Exploring Protein Sequence Space Using Computationally Directed Recombination

Thesis by

Michelle Margaret Meyer

In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

California Institute of Technology
Pasadena, California, USA
2006

(Defended May 24, 2006)

Acknowledgements

My time at Caltech has been a gift. I have had the opportunity to work with many great people who have given me much in terms of both personal and professional development. First I would like to thank my advisor, Frances Arnold. She has given me guidance and support throughout this process but also allowed me the freedom to make my own mistakes as well as achieve my own victories. She has also been very understanding in times of great stress, and truly a champion when it comes to giving feedback quickly. This work would not have been possible without her, and certainly would never have been done in time. My other committee members, Steve Mayo, Doug Rees, and Rob Phillip, have provided many comments that have guided the direction of this research.

The work described here was not conducted in isolation, but with many incredible collaborators. Joff Silberg was an important guide during my first year here, and Allan Drummond, Lisa Hochrein, and Zhongyi Lu have helped me with various experiments. While only the mutagenesis experiments Lisa started made it into this document, all the experiments were essential for realizing what the right direction should be. I also need to thank Jennifer Keefe in the Mayo group for her help in obtaining the circular dichroism measurements.

Aside from all the people that have worked on experiments with me, I also owe a great deal to a group of less experimentally inclined collaborators; the “theoretical people,” who made all of this work possible. Christopher Voigt first told me to write my own code, and made me aware that I could. Jeff Endelman allowed much of this work to take place developing many improvements in our design and analysis methodologies.

Allan Drummond is always willing to talk about crazy ideas. He continues to give me confidence that when I play with numbers, I am making a worthwhile contribution.

Aside from all of my immediate collaborators there have been discussions with many people over the years that have contributed to not just to science but also to maintaining sanity. Karen Wawrousek, Jenny Witman, and Lisa Hochrein are not only good friends but also kept me from falling apart during a difficult time. Chris Otey once reminded me that it is ok, and sometimes better, to just let go of the problem. Jesse Bloom has been a good friend and a great person to hang out with in the lab, especially since we have the same taste in radio stations. Marco Landwehr continues to remind me that I cannot work all the time, so computer games do still have a place. I have had many morning coffee breaks with Jorge Rodriguez and Geethani Bandara that have served as a welcome reason to procrastinate. Finally, conversations with my other officemates not yet named, Mike Chen and Alex Tobias, have reminded me from time to time exactly why we keep coming in every day.

My path, however, did not start when I reached Caltech. There were many people that helped bring me here. My high school biology teacher Brian Pelkey always raised the bar and demanded true excellence. In doing so he showed me how much more there was to science and introduced me to what I consider a fantastic puzzle: how life works. My undergraduate research advisor at Rice University, Seiichi Matsuda not only introduced me to the work that I love, but also taught me to observe carefully and to remember that there is usually a reason *why* an experiment does not work.

My parents, Robert Meyer and Susan Conry Meyer, have always encouraged me to dream big dreams. They have been a constant source of support through this entire

process. My husband Matthew has, in addition to answering the occasional math or programming question, often provided a surrogate sense of humor when mine was sadly absent. He has also served to remind me how important it is to do what you love; otherwise life is just not the same.

Abstract

Evolution has provided us with many protein sequences. However, these sequences represent a very small fraction of the possible sequences. In the laboratory, scientists have explored areas of sequence space not represented by natural proteins both to better understand natural proteins, and to create new proteins with desirable properties. The principle mechanism used to explore protein sequence space is mutagenesis. However, recombination of homologous genes can also explore regions of sequence space rich with folded and functional proteins.

In this work we demonstrated using a β-lactamase model system that a computation energy function (SCHEMA) can predict which of the chimeras made by recombining distantly related proteins are likely to fold. SCHEMA uses protein sequence and structure information to identify pairwise amino acid interactions disrupted by recombination. Using SCHEMA we designed libraries of chimeric β-lactamases. These libraries were intended to have a high fraction of folded variants, while incorporating many amino acid substitutions compared with the parental proteins. The chimeras in these libraries were characterized to determine whether they retain the parental function and what new substrate specificities could be obtained.

To identify critical variables for determining whether a chimera functions, we used logistic regression analysis to analyze functional and nonfunctional chimeras. From this analysis it is apparent that both two-body (pairwise) and one-body terms play a significant role in determining whether a chimera functions. We also used random mutagenesis to restore functionality to nonfunctional chimeras showing that a thermostabilizing mutation can rescue approximately 5% of the nonfunctional chimeras.

The one-body terms that appear significant for determining whether a chimera functions are not explicitly counted by SCHEMA when predicting chimera folding. To estimate the effects on chimera folding represented by the one-body terms, we developed an additional measure to predict chimera folding based on just the chimera amino acid sequence and a multiple sequence alignment of homologous proteins. This measure is predictive of chimera folding alone, and when combined with the pairwise SCHEMA energy increases the accuracy of the folding predictions compared to SCHEMA.

Table of Contents

List of Tables

List of Figures

Chapter I: Recombination as an Evolutionary Search Strategy

Mechanisms for Evolution

The space of all proteins is very large, containing more possible sequences than there are atoms in the universe. The protein sequences we observe today were created by evolution over millions of years, yet all the proteins explored by evolution represent a very small proportion of the possible sequence space. The frequency of functional proteins among random sequences is estimated between 1 in 10^{11} (Keefe and Szostak 2001) and 1 in 10^{77} (Axe 2004). However, nature has managed to give us thousands of proteins with a plethora of different functions principally using two mechanisms to explore possible protein sequences: mutation and recombination.

The effects of mutation on proteins have been well studied, partially because mutagenesis has been used extensively to perturb protein sequence in order to study the relationship between sequence and structure or function. Most proteins are robust to random mutations and show a great deal of tolerance for single amino acid substitutions (Rennell et al. 1991; Markiewicz et al. 1994). However, as additional random substitutions are incorporated, the probability of a protein retaining function declines exponentially with the number of substitutions (Daugherty et al. 2000; Guo et al. 2004). Simulations using model proteins have suggested that the mutational tolerance of a natural protein is a property arising from evolutionary history (Taverna and Goldstein 2002). Evolved populations of model proteins tend to form “neutral networks” of proteins clustered around a sequence that is most mutationally robust or the “prototype sequence” (Bornberg-Bauer and Chan 1999). These sequences are related by single point mutations and are selectively neutral. Some structures have larger neutral networks than others, and

are considered more designable. Such highly designable sequences are robust to mutation and are thought to correspond to natural protein folds (Li et al. 1996; Govindarajan and Goldstein 1996).

The effects of recombination on protein structure and function are much less well explored. Recombination has not been extensively used to perturb existing natural proteins, thus there is much less data available on the effects of recombination on protein function. Genetic studies have shown that recombination can occur within protein encoding regions (Dooner and Martinez-Ferez 1997; Feil et al. 1999; Feil et al. 2001; Fu et al. 2002). However, all of these data are drawn from genomic sequences and thus reflect extensive natural selection of the organism in addition to selective pressure at the single protein level. Studies of recombination with model proteins show nonhomologous recombination can lead to more efficient searching of sequence space than point mutation alone (Bogarad and Deem 1999; Cui et al. 2002), and that populations evolved allowing recombination are more centered around the prototype sequence (Xia and Levitt 2002).

Laboratory Evolution

Both recombination and mutation are used as tools to engineer new proteins with desirable properties. While evolution of natural proteins has taken place with large populations over millions of years, the situation in the laboratory is more limited. In laboratory evolution an improved variant must be identifiable within a population of variants, and experimental constraints limit the number of variants that can be searched in a single generation (Voigt et al. 2001a). Thus finding the improved variants within a population is critical for success. Recombination and mutation allow access to very

different areas of sequence space, and depending on the desired properties one may be more beneficial than the other.

Random mutagenesis using error-prone PCR is by far the easiest method for generating diversity in directed evolution. It is inexpensive, requires no information beyond a starting protein sequence, and is usually effective at producing variants with modified properties such as increased thermostability or altered substrate specificity. However, as structural and sequence information have accumulated, site-saturation mutagenesis has become more widely used for generating diversity. Site-saturation mutagenesis seeks to bias the population of variants generated to specific areas of sequence space in order to increase the probability of identifying variants of interest. It also allows access to some mutations that cannot be searched using random mutagenesis due to the conservative nature of the genetic code.

Several different studies have shown that variant populations incorporating many mutations simultaneously are better at adaptive evolution than population of variants incorporating only a few mutations (Crameri et al. 1998; Zaccolo and Gherardi 1999; Daugherty et al. 2000). However, additional diversity generated by increasing the number of random mutations comes at a cost to the fraction of functional variants. Due to the exponential decline in functional proteins as the number of random mutations increases, the majority of sequences constructed with a high mutation rate are nonfunctional and likely not folded. Without very high throughput screening or selection techniques, identifying an improved variant in the large population is impossible.

Recombination between protein sequences makes it possible to take larger steps in sequence space without sacrificing the fraction of folded variants. Recombination of

proteins sharing the same fold allows only mutations that are compatible with the backbone characteristics of the fold. Because the amino acids introduced by recombination are unlikely to have deleterious interactions with the backbone, the major contributions to whether a chimera retains function are the pairwise interactions between the amino acid substitutions introduced. Thus, mutations introduced by recombination are less likely to disturb the protein structure than random mutations (Drummond et al. 2005). Recombination of homologous proteins mimics the accumulation of neutral mutations that occur in nature, but on a shorter time scale.

Homologous recombination has been a successful strategy for creating protein variation in directed evolution studies. Proteins isolated from DNA shuffling experiments have properties that are due to the action of several mutations working synergistically (Stemmer 1994; Crameri et al. 1998). However, DNA shuffling experiments are still limited in the area of sequence space that they can explore. The variants produced are usually not significantly different than their parents, displaying more than 85% sequence identity to the closest parent (Arnold 2000). While this represents a significant step in sequence space, the diversity of natural proteins suggests that even larger jumps are possible without disrupting protein structure.

DNA shuffling is an effective strategy for recombining closely related sequences. However, as the sequence identity between the DNA sequences recombined decreases below approximately 70%, DNA shuffling becomes much less effective as the number of potential crossover sites decreases and parental genes are recovered more frequently from the reaction. In addition to limiting the sequence diversity of the proteins recombined, annealing-based methods also bias the positions of recombination points and the

incorporation of different parent sequences (Joern et al. 2002). Therefore, to increase the possible sequence diversity incorporated by recombination it is necessary to find alternative recombination methods to DNA shuffling.

To overcome the limitations of annealing-based recombination methods, homology-independent techniques have been developed that can recombine sequences of any identity (Ostermeier et al. 1999a; Ostermeier et al. 1999b; Lutz et al. 2001; Sieber et al. 2001; Kawarasaki et al. 2003). However, these techniques often result in a population of chimeras that is largely unfolded. Chimeras are unfolded for two reasons in these experiments: First the methodologies incorporate a significant number of frameshifts, deletions and insertions because recombination sites are randomly generated in the DNA sequence. This results in a large proportion (>66% statistically) of nonfunctional proteins in the library. Second, even using recombination, additional diversity introduced by recombining more distantly related sequences comes with a decrease in the fraction of folded variants (Meyer et al. 2003; Ostermeier 2003); recombining more diverse sequences results in an increased number of potentially deleterious pairwise interactions (Drummond et al. 2005).

Site-Directed Recombination

Site-directed recombination seeks to circumvent many of the problems in both annealing-based and homology-independent recombination methods by choosing specific sequence blocks of the parental genes, and reassembling these blocks combinatorially to create a library of variants (Figure I-1). Genes of any sequence identity can be used, and the assembly technique incorporates few insertions, deletions and frame shifts because

the recombination sites are chosen before construction and not randomly determined during the experiment (Hiraga and Arnold 2003). Additionally, site-directed recombination can be used to bias the sequence space examined to areas that are more likely to have folded sequences. Similar to the way that site-saturation mutagenesis limits the pool of possible mutants to those expected to have certain properties, site-directed recombination limits the chimeras constructed to a pool with expected properties. Structural or previous experimental data are typically used to determine sites for site-saturation mutagenesis. For site-directed recombination, a chimera's probability of retaining fold and function can be assessed *in silico* prior to construction.

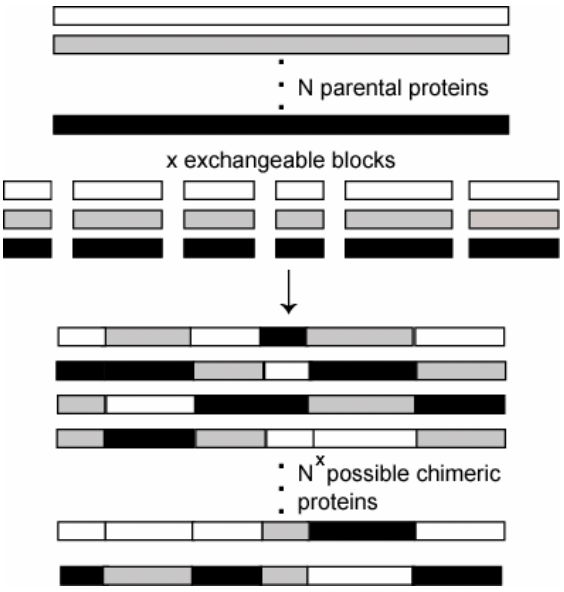


Figure I-1. Site-directed recombination recombines discrete blocks of sequence combinatorially to create a library of chimeric proteins.

Several different computational energy functions have been developed to evaluate *in silico* the probability of chimeric proteins folding (Voigt et al. 2002; Moore and Maranas 2003; Saraf and Maranas 2003; Saraf et al. 2004; Hernandez and LeMaster 2005). All of these energy functions strive to predict chimera folding or functionality based on pairwise interactions. This is compatible with the understanding that pairwise

side chain interactions are the largest contributor to chimera misfolding. Many energy functions identify amino acid residue pairs using three-dimensional structure information. However, the treatment of these pairs varies considerably, from simply counting clashes (Voigt et al. 2002), to mediating such clashes with biophysical information (Saraf and Maranas 2003), or mean-field calculations (Moore and Maranas 2003). Additionally, paired residues have been identified using multiple sequence alignment conservation (Saraf et al. 2004).

The scoring functions for chimeric proteins have not been tested against large, well-characterized data sets of chimeric proteins. Several have been scored against chimeras derived from directed evolution experiments (Voigt et al. 2002; Moore and Maranas 2003; Saraf and Maranas 2003). However, the naïve populations from which the chimeras were selected are typically not well characterized. The lack of characterization of the naïve populations makes it difficult to determine if trends observed in functional chimeras are a result of the functional selection, or trends within the naïve population. To better understand how recombination can be used as an effective mechanism for searching sequence space, both in the laboratory and in nature, larger and better-characterized data sets are required for analysis. In this work we describe the design and creation of several large libraries of chimeric proteins. Analysis of the chimeras produced allows us to examine attributes of chimeras that contribute to folding and function, and to evaluate the tools used to predict chimeric protein folding.

Chapter II: Library Analysis of SCHEMA-Guided Recombination

Portions of this chapter are reproduced from Meyer et al. 2003 "Library Analysis of SCHEMA-Guided Recombination" *Protein Science* **12**: 1686-1693.

Introduction

Recombination is an effective strategy for exploring protein sequences that differ significantly from those found in nature but maintain folded and functional structures. However, as the sequence identity between the proteins to be recombined decreases, the fraction of folded variants created also decreases (Ostermeier 2003). Several computational energy functions have been developed to predict which chimeras are likely to fold and function (Voigt et al. 2002; Moore and Maranas 2003; Saraf and Maranas 2003; Saraf et al. 2004; Hernandez and LeMaster 2005). These scoring functions examine potential pairwise clashes between amino acids introduced from different parents. The residue-residue interactions are predicted to be the dominant contributors to whether a chimera retains the parental structure (Drummond et al. 2005). However, most energy functions are typically tested using small and incompletely characterized data sets, making it difficult to determine how well the energy function is performing.

In this work we examine the pairwise scoring function SCHEMA that predicts which fragments of homologous proteins can be recombined without disturbing the integrity of the structure (Voigt et al. 2002). This is by far the simplest scoring function described and makes few assumptions. Based on a three-dimensional structure of a parent protein, SCHEMA identifies pairs of amino acids that are interacting, defined as those residues within a cutoff distance of 4.5Å, and determines the net number of interactions

broken when a chimeric protein inherits portions of its sequence from different parents (defined as *E*). A pair of residues whose identities do not change upon recombination cannot be broken by the recombination event.

Because calculating *E* (see methods) for all possible combinations of recombination sites, or crossovers, is computationally intractable, it is difficult to identify optimal crossovers that yield folded chimeras. The SCHEMA profile proposed by Voigt et al. (2002) circumvents this computational difficulty by finding compact, contiguous polypeptides with the largest number of intra-block interactions – these polypeptides correspond to fragments which, in theory, can be swapped with minimal cost. This is achieved by scanning the protein sequence with a window of defined size to create a disruption profile whose minima are predicted to represent crossover locations that preserve more interactions. It was proposed that the resulting fragments, or schemata, could be recombined using available laboratory recombination methods (Horton et al. 1989; Solaiman et al. 2000; Gibbs et al. 2001; O'Maille et al. 2002) to generate novel mosaic sequences that retain the parental structure.

A strong correlation exists between SCHEMA profiles and existing experimental data on chimeras from site-directed recombination and DNA-shuffling experiments. In particular, the vast majority of the crossovers found in functional chimeras containing 1 or 2 crossovers appear in or near the minima of their calculated disruption profiles (Voigt et al. 2002), suggesting that crossovers at other locations (e.g., profile maxima) are unfavorable. Furthermore, functional analysis of twelve lactamase chimeras revealed that proteins tolerate a limited level of *E*; only those with $E \leq 26$ were functional (Voigt et al. 2002). However, the small numbers of functional and nonfunctional chimeras analyzed in

these studies and the small number of crossovers incorporated make it difficult to determine just how SCHEMA predictions correlate with functional and structural disruption. We would like to know whether chimeras with low *E* have a higher probability of retaining parental function than those with the same effective level of mutation but chosen at random. We would also like to know whether the minima in the profile still correspond to the best recombination sites when multiple crossovers are allowed.

To address these questions we created a large library of chimeras with a broad range of *E* and examined which recombination events conserved function. For this test we chose to recombine two class A β-lactamases, TEM-1 and PSE-4, that share 40% sequence identity. The class A β-lactamases represent an ideal model system because functional chimeras are easily identified using antibiotic selection. Additionally, due to their medical significance there is a great deal of structural and sequence information available for class A β-lactamases. There are hundreds of β-lactamase sequences available in the database sharing between 99% and 15% sequence identity (Bateman et al. 2004) and twelve class A β-lactamases have been crystallized. The crystallized proteins share between 70% and 23% identity, and despite highly diverged sequences they have nearly identical structures with no more than 3.5 Å RMSD over all backbone atoms (Figure II-1) (Dideberg et al. 1987; Herzberg 1991; Knox and Moews 1991; Swarent et al. 1998; Ibuka et al. 1999; Kuzin et al. 1999; Tranier et al. 2000).

Sequence identity (%)

	1BSG	1BTL	1BUE	1BZA	1DY6	1E25	1G68	1MFO	1SHV	3BLM	4BLM
1BSG		37	41	41	40	23	30	42	40	31	41
1BTL	2.3		33	36	34	24	41	41	67	33	37
1BUE	1.8	2.2		47	74	23	34	38	35	34	42
1BZA	1.2	2.4	1.3		47	25	37	44	38	34	40
1DY6	1.6	2.0	0.5	1.2		26	36	36	39	36	44
1E25	3.2	3.0	2.9	3.1	2.9		23	26	25	21	24
1G68	2.5	1.4	1.9	1.9	1.9	2.7		34	45	36	28
1MFO	2.0	2.4	1.6	1.5	1.6	3.5	2.1		39	32	39
1SHV	2.9	1.3	2.5	2.8	2.4	2.7	1.6	2.7		30	34
3BLM	2.1	2.5	2.2	2.2	2.1	3.9	2.5	2.5	2.6		43
4BLM	1.4	2.3	1.6	1.2	1.6	3.3	1.7	1.7	2.8	1.4	

RMSD over backbone atoms (Å)

Figure II-1. Pairwise sequence identity and RMS deviation over the backbone atoms of all distinct class A β-lactamase structures designated by their protein data bank (pdb) code. Despite the highly diverged sequences, all lactamases crystallized have very similar structures.

Results

Library Design and Characterization

The SCHEMA-calculated profile shown in Figure II-2 was used to guide the creation of a diverse library of lactamase chimeras exhibiting a broad range of disruption. Eight major peaks in the profile correspond to eight polypeptides with the largest number of intra-block interactions. We allowed recombination at seven minima and six maxima of the disruption profile, yielding a library containing 2^{14} (16,384) possible unique chimeras. By calculating the exact disruption (E) of every sequence, we determined that the library contains chimeras with disruption values ranging from 7 to 113. Additionally, the chimeras display a broad range of effective mutations (m), from 7 to 75 amino acid substitutions relative to the closest parent.

