich formed by the N- and C-termini and an α-helical subdomain. Recombination sites that appear in many RASPP-identified libraries lie at the boundaries of these subdomains {approximately residues 63 and 216 [Ambler standard numbering (Ambler *et al*., 1991)], see Figure 1}, similar to what was observed for the SCHEMA library design for cytochrome P450 (Otey *et al*., 2006). The C-terminal subdomain boundary (residue 216) was chosen for the new N- and C-termini in a functional circularly permutated TEM-1 (Osuna *et al*., 2002). The third recombination site at residue 150 that appears in many of the RASPP libraries does not correspond to any previously identified subdomain boundaries.

The libraries identified by RASPP fall into three categories. The first group of libraries (Figure 1, black), at low <*m*>, have chimeras composed of a single large block with most of the recombination sites pushed toward the termini. These are the libraries with lowest <*E*> given the small fragment size ($L = 5$ or 6) allowed. While a large proportion of chimeras in these libraries are predicted to fold correctly, they are not very different from one another or from the parental sequences. The second group of libraries (Figure 1, red) has recombination sites that are distributed over the protein, yielding more diverse populations of chimeras. The <*E*> of these libraries are not significantly larger than those of the previous group. However, the blocks produced by the recombination sites vary considerably in size. The third group (Figure 1, green) has recombination sites that are well distributed over the protein

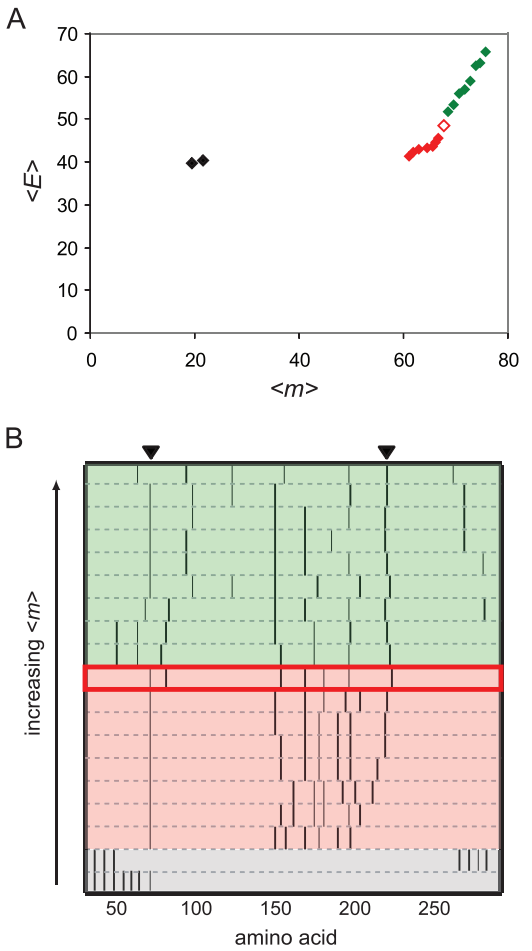

Fig. 1. The RASPP optimization algorithm produces a set of libraries with the lowest <*E*> at a range of <*m*> values. The libraries separate into three groups, designated by different colors (see text). (**A**) Plot of <*E*> versus <*m*> for RASPP libraries. The large gap in <*m*> is due to the difference between minimum fragment length, *L*, and <*m*> as measures of library diversity (Endelman *et al*., 2004). In the region of <*m*> between 25 and 55 there exist no libraries that have lower <*E*> than libraries identified at higher <*m*>. (**B**) The recombination sites of the RASPP libraries shown in order of decreasing <*m*> for the library. Black triangles indicate subdomain boundaries. The library chosen for construction is indicated by an open diamond on (A) and by a border on (B).

and blocks that are relatively uniform in size, yielding libraries of chimeras with high <*m*> and high <*E*>. Based on previous experiments with β-lactamases (Hiraga and Arnold, 2003; Meyer *et al*., 2003), the vast majority chimeras in this third group of libraries are predicted to be unfolded.