Twenty-eight gene modules were synthesized chemically or by PCR (fourteen for each parent). Gene modules encoding structurally related elements contained identical unique 5' overhangs, but the sequences of the overhangs at each module boundary were

distinct and nonpalindromic. Each parental gene was assembled to confirm that no mutations were present in the modules and to validate that full-length genes could be created. Because ligation efficiency decreased as the number of fragments increased, we used a serial assembly protocol. Two or three adjacent gene fragments were ligated and purified using an agarose gel to create six distinct sets of products. This process was repeated using the ligated products until the full-length genes were assembled.

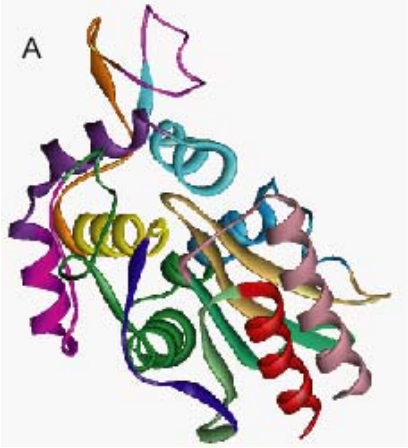
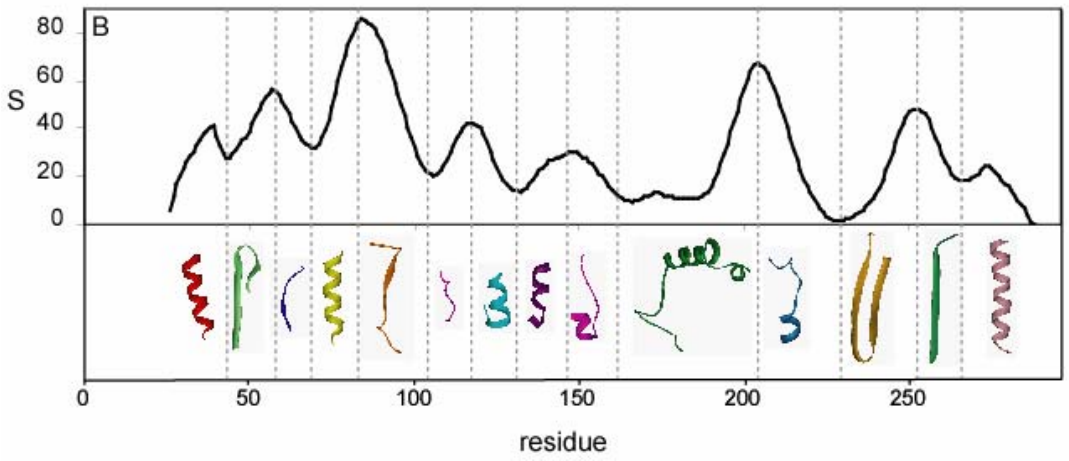


Figure II-2. Polypeptides recombined between TEM-1 and PSE-4. A: Polypeptide modules swapped between lactamases are mapped onto the structure of TEM-1. B: Profile disruption S was calculated for recombination of TEM-1 and PSE-4 using the crystal structure coordinates for TEM-1 and a window size of 14 (see methods). Residues are numbered based on the sequence of TEM-1. Vertical dashed lines represent crossover sites. (This figure is reproduced from Meyer et al. 2003 *Protein Science* **12**: 1686-1693).



To create the library of chimeric lactamases, equimolar mixtures of modules from each parent were mixed and ligated using a procedure similar to that for assembling the parental genes. *E. coli* were transformed with this library, and thousands of variants were plated on nonselective medium, i.e., LB-agar plates containing kanamycin. To determine

if the library contained any significant sequence biases, we measured the distribution of *pse-4* and *tem-1* modules in 79 randomly chosen chimeras using oligonucleotide probe hybridization (Meinhold et al. 2003). Figure II-3 shows the incorporation of the different parental sequences at seven positions throughout the genes and the frequency of crossovers between the modules probed, i.e., how often adjacent probed positions had sequence from the same parent. All chimeras exhibited a near-random crossover frequency between the probed modules (46 ± 5%), and the average frequency of observing the rarer of the two parents at each position was 40 ± 6%. Sequencing of unselected chimeras shows that up to 25% of clones may contain a single basepair deletion incorporated in the oligonucleotides used for construction. However, these deletions occur throughout both parental sequences so it is likely that the sampled portion of the library is reflective of the entire library. Assuming the (small) sequence bias arises from systematic errors in the assembly process, the average module bias can be used to calculate the probabilities of finding each chimera in our library. This type of analysis indicates that >90% of the unique chimeras occur with a probability $\geq 5.3 \times 10^{-6}$ at a confidence of 90%, and suggests that a sample of 150,000 unselected variants contains ≥65% of the unique sequences.

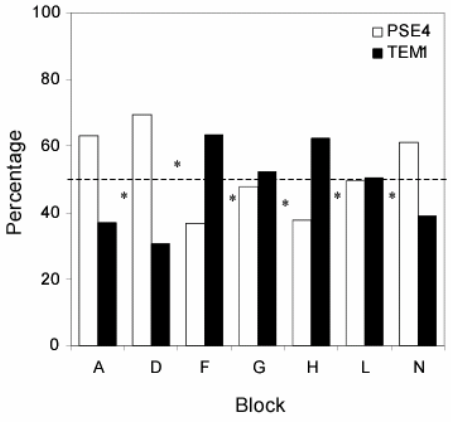


Figure II-3. Incorporation of *tem-1* and *pse-4* at different sequence positions in the unselected library. The presence of sequence from *tem-1* and *pse-4* at seven different module positions in 79 randomly picked unselected chimeras was determined using oligonucleotide probe hybridization. Asterisks represent the percentage of chimeras with crossovers occurring between adjacent probed positions. The dashed line represents the expected percentage of genes and crossovers in an unbiased library (50%).

Functional Chimera Characterization

Approximately 200,000 variants were plated on selective medium, LB-agar containing kanamycin and 20 μg/mL ampicillin. More than 100 colonies were observed, and sequencing these clones identified thirty unique functional lactamase chimeras, in addition to PSE-4 and TEM-1. Identification of the parental clones is consistent with predictions from hybridization results that suggest more than half of the chimeras were analyzed. Despite the PCR steps involved in library construction, the selected library displays a low point mutagenesis rate (0.005%). Only one chimera, the third sequence shown in Figure II-4, has amino acid substitutions. In this chimera, PSE-4 residues 265 and 266 are mutated from glutamine to histidine and threonine to serine, respectively. Examination of the TEM-1 and PSE-4 crystal structures reveal that these residues are both on the surface of the protein, and neither is in the active site (Jelsch et al. 1993; Lim et al. 2001a).

As shown in Figure II-4, the functional chimeras are highly mosaic, with 1, 2, 3, 4, 5, 6, or 7 modules swapped, and have between 7 and 67 effective mutations per chimera; the maximum possible in the library is 75. Furthermore, selected chimeras exhibited an average of 3.8 ± 2.0 crossovers, significantly lower than that expected from a random library (6.5 ± 1.8), and all chimeras have an even number of crossovers (2, 4, or 6), i.e., each functional chimera derives the A and N modules from the same parent. Modules A and N modules are derived from different parents in 41% of the clones in the unselected library.

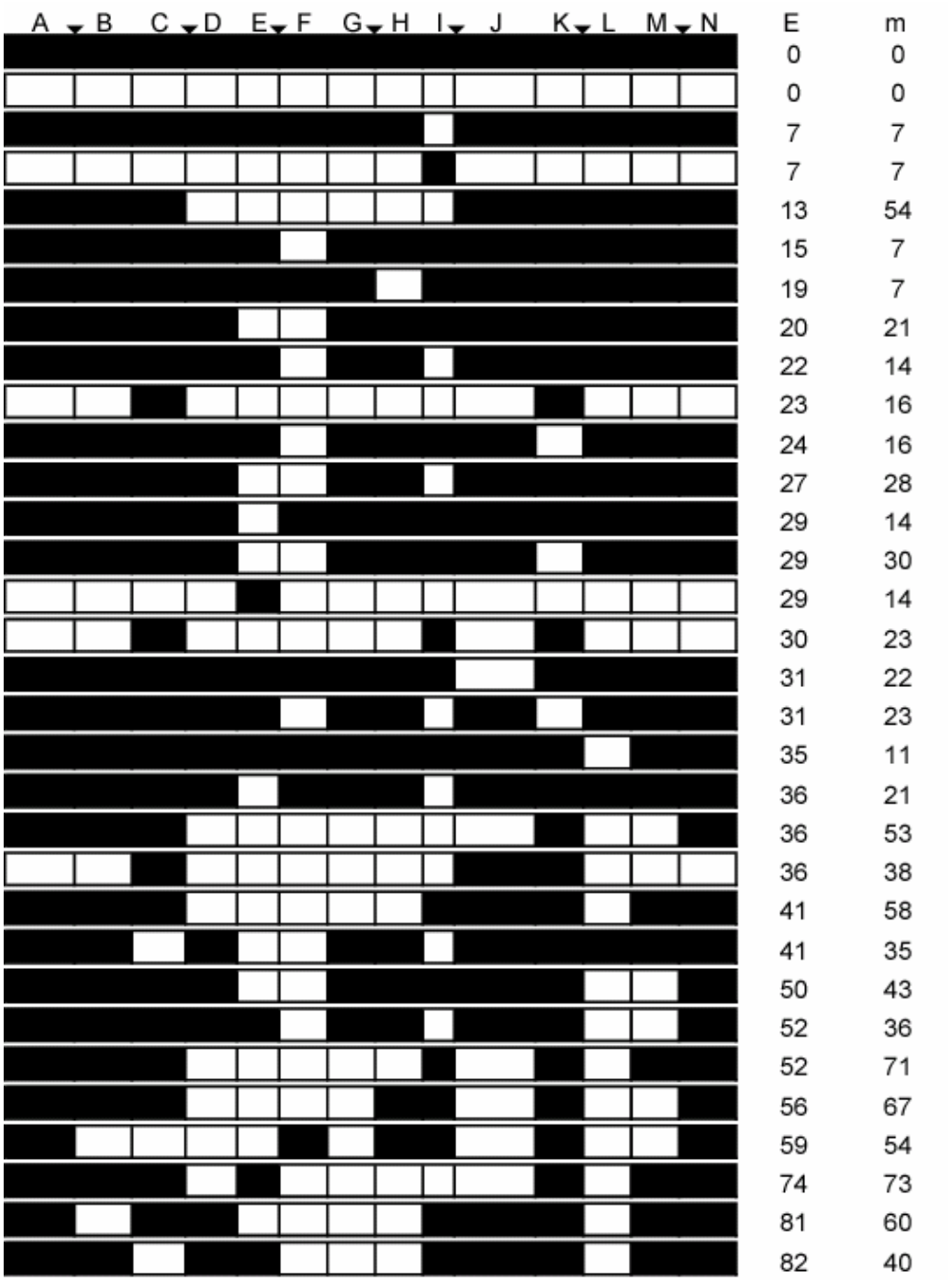


Figure II-4. Sequences, calculated disruption, and effective level of mutation of functional lactamases. Closed triangles indicate profile minima, and filled and open blocks represent TEM-1 and PSE-4 sequences, respectively. The calculated disruption *E* represents the number of interactions broken by recombination. Effective level of mutation (*m*) is the minimum number of mutations required to convert a chimera into one of its parents at only those residues recognized by SCHEMA, i.e., residues whose coordinates are defined in the TEM-1 structure (Jelsch et al. 1993).

Of the functional lactamases, only six derived both terminal fragments from TEM-1, five chimeras and one TEM-1. This indicates that chimeras which derive sequence from opposite parents at each position (chimera mirrors) are not functionally equivalent, even though SCHEMA does not distinguish them. Sequence analysis of randomly picked clones from the unselected library showed that 34% of the clones which acquire the A and N modules from the same parent contain TEM-1 at these positions. This small bias in the library does not account for the low level of TEM-1 terminal modules in functional chimeras (18%). The enrichment of one chimera from a mirror pair may arise because functional chimeras with TEM-1 terminal modules exhibit lower activity than those with PSE-4 at those positions. In fact, functional chimeras with TEM-1 terminal modules exhibit a significantly lower average MIC (250 μg/mL) than those with PSE-4 termini (1,400 μg/mL).

To determine if conservation of function corresponds to low *E*, we compared the distribution of *E* for the functional sequences with every theoretically possible unique chimera in our library. Figure II-5A shows the distributions of disruption for all chimeras in the selected and theoretical unselected libraries. The average *E* observed for functional clones (34 ± 21) is significantly lower than that calculated for the entire library (72 ± 16), indicating a strong association of low levels of disruption with maintenance of function. More than 85% of the functional chimeras have $E \leq 54$, while only 14% of the chimeras in the theoretical library fall below this threshold. We quantified the fraction of functional chimeras at each *E* in Figure II-5A by dividing the number of different functional sequences by the number of different sequences in the unselected library at

each *E* (Figure II-5B). This analysis reveals that the fraction of chimeras that retain lactamase activity decreases exponentially with increasing disruption.

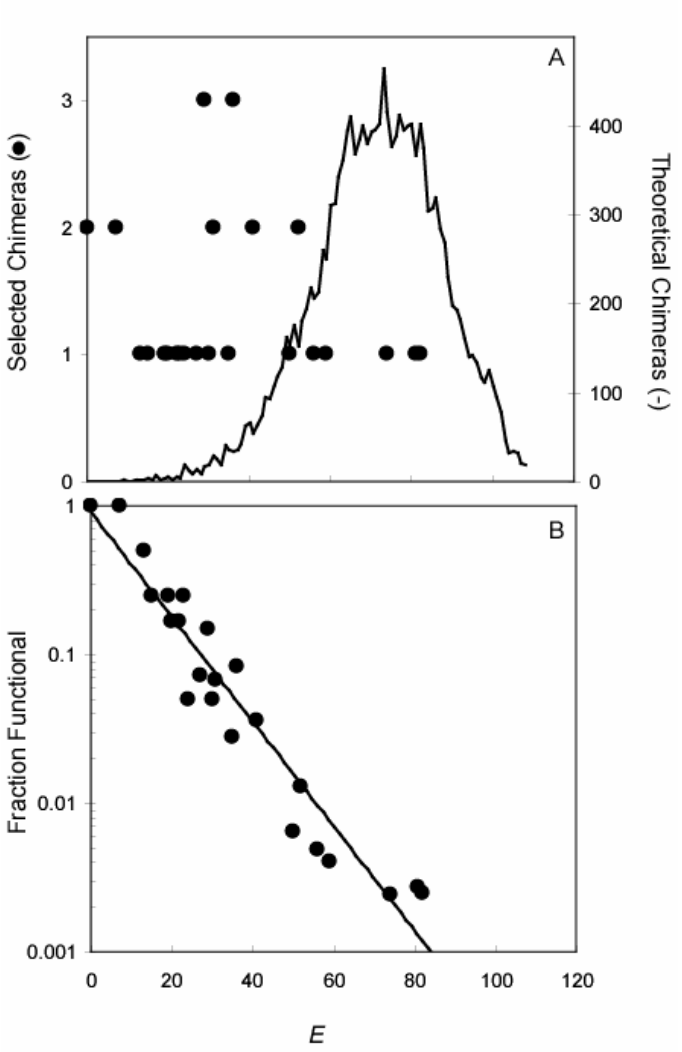


Figure II-5. Relationship between *E* and chimera function. A: The disruption distribution of all possible chimeras (solid line) is compared with those discovered in the selection for activity (●). B: The fraction of theoretical chimeras identified as functional is shown for each *E*. The data were fit to Equation (II-1) using N = 322 to obtain the probability that a disruption leads to a nonfunctional chimera, $f_d = 0.083$.

The fraction of chimeras in our library that retain function also depends on the level of mutation (Figure II-6), which raises the possibility that the low average *E* of functional chimeras could arise because low *E* corresponds to a lower average number of mutations. To investigate this, we calculated the relative difference $(E_{selected} - \langle E \rangle)/\langle E \rangle$ for each functional chimera, where $E_{selected}$ is the disruption of the functional chimera, and $\langle E \rangle$ is the average disruption of all chimeras in the theoretical library with the same

effective level of mutation (Figure II-7). The average relative difference for all functional chimeras in our library is -17.3%, suggesting that functional chimeras have lower disruption than those chosen at random with the same level of mutation. We then applied the Wilcoxon signed-rank test to evaluate the significance of these relative differences (Bernstein and Bernstein 1999). The Wilcoxon analysis yielded a ≥99% probability that the relative difference for all functional chimeras in any library is <0. Thus, chimeras that minimize *E* will have a greater likelihood of exhibiting undisturbed function than those chosen at random with the same level of mutation.

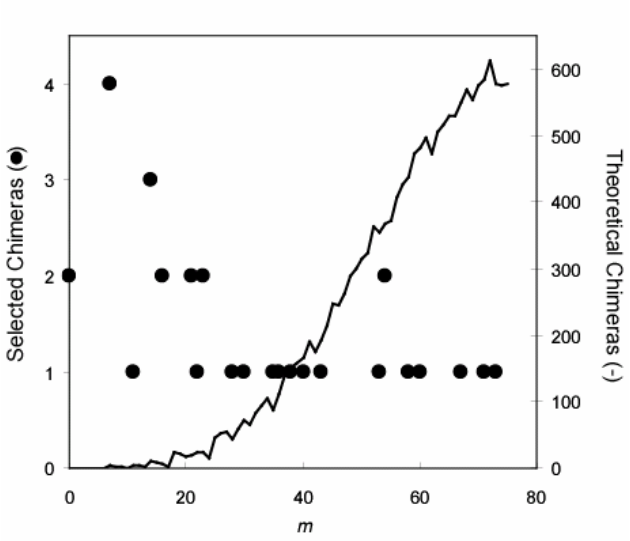


Figure II-6. Relationship between level of mutation and chimera function. The underlying distributions for the number of effective mutations (*m*) of all possible chimeras (solid line) and selected (●) chimeras are shown.

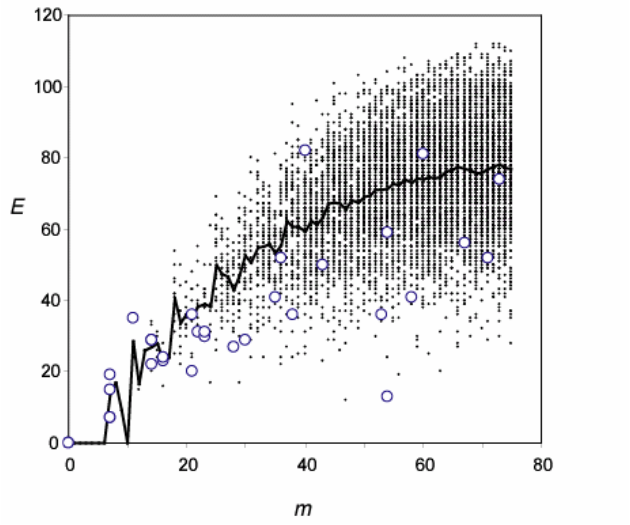


Figure II-7. *E* and *m* for all possible chimeras. At each level of mutation (*m*) where functional chimeras were obtained, the possible *E* values (•), the mean *E* for all possible chimeras (solid line), and the *E* of functional chimeras (○) are shown. Highly mutated chimeras have significantly lower disruption than the mean.

Discussion

Probabilistic Model for Chimera Function

Our results demonstrate that SCHEMA-calculated disruption (*E*) is a good metric for predicting functional conservation upon recombination. Sequence analysis of functional lactamases selected from a large library shows that chimeras with low *E* have a higher probability of retaining function than do chimeras with the same effective level of mutation but chosen at random. Our results also show that functional conservation decreases exponentially as *E* increases. This complements our previous finding, based on a small number of chimeras, that recombination disrupts protein function when it breaks many contacts in the three-dimensional structure (high *E*) (Voigt et al. 2002).

A simple probabilistic model can be invoked to anticipate the likelihood that lactamase chimeras will retain function. Assuming all contacts defined by SCHEMA are statistically independent, the fraction of possible recombinants at each *E* that retain function P_f is

$$P_f = (1 - f_d\, E/\mathrm{N})^{\mathrm{N}}, \qquad \text{(II-1)}$$

where N is the total number of interactions in the parental structures that can be disrupted upon recombination, and f_d is the probability that a disrupted contact yields a nonfunctional chimera. When N is large, as it is for proteins, this model yields a P_f that decays exponentially with *E*. Fitting Equation (II-1) to our data yields $f_d = 0.083$ (Figure II-5B). Because the experiment selected for functional chimeras, it could not uncover nonfunctional proteins that nonetheless retain proper fold. Furthermore, our use of a weak constitutive lactamase promoter to express chimeras in *E. coli* limits our ability to

identify lactamases with very low activity. Therefore, this value for f_d should be considered an upper bound on the probability that a disrupted contact yields unstructured or misfolded proteins, and the value of P_f that we calculate from f_d is therefore a conservative estimate of the probability that a protein structure will not be disrupted by recombination.

Identification of Optimal Crossover Locations

To simplify the identification of chimeras with low disruption, the SCHEMA algorithm generated a disruption profile such as shown in Figure II-2 by calculating the contribution each residue makes to the internal interactions within a fragment covered by a sliding window of a given size. We previously found that nondisruptive crossovers frequently occur in or near minima of SCHEMA profiles in chimeras with 1 or 2 crossovers, suggesting these minima may be a useful guide for generating folded and functional chimeras (Voigt et al. 2002). Interestingly, crossovers in functional lactamase chimeras from our library did not occur predominantly at these minima. Almost half of all crossovers in the functional lactamases occurred at the sites corresponding to profile maxima (Figure II-4). In addition, no functional chimeras were found with an odd number of crossovers: only 2, 4, and 6 crossovers generated functional chimeras. This crossover distribution is similar to that predicted for chimeras with a ≥10% probability of exhibiting undisturbed function ($E \leq 24$; Figure II-5B); of these chimeras, 88% have even numbers of crossovers and almost half of the crossovers (46%) occur at maxima. These findings suggest that interactions between polypeptides distal in the primary sequence, i.e., those not included in the profile calculation, should be considered when choosing

crossover locations. In other words, profile minima become a poor guide for predicting nondisruptive crossover locations when many crossovers can take place.

A better way to identify crossover points that minimize functional disruption is to determine which chimeras have the lowest *E*. But, because crossovers that do not lead to mutation will always minimize *E*, we also have to maintain a desired level of mutation. For chimeras arising from a small number of crossovers, it is easy to enumerate *E* for all possible chimera and identify crossover locations that minimize disruption. However, complete enumeration becomes impossible when multiple crossovers are allowed. For example, it is computationally intractable to calculate *E* for all possible seven crossovers between PSE-4 and TEM-1 and identify which seven-crossover library encodes chimeras with the lowest average *E* values, among libraries encoding chimeras with similar average levels of mutation. However, it is not difficult to evaluate thousands of randomly chosen seven-crossover libraries using SCHEMA to determine which ones encode chimeras with lower than average *E*. We find that this type of analysis is better than using profile minima to choose nondisruptive crossover locations for multiple-crossover libraries. For example, a PSE-4 and TEM-1 recombinant library made by allowing crossovers at the seven profile minima of Figure II-2 is predicted to encode 10 times fewer functional chimeras ($\langle E \rangle = 52 \pm 17$) than the best library found by searching 10,000 randomly generated libraries with seven crossovers ($\langle E \rangle = 33 \pm 10$), even though both libraries encode chimeras with similar levels of mutation.

Methods

Materials

E. coli XL1-Blue was from Stratagene (La Jolla, CA). Enzymes for DNA manipulations were obtained from New England Biolabs (Beverly, MA), Roche Biochemicals (Indianapolis, IN), or United States Biochemical Corp (Cleveland, OH). Synthetic oligonucleotides were obtained from Invitrogen (Carlsbad, CA). DNA purification kits were from Zymo Research (Orange, CA) and Qiagen (Valencia, CA), and other reagents were from Sigma Chemical Co (St Louis, MO) or Fisher Scientific (Pittsburgh, PA).

Calculations

For hybrids in which fragment(s) α and β are inherited from PSE-4 and TEM-1, respectively, the disruption (E) of the hybrid was calculated using Equation (II-2), where $c_{ij} = 1$ if residues are contacting (otherwise $c_{ij} = 0$), and $\Delta_{ij} = 0$ if i or j are identical in PSE-4 and TEM-1 (otherwise $\Delta_{ij} = 1$) (Voigt et al. 2002). Two residues were considered contacting if any atoms in the TEM-1 structure (1BTL) (Jelsch et al. 1993), excluding hydrogens, backbone nitrogens, and backbone oxygens, were within 4.5Å. Software to calculate the SCHEMA disruption *E* of protein chimeras is available on the web at http://www.che.caltech.edu/groups/fha.

$$E = \sum_{i \in \alpha} \sum_{j \in \beta} c_{ij} \Delta_{ij} \qquad \text{(II-2)}$$

To calculate the SCHEMA profile, a window of *w* residues was defined, and the number of intra-window interactions was counted. The profile disruption (S_i) of all residues in this window was incremented by the number of contacts within the window. The window was then slid along the protein sequence, and a profile was generated by incrementing the disruption of each residue (S_i) for all windows in which it resides. The numerical value of the SCHEMA-profile function S at residue i was defined by Equation (II-3); the magnitude of S_i corresponds with the level of predicted structural disruption for a crossover at a residue. A window of 14 residues was used to calculate the profile in Figure II-2.

$$S_i = \left(w^{-1/2}\right) \sum_{j=i-w+1}^{i} \sum_{k=j}^{j+w-2} \sum_{l=k+1}^{j+2-1} c_{kl} \Delta_{kl} \qquad \text{(II-3)}$$

Vectors

Lactamases were cloned into the vector pMon·1A2, which was created by cloning the gene encoding the heme domain of cytochrome P450 1A2 into pMon711 (Sabbagh et al. 1998). This vector was used for all selections. However, since this vector yields high background in oligonucleotide probe hybridization experiments, chimeras were cloned into pBC KS+ (Stratagene; La Jolla, CA) for these studies. *Escherichia coli* XL1-Blue transformed with these vectors were used for all analysis.

Library Construction

Twenty-eight gene modules were created to assemble the lactamase genes (fourteen for each parent). The protein modules 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 (Guex and Peitsch 1997). All modules used in assembly were double-stranded and contained unique nonpalindromic overhangs that allow for specific sequential ligation without concatamer production. Silent mutations were introduced into both genes at module boundaries (overhangs) to allow for facile assembly.

Chemically synthesized oligonucleotides used to create modules B, C, D, E, F, G, H, I, K, and M were phosphorylated using T4 polynucleotide kinase, and double-stranded modules were created from these by heating a reaction mixture containing 2.5 μM of complementary oligonucleotides, 10 mM Tris pH 8.0, 1 mM EDTA, and 50 mM NaCl at 95 °C for 2 min and subsequently cooling the reaction to room temperature at a rate of 0.1 °C per second. Modules larger than 70 basepairs (A, J, L, and N) were amplified with Vent DNA polymerase using primers containing SapI restriction sites; this allowed for rapid generation of complementary overhangs after amplification. Primers that amplified the terminal modules had a single SacI or HindIII site to allow for subsequent cloning. Amplified modules were purified by agarose gel electrophoresis, each (200 ng) was cut with 10 units of SapI at 37 °C for 24 hours, and digested modules were purified using agarose gel electrophoresis before assembly.

T4 DNA ligase was used to assemble *pse-4*, *tem-1*, and chimeric genes through a sequential process where pairs of adjacent modules were ligated, purified by agarose gel electrophoresis, and subsequently ligated to other assembled modules. Gene fragments composed of modules AB, CD, EFG, HIJ, KL, and MN were created in the first ligation reactions. For reactions in which the chimeric library was assembled, equimolar mixtures

of modules derived from each parent were used in this step. The ligated module dimers and trimers were further assembled, using the ligated fragments which had been purified with an agarose gel, to construct ABCDEFG and HIJKLMN using T4 DNA ligase. Because yields were low, ABCDEFG and HIJKLMN were amplified using Vent DNA polymerase and cleaved by SapI prior to assembly of full-length lactamases in a third ligation step; SapI created complementary overhangs at the G and H termini. Full-length constructs were treated with SacI and HindIII, purified using a Zymo DNA Clean and Concentrator Kit and ligated into pMon·1A2 and pBC KS(+), which were prepared similarly, to create the chimeric library.

Oligonucleotide Probe Hybridization

Sequences of 79 randomly selected chimeras from the unselected library in pBC KS+ were determined for 7 modules (A, D, F, G, H, L, and N) using oligonucleotide probe hybridization (Meinhold et al. 2003). The fraction of unique chimeras (*f*) found in a sample size *m* is

$$f \geq (1 - e^{m*\ln(1-v)}), \qquad \text{(II-4)}$$

where v is the probability of finding each sequence obtained from probe hybridization data.

Functional Chimera Selection

The minimal inhibitory concentration (MIC) of ampicillin for XL1-Blue *E. coli* containing pMon·1A2 is <5 µg/mL on LB-agar medium containing 10 µg/mL kanamycin. Therefore, functional selections using the pMon plasmid were performed

under conditions that gave no background, i.e., 20 μg/mL ampicillin and 10 μg/mL kanamycin. XL1-Blue were transformed with the unselected library using a heat-shock protocol recommended by the supplier, plated on selective medium, and incubated at 37 °C for 24 hours. Plasmid DNA was purified from all functional clones and digested with HindIII and SacI to confirm *pse-4* and *tem-1* length inserts (*ca*. 1 kb) were present. In addition, XL1-Blue were transformed with the purified DNA to verify the purified vectors conferred the ampicillin resistance. A majority of the clones had plasmids with an appropriate-size insert and conferred resistance in a second selection; fifty of these were sequenced.

Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is a nonparametric technique for investigating hypotheses about the median of a population (Bernstein and Bernstein 1999). While this test has less power than a *t* test for small sample sizes, i.e., is less likely to yield as dramatic a P value, we used this method because it makes no assumptions about the data being sampled from a normal distribution. To calculate the test statistic (W), we ranked the relative differences, $(E_{selected} - \langle E \rangle)/\langle E \rangle$, at each level of mutation according to their absolute magnitude and summed the rank scores according to the sign of the relative difference. This yielded W+ and W- values of 104 and –361, respectively.

Chapter III: Design of Site-Directed Recombination Libraries

Introduction

One of the strengths of directed evolution is that very little information is required for success. The less information incorporated into the experimental design, the less concern whether this information is useful or not (Arnold 1998). However, one of the biggest challenges in directed evolution remains that a limited number of variants can be screened, and a variant of interest must be within this population (Voigt et al. 2001a). As directed evolution has matured and is tasked with more challenging problems, the variants that solve these problems often contain more than a few mutations. Libraries containing more highly mutated variants were more effective for adaptive enzyme evolution in several different studies (Crameri et al. 1998; Zaccolo and Gherardi 1999; Daugherty et al. 2000). However, because most mutations are neutral or deleterious to protein structure the fraction of folded variants in a population decreases exponentially as additional random mutations are introduced (Bloom et al. 2005b). Thus additional diversity comes at the cost of a much lower fraction of folded variants if mutations are made randomly.

In order to overcome this problem many strategies have been developed to bias the variants used for directed evolution toward regions of sequence space that are more likely to contain the variant of interest (Patrick and Firth 2005). Such strategies include both intensively mutating specific sites identified from structural studies, as well as trying to limit mutations introduced to those that are less likely to disrupt the folded protein structure (Voigt et al. 2001b). Some of these strategies, or library designs, have proven successful, and there is a continuous push to increase the number of mutations that can be

incorporated while maintaining a high fraction of folded variants in a library. Homologous recombination of very distantly or even nonrelated proteins is one strategy that reaches toward this goal. However, as the sequence identity between the recombined proteins decreases, the fraction of folded variants also decreases (Ostermeier et al. 1999b; Sieber et al. 2001).

We have developed a metric, SCHEMA disruption, which can evaluate chimeric proteins *in silico* before they are constructed in the laboratory, allowing structure and sequence information to be incorporated into a library design (Voigt et al. 2002). We have shown that SCHEMA disruption (*E*) is a good metric for determining whether a chimeric protein will retain its fold and function (Meyer et al. 2003). However, how exactly this understanding translates into a library of proteins that is both diverse and contains a high fraction of folded variants is still unclear. Mutation and disruption are correlated; the more mutations a chimera contains, the higher its *E* is likely to be. Balancing these two parameters and finding a good trade-off between them is critical to designing a library that meets the desired goals.

Current construction methodology limits the libraries that can be created to combinatorial libraries with a fixed set of recombination sites or crossovers. This restriction makes the task of library design more manageable because it limits the search space. However, there are still a very large number of libraries that can be constructed. For a 300 amino acid protein with seven possible recombination sites, there are 6×10^{19} possible libraries. Numerically evaluating all of them is unfeasible even if the search space is decreased by placing restrictions on the size of the sequence blocks between recombination sites. One solution to this problem is to find the global optimum without

exhaustive enumeration. “Recombination as a Shortest Path Problem” (RASPP) is an optimization function that identifies libraries at the diversity/<*E*> trade-off curve (Endelman et al. 2004). Using an optimization function limits the design options slightly but may confer a large advantage compared to randomly enumerating many libraries and picking the best one.

This chapter addresses different strategies for designing recombination libraries between distantly related β-lactamases. We ask the following questions in the course of designing two libraries using different strategies: (1) What measures should be used to evaluate libraries of chimeras to identify those with the desired features? (2) What are ways to balance the fraction of folded variants with diversity? (3) How well does RASPP identify libraries that meet the stated goals of a high level of diversity and a large fraction of folded variants?

Methods for Library Design

There are essentially two ways to identify the crossover locations that lead to a good recombination library. The first is to randomly enumerate a large number of libraries, evaluate them based on some parameters, and choose the library that performs the best. This process is computationally intensive, and there is no guarantee that the best library identified by random enumeration will be anything close to the best library that could be made. However, random enumeration has the advantage that any calculable parameter can be used to evaluate the libraries, and very specific requirements can easily be incorporated into the design.

The second method of identifying a good recombination library is to use an optimization function. An algorithm called “Recombination as a Shortest Path Problem” (Endelman et al. 2004) was developed specifically to generate a list of libraries at the optimum diversity/fraction folded trade-off. RASPP identifies these libraries by determining which library has the lowest <*E*> subject to constraints on the minimum length of the sequence blocks. By iterating over a series of different length constraints optimal libraries are generated at varying levels of diversity. To make comparison of the libraries more intuitively understandable and to remove redundancies, the libraries are binned by <*m*>, and the library with the lowest <*E*> is reported. An example of this “RASPP curve” for libraries made with three β-lactamases is shown in Figure III-1. There are often levels of <*m*> for which no library is identified, resulting in some gaps in the curve. This occurs because block minimum length is used as a measure for diversity. There are regions in the space of all possible *X*-crossover libraries, where *X* is the desired number of crossovers, where *E* and *m* are not well correlated and libraries with higher *m* have lower *E*. These regions of *m* are skipped by the RASPP curve. The *m* bin sizes and the number of recombination sites are both user-adjustable parameters. RASPP is much faster computationally than random enumeration, and the best libraries are guaranteed to be identified. However, RASPP is limited because it uses specific parameters (discussed below) to evaluate libraries in identifying the trade-off curve. The parameters may or may not accurately reflect the desired properties of the library.

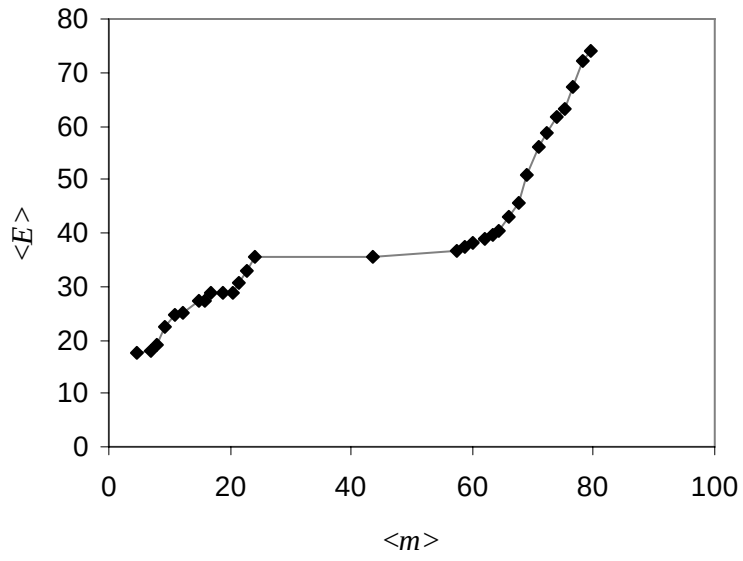


Figure III-1. A RASPP curve for nine crossover libraries made by recombining PSE-4, TEM-1 and SED-1. The minimum length constraint L=1, and the bin size set to 1 *m.*

Parameters for Evaluating Library Fraction Folded

In order to construct a library which meets the goals of having both a high fraction of folded chimeras and chimeras that are diverse, there must be some criterion that can be easily calculated to evaluate the library. The simplest surrogate of the fraction of folded variants is the average disruption <*E*> of chimeras in the library. It is easy to calculate and gives a general idea of the library properties. RASPP utilizes <*E*> as its parameter for fraction folded. Yet, it is not known how effective this metric is for determining the lactamase library with the most folded variants, given that the relationship between *E* and the probability that a chimera will fold, P_f, is nonlinear (Voigt et al. 2002; Meyer et al. 2003; Otey et al. 2004). The fraction of folded variants F_{folded} can be evaluated using Equation (III-1) directly if P_f is known.

$$F_{folded} = \frac{\sum_i P_{f\,i}}{N}, \qquad \text{(III-1)}$$

where the probability of folding $P_{f\,i}$ is determined for each chimera *i*, summed over all the chimeras in the library, and divided by the total number of chimeras in the library, N. However P_f is not usually known *a priori* and has varied considerably between different experiments, not just in value but also in form (exponential vs. sigmoidal).

To address whether the lowest <*E*> is a good surrogate for identifying a library with the highest F_{folded} and how library ranking by F_{folded} changes with variation in P_f, the F_{folded} of RASPP lactamase libraries (304 nonredundant before binning by <*m*>, see methods) was calculated using both the exponential function described in Chapter II and a sigmoid function that reflects results obtained for the lactamases by Voigt et al. (2002) and cytochromes P450 (Otey et al. 2004). To compare how the libraries would be perceived by the library designer, they were ranked with respect to F_{folded} calculated with the two different forms of P_f. Ranking the libraries is more relevant to the situation faced by the library designer than examination of values directly. There is a strong correlation (R^2=0.9936) between libraries ranked with F_{folded} calculated using an exponential P_f and libraries ranked with F_{folded} calculated using a sigmoidal P_f (Figure III-2) . This suggests that potential variability of P_f is not likely to change the rank ordering of the libraries. The F_{folded} values may differ greatly, but the best library is probably the same using either function. Furthermore, rank ordering the libraries with respect to their <*E*> shows strong correlation with respect to F_{folded} calculated using either form of P_f (R^2=0.9485 or 0.9149). This indicates that low <*E*> is a good surrogate for identifying libraries with a high fraction of folded chimeras. It may not give the same easily interpretable information as F_{folded}, but rank ordering libraries by <*E*> is effective for a range of different P_f behaviors.

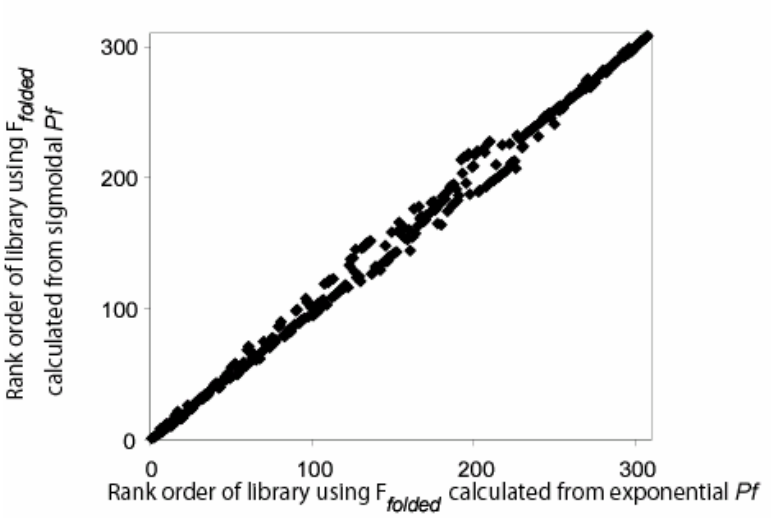


Figure III-2. The rank order of test libraries by F_{folded} using two different P_f functions (see methods) is strongly correlated indicating that rank ordering by F_{folded} is not sensitive to P_f.

Parameters for Evaluating Library Diversity

There are many ways to measure library diversity. The most intuitive measure is <*m*>, the average number of amino acid substitutions from a chimera to its closest parent. However, this measure is not necessarily the most useful when trying to set parameters on library design. To ensure a library with a certain level of diversity, putting length constraints on the blocks is far easier. Placing constraints on the minimum block size reduces the search space for random enumeration, and RASPP uses a minimum length constraint to identify libraries with the lowest <*E*> at a range of different diversity levels.

The combination of <*E*> and <*m*> does not necessarily describe whether a library meets the stated goals. It is possible to design a library that has both low *E* chimeras and high *m* chimeras, but these populations may have little overlap. The chimeras of greatest interest are the low *E*, high *m* chimeras and it is necessary to ensure that they exist within a library. One way to do this is to calculate the *m* for chimeras below a certain threshold of *E*. However, this reflects only a sigmoidal shape for P_f and not an exponential one.