The second group was inspected further because these libraries are likely to yield diverse chimeras with relatively low *E*, and therefore a high fraction of folded proteins. From this group, the library with the greatest number of mutations per disruption (<*m*/*E*>) was chosen for construction. Two of the recombination sites were shifted by 1 or 2 amino acids from the recombination sites generated by RASPP to accommodate limitations of the construction protocol (Hiraga and Arnold, 2003). The shifted recombination sites do not change the overall characteristics of the library significantly. The library that was constructed recombines gene fragments of TEM-1, SED-1 and PSE-4 corresponding to the following blocks of amino acids [Ambler standard numbering (Ambler *et al*., 1991)]: 1–65, 66–73, 74–149, 150–161, 162–176, 177–190,

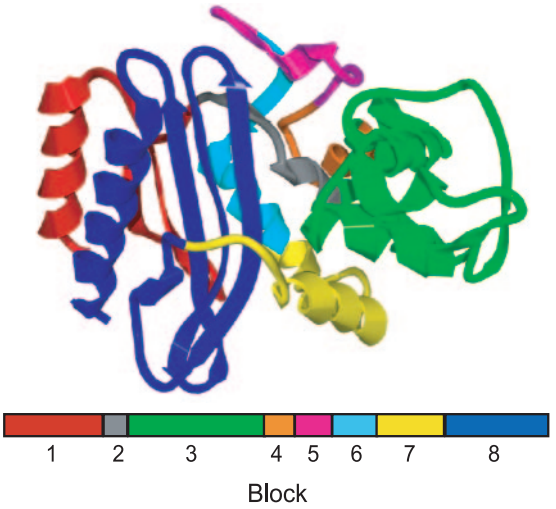

Fig. 2. Structure of TEM-1 (1BTL) showing positions of the sequence blocks that were recombined, colored coded as indicated, to generate the library.

191–218, 219–290. The corresponding structural elements are shown in Figure 2.

The β-lactamase signal sequence is included as part of the first block. However, all *E* and *m* calculations take into account only the mature proteins. The catalytically important residues are distant in the linear sequence and are therefore found on several different blocks, including blocks 2, 3, 5 and 8. Blocks 1 and 8 together comprise almost half the protein, consisting of the N- and C-terminal α+β subdomain. The final library design balances high <*m*> (66) for a diverse population with relatively low <*E*> (44) to ensure a large proportion of folded chimeras. Using previous data from a much smaller set of β-lactamase chimeras (Hiraga and Arnold, 2003), we estimated the probability of chimera folding based on *E* and predicted that ∼10% of chimeras in the library should retain fold and function.

The gene fragments from the three parental proteins corresponding to each block were combinatorially assembled using SISDC (Sequence Independent Site-Directed Chimeragenesis) (Hiraga and Arnold, 2003) to create a library of 6561 possible chimeric sequences. These genes were expressed in *E.coli*, and the sequences and functional status were determined by high-throughput probe hybridization and functional screening.

Sequence analysis of chimeras

The DNA sequences of 553 unique sequences were obtained by probe hybridization sequencing (Meinhold *et al*., 2003) of 1100 randomly selected clones. To determine the accuracy of the probe hybridization, we completely sequenced 48 randomly chosen chimeras. Comparison of the sequences with the block sequences determined by probe hybridization showed that the probe hybridization accurately determined the sequences of 47 of 48 chimeras. In the same group of chimeras we found two point mutations and ten deletions affecting 11 of the 48 chimeras. Of the 10 deletions, 3 were found at segment junctions and the remaining seven were found in regions within PCR primers used during construction, usually at the N-terminus.

Examining the sequence composition of the characterized chimeras on a ternary diagram shows that the characterized library does not have equal representation of the different parents (Figure 3A). In particular, many chimeras similar to