A score for diversity that penalizes chimeras that are unlikely to fold will more accurately reflect the desired result. While *E* penalizes unfolded chimeras, *E* is not a measure of diversity and is usually negatively correlated with diversity. However, *E* can be incorporated into a measure of diversity. The m_{folded}, or mutation of the fraction folded, is a measure that weights a chimera's probability of folding with respect to *E*, P_f, into the <*m*> calculation.

$$m_{folded} = \frac{\sum_i P_{f\,i} m_i}{\sum_i P_{f\,i}}, \qquad \text{(III-2)}$$

where m_i is the number of mutations from chimera *i* to its closest parent. However, this measure also requires $P_{f,}$ which may not be known. To examine how library choice would be affected by variation in P_f, m_{folded} was calculated for the test libraries described above. While the rank ordering of libraries with respect to F_{folded} is not sensitive to P_f, rank ordering of libraries with respect to m_{folded} is strongly affected by P_f. Rank ordering of the test libraries by m_{folded} calculated using the an exponential and a sigmoidal P_f show correlation (R^2=0.4698). This much weaker correlation indicates that variation in P_f has a significant effect on the ranking of the libraries, making m_{folded} a less useful metric (Figure III-3). Due to the dependency on P_f, m_{folded} is not a function that is always readily applicable to library design, despite its advantages over <*m*> in understanding the balance between diversity and fraction folded. Given these issues, nothing replaces examination of an *E* vs. *m* plot for a given library. Examining such a plot easily identifies libraries with desirable or undesirable properties. However, such examination is qualitative in nature and does not provide a quantitative measure than can be used to rank libraries so that many can be compared and evaluated quickly.

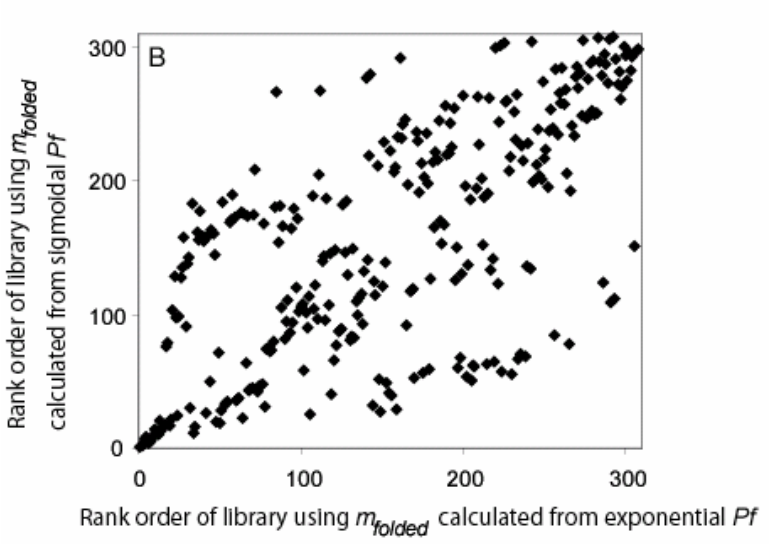


Figure III-3. The rank order of test libraries by m_{folded} using an exponential and a sigmoidal P_f function (see methods) shows that m_{folded} is very sensitive to P_f. This indicates that m_{folded} is not likely a good measure to use for evaluating libraries if P_f is not known.

Balancing Diversity and Fraction Folded

One of the challenges with evaluating diversity in recombination libraries is that diversity and fraction folded are inversely related. In situations where P_f is known, maximizing the product of F_{folded} and m_{folded} can be an adequate metric for choosing a library. When the relationship between E and probability of chimera folding P_f is unknown, recognizing the best trade-off between diversity and E is difficult. One way to evaluate a library is to determine the average number of mutations per disruption or $\langle m/E \rangle$. This measure effectively identifies libraries with the most mutations per disrupted contact. Examining plots of $\langle m/E \rangle$ vs. $\langle m \rangle$ and $\langle E \rangle$ vs. $\langle m \rangle$ for the test libraries described above shows that $\langle m/E \rangle$ reaches a maximum that roughly corresponds to the plateau region of the RASPP curve. The libraries that score best with this measure balance low E with high m (Figure III-4).

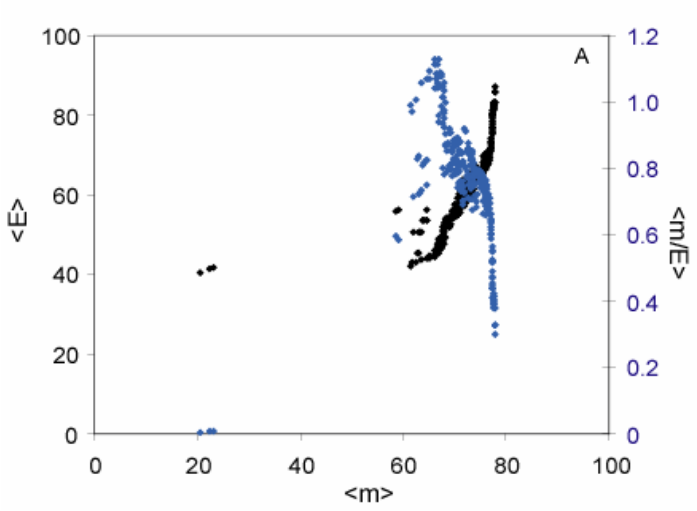


Figure III-4. Comparing <*E*> vs. <*m*> (black points) and <*m/E*> vs. <*m*> (blue points) shows that for RASPP test libraries there is a maximum <*m/E*> that corresponds to a plateau in the <*E*> vs. <*m*> curve.

Given that <*m*> may not necessarily be the only, or best, metric for diversity in a library, it becomes more difficult to justify removing >90% of the libraries from consideration during RASPP's binning by <*m*>. This binning removes libraries that may be more desirable based on some other metric. However, one of the features of RASPP is that it provides a tractable number of distinct choices. Without this binning, there are too many libraries to effectively examine. It is likely worthwhile to calculate <*m/E*> of all libraries near the region of interest, or to use <*m/E*> to identify which regions should be of interest on the RASPP curve.

Diversity Among Chimeras

All of the metrics discussed measure the diversity of the library based on the sequence distance of the chimeras from the starting proteins. We have not developed an effective diversity measure that compares how different the chimeras are from one another. If all the chimeras are distinct from the parents (the lowest *m* in the library is relatively high), then the resulting population of chimeras will tend to be very different from one another as well. This occurs because the smallest possible difference between

individual chimeras corresponds to the smallest possible difference between a chimera and the parents. However, a library that looks diverse using the measures discussed above may still have a small population of chimeras with very few mutations, due to a single sequence block that contains few amino acid changes. While the chimeras with very low *m* are treated appropriately by all the measures of diversity discussed above, the clusters of chimeric sequences that are also separated by only a few mutations are not handled in any special way. It is not obvious how best to quantitatively measure this effect so that it can be taken into account during library design. The best approach is likely to enforce length constraints on the sequence blocks so that very small blocks with few mutations are not used to construct the library.

Choosing Parental Sequences

An essential component in the design of a recombination library is the choice of parental starting sequences. The divergence of the parental sequences dramatically affects the fraction of folded chimeras as well as the diversity of the chimeras. In this work we are striving to push the boundary of effective homologous recombination to sequences that share little identity. Because of this, six trial sequences ranging from 25% to 45% identity to both TEM-1 and PSE-4 were examined: AST-1, CFX-A2, FAR-1, KLUC-1, SED-1, and VHW-1 (Laurent et al. 1999; Teo et al. 2000; Decousser et al. 2001; Madinier et al. 2001; Petrella et al. 2001; Poirel et al. 2001). To identify the parents that introduce the most diversity, but yield the lowest *E* in chimeras formed when recombined with PSE-4 and TEM-1, we randomly enumerated 500 three-parent libraries and examined the <*E*> of the libraries produced. Another way to evaluate potential parents is

to generate RASPP curves using different parent sets and examine which parents produce the best trade-off curve. The two sequences introducing the least calculated structural disruption were SED-1 and AST-1 (Petrella et al. 2001; Poirel et al. 2001). AST-1 is an inhibitor-resistant lactamase isolated from *Nocardia asteroides*, and SED-1 is a lactamase displaying CTX-M type extended spectrum activity isolated from *Citrobacter sedlakii* that hydrolyzes atreonam and first-generation cephalosporins. Structural information is available for neither of these proteins. However, because all class A lactamases share high structural identity (Figure II-1) and there are no significant gaps within the sequence alignment, it is likely that they are similar in structure to TEM-1 and PSE-4. SED-1 and AST-1 introduce the least disruption when recombined with PSE-4 and TEM-1 because the sequence identity between the sequences chosen occurs at positions that are more likely to have large numbers of contacts compared to the other sequences tested.

Library Design by Random Enumeration

In previous studies we have observed that the N- and C-termini of functional β-lactamase chimeras nearly always (>95%) originate from the same parent (Hiraga and Arnold 2003; Meyer et al. 2003). Additionally, recombining the termini introduces a great deal of disruption (~30 *E*). We designed a library by evaluating the properties of many random libraries to meet the specific requirement that the N- and C-terminal blocks always originate from the same parent. Potential crossover sites were chosen by random number generation with the minimum block size constrained to 15 amino acids. The *E* and *m* of all possible chimeras in a library resulting from each set of crossover points were calculated and then the metrics discussed above determined. To enforce the

constraint that the N- and C-termini originate from the same parent, only chimeras with this property were included in the calculations. RASPP cannot be used to restrain noncontiguous portions of the sequence to the same parent because there is no mechanism to implement this constraint (Endelman et al. 2004). However, with random enumeration this specification is easy to implement. This library was intended to be large (4^9=262,144 members), with four parents (TEM-1, PSE-4, AST-1, and SED-1) and 9 exchangeable sequence blocks (counting the N- and C-termini as a single block).

To choose the recombination sites, approximately 3,000 randomly generated libraries with 9 recombination sites were evaluated using three of the four parents to minimize computation time. Because previously obtained data allowed calculation of a P_f (Meyer et al. 2003), the libraries were evaluated based on the F_{folded} and m_{folded}. The best 22 libraries were ranked by the product of F_{folded} and m_{folded}, and the recombination sites were shifted to make them experimentally feasible (2-3 bp identity at each recombination site) (Figure III-5). Only a few recombination sites appear to be used in several of the libraries, and all of the libraries are have fairly-well spaced blocks due to the stipulated minimum fragment size (15 residues). The libraries were evaluated using all four parents, and the best of those libraries (determined by the maximum product of F_{folded} and m_{folded}) was selected.

The library chosen for construction (RandE:APST, for random enumeration, with parents AST-1, PSE-4, SED-1 and TEM-1) has the following independently exchangeable blocks of sequence (Ambler standard numbering (Ambler et al. 1991)) 66-80, 81-100, 101-116, 117-138, 139-155, 156-175, 176-195, 196-210, and the N- and C-termini (beginning-65 and 210-end). The library's characteristics, as measured by the

parameters discussed above, can be found in Table III-1. It is has lower <*E*> than the other libraries examined, but sacrifices some diversity. The <*m*> is also lower than the other libraries. The largest block consists of N- and C-termini and accounts for all of the α/β domain. The ω-loop (residues 160-181), a motif important for substrate binding and specificity (Petrosino and Palzkill 1996; Therrien et al. 1998; Sanschagrin et al. 2000), is split over two blocks (Figure III-6A) and five blocks contain residues with 5 Å of a bound inhibitor.

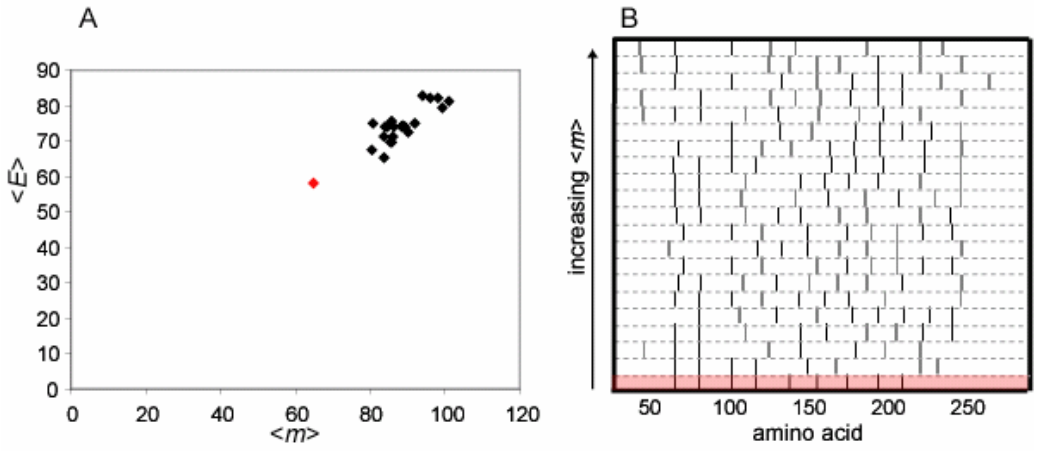


Figure III-5. <*E*> vs. <*m*> for the top 22 of 3,000 randomly enumerated libraries. The library shown in red was chosen for construction. B: The crossover locations for all the libraries shown to the right; the highlighted library was constructed. For all libraries the N- and C-termini together are considered a single block.

During construction, one of the parental sequences, AST-1, proved problematic. AST-1 was originally cloned with a GTG start codon (Poirel et al. 2001). Once the clone was obtained and placed into the expression system used for this work, lactamase activity was much lower than that of the other three parents using either ATG or GTG start codons. Additionally, the PCR conditions necessary to amplify AST-1 were significantly different from those of the other three parents. The AST-1 gene is 71% GC and required extreme PCR conditions for successful amplification.

Due to the problems encountered with AST-1, it was dropped from the library and the three-parent library without AST-1 was constructed (RandE:PST, random enumeration with parents PSE-4, SED-1 and AST-1). The three-parent library created without AST-1 is different than the one originally designed. This library is not specifically designed to be optimal because it is missing one of the original parent sequences. It is significantly smaller (3^9=19,683 vs. 4^9=262,144) and less diverse (<*m*> 52 vs. <*m*>= 59, Table III-1) than the designed library (RandE:APST). However, the <*E*> is lower, resulting in a higher F_{folded} than the larger library. This occurs because fewer chimeras with very deleterious combinations of blocks are created. However, the lower *E* comes at the cost of diversity as noted above. The biggest price to dropping AST-1 to make the RandE:PST library is the number of potential chimeras created. The trade-off between diversity and folding is about the same for both libraries (<*m/E*> remains about the same).

Table III-1. Characteristics of the Libraries Constructed

	RandE:APST	RandE:PST	RASPP:PST
	Random Enumeration		(RASPP)
<*E*>	59 ± 12	52 ± 12	45 ± 15
<*m*>	60 ± 13	53 ± 14	66 ± 21
F_{folded}	1.9%	2.8%	6.3%
m_{folded}	52	46	53
<*m/E*>	1.04	1.03	1.58

RandE:APST was designed to incorporate 9 blocks with parents AST-1, SED-1, PSE-4 and TEM-1 using random enumeration. RandE:PST has the same recombination sites as RandE:APST but considers only those chimeras that do not inherit any blocks from AST-1 and was only constructed because of problems with AST-1 after the design process was complete. RASPP:PST was designed to incorporate 8 blocks using RASPP with PSE-4, TEM-1 and SED-1. All the parameters listed are directly comparable and were calculated with the following assumptions where necessary: $P_f = (1- (f_d E/n))^n$, where n is total number of contacts (322), and f_d=0.075 (Meyer et al. 2003).

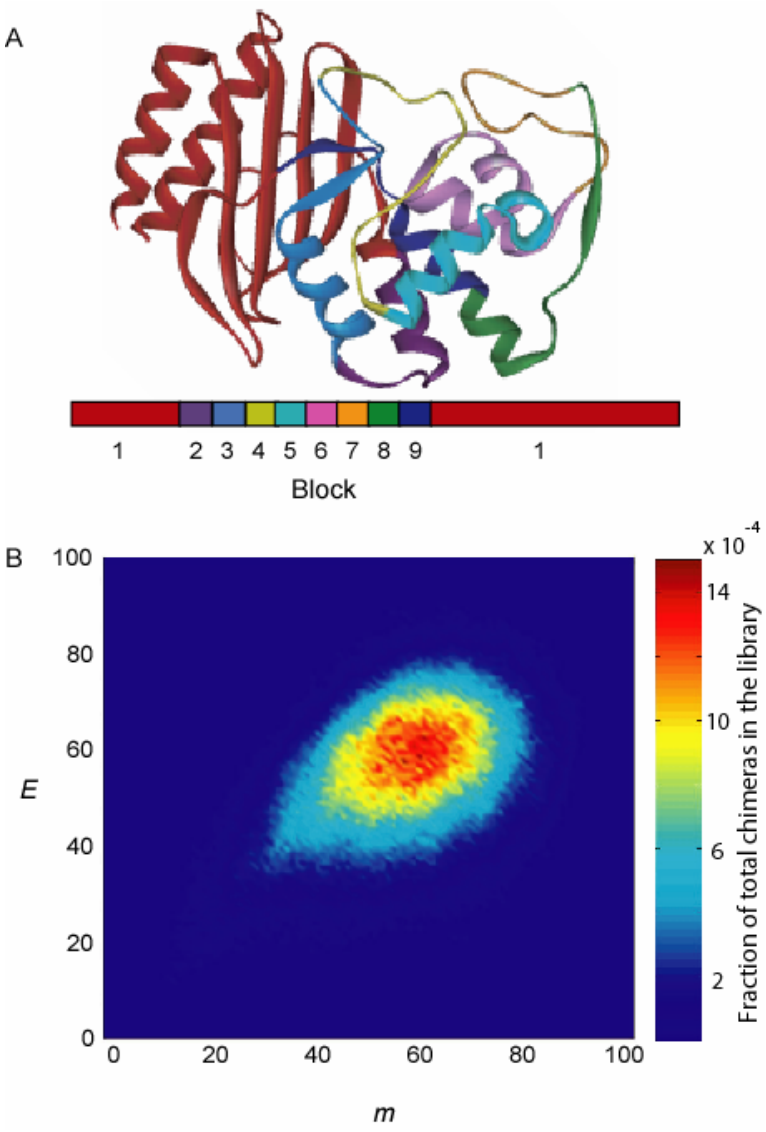


Figure III-6. An overview of the RandE:APST library. A: The differently colored sequence blocks are mapped to the structure of TEM-1. B: The *E* vs. *m* density plot of chimeras in the library shows a single large peak in the population with a slight tail toward low *E* and *m*.

RASPP Library Design

In addition to the library described above, a second library was designed using RASPP. RASPP identifies libraries at the optimal diversity/fraction folded trade-off and a library designed with RASPP is likely better than a library designed with random enumeration, even if the N- and C-termini cannot be constrained. The library is designed to be smaller, containing only three parents (TEM-1, PSE-4 and SED-1). The N- and C-termini cannot be fixed to the same parent using RASPP, but the globally optimal libraries are identified rapidly. To examine whether the trade-off between fraction folded

and diversity was altered by changing the number of crossovers, RASPP was run stipulating 7, 8 or 9 crossovers (Figure III-7). The curves directly overlay, indicating that additional crossovers do not produce a significant gain in fraction folded at the same level of diversity. All libraries represented on these curves have significantly lower disruption at similar levels of mutation than the RandE:APST library designed by random enumeration.

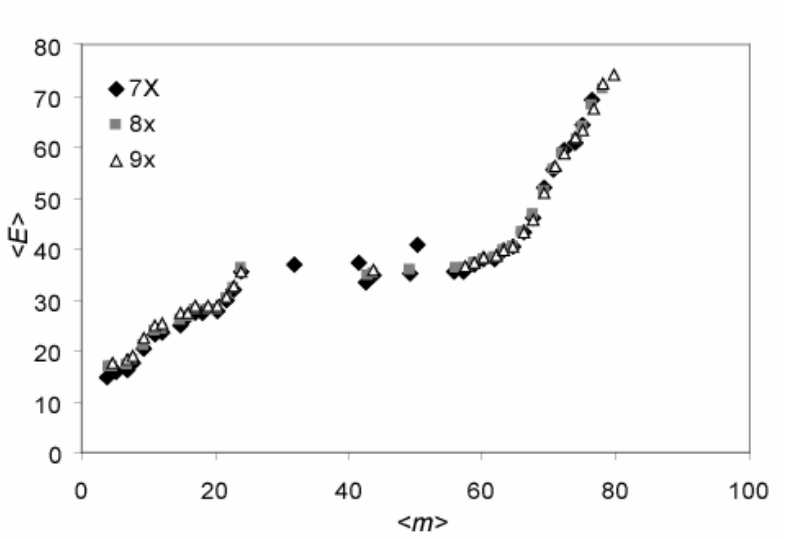


Figure III-7. RASPP curves generated for TEM-1, PSE-4, and SED-1 using 7, 8, or 9 crossovers show that there is no gain in fraction folded at a given level of diversity associated with 8 or 9 crossovers vs. 7.