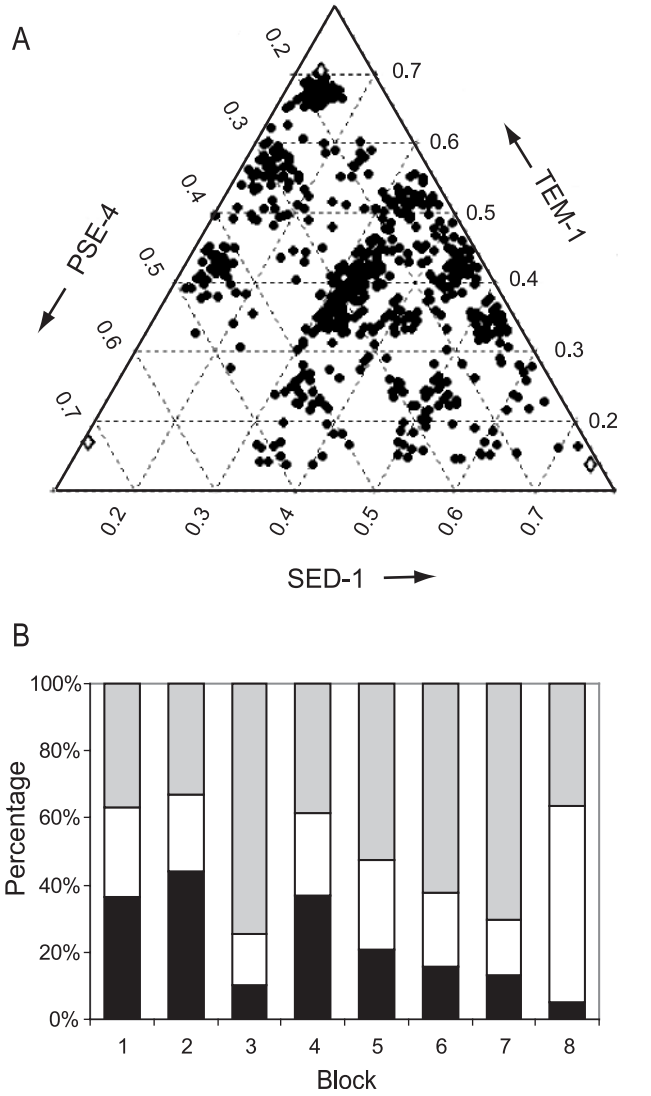

Fig. 3. The composition of characterized chimeras. (**A**) Ternary diagram showing the composition of unique characterized chimeras. Each chimera is represented by a data point whose position is determined by the sequence identity of the chimera to each parental sequence, not including residues shared by all three parents. The parents (open diamonds) are not at the corners of the diagram because each parent shares some identity with the other two. (**B**) The percentage parental sequences at each block for all characterized chimeras. Black represents PSE-4, white SED-1 and gray TEM-1. For the characterized library to perfectly reflect the theoretical library the percentage of each parent at each block should be 33%.

TEM-1 were characterized; only two inherit no block from TEM-1. The proportion of the different parents at each position shows that PSE-4 was severely underrepresented at block 8 (Figure 3B). This is due to an error in construction: a restriction site within block 8 from PSE-4 was used in the construction process. Chimeras that do contain PSE-4 at block 8 are a result of incomplete cleavage of the site. Examination of the *E* and *m* distributions of the characterized chimeras shows that the characterized library has disruption and mutation levels similar to the designed library, despite its biases ($E = 44 \pm 17$ and $m = 66 \pm 22$ for chimeras in the designed library versus $E = 45 \pm 17$ and $m = 66 \pm 24$ for chimeras in the characterized library).

Functional analysis of chimeras

In contrast to the cytochromes P450 previously studied by SCHEMA-guided recombination (Otey *et al*., 2006), there is

no simple assay for the folding status of β-lactamases. Consequently, we used a low stringency screen for catalytic activity to assess which chimeras retained basic catalytic function and thus a folded structure. Chimeras were screened for the ability to confer ampicillin resistance, a function shared by all three parental proteins. The screen was conducted at very low stringency (>500× lower concentration of ampicillin than the wild-type MIC) to capture chimeras with even very minimal activity. Sequences and functional status of the 553 unique characterized chimeras are presented in Supplementary Table S1 available at *PEDS* online. Of the 553 unique sequences tested, 111 (20%) conferred resistance to ampicillin and are considered functional β-lactamases (See Supplementary Table SI available at *PEDS* online). Of the functional chimeras, 57% conferred an MIC of 2000 μg/ml ampicillin or greater, indicating approximately wild-type activity (∼5000 μg/ml for all three parents) and 15% of functional chimeras were weakly active, displaying a MIC of 50 μg/ml or below. Chimeras that did not confer resistance to ampicillin may not fold, may not be well-expressed or may be folded but not catalytically active.