So that a significant proportion of the library could be characterized, we chose to maintain a relatively small library size and construct it with 7 crossovers (8 blocks). The libraries RASPP identified fall into three general groups (Figure III-8A). The first group of libraries has relatively low <*E*> and low <*m*>. The crossovers predominantly occur at the termini of the protein sequence, producing chimeras with one very large piece and many small chips at termini (gray in Figure III-8B). Most of these chimeras are not significantly different from the three parents or from one another. The next group of libraries has slightly higher <*E*> than the first group, but <*m*> is significantly higher, making these libraries attractive choices for construction (red or blue in Figure III-8).

The crossovers occur throughout the protein; however, the blocks produced are somewhat uneven in size. The third group of libraries has increasingly high <*E*> and <*m*> (green in Figure III-8), and the crossovers are progressively more spread out over the protein sequence, generating blocks that are all approximately the same size. Due to the high <*E*> of these libraries, most chimeras are probably not folded.

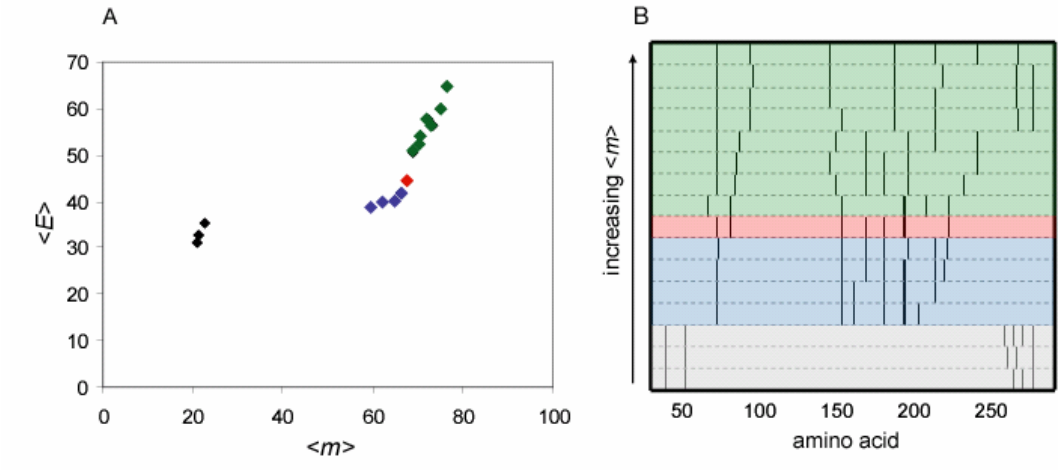


Figure III-8. A: The <*E*> vs. <*m*> RASPP curve generated for TEM-1, PSE-4 and SED-1 using seven crossovers. The libraries break into three regions that are colored black, blue and green. The red point represents the library chosen for construction. B: The crossover locations for the RASPP libraries shown in A. The coloring matches the plot and highlights libraries with similar characteristics. The red library was chosen for construction (RASPP:PST).

The second group of libraries (red or blue in Figure III-8) with midrange <*m*> and <*E*> was further inspected because these libraries are in the plateau region of the curve <E> vs. <m> curve (i.e., increase <*m*> with little cost to <*E*>) and have significantly higher <*m/E*> than the other two groups. From this group the library, RASPP:PST (RASSP designed with parents PSE-4, SED-1 and TEM-1) with the following blocks was chosen for construction (Ambler standard numbering (Ambler et al. 1991)): 1-65, 66-73, 74-149, 150-161, 162-176, 177-190, 191-218, 219-290 (Figure III-9A). Two of the recombination sites were shifted by 1 or 2 amino acids from the recombination sites

generated by RASPP to accommodate the 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 balances high <*m*> (66 ± 21) for a diverse population with low <*E*> (45 ± 15) (Table III-1) to ensure a large proportion of folded chimeras. The <*m/E*>=1.58. The average <*m/E*> for libraries in this region (<*m*> between 60 and 70) is 0.92 ± 0.13. Unlike the larger library where all the chimeras were focused into a relatively small area of the *E* vs. *m* graph, the chimeras in this library are diffusely spread over a large region and the distribution of chimeras is bimodal in both dimensions (Figure III-9).

This library was chosen because the active site Ser70 and Lys73 (Block 2) are divided from the large internal block (block 3), which comprises nearly 25% of the protein (Figure III-9). This separates the active site from the largest single block, allowing them to be inherited from different parents so that properties of the protein that are potentially specific to the active site can be inherited independently from bulk of the protein. The ω-loop is split between blocks 5 and 6. The library also has crossovers that are pushed toward the C-terminus, reducing the size of block 8. Blocks 1 and 8 together comprise almost half the protein, consisting of the N- and C-terminal helices and the entire β-sheet beneath them (Lim et al. 2001b). The last crossover at 218 is very close to position 216 chosen for the new N- and C-termini of a circularly permutated TEM-1 (Osuna et al. 2002). This indicates that this particular crossover location is likely a good place to divide the protein with minimal impact on folding.

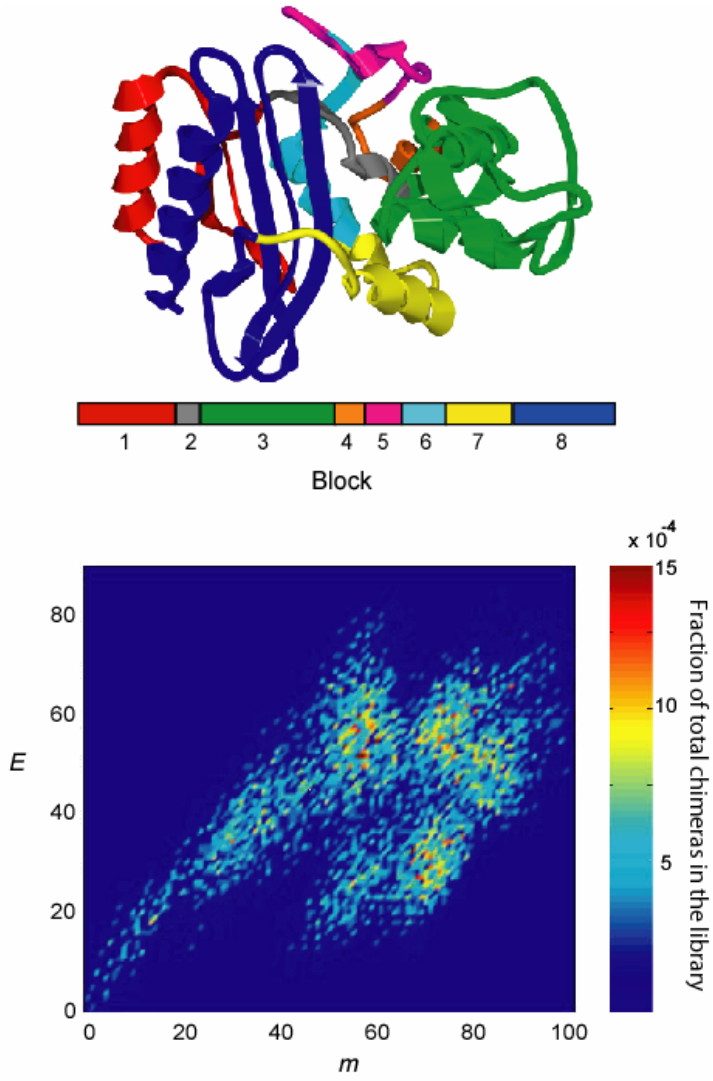


Figure III-9. RASPP:PST library chosen for construction. A: The differently colored sequence blocks are mapped to the structure of TEM-1. B: The *E* vs. *m* plot of chimeras in the library shows a relatively diffuse population which has bimodal properties in both the *E* and the *m* dimensions.

Conclusions

The two library designs described here are not directly comparable. The libraries were designed for different purposes, are different sizes, and have different input parameters. Furthermore during design, they were evaluated and chosen based on different metrics. The library generated by RASPP has lower $\langle E \rangle$, a higher F_{folded}, and a higher $\langle m \rangle$ than the library identified using random enumeration. The RASPP library is better using all of measures of library fitness (Table III-1). However, it is important to

remember that the random enumeration was restricted to libraries with blocks containing at least 15 amino acids. Three of the blocks in the RASPP library (RASPP:PST) are smaller than 15 amino acids, one of them significantly so (8 amino acids). If the 15 amino acid limitation were relaxed, the libraries produced by random enumeration might be as good as or better than the RASPP library (because the N- and C-termini are always retained from the same parent). RASPP does a much better job of identifying a range of libraries from which to choose than random enumeration. With random enumeration, finding one good library is an achievement. Identifying more than one, so that there are many good choices, is much more difficult. RASPP effectively identifies many good libraries with a range of different properties.

While RASPP has its limitations, it is a very effective tool for library design. It may not allow nonconsecutive portions of sequence to be fixed to the same parent, but by identifying libraries at the global minimum RASPP may be able to compensate for the disruption caused by not allowing such noncontiguous blocks. The $\langle E \rangle$, which RASPP uses as its minimization criterion is a good surrogate for the fraction of folded variants. The binning of RASPP libraries by $\langle m \rangle$ may not be the best practice because some libraries with better characteristics are eliminated. However, this is easy to circumvent if desired by setting the bin size to 0. Finding the right balance between fraction folded and diversity will always be a challenge, but RASPP identifies libraries that are on the diversity/fraction folded optimum trade-off curve to give a choice of libraries that have different properties along this curve.

Methods

E Calculations

To obtain a sequence alignment for computing the SCHEMA disruption the sequences of TEM-1, SED-1, and PSE-4 were aligned using CLUSTALW (Chenna et al. 2003). This alignment has shows no differences from a structural alignment between TEM-1 (1BT5) (Maveyraud et al. 1998) and PSE-4 (1G68) (Lim et al. 2001b) generated in Swiss-pdb viewer (Guex and Peitsch 1997). The structure of PSE-4 was used to calculate the contact map necessary for computing SCHEMA disruption; using the TEM-1 structure causes only slight changes. The SCHEMA disruption (E) is

$$E = \sum_{i} \sum_{j>i} C_{ij} \Delta_{ij} \,, \qquad \text{(III-3)}$$

where C_{ij} =1 if any side-chain heavy atoms or main-chain carbons in residues i and j are within 4.5 Å. The Δ_{ij} function is based on the sequences of the parental proteins. $\Delta_{ij} = 0$ if amino acids i and j in the chimera are found together at the same positions in any parental protein sequence, otherwise $\Delta_{ij} = 1$. Python scripts for calculating E are available on the Arnold lab website http://www.che.caltech.edu/groups/fha/.

Testing Library Scoring Parameters

The test libraries scored using different measures of fitness were generated by running RASPP to create a three-parent, seven-crossover library using the structure of PSE-4 (1G68). The lactamase parents were PSE-4, TEM-1 and SED-1 and the minimum block size L was 5 amino acids. The <m> bin size was set to 0 to ensure that all nonredundant libraries were reported. Two separate P_f functions were used to calculate F_{folded} and m_{folded} for each library. The first is the exponential decline described for

lactamases by Meyer et al. (2003) of the form $(1- (f_dE/n))^n$, where n is total number of contacts (322), and f_d=0.075. The second is a sigmoid function of the form $1/(c+e^{bE+a})$, where a=-3.6, b=-0.12 and c=1.0. This function was derived from an analysis of lactamase data (Chapter VII), but reflects sigmoidal characteristics of other data as well (Voigt et al. 2002; Otey et al. 2004). C++ code to perform this analysis can be found in Appendix I. To calculate the F_{folded} and m_{folded} for designed libraries, f_d=0.075.

Random Enumeration

Lists of 9 crossovers were generated by picking random numbers. The minimum block size was set to 15 amino acids to prevent the creation and analysis of libraries containing trivial changes. $<E>$,$< m>$, and the other library parameters described were calculated by a C++ program written for this purpose (see Appendix I).

RASPP

The RASPP curves for the proteins were generated with a minimum block length L of 5 amino acids (Endelman et al. 2004), and a $<m>$ bin size of 1 during library design and 0 to generate a set of test libraries. Python scripts to perform RASPP can be found at the Arnold lab website http://www.che.caltech.edu/groups/fha/.

Chapter IV: Construction and Characterization of Site-Directed Recombination Libraries

Introduction

There are several different methods for creating protein chimeras. By far the simplest technique is to take advantage of existing restriction sites to swap portions of DNA sequence. This is often augmented by PCR overlap-extension to generate a small populations of protein chimeras that are used for structure/function experiments (Back and Chappell 1996; Kushiro et al. 1999). However, constructing chimeras individually is time consuming and is not practical for creating a library of chimeras.

To create large numbers of protein chimeras there are variety of techniques that allow proteins to be recombined randomly. These methods fall into two general categories: homology-dependent and homology-independent. The homology-dependent methods include methodologies similar to DNA shuffling that create random gene fragments and reassemble them through PCR. Because these methods are annealing-based they rely on high sequence identity between the parental sequences for successful assembly. This limits them to recombining genes that share more than about 70% sequence identity. The homology-independent methods are capable of recombining distantly- or even nonrelated proteins. However, many of these techniques do not maintain the reading frame and allow inserts and deletions to occur at the recombination sites. Because the reading frame is not maintained, 2/3 of variants are out of frame and therefore not encoding useful proteins.

We have extended site-directed chimera construction to combinatorial assembly of gene fragments using ligation (Hiraga and Arnold 2003). This technique allows

nonhomologous genes to be recombined, but maintains the reading frame. Similarly to constructing individual chimeras using restriction sites, combinatorial ligation utilizes blocks of sequence with specific basepair overhangs. These overhangs are the same in all the parental sequences, but different at each recombination site, and allow specific ligation of the sequence blocks in the correct order. Because the overhangs are the same for each parental sequence, the blocks from different parents are freely exchangeable during construction (Figure IV-1)

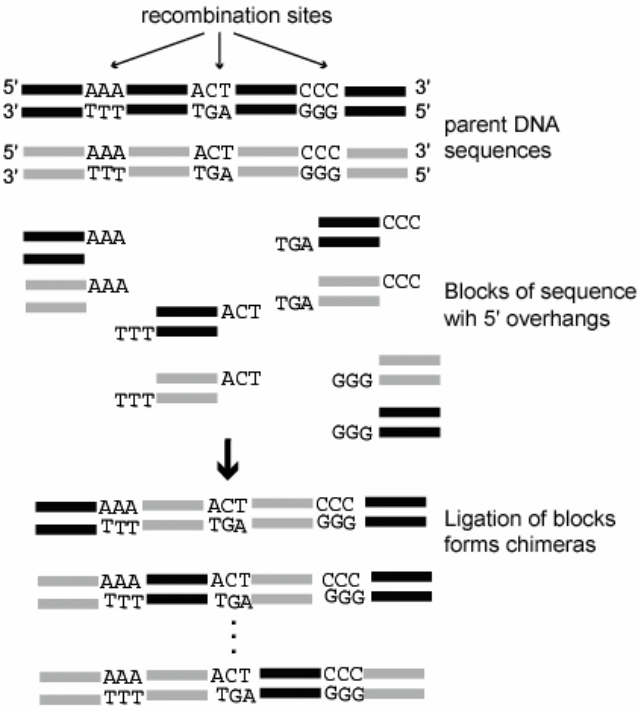


Figure IV-1. Overview of combinatorial gene assembly using sequence blocks of any sequence identity.

There are several potential methods for creating the necessary sequence blocks. As the cost of DNA synthesis has decreased, the easiest method is to simply order the gene fragments as oligonucleotides. These oligonucleotides can be phosphorylated, annealed and used directly without further modification. However, the blocks of sequence must be relatively small (<20-25 amino acids) so that the DNA segments

purchased are not too long. Longer oligonucleotides have a greater probability of incorporating a deletion. This synthetic methodology was used in the construction of the library described in Chapter II (Meyer et al. 2003). Alternatively, the gene fragments can be PCR amplified and separated with tag sequences that can generate any desired overhang when cleaved with a Type IIB restriction enzyme (cleaves outside its recognition site) (Hiraga and Arnold 2003). This methodology permits the incorporation of larger gene fragments, and has the advantage errors can be minimized by cloning and sequencing the PCR amplified gene segments prior to library construction.

In this chapter we describe the construction and characterization of the lactamase libraries designed in Chapter III. They were constructed using the two different variations on combinatorial ligation described above. Once a library is constructed, information about the chimeras must be obtained in a high-throughput manner so that a large number of chimeras can be assessed. The goals of this work include exploring what altered substrate specificities might be obtained in chimeric proteins, as well as investigating the properties of folded and functional chimeras. To meet these goals it is necessary to obtain sequence, function, and folding information for a large number of the chimeras created. High-throughput techniques exist for gathering much of this information. However, several methods required adaptation for this particular system.

Results

Construction of the RandE:APST lactamase library

The RandE:APST lactamase library was designed in Chapter III using random enumeration. It consists of four parents (TEM-1, PSE-4, AST-1 and SED-1) with nine

exchangeable fragments. The N- and C-termini are fixed to the same parent, and there are eight blocks between them. This library was constructed through sequential ligation of purchased oligonucleotides. All of the library blocks with the exception of the N- and C-termini are less than 21 amino acids. Since the N- and C-termini are always fixed to the same parent, a separate plasmid containing the termini was constructed for each parent (Figure IV-2A). Thus to construct the complete library, a set of reconstructed library blocks (2-9) must be ligated into each parental plasmid. Between the termini there is a cassette that contains a stop codon in each frame to prevent any translation before the complete library is added. This cassette is removed to construct the final libraries.

The purchased oligonucleotides were phosphorylated, annealed, and subsequently ligated in a sequential scheme, outlined in Figure IV-2B. There was some difficulty obtaining sufficient material for ligation into the final plasmid, which was remedied by PCR amplification of the reconstructed blocks prior to the final ligation. A second difficultly was encountered with one of the parent proteins, AST-1. AST-1 was originally cloned with a GTG start codon (Poirel et al. 2001). In our hands, the clones failed to confer resistance in our expression system. Additionally, AST-1 required extreme PCR conditions compared with the other parents (see methods) which might result in biases in the final library. As a result, AST-1 was dropped from the library. The library actually constructed therefore consisted only of parents TEM-1, SED-1 and PSE-4 and was known as RandE:PST. As discussed in Chapter III, this library is significantly smaller than the designed library (3^9= 19,653 vs. 4^9=262,144). The trade-off between diversity and fraction folded, however, remained similar to that of the original library (Table III-1).

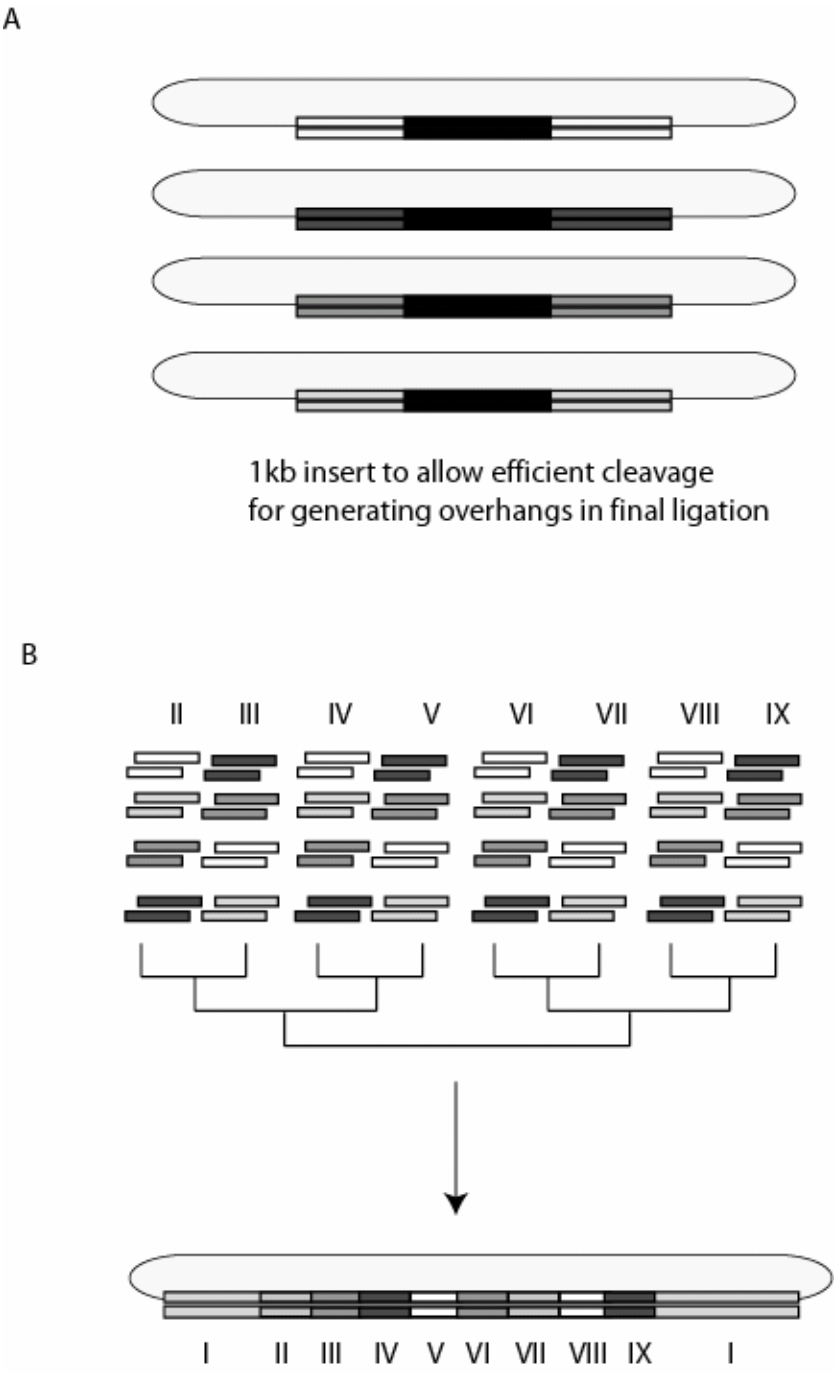


Figure IV-2. Overview of construction methodology for RandE:APST, and RandE:PST libraries. A: The N- and C-termini for each of the four parents were placed into a plasmid with a stop cassette between them. B: Annealed and phosphorylated oligonucleotides were ligated in series (see methods) to generate a full-length insert. This insert is ligated between the N- and C-termini from each parent to generate the full-size library.