The functional β-lactamases are highly mosaic and have up to 86 mutations to the closest parental sequence (Figure 4). Most (75%) functional chimeras contain blocks from all three parents. Similar to previous observations for β-lactamase chimeras (Hiraga and Arnold, 2003; Meyer *et al*., 2003), the majority of the functional chimeras (80%) retain the N- and C- terminal fragments from the same parent. The functional β-lactamases have lower SCHEMA disruption than the nonfunctional β-lactamases ($E = 23 \pm 17$ versus $E = 49 \pm 14$) and fewer mutations ($m = 44 \pm 29$ versus $m = 71 \pm 31$). Examination of the E and m for functional and nonfunctional chimeras in the library shows that, at the same level of mutation, chimeras with lower E are much more likely to function and fold (Figure 5A).

Altering the MIC of ampicillin used as the functional cut-off does not significantly change the disruption distribution of the functional chimeras. However, nearly half of the chimeras with very high m (>75) are marginally functional (MIC $\leqslant$ 50 μg/ml). Removing the marginally functional chimeras from the population leaves 96 chimeras with $E = 23 \pm 11$ and $m = 42 \pm 27$. Allowing only chimeras with approximately wild-type activity (MIC $\geqslant$ 2000 μg/ml) results in a population of 63 chimeras with $E = 21 \pm 10$ and $m = 41 \pm 27$.

Rescue of nonfunctional chimeras

While chimeras with low E are more likely than chimeras with high E to retain at least weak catalytic activity, there remain many low-E chimeras that are nonfunctional. To examine whether and how nonfunctional chimeras could be rescued, we individually randomly mutated 10 low-E chimeras ($E < 35$) using error-prone PCR and selected clones conferring resistance to ampicillin. Of the ten chimeras, eight were rescued, most of them by a single mutation (Table I). There are 177 characterized chimeras with $E < 35$, of which 78 are nonfunctional. It is likely that many of these low-E chimeras can also be rescued. To examine whether all nonfunctional chimeras are as easily rescued by random mutagenesis, we chose an additional 12 chimeras with higher E ($E > 40$) and randomly mutated them (Supplementary Table SII available at *PEDS* online). None of these chimeras was rescued (Figure 5B).

Table I lists the mutations that rescue the eight low-E chimeras. About half change a single amino acid to an amino acid found in one or both of the other parents. This is not surprising for several reasons. First, the residues in the other parents are more likely to appear upon random nucleotide mutation due to conservation in the genetic code. Second, changing a residue to match one found in another parent may correct a beneficial interaction that was disrupted in the chimera. The mutations that introduce an amino acid found in a parental protein sequence are twice as likely to occur in interface or buried positions (<50% solvent exposed surface area) than on the surface of the protein compared with the mutations that introduce an amino acid not observed in the parental sequences. Two of the mutations have been described previously: H153R and M182T in TEM-1 not only revert to

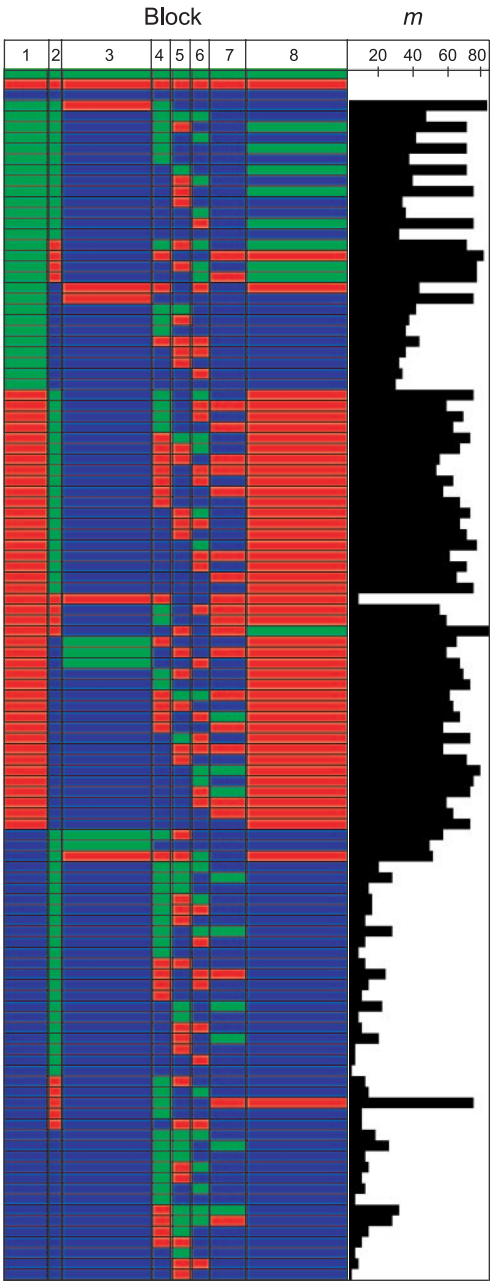

Fig. 4. The sequences of functional chimeras indicated by the parent from which each block is inherited, green for PSE-4, red for SED-1 and blue for TEM-1. The number of mutations to the closest parent (m) for each chimera is indicated by the length of the black bar and ranges between 0 for the parental sequences to 86.

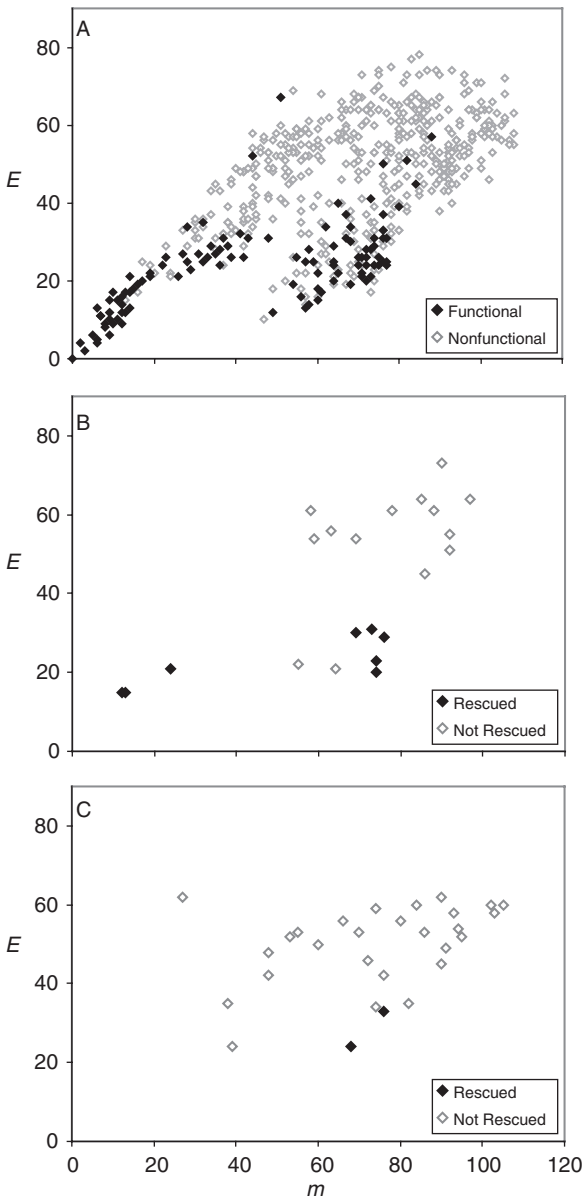

Fig. 5. Disruption (*E*) and mutations to the closest parent (*m*) for characterized chimeras shows that lower-*E* chimeras are more likely to be functional (**A**) and also are more likely to be rescued by random mutagenesis (**B**) or by the stabilizing mutation TEM-1 M182T (**C**).

the amino acids found in PSE-4 or SED-1 but are also known stabilizing mutations frequently identified in extended-spectrum TEM-1 variants (Knox, 1995). Most of the remaining mutations are on the protein surface or in interface regions, and the rescue mechanism is not immediately apparent. There is, for example, no trend to replace an amino acid residue with one that appears more frequently in the β-lactamase PFAM seed alignment (Bateman *et al*., 2004). All positions identified were shown to be tolerant to mutation in a site-saturation study of TEM-1 (Huang *et al*., 1996).

The TEM-1 M182T mutation was identified in half of the chimeras rescued, and it was the most frequently observed. It has been shown to suppress the effects of other deleterious mutations by increasing the stability of TEM-1 by 2.7 kcal/mol (Wang *et al*., 2002) and most likely has the same effect in the chimeric proteins. To examine whether TEM-1 M182T could rescue other chimeras, we introduced it into 29 nonfunctional chimeras with a range of disruption levels (Supplementary Table SII available at *PEDS* online). Of the 29 chimeras, two were rescued by this single mutation. Similarly to chimeras rescued by random mutation, chimeras with low *E* appear more likely to be rescued (Figure 5C). Both of the chimeras rescued by M182T have $E < 35$ and the N- and C-termini from the same parent.