To examine the quality of the RandE:PST library, 20 chimeras were completely sequenced. This sequencing revealed 2.5-3 basepair mutations per chimera and 0.6 deletions per chimera on average. These events were spread throughout the oligonucleotide-derived portion of the library. Due to the high deletion rate, a large proportion (~40%) of the library was out of frame. The oligonucleotides used for construction were on average 53.4 bp, not significantly longer than those commonly used for many molecular biology applications. However, to construct a perfect chimera with no mutations or deletions, 16 such oligonucleotides must all be perfect. The nucleotides purchased were cartridge-purified, which is not sufficient for this application. This library was not characterized further due to its high deletion rate.

Construction of RASPP:PST lactamase library

The RASPP:PST library described in Chapter III (Table III-1) designed using RASPP for three parents (TEM-1, PSE-4 and SED-1) with eight exchangeable fragments was constructed using Sequence Independent Site-Directed Chimeragenesis (SISDC) (Hiraga and Arnold 2003); an overview is in Figure IV-3. This method involves PCR amplifying the gene fragments to insert sequence tags. These tags are later removed using a type IIB restriction endonuclease (BsaX1) to generate the specific basepair overhangs necessary for ligation. The only problem encountered with this methodology is that one of the gene fragments was small (<30 bp) and was consistently lost during one phase of the procedure (see methods). To remedy this, oligonucleotides corresponding to the gene fragments were purchased and added to the ligation reactions as phosphorylated and annealed gene fragments. The oligonucleotides were short (24 bp) and PAGE purified.

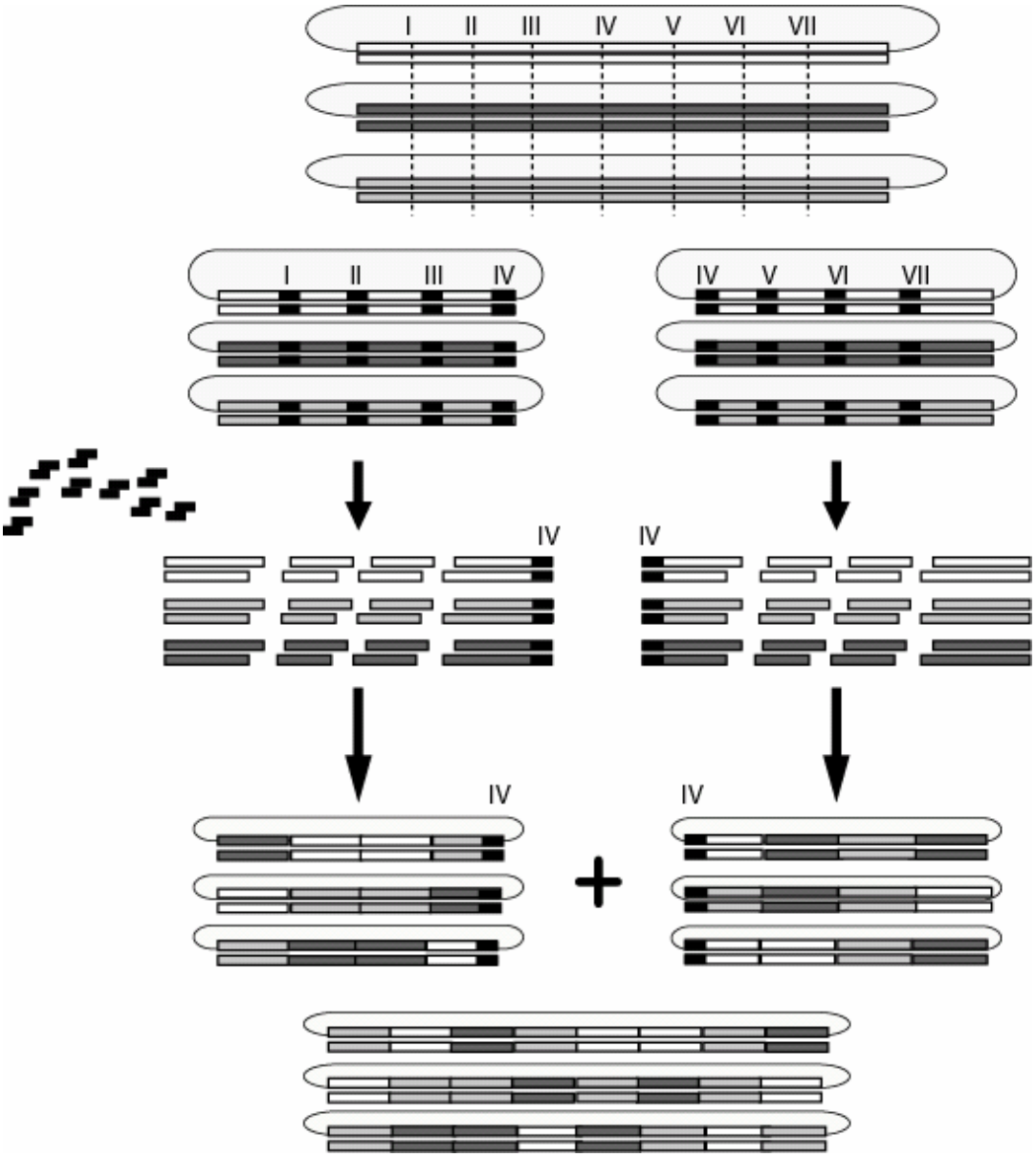


Figure IV-3. Overview of construction methodology for the RASPP:PST library. Tag sequences that will allow specific overhangs to be generated are inserted into the genes using overlap-extension PCR. The tag-inserted genes are cut with a type IIB restriction enzyme to expose the DNA fragments with desired overhangs, and the tag sequences are removed. The DNA fragments are then ligated together to form two minilibraries which are cloned individually. Finally, the two mini-libraries are then ligated to form full-length genes. Sequences cloned and transformed into *E. coli* are shown with the plasmid backbone.

To assess the quality of the library, 48 randomly chosen chimeras were sequenced. The rates of single-base mutation and basepair deletion are much lower than those observed for the RandE:PST library. However, with 2 mutations and 10 deletions affecting 11 of 48 chimeras, deletions are still prevalent. Of the 10 deletions, 3 were found at segment junctions and the remaining 7 were found in regions within PCR primers used after the block assembly during construction, usually at the N-terminus. None of the deletions or mutations was found within the small block added as an oligonucleotide. Additionally no deletions were detected while sequencing half-length chimeras generated during the construction procedure (see methods).

The high rate of single basepair deletion observed in 19% the full-length chimeras may occur because producing no protein (frame shift in first few amino acids) is more favorable than producing large amounts of unfolded protein. Thus chimeras with deletions are slightly favored over those without under nonselective conditions. We have previously observed that expression of lactamase chimeras under nonselective conditions can affect fragment biases in the library, presumably because some fragments are potentially deleterious (Hiraga and Arnold 2003). As a consequence of the deletion rate, there are potentially a large number of false negatives (up to 4% of characterized chimeras). However, 23% of the functionally characterized chimeras were observed multiple independent times. Any contradictions in functionality assignments were explicitly examined, further reducing the number of erroneous functionality assignments. Anecdotally, several chimeras that confer resistance to ampicillin contain deletions in the first few amino acids. This may occur because the first 24-30 amino acids comprise a

periplasmic targeting sequence and a downstream ATG permits sufficient expression to confer resistance.

High-Throughput Sequencing

The DNA sequences of 811 randomly chosen chimeras were determined by DNA probe hybridization, to obtain 553 unique sequences (Meinhold et al. 2003). To assess the error rate in the hybridization, 48 randomly chosen chimeras were sequenced. The probe hybridization is accurate, with 47 of 48 sequences correctly assessed. Examining the composition of the characterized sequences on a ternary diagram shows that the characterized library does not have equal representation of the different parents (Figure IV-4A). In particular, few chimeras similar to PSE-4 and many similar to TEM-1 were characterized. The proportion of the different parents at each position shows that PSE-4 was severely underrepresented at block 8 (Figure IV-4B). This discrepancy is due to an error in construction; a restriction site in 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. The library properties do not change significantly if all chimeras containing PSE-4 at block 8 are omitted ($<m>=67 \pm 36$ and $<E>=43 \pm 21$ for all chimeras lacking PSE-4 at block 8 vs. $<m>=66 \pm 21$ and $<E>=45 \pm 15$ for all chimeras in library). However, the library size is reduced by 1/3. Additionally TEM-1 is favored at most positions, block 3 most strongly. Because one parent (TEM-1) is more likely to be found at all of the blocks whose frequencies were determined from the original DNA mixing (3, 6, and 7), it is possible that the unequal representation of the different parents in the characterized library is due to improper quantification of the DNA during the initial

stages of construction. It is also possible that some clones do not survive to the characterization stage because the expressed protein is deleterious. Examination of the *E* and *m* distributions of the characterized chimeras shows that the characterized library has roughly the same distributions as the theoretical library despite its biases (<*E*> = 44 ± 17 and <*m*> =66 ± 22 for theoretical library, <E>=45 ± 15and <m>=66 ± 21, for characterized library, Figure IV-5)

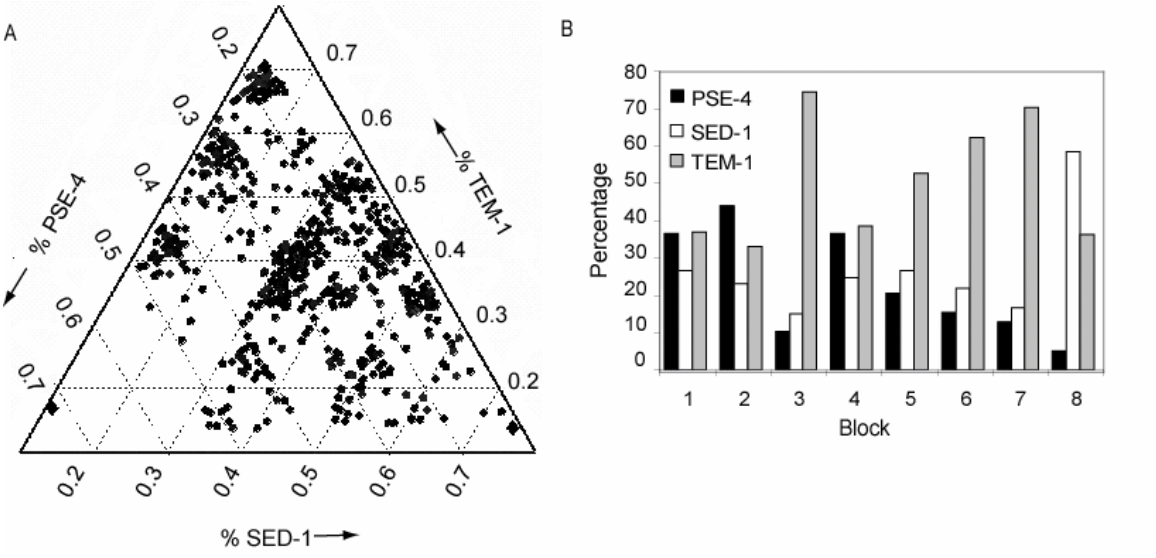


Figure IV-4. A: Ternary diagram showing the compositions of the 553 characterized chimeras. Characterized chimeras do not evenly populate the available sequence space, but are biased toward some areas. The position of each point is determined by the relative similarity of the chimera to each of the parents. To establish the location of a point on the ternary diagram the number of amino acids a chimera shares with each parent sequence is determined. Positions where there is no variation among the three parents are not included. Including such positions does not change the qualitative representation but merely shrinks the diagram into a smaller spread of space. The similarity of the chimera to each parent is then normalized by dividing by the sum of the similarities to each parent. B: The proportion of each parent protein at each block of the library. A perfectly balanced library would have 33% of each parent at each position.

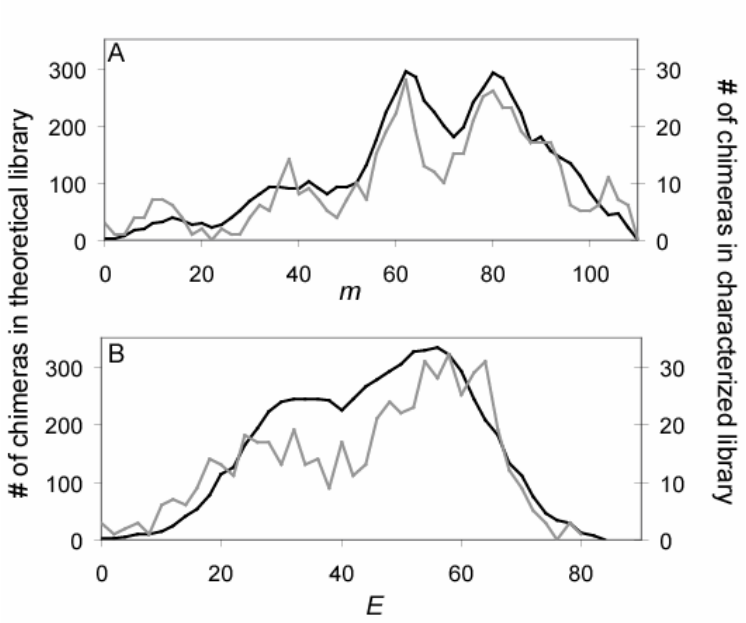


Figure IV-5. Distribution of chimeras with respect to *m* (A) and *E* (B) in the theoretical (black line) and characterized (gray line) libraries shows that the characterized library has approximately the same distribution as the theoretical library.

Evolution of New Function

One of the goals of this project is to determine whether site-directed recombination can generate chimeras with new functionality. We searched for chimeras with resistance to extended-spectrum antibiotics using functional selections. Unfortunately, there were two confounding factors in this process. First, SED-1 in our hands was significantly more resistant to most cephalosporins than originally described in the literature (Petrella et al. 2001); this has since been reexamined by the authors (Petrella et al. 2004). Second, most β-lactam based antibiotics are cell density dependent. This property makes it very easy to isolate false positives. The lactamase parents were tested against 11 different antibiotics (Table IV-1). For many of these, SED-1 displayed significantly more activity than TEM-1 or PSE-4. The RASPP:PST library was tested using antibiotics to which the lactamases displayed relatively low resistance (see methods). Typically the antibiotic concentration was lowered to the point where false

positives were isolated due to the density dependency of the antibiotic. No chimeras with significantly improved resistance to any of these antibiotics were isolated.

Table IV-1. MICs (μg/mL) of TEM-1, PSE-4, and SED-1 on β-lactam and Cephalosporin Antibiotics

antibiotic	TEM-1	PSE-4	SED-1
ampicillin	>2000	>2000	>2000
cefamandole	>2000	500	>2000
cephalothin	2000	1000	>2000
ceftazidime	<1	<1	200
cefoxitin	100	200	200
cefoperazone	>2000	1000	>2000
cefotaxime	1	20	>2000
ceftriaxone	2	40	>2000
cefsulodin	1000	500	>2000
carbenicillin	>2000	>2000	>2000
moxalactam	1	2	10
aztreonam	1	2	>50

It is somewhat surprising that no chimeras with increased resistance to cephalosporins were identified. Previous studies with TEM-1, and the profusion of natural TEM-1 variants, indicate that it is relatively easy to obtain extended-spectrum activity in TEM-1. However, many of the single mutations introduced to give TEM-1 extended-spectrum activity are not incorporated by our recombination. Additionally, because SED-1 already confers a higher level of resistance to many of these antibiotics, it is more difficult to identify variants with increased resistance. SED-1 limits the number of possible substrates, and the baseline antibiotic concentrations used in the selections are much higher than they would be for PSE-4 and TEM-1 alone (Table IV-1). It may be that recombination is not a good strategy for improving functions, but rather is more

exploratory for finding different functions. It is also possible that the sequences used in this work do not have the "right" sequence diversity within them to increase activity toward cephalosporins.

Folding

There are several different high-throughput methodologies in the literature that purport to detect a correctly folded protein (Maxwell et al. 1999; Waldo et al. 1999; Philipps et al. 2003). Most of them actually measure the amount of soluble protein in the cell because this corresponds well with properly folded protein. One potential complication in using many of these methods is that β-lactamases are exported to the periplasm and have an N-terminal signal sequence. We chose to implement one of these methods by fusing GFP to the N-terminus of the lactamases. Good signal differences between positive and negative controls were achieved under high-throughput conditions (Figure IV-6). However, when a library of clones was examined, the distribution of values obtained made it difficult to assess where the line between folded and unfolded should be drawn. Additionally there were several sequences which displayed very low fluorescence (less than negative control) but retained resistance to ampicillin (Figure IV-7), indicating chimeras may still be capable of catalysis, but not accumulate large quantities of protein in the cell. This observation indicates that any folding assay based on measuring soluble protein may not accurately describe those proteins that are folded enough to maintain catalytic activity but do not accumulate in the cell. Therefore, rather than continue to pursue a folding assay, we concentrated on measuring base line catalytic function.

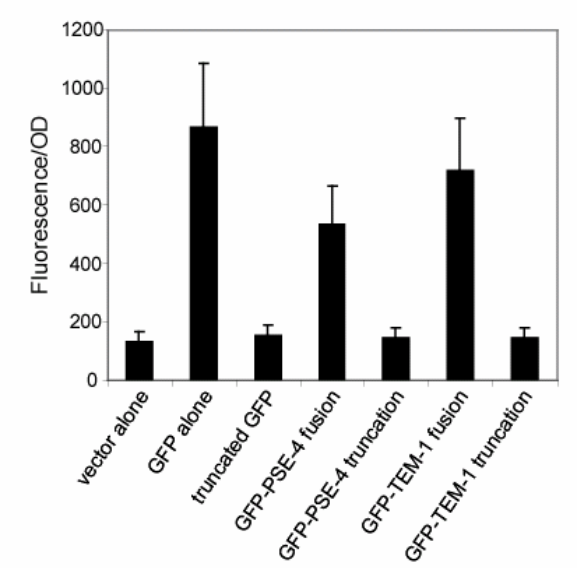


Figure IV-6. Fluorescence measurements of control strains used for GFP folding screen showing a good difference between positive and negative controls.

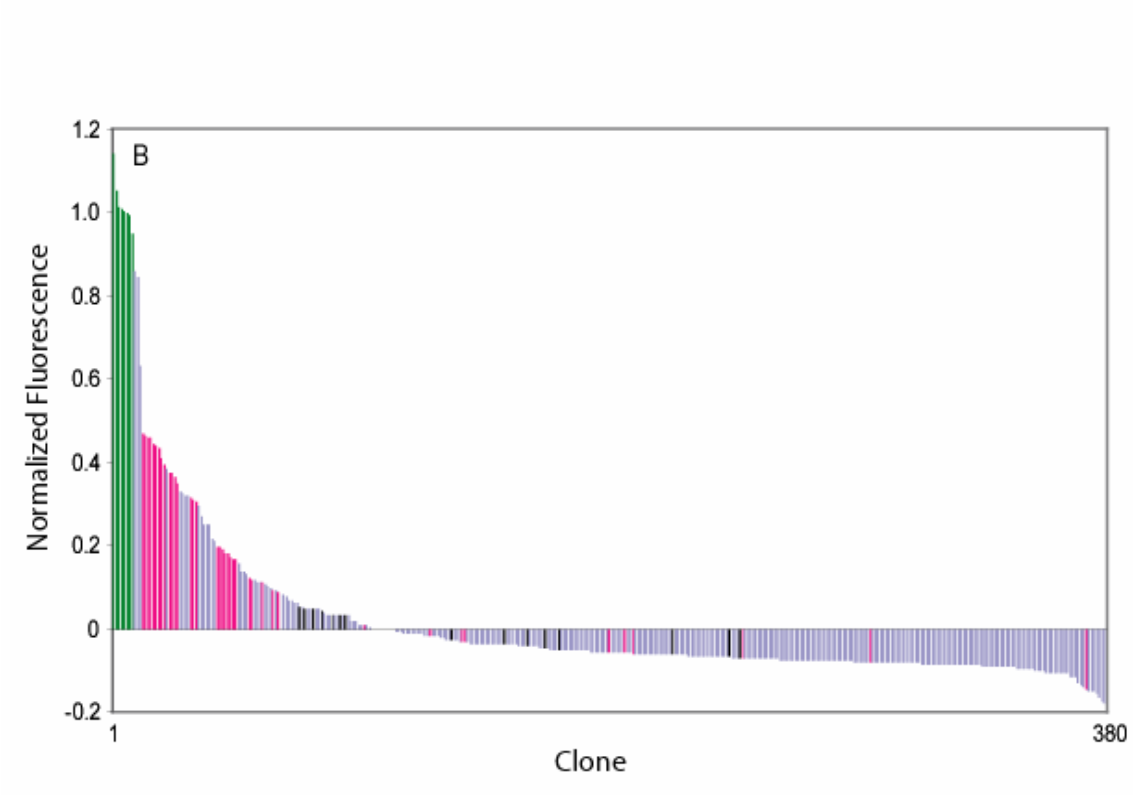


Figure IV-7. Normalized fluorescence from a library of different clones in the GFP screening system. Green bars are positive controls, black bars are negative controls, and pink bars are ampicillin resistant clones. All values are normalized between zero (average of negative controls) and one (average of positive controls). The distribution of values obtained for chimeras makes it difficult to draw a line between "folded" and "unfolded." Additionally some chimeras displaying less fluorescence than negative controls conferred resistance to ampicillin.