Discussion

There are 1785 β-lactamase sequences in the PFAM database for protein families (Bateman *et al*., 2004), of which at least 450 are class A β-lactamases by phylogenetic analysis. However, many of the characterized β-lactamases are minor variants of a few very prevalent sequences. For example, there are over 100 characterized variants of TEM-1 that differ from TEM-1 by only a few amino acids (Jacoby and Bush, 2005). The β-lactamase structure is relatively tolerant to mutation: 220 of 263 positions in TEM-1 accept at least one other amino acid when mutated (Huang *et al*., 1996), and several other experiments indicate that PSE-4 and TEM-1 can easily tolerate minor modifications (Petrosino and Palzkill, 1996; Matagne *et al*., 1998; Sanschagrin *et al*., 2000; Osuna *et al*., 2002). The robustness of the β-lactamase structure is also apparent from the work performed here. We have identified >100 new β-lactamases which share as little as 70% sequence identity with any known sequence. While many of the chimeras are quite similar to one of the parental sequences, the majority have 45 or more sequence changes compared to the closest parent. The library contains many hundreds more new functional β-lactamases.

In contrast to our previous work with β-lactamase chimera libraries (Hiraga and Arnold, 2003; Meyer *et al*., 2003), this library was specifically designed to minimize the average disruption (<*E*>) of the population of chimeras. While the chimeras analyzed in the Meyer *et al*. study are not directly comparable due to differences in the experimental system used to define a functional β-lactamase, the chimeras in the Hiraga *et al*. study are directly comparable. The library described in this work contains approximately a 4-fold greater fraction of functional chimeras while maintaining a higher average level of mutation ($m = 66 \pm 24$ in this work versus $m = 52 \pm 16$ for the Hiraga *et al*. library). The increase in fraction of folded chimeras is a reflection of the lower *E* of chimeras in the library described here ($E = 44 \pm 17$) compared with the Hiraga *et al*. library ($E = 54 \pm 17$).

Maranas and co-workers have proposed a computational procedure for library design, OPTCOMB, which permits leaving out specific parental fragments at key positions in order to reduce the disruption caused by recombination (Saraf *et al*., 2005). In this work we observed that functional chimeras tend to have the N- and C- termini from the same parent. The population of 2187 chimeras in the library whose N- and C- termini originate from the same parent in fact have a much lower *E* (27 ± 7). However, there is currently no good method for constructing such a constrained library.

We observed that ∼20% of characterized chimeras in the library retained function. The true fraction of folded chimeras is most likely higher because there are false-negative signals resulting from the single base-pair deletions. The SCHEMA-guided library of cytochrome P450 heme domains described previously (Otey *et al*., 2006) contains a significantly higher

Table I. Mutations that rescue low-*E* chimeras

Position	8[a]	22[a]	27[a]	63	72	99	100	114	120	147	153	171	174	182	191	193	224	261
PSE-4				N	F	K	A	G	D	G	R	E	L	T	N	F	V	V
SED-1	Q		H	E	S	K	A	G	A	N	R	T	P	S	R	L	G	L
TEM-1		F		D	F	Q	N	T	R	E	H	E	P	M	R	L	A	V
STSPTSS					F-S													
	Q-L				F-S													
TSTSSTTT														M-T				
												T-S		M-T				
TTTSSTPT														M-T				
						Q-R								M-T				
														M-T			A-T	
									R-G	E-G					N-Y			
																F-L		
		F-L						T-S								F-L		
PTTTPTTP														M-T				
SPTTPSTS													L-P					
													L-P					L-A
				E-G									L-P					
SPPTTSTS	Q-L		H-L								H-R							
PSTTSTTP														M-T				
SSTTTPTS	Q-L																	
	Q-L									E-G								
							N-S											L-A
TPSPTTTT																		
TSSSSSTT																		

[a]Residue found in the signal sequence of the protein. This region is not included in the protein alignment.

fraction of folded chimeras (47%). The lower β-lactamase functional fraction likely reflects the greater divergence of the β-lactamase parental sequences (34–42% versus ~61–63% for the cytochrome P450 heme domains), which results in more disruption in the chimeras. Although the β-lactamase is considerably smaller than the cytochrome P450 heme domain (~265 versus ~460 amino acids), the SCHEMA disruption is higher for the β-lactamase chimeras than for the cytochrome P450 chimeras ($E = 44 \pm 17$ versus $E = 32 \pm 10$). Individual mutations may also be inherently more disruptive for the more diverged sequences, because these sequences accumulate more mutations in core regions.