Retention of Function

Because the folding assays were unable to distinguish folded and unfolded chimeras, a low stringency functional screen was used to assess which chimeras retained basic catalytic function, and thus a folded structure. Chimeras in the RASPP:PST library were screened for a function shared by all three parents, the ability to confer ampicillin resistance. The screen was conducted at very low stringency (>500x lower concentration of ampicillin than the wild-type MIC) to capture chimeras with even very minimal activity. Of the 554 unique sequences tested, 20% (111) conferred resistance to ampicillin and are considered functional lactamases. An additional 51 unique functional lactamase sequences were obtained by selecting functional clones prior to probe hybridization sequencing, giving a total of 162 functional lactamases. A complete listing of all chimera sequences and their functionality status can be found in Appendix III. Of the functional chimeras, 51% conferred a MIC of 2,000 μg/ml ampicillin or greater, indicating approximately wild-type activity (~5,000 μg/ml for all three parents). 10% of chimeras displayed a MIC of 50 or below, indicating weak activity**.** Chimeras that did not confer resistance to ampicillin may be not folded, may not be well expressed, may be folded but not catalytically active, or may have a combination of these properties. Because the screen was very low stringency, chimeras that are well-expressed, folded proteins with any catalytic activity are likely to have been identified.

Discussion

Examining the naïve data set of 553 chimeras of which 111 (20%) are folded shows that, like the previous lactamase chimera library (Meyer et al. 2003), chimeras

with low E are more likely to function than chimeras with high E (Figure IV-8). Unlike the library described in Chapter II, the decline in probability of retaining function with respect to E is not exponential (Meyer et al. 2003), but instead is more reminiscent of the sigmoidal function originally described by Voigt et al. (2002). The difference in the form of P_f is not surprising. This is a designed library, where the distribution of chimeras is skewed toward those that are likely to fold. The distribution of chimeras in the library affects the form P_f with respect to E can take when measured with any given library. This is also a more accurate estimation of P_f because all the chimeras used in the calculation have been explicitly observed. In the previous work (Chapter II) we observed only functional chimeras directly and assumed that most other were nonfunctional. This assumption can lead to a less accurate description of P_f. Chimera probability of functioning decreases with increasing m (Figure IV-8). However, there is not a simple function that describes the behavior P_f with respect to m. The slightly bimodal behavior is a result of the underlying bimodal distribution of m of the chimeras in the characterized library (Figure IV-5A).

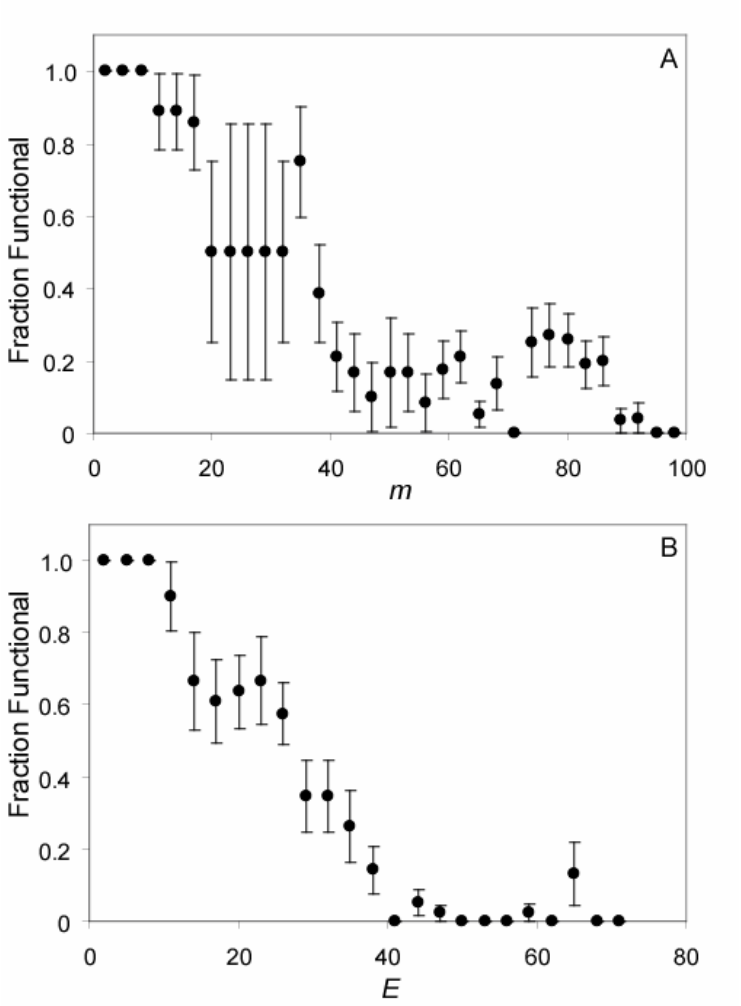


Figure IV-8. A: Fraction of functional chimeras with respect to *m.* B: Fraction of functional chimeras with respect to *m.*
Chimeras were put into bins of 3 *E* or *m* to ensure representation in all bins and the fraction functional *(*$F_{functional}$*)* is = $N_{functional}/N$ where N is the total number in the bin and $N_{functional}$ is the number functional. The error was approximated by the standard error on an estimate of the binomial proportion (sqrt (($F_{functional}$*(1-$F_{functional}$)/N))).

The 162 functional chimeric lactamases span a range of mutation levels compared to the parental proteins, and contain up to 80 mutations from the closest parent. One-third of active chimeras displayed ≤75% sequence identity to any known lactamase. Five concentrated clusters of sequences account for 77% of the functional chimeras (Figure IV-9). Within these clusters, sequences share on average 95% identity. The clusters result from several different factors including uneven block sizes, sparse sampling of the theoretical library, and favorable or unfavorable block interactions. It is likely that there are other such clusters which are not observed due to differences between the characterized and theoretical library.

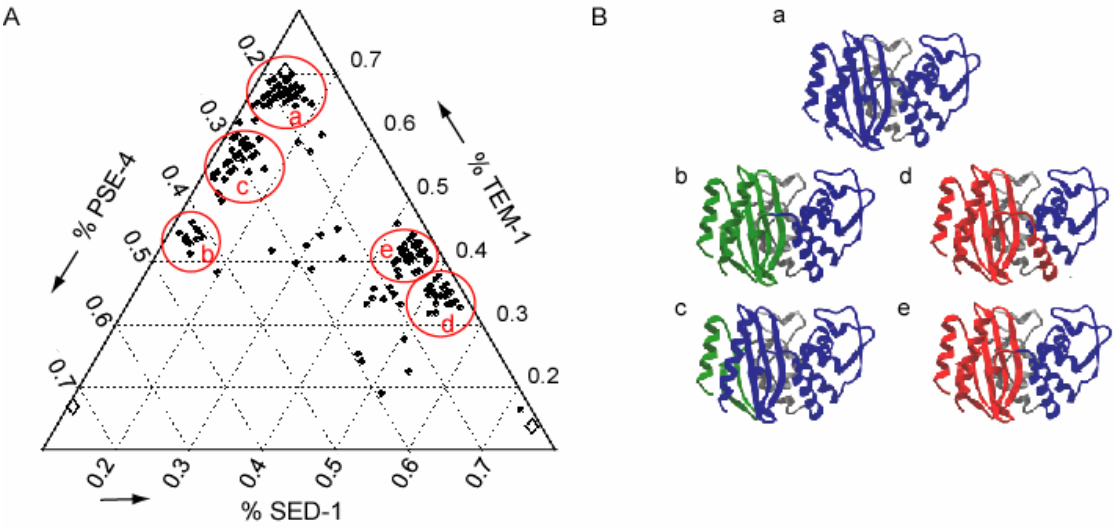


Figure IV-9. Ternary diagram similar to the one in Figure IV-4 representing only chimeras that display ampicillin resistance. The chimeras fall into five main clusters which can be described by which parents the four largest sequence blocks (1, 3, 7, and 8) are inherited from. The structures show which parents each cluster inherits its major blocks from: green ribbons indicate PSE-4, blue TEM-1 and red SED-1, gray blocks are variable within each cluster.

Chimeras in the first cluster (Figure IV-9, cluster a) have all the large blocks (1, 3, 7, and 8) from TEM-1. On average these chimeras differ from TEM-1 by only 12 mutations. Chimeras in the second cluster (Figure IV-9, cluster b) have the N- and C-termini from PSE-4, and blocks 3 and 7 from TEM-1. These chimeras are significantly different from both TEM-1 and PSE-4, and an average of 74 mutations from the closest parent. Chimeras in the third cluster (Figure IV-9, cluster c) have the N-terminus from PSE-4 and the remaining large blocks (3, 7, and 8) from TEM-1. These chimeras have on average 37 mutations to TEM-1. The last two clusters are sequences comprised mostly of TEM-1 and SED-1. Chimeras in the fourth cluster (Figure IV-9, cluster d) of these have the N- and C-termini as well as block 7 from SED-1 and block 3 from TEM-1. These chimeras have on average 60 mutations to SED-1. The fifth cluster (Figure IV-9, cluster e) has the N-and C-terminus from SED-1 and blocks 3 and 7 from TEM-1. These

sequences are the most distant from any of the parents, with an average 78 mutations to the closest parent.

There are 1785 lactamase sequences in the PFAM database for protein families (Bateman et al. 2004), at least 450 of which are class A lactamases by phylogenetic analysis. New lactamase sequences continue to be identified. However, many of the characterized lactamases are minor variations of a few very prevalent sequences. For example, there are over 100 characterized variants of TEM-1, differing from TEM-1 by only a few amino acids (Jacoby and Bush 2005). The lactamase skeleton seems relatively tolerant to mutation. 220 of 263 positions in TEM-1 accept at least one other amino acid when mutated in isolation (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 clusters of functional lactamases observed here represent regions of sequence space that are not populated by known natural β-lactamases. Not including the cluster most similar to TEM-1, the cluster consensus sequences range from 72% to 80% identity to any natural lactamase. However, these areas appear relatively densely populated with catalytically active and folded proteins. While functional lactamases cluster toward some areas of sequence space, it is not known based on these simple observations whether they cannot occupy others.

Methods

All enzymes used were purchased from New England Biolabs and all chemicals were purchased from Sigma unless otherwise indicated.

Construction of RandE:APST and RandE:PST Libraries (purchased oligonucleotides)

Four parent plasmids were constructed by PCR amplifying the N- and C-termini from each parent separately and combining them by overlap extension PCR with a cassette that consisted of stop codons in each frame and an out of frame segment from P4501A2. TEM-1, SED-1 and PSE-4 all amplified under standard PCR conditions. AST-1 required the addition of 5% DMSO as well as a denaturation temperature of 98 °C. These constructs were placed into the expression plasmid pProTet E.333 (Clontech). To release the correct overhangs, the final plasmids were digested with SapI, and treated with alkaline phosphatase. Doubly cut plasmid was separated from linearized plasmid by agarose gel electrophoresis. The desired product was recovered from the gel and purified using a Zymocan DNA gel recovery kit.

Oligonucleotides that compose the smaller sequence blocks for the library described in Chapter III were ordered from Invitrogen as cartridge purified stocks (Table AII-1). Complementary oligonucleotides were annealed (1.25 mM each primer, 50 mM NaCl, heat to 95 °C for 2 min and ramp 1 °C/s to 4 °C) and then phosphorylated (10U T4 polynucleotide kinase, 37 °C, 1 hr in T4 ligase buffer). The annealed and phosphorylated oligonucleotides from all parents for each sequence block were mixed and ligated together in pairs (2-3, 4-5, 6-7, 8-9) using T4 ligase at 16 °C for 5 hours. Following ligation, the correct product was isolated by gel electrophoresis and purified using a

Zymoclean DNA gel recovery kit. The ligation process was repeated with the products until full length inserts (8 pieces total) were obtained. This product was PCR amplified using *Pfu* polymerase (Stratagene) with primers from all possible parent pairs (Table AII-2). The final product was cut with Sap1 to generate the overhangs and ligated as above into the four parent plasmids. The number of clones obtained for each parent plasmid varied between 5,000 and 13,000.

Construction of RASPP:PST Library (SISDC)

Each fragment was PCR amplified (*Pfu* Ultra, Stratagene) with primers to introduce a tag sequence at all internal junctions (see Table AII-3) for primers and sequence fragments). All tag sequences contained BsaX1 and Nde1 restriction sites, as well as a unique region for each junction. Additionally each junction was designed to have a unique 3bp sequence found in all three parents which is released upon BsaX1 cleavage (See Table AII-3). There is no tag sequence between segment 4 and 5, but rather a Sap1 restriction site accompanied by a Pst1 restriction site (fragment 4) or Sal1 restriction site (fragment 5) was inserted. The first and last four segments of each parent were separately reconstructed using overlap extension from PCR amplified segments, each resulting in a half-length gene product with tag sequences inserted at the junctions. The N-terminal half was cloned into pBC (SK+) with Sal1 and Pst1 sites on the forward and reverse primers, and the C-terminal half with Xho1 and Pst1 sites. The primer sequences for PCR reactions for tag insertion can be found on Table III-1.

The DNA sequence of each half-library parent was confirmed. The DNA for each half library was mixed in equal proportions based on spectrophotometric quantification,

then cut with Sal1/Pst1 (front) or Xho1/Pst1 (back), dephosphorylated, and the insert purified by agar gel electrophoresis. This insert was then cut with BsaX1 to remove the tag sequences and generate 3 bp overhangs. Following column purification (Zymogen Clean and Concentrate) to remove the tag sequences, fragments for block 2 were added as annealed and phosphorylated oligonucleotides, and the mixture was ligated using T4 ligase for 5 hours at 16°C. The ligation was column purified, cut with Nde1 and BsaX1 to remove any incompletely cleaved tag sequences, and then PCR amplified with 9 primer sets (see Table AII-4) to generate the complete library. The PCR reactions were mixed in equal proportions based on agarose gel quantification. The N-terminal half-library was cut with Xho1/HindIII and ligated into pBC cut with the same enzymes. The C-terminal half-library was cut with Sal1/Pst1 and ligated into pProTet (Clontech) cut with the same enzymes. The resulting DNA was transformed into DH5αPRO to prevent expression from the Tet promoter on pProTet.

A few thousand clones were obtained for each half-library, sufficiently higher than the expected complexity (81 sequences) of each half. These colonies were pooled, and the DNA was purified from them (Qiagen midi-prep). The N-terminal half-library was removed from pBC using Kpn1 and Sap1 restriction sites and inserted into the C-terminal half-library cut with the same two enzymes to reconstruct full-length genes. This ligation was transformed into XL-l Blue (Stratagene) and the clones used directly for analysis.

High Throughput Sequencing

To determine the block sequence of the chimeras, ~1100 clones were picked and grown overnight to saturation in 384-well plates containing 70 µl of LB + 35 µg/ml chloramphenicol. Each plate contained four samplings from each parent as well as the expression plasmid (pProTet) containing no lactamase insert. The plates were stamped onto N+ Hybond membranes (Amersham) layered onto 2% agar LB plates and allowed to grow at 37 °C for 18 hours. The membranes were removed from the plates, the cells lysed according to Meinhold et al. (2003) and the DNA attached to the membrane through UV cross-linking. The membranes were dried and stored at 25 °C for up to 1 month.

Probes for used for hybridization are listed in Table AII-5 and were labeled with DIG-dUTP using Roche DIG Oligonucleotide 3'-End labeling Kit, 2nd Generation according to the manufacturer's instructions. Hybridization was performed at 58 °C, and the stringency washes carried out at 53 °C in 2x, 1x or 0.5 x SSC +0.1% SDS depending on the probe (Table AII-5). Probes were detected using a Roche DIG nucleic acid detection kit according to the manufacturer's instructions and visualized using Kodak MRX film.

New Antibiotics

The MICs of 11 different antibiotics was determined for TEM-1, SED-1 and PSE-4 by spotting saturated culture onto an agar plate containing the antibiotic and 35 ug/mL chloramphenicol (Table IV-1). The MIC was measured as concentration of antibiotic which prevented visible growth. These conditions simulate antibiotic screening rather

than selection. For antibiotic library selections (moxalactam, ceftazidime, and cefoxitin) the MIC determined above was used as the starting concentration of antibiotic. The concentration was progressively decreased until colonies were observed on a negative control plate. To search for chimeras with increased resistance to these antibiotics, library plasmid DNA was transformed into XL-1 Blue (Stratagene) according to the manufacturer's instructions. 200 uL of cells was spread on each 100 x 15 mm plate LB agar plate containing antibiotic, and 10 uL was spread on a nonselective plate to determine the approximate number of colonies obtained. The plates were grown for 18 hours at 37 °C. Colonies were restreaked onto plates with the same antibiotic concentration to verify resistance.

Folding

The GFP folding assay was implemented similarly to Waldo et al. (Waldo et al. 1999). Briefly GFPuv (Clontech, from pGFPuv) was placed N-terminal to the lactamase in pProTet. The signal sequences were removed to residue 24 for PSE-4 and 26 for TEM-1. SED-1 was never tested in the folding assay because the assay failed to distinguish between folded and unfolded chimeras cleanly. For PSE-4 the linker was Gly-Ser-Ala-Gly-Ser-Ala-Asn-Ala-Ser-Gly, an additional Ser-Gly was added directly before TEM-1. An NsiI restriction site was incorporated within the linker. To place a library into the expression system, a negative control protein was removed and the PCR amplified library incorporated using NsiI and PstI. Chimeras expressed in BL-21 were grown in deep-well plates containing 1 mL M9 medium with 35 ug/mL chloramphenicol at 30 °C, 220 rpm, 80% humidity for 18 hours. The plates were then centrifuged to pellet cells and stored at

4 °C for 24 hours. Cells were rinsed with 500 uL PBS and then resuspended in 300 uL PBS. OD_{600} and fluorescence (excite 395 nm, emit 509 nm) were measured.

Ampicillin Activity 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 the 384-well plates used for hybridization and allowed to grow at 37 °C for 18 hours 220 rpm 80% humidity. Approximately 2 μl aliquots of each culture were transferred to LB agar plates containing varying concentrations of ampicillin (0, 5, 10, 25, 50, 100, 250, 500, 1,000, 2,000 μg/mL) using the 96-well stamp and allowed to grow for 18 hours. Duplicate plates were generated at each concentration. After 18 hours the plates were observed for growth. Chimeras growing at concentrations of ampicillin 10 μg/ml or greater were considered positive. XL-1 containing pPro with no lactamase insert survive to 5 μg/ml ampicillin in this assay. The concentration of ampicillin necessary to prevent growth was recorded as the MIC. Chimeras that grew on the 2,000 μg/mL plates are recorded at 2,000+.

Chapter V: Using Chimeras to Identify Determinants of β-lactamase Function

Introduction

The most informative techniques for probing the relationships between protein sequence, structure and function are those that perturb a natural protein sequence to examine the properties of the new protein. Site-directed mutagenesis has become a standard tool for determining if a particular amino acid is necessary for a specific protein property, whether the property is folding, substrate specificity or catalytic activity. However, using mutagenesis alone it is difficult to explore properties that are not specifically tied to one or a few amino acids such as dynamics or allostery. Multiple sequence alignment (MSA) analysis of protein families has allowed the identification of energetic coupling within proteins (Lockless and Ranganathan 1999). However, natural sequences are under selection for additional properties besides the property under investigation and it can be difficult to discern which attributes are responsible for the property of interest.

Recombination of homologous proteins allows construction of proteins that are significantly different from natural proteins. This allows differences between homologous proteins as well as the determinants for a particular protein fold to be examined. Characterization of chimeric proteins in small studies has contributed to understanding product or substrate specificity (Kushiro et al. 1999; Nicot et al. 2002), as well as key elements for folding (Morimotoa and Tamura 2004). However, these data sets are invariably small and conclusions are drawn based on only a few chimeric sequences.