We have also shown that at least some nonfunctional chimeras can be rescued by point mutations. The most common mutation observed to rescue β-lactamase function, TEM-1 M182T, is a well-known stabilizing mutation, which suggests that many chimeras fail to function due to loss of stability. We have recently shown that more stable proteins are more tolerant to random mutations (Bloom *et al*., 2005) and therefore have a greater capacity to evolve functionally because they can accept more destabilizing mutations (Bloom *et al*., 2006). More stable proteins will also be more robust to the mutations introduced by recombination.

SCHEMA-guided recombination is an effective way to generate synthetic protein families with broad sequence diversity while maintaining a relatively high percentage of folded and functional proteins. Furthermore, the proportion of folded variants can probably be increased through simple solutions such as utilizing stabilized parental sequences. Large datasets are generated by characterizing these libraries, and, unlike natural protein families, these sets include both functional and nonfunctional sequences that can be queried for specific properties in high throughput formats. The value of this resource for sequence-structure-function analyses was recently demonstrated by Li, Y., Drummond, D.A., Otey, C.R., Landwehr, M. and Arnold, F.H. (unpublished data) who showed that folding status and thermostability can be predicted from analyzing the multiple sequence alignments of folded and not-folded chimeras.

Acknowledgements

We thank Costas Maranas, Brian Shoichet and Joelle Pelletier for their comments. The *sed-1* gene was a gift from S. Petrella and W. Sougakoff. This work was supported by NIH R01 GM068664, an HHMI predoctoral fellowship (to M.M.M.), and a NSF graduate research fellowship (to L.H.).

References

Ambler,R.P., Coulson,A.F.W., Frere,J.-M., Ghuysen,J.-M., Joris,B., Forsman,M., Levesque,R.C., Tiraby,G. and Waley,S.G. (1991) *Biochem. J*., **276**, 269–272.

Bateman,A. *et al*. (2004) *Nucleic Acids Res*., **32**, D138–D141.

Bloom,J.D., Silberg,J.J., Wilke,C.O., Drummond,D.A., Adami,C. and Arnold,F.H. (2005) *Proc. Natl Acad. Sci. USA*, **102**, 606–611.

Bloom,J.D., Labthavikul,S.T., Otey,C.R. and Arnold,F.H. (2006) *Proc. Natl Acad. Sci. USA*, **109**, 5869–5874.

Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) *Nucleic Acids Res*., **31**, 3497–3500.
Doolittle,R.F. (1986) *Of URF and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences*. University Science Books, Mill Valley, CA, USA.
Drummond,D.A., Silberg,J.J., Meyer,M.M., Wilke,C.O. and Arnold,F.H. (2005) *Proc. Natl Acad. Sci. USA*, **102**, 5280–5385.
Endelman,J.B., Silberg,J.J., Wang,Z.-G. and Arnold,F.H. (2004) *Protein Eng. Des. Sel*., **17**, 589–594.
Guex,N. and Peitsch,M.C. (1997) *Electrophoresis*, **18**, 2714–2723.
Hiraga,K. and Arnold,F.H. (2003) *J. Mol. Biol*., **330**, 287–296.
Huang,W., Petrosino,J., Hirsch,M., Shenkin,P.S. and Palzkill,T. (1996) *J. Mol. Biol*., **258**, 688–703.
Jacoby,G. and Bush,K. (2005) Lahey Clinic page on *''Amino Acid Sequences for TEM, SHV and OXA Extended-Spectrum and Inhibitor Resistantbeta-lactamases''* http://www.lahey.org/Studies/.
Jelsch,C., Mourey,L., Masson,J.M. and Samama,J.P. (1993) *Proteins*, **16**, 364–383.
Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure*, **5**, 1093–1108.
Knox,J.R. (1995) *Antimicrob. Agents Chemother*., **39**, 2593–2601.
Lim,D., Sanschagrin,F., Passmore,L., De Castro,L., Levesque,R.C. and Strynadka,N.C.J. (2001) *Biochemistry*, **40**, 395–402.
Lutz,S. and Patrick,W.M. (2004) *Curr. Opin. Biotechnol*., **15**, 291–297.
Matagne,A., LaMotte-Brasseur,J. and Frere,J.-M. (1998) *Biochem. J*., **330**, 581–598.
Maveyraud,L., Mourey,L., Pedalacq,J.-D., Guillet,V., Kotra,L.K., Mobashery,S. and Samama,J.P. (1998) *J. Am. Chem. Soc*., **120**, 9748–9752.
Meinhold,P., Joern,J.M. and Silberg,J.J. (2003) In Arnold,F.H and Georgiou,G. (eds), *Directed Evolution Library Creation*. Humana Press, Totowa, New Jersey, pp. 177–187.
Meyer,M.M., Silberg,J.J., Voigt,C.A., Endelman,J.B., Mayo,S.L., Wang,Z.-G. and Arnold,F.H. (2003) *Protein Sci*., **12**, 1686–1693.
Meyer,M.M., Hiraga,H. and Arnold,F.H. (2006) In Coligan,J.E., Dunn,B.M., Speicher,D.W. and Wingfield,P.T. (eds), *Current Protocols in Protein Science*. John Wiley & Sons, Hoboken, NJ, pp. 26.22.21–26.22.17.
Moore,G.L. and Maranas,C.D. (2003) *Proc. Natl Acad. Sci. USA*, **100**, 5091–5096.
Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol*., **247**, 536–540.
Ness,J.E., Welch,M., Giver,L., Bueno,M., Cherry,J.R., Borchert,T.V., Stemmer,W.P.C. and Minshull,J. (1999) *Nat. Biotechnol*., **17**, 893–896.
Ostermeier,M. (2003) *Trends Biotech*., **21**, 244–247.
Ostermeier,M., Shim,J.H. and Benkovic,S.J. (1999) *Nat. Biotechnol*., **17**, 1205–1209.
Osuna,J., Perez-Blancas,A. and Soberon,X. (2002) *Protein Eng*., **15**, 463–470.
Otey,C.R., Landwehr,M., Endelman,J.B., Hiraga,K., Bloom,J.D. and Arnold,F.H. (2006) *PLoS Biol*., **4**, e112.
Petrella,S., Clermont,D., Casin,I., Jarlier,V. and Sougakoff,W. (2001) *Antimicrob. Agents Chemother*., **45**, 2287–2298.
Petrosino,J.F. and Palzkill,T. (1996) *J. Bacteriol*., **178**, 1821–1828.
Poteete,A.R., Rennell,D., Bouvier,S.E. and Hardy,L.W. (1997) *Protein Sci*., **6**, 2418–2425.
Rost,B. (1999) *Protein Eng*., **12**, 85–94.
Sanschagrin,F., Theriault,E., Sabbagh,Y., Voyer,N. and Levesque,R.C. (2000) *Antimicrob. Agents Chemother*., **45**, 517–519.
Saraf,M.C. and Maranas,C.D. (2003) *Protein Eng*., **16**, 1025–1034.
Saraf,M.C., Horswill,A.R., Benkovic,S.J. and Maranas,C.D. (2004) *Proc. Natl Acad. Sci. USA*, **101**, 4142–4147.
Saraf,M.C., Gupta,A. and Maranas,C.D. (2005) *Proteins*, **60**, 769–777.
Shortle,D. and Lin,B. (1985) *Genetics*, **110**, 539–555.
Sieber,V., Martinez,C.A. and Arnold,F.H. (2001) *Nat. Biotechnol*., **19**, 456–460.
Voigt,C.A., Kauffman,S. and Wang,Z.G. (2001) In Arnold,F.H. (ed.), *Advances in Protein Chemistry, Vol 55,* Academic Press, pp. 79–160.
Voigt,C.A., Martinez,C., Wang,Z.-G., Mayo,S.L. and Arnold,F.H. (2002) *Nat. Struct. Biol*., **9**, 553–558.
Wang,X., Misasov,G. and Shoichet,B. (2002) *J. Mol. Biol*., **320**, 85–95.

Received September 20, 2006; accepted September 27, 2006

Edited by Stephen Mayo