We have created and characterized a large number of chimeras made by recombining distantly related β-lactamases TEM-1, PSE-4 and SED-1. By examining the functional chimeras we can explore which portions of SED-1 are key contributors to the altered substrate specificity that corresponds with extended-spectrum activity. Mutagenesis studies and analysis of multiple sequence alignments and several crystal structures have generated many hypotheses about the sequence determinants of this altered substrate specificity, but few concrete answers.

We have previously observed that the functional chimeras cluster into a few areas of the possible sequence space (Figure V-1). However, examination of the functional chimeras alone does not provide enough information to determine if this is due to our sparse sampling of the theoretical library, or whether some areas are not compatible with functional lactamases. In addition to the many functional lactamase chimeras, we have also generated and characterized a large number of nonfunctional chimeras. Using both sets of sequences we can determine whether the clusters of sequences we observe are caused by inherent limitations of the protein fold, or by our sparse sampling theoretical library.

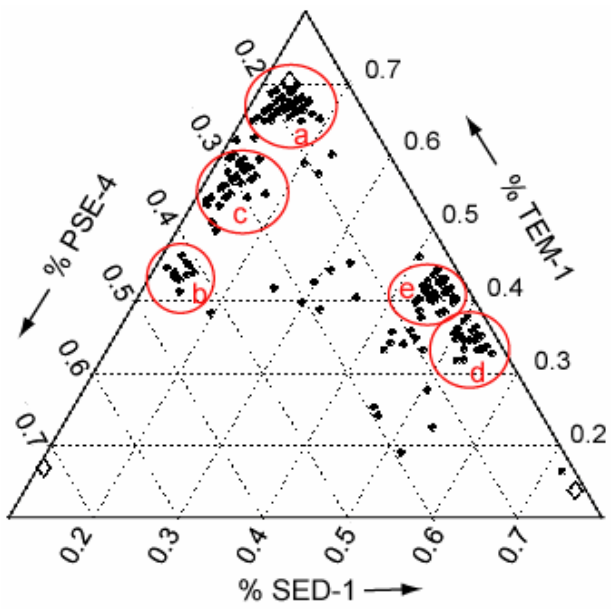


Figure V-1. Ternary diagram showing lactamase chimeras that display ampicillin resistance. The position of each point is determined by the relative similarity of the chimera to each of the parents. To establish the location of a point on the ternary diagram the number of amino acids a chimera shares with each parent sequence is determined. The similarity of the chimera to each parent is then normalized by dividing by the sum of the similarities to each parent. The chimeras fall into five main clusters (a, b, c, d, and e) which can be described by which parents the four largest sequence blocks (1, 3, 7, and 8) are inherited from. The ternary diagram represents compositional space. However sequences clustered on the ternary diagram tend to be clustered based on sequence identity as well. Cefotaxime resistant chimeras fall into cluster d.

Results and Discussion

Determinates of Cephalosporin Resistance

SED-1 is an extended spectrum CTX-M type lactamase that has significant activity toward cefotaxime, while TEM-1 and PSE-4 do not. CTX-M type lactamases are a class of extended-spectrum lactamases that have recently been isolated that are not similar to extended-spectrum TEM-1 variant and they do not simply widen the active site to alter substrate specificity (Orencia et al. 2001). Instead the source of the altered substrate specificity appears to originate from many sequence changes, and remains difficult to pinpoint. To identify which portions of the protein are critical for cefotaxime

resistance we screened characterized chimeras for resistance to cefotaxime. Twenty sequences displayed >10-fold higher cefotaxime resistance (0.1 μg/ml) compared to PSE-4 and TEM-1 and were considered positive on cefotaxime (Table V-1). The cefotaxime resistant sequences appear similar upon inspection; they share blocks 1, 7, and 8 with SED-1. These sequences are nearly all (18 of 20) from one of the clusters of sequences identified in Chapter IV (Figure V-1). All of them conferred resistance to ampicillin, usually to high levels of ampicillin (>1,000 μg/ml). Of the proteins conferring resistance to ampicillin, 83% of those with blocks 1, 7 and 8 from SED-1 confer resistance to high levels of cefotaxime. Those chimeras with blocks 1, 7 and 8 from SED-1 that do not show resistance to cefotaxime have low ampicillin resistance (<250 ug/ml), which indicates they may suffer from marginal stability or poor expression, rather than lack the ability to hydrolyze cefotaxime (Table V-1). In addition, there are four sequences which share this block pattern that did not confer resistance to either ampicillin or cefotaxime. These sequences are likely unfolded or not expressed (Table V-1).

Table V-1. Characterized Chimeras Inheriting Blocks 1, 7, and 8 from SED-1.

Cefotaxime resistant chimeras								CTX	AMP
S	P	T	P	T	S	S	S	0.2	1,000
S	P	T	P	T	S	S	S	0.2	2,000
S	P	T	S	S	T	S	S	0.2	1,000
S	P	T	S	T	S	S	S	2	1,000
S	P	T	S	T	T	S	S	1	2,000
S	P	T	T	S	S	S	S	2	1,000
S	P	T	T	T	S	S	S	0.2	2,000
S	P	T	T	T	T	S	S	0.2	2,000
S	S	S	S	S	S	S	S	>50	2,000
S	S	S	S	T	T	S	S	0.2	1,000
S	T	T	P	T	T	S	S	1	1,000
S	T	T	S	P	P	S	S	5	2,000
S	T	T	S	S	S	S	S	1	2,000
S	T	T	S	T	S	S	S	2	2,000
S	T	T	S	T	T	S	S	1	2,000
S	T	T	T	P	S	S	S	10	2,000
S	T	T	T	S	S	S	S	0.2	2,000
S	T	T	T	T	P	S	S	10	2,000
S	T	T	T	T	S	S	S	5	1,000
S	T	T	T	T	T	S	S	0.2	2,000
Ampicillin resistant, cefotaxime sensitive chimeras									
S	S	T	P	T	T	S	S	<0.1	100
S	S	T	P	T	S	S	S	<0.1	100
S	T	P	T	S	T	S	S	<0.1	250
S	P	T	T	T	P	S	S	<0.1	50
Ampicillin and cefotaxime sensitive chimeras									
S	S	T	P	P	T	S	S	<0.1	<10
S	S	T	S	T	P	S	S	<0.1	<10
S	T	S	P	T	S	S	S	<0.1	<10
S	T	P	T	T	T	S	S	<0.1	<10

Chimera sequences are represented by the parent each block is inherited from P for PSE-4, S for SED-1, and T for TEM-1. MICs for cefotaxime (CTX) and ampicillin (AMP) are given in μg/mL.

Based on various crystal structures of CTX-M type lactamases, the extended-spectrum activity cannot be attributed to active-site widening (Ibuka et al. 1999; Shimamura et al. 2002; Chen et al. 2005). Instead it is credited to several different factors including specific amino acid interactions with the substrate. Asn104 and Ser237 are

residues conserved in CTX-M type lactamases, but not in their narrow-spectrum relatives. Specific interactions between these residues and the carboxylate group and the acylamide side-chain of cefotaxime have been reported (Shimamura et al. 2002). These specific interactions are hypothesized to bind the substrate tightly into the active site. Another factor hypothesized to allow efficient cefotaxime hydrolysis is the position of the ω-loop. The ω-loop has significant affects on substrate specificity when altered in TEM-1 and PSE-4 (Petrosino and Palzkill 1996; Therrien et al. 1998; Sanschagrin et al. 2000). In CTX-M type lactamases there are fewer hydrogen bonds both within the ω-loop (residues 160-181) and between the ω-loop and the third strand of the β-sheet, β3 (residues 229-238). The altered hydrogen bonding pattern results in a change in the position of the ω-loop compared with TEM-1 (Shimamura et al. 2002). However, it also indicates that β3 is less restricted by hydrogen bonds (Ibuka et al. 1999). The third major hypothesis is that movements of β3 allow larger substrates to be accommodated by the active site (Chen et al. 2005). Comparison of anisotropic temperature factors for several CTX-M crystal structures shows that in broader spectrum CTX-M variants there is increased mobility of β3 (Chen et al. 2005). An engineered disulfide to restrict the movement of β3 can reduce the rate of cefotaxime hydrolysis of CTX-M type lactamase TOHO-1 (Shimizu-Ibuka et al. 2004), further supporting the importance of β3 movement for efficient cefotaxime hydrolysis.

Our results suggest that SED-1 blocks 1, 7 and 8 (residues 1-64 and 190-290) contain the necessary components to confer cefotaxime hydrolysis. The ω-loop is composed of blocks 4 and 5, and from this work it appears that they can originate from any of the parents. Additionally, inheriting the ω-loop from SED-1 does not confer

resistance to cefotaxime. The substitutions responsible for disrupting the hydrogen bonds between the ω-loop and β3 occur in both the ω-loop and within β3. Interactions between Asn104 and cefotaxime also do not appear critical for cefotaxime hydrolysis. Asn104 is found in block 3 which is inherited from TEM-1 in most of the chimeras identified.

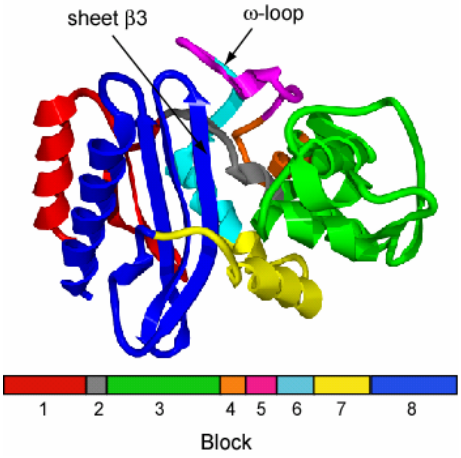


Figure V-2. Structure of TEM-1 with the exchangeable sequence blocks indicated by different colors. The sheet β3 and the ω-loop proposed to be important for cefotaxime resistance are marked.

The remaining hypothesized determinates of CTX-M extended-spectrum substrate specificity, strand β3 including Ser 237, are within block 8. The apparent necessity of inheriting block 7 from SED-1, which is an α-helix that packs against the β-sheet, for cefotaxime hydrolysis is somewhat more surprising. No amino acids within block 7 are near the active site. It is possible that block 7 is constraining the movement of β3 when inherited from TEM-1 or PSE-4. Unfortunately, there is no structural information available for SED-1, so it is difficult to determine what role that particular sequence block is playing. Block 1 may or may not be directly involved with the differences in substrate specificity. As will be discussed shortly, its presence may be necessary for forming a folded protein in conjunction with block 8.

Logistic Regression Analysis of Multiple Sequence Alignment

We previously observed that most functional chimeras cluster into five areas of sequence space (Chapter IV) (Figure V-1). While the sequences cluster on a ternary diagram used to represent the composition of the chimeras, these sequences also cluster based on pairwise sequence identity. To probe whether this clustering is due the sparse sampling of theoretical library, or whether it indicates that some regions of sequence space are unlikely to yield functional lactamases, we examined the entire dataset of 664 functional and nonfunctional chimeras (Appendix III). These data cannot be evaluated by eye like the smaller cefotaxime resistant data set not only because there are many more chimeras, but also because the characterized library is not a random sampling of the theoretical library due to biases introduced during construction. Therefore it is necessary to use an analysis methodology that compares the folded and unfolded chimeras, rather than an analysis method that implicitly assumes an even distribution of possible sequences.

Due to the binary nature of our data (1 for functional, 0 for nonfunctional), we can use logistic regression, an analog of linear regression, to analyze the data. Using logistic regression we fit the folding data to an energy model containing one-body ($\varepsilon_1(i.x)$) and two-body terms ($\varepsilon_2(i.x, j.y)$), that correspond to intra- and inter-block contributions to chimera folding (Equation (V-1)).

$$E = \varepsilon_0 + \sum_i \varepsilon_1(i.x) + \sum_i \sum_{j>i} \varepsilon_2(i.x, j.y) \qquad \text{(V-1)}$$

This method has previously been used to accurately infer interactions from an alignment containing folded and unfolded cytochrome P450s (Otey et al. 2006). The intra-block terms correspond to interactions between the amino acids and the solvent or the main

chain atoms as well as interactions between conserved residues. The inter-block terms correspond to pairwise interactions between the blocks.

Logistic Regression Analysis (LRA) of the β-lactamase data (Appendix III, entire data set including extra positive chimeras), the results of which are shown in Figure V-3, identified five variables as strongly significant (blocks 1, 2, 3, 8, 1-8) (($p << 10^{-6}$)) and three others as marginally significant (5, 1-7, 2-8) ($p \sim= 10^{-4}$) (Figure V-3). When the p-values of blocks 1 and 8 were recalculated relative to a model that includes pair 1-8, their significance diminished considerably (p=0.5 and $4x10^{-3}$ respectively) indicating that the pairwise interaction is the important determinant for folding rather than individual one-body terms. Blocks 2, 3, and the interaction between blocks 1 and 8 remained significant after the second round of p-value testing. The remaining block identities do not seem to have a significant impact on whether a chimera functions.

Inter-Block Interaction between 1 and 8 is Important for Function

The interaction of blocks 1 and 8 is the most significant determinant for retaining functionality (ampicillin resistance) according to the energy model derived by LRA. The diagonal entries corresponding to wild-type interactions are the most favorable (Table V-2), indicating that chimeras inheriting the blocks from the same parent are more likely to function than those inheriting the blocks from different parents. Additionally, chimeras inheriting block 1 from PSE-4 and block 8 from TEM-1 are more likely to function than any other mismatched pairing of the blocks. The importance of the interaction between blocks 1 and 8 was observed in previous experiments where the N- and C-terminal

fragments of the β-lactamase were almost always found from the same parent in functional chimeras (Hiraga and Arnold 2003; Meyer et al. 2003).

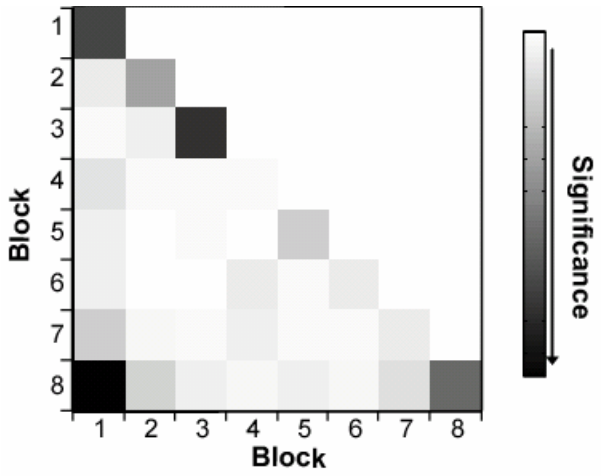


Figure V-3. Logistic regression analysis of functional and nonfunctional chimeras shows that some individual blocks (diagonal), or pairs of blocks are more significant than other for determining whether a chimera will function. The significant terms affecting chimera function are the interaction between blocks 1 and 8, and blocks 2 and 3.

Table V-2. Energies Assigned to Important Interactions by Logistic Regression Analysis

Two-body Terms	Parent at Block 8		
Parent at Block 1	PSE-4	SED-1	TEM-1
PSE-4	-1.2	1.7	-0.5
SED-1	-0.1	-2.8	2.8
TEM-1	1.3	1	-2.3
One-body Terms	Parent		
Block	PSE-4	SED-1	TEM-1
2	-0.5	1.1	-0.5
3	-0.6	1.1	-1.7

An energy value is assigned to each possible parent combination for the pairwise interaction between blocks 1 and 8, and to each parent for blocks 2 and 3. A more negative energy value indicates the block is more likely to be found in functional chimeras.

Blocks 1 and 8 together form almost half the protein; they also have the largest number of structurally contacting residues between two blocks (Table V-3). Block 1 is the most diverse: here the parents share on average only 25% sequence identity. Block 8 is not as diverse as block 1, but SED-1 is significantly more diverged from TEM-1 and PSE-4 than they are to each other in block 8 (Table V-4). The differences between SED-1

and TEM-1 or PSE-4 in block 8 account for most of the increased divergence between them.

Table V-3. Residue-Residue Contacts between Block Pairs and Within Each Block

Block	1	2	3	4	5	6	7	8
1	108	5	0	1	2	19	2	41
2	-	14	15	1	10	8	0	22
3	-	-	258	15	15	2	28	4
4	-	-	-	30	2	9	5	0
5	-	-	-	-	32	15	0	8
6	-	-	-	-	-	32	10	12
7	-	-	-	-	-	-	69	25
8	-	-	-	-	-	-	-	221

Residue-residue contacts between block pairs and within each block (diagonal entries). A residue-residue contact is defined as two amino acids that have any heavy atom, excluding the main chain N and O, within 4.5 Å.

Table V-4. Length and Sequence Identity between Each Pair of Parent Proteins for Each Block

		Sequence Identity		
Block	Length (aa)	PSE-4/TEM-1	TEM-1/SED-1	PSE-4/SED-1
1	40	28%	23%	28%
2	8	75%	63%	50%
3	76	38%	37%	39%
4	11	45%	36%	45%
5	15	60%	80%	60%
6	15	67%	73%	60%
7	27	30%	37%	26%
8	73	51%	30%	32%

The N- and C-termini of the lactamases have diverged significantly so that there are many substitutions in these regions. Analysis of these areas in a multiple sequence alignment shows that they are widely variable: many alignments in fact truncate the N-terminal helix because the sequence identity is nearly undetectable and the start of the mature protein is often uncertain (Bateman et al. 2004). Yet, inheriting residues at the N-

and C-termini from the same parent is almost essential to maintaining a functional protein. Using two different algorithms, Statistical Coupling Analysis (SCA) (Lockless and Ranganathan 1999) and McLachlan Based Substitution Correlation (Gobel et al. 1994; Olmea et al. 1999), to examine the evolutionary covariation of amino acids at the N- and C-terminal helices shows few or no significant interactions (2 or 5 in the top 1% (248) of possible interactions). This is surprising given our results and may indicate that strict covariation is not necessary. The detrimental effect of altering these residues has previously been shown in TEM-1. In a site-saturation study of TEM-1, 18 of the 30 residues which are variable in multiple sequence alignments of class A lactamases and invariable in TEM-1 are within blocks 1 and 8 (Huang et al. 1996). Fourteen of these residues are variable among the three parents studied here. Changing any one of these amino acids could potentially cause the protein to not function correctly. Despite the diversity of sequences at the N- and C-termini of lactamases and the apparent lack of covariation at the individual amino acid level, maintaining the contacts between these two blocks is nearly essential to maintaining a functional lactamase.

Intra-Block Interactions at Blocks 2 and 3 are Important for Function

In addition to the critical interaction of blocks 1 and 8, there are two intra-block variables that are important for determining chimera function. The more significant of these is block 3, where TEM-1 is favored in functional chimeras (Table V-2). TEM-1 at block 3 is found in more of the characterized chimeras than the other two parents due to the biased construction of the library, where 61% inherit this block from TEM-1 and only 17% inherit it from PSE-4 and 21% from SED-1. Because the LGA analysis takes into

account both functional and nonfunctional chimeras in determining the important contributions to chimera folding, this bias only affects whether or not we detect all of the significant variables, not the significance of the variables we do observe.

Block 3 is the largest segment and has the most internal structural contacts, both absolute number and per amino acid (Tables V-3 and V-4). Understanding why block 3 is so strongly favored is difficult due to its large size (76 amino acids) and the small number of functional chimeras (13) that do not have TEM-1 at block 3. There is a disulfide bond within block 3 in TEM-1 (Cys 77 to Cys 123). While this disulfide is not found in SED-1 (the residues are Ala and Ser), it is also present in PSE-4.

The second intra-block term that affects which chimeras are functional is the identity of the parent at block 2. At block 2 SED-1 is disfavored (only 20 of the 143 chimeras with SED-1 at block 2 are functional). In contrast to block 3, block 2 is the smallest segment, with only 8 amino acids, and incorporates at most 4 amino acid changes because the remaining 4 amino acids are conserved in all three parents (Figure V-4). Block 2 contains the active site residues Ser70 and Lys73, and altering any amino acids within it may have a large impact on the activity. SED-1 contains an Ala at position 67 and a Ser at position 72; these positions are Pro and Phe, respectively, in the other parents. In a site-saturation study of TEM-1, Pro67 was found to be invariable, despite the Ala found at this position in multiple sequence alignments of β-lactamases (Huang et al. 1996). In the same study, Phe72 allowed some variation. However, Ser was not one of the identified amino acids. Block 2 is a much more tractable target than block 3 for analyzing the basis for effects of one-body terms on chimera function.

Figure V-4. Sequence alignment of the three parents for block 2 shows only four differences between SED-1 and PSE-4 or TEM-1.

Biophysical Analysis of Block 2

To further investigate the significance of block 2, we examined all 20 functional chimeras that inherited this block from SED-1. The MICs of these chimeras are significantly lower than the MICs of the remaining 162 functional chimeras (Figure V-5). Chimeras with block 2 from SED-1 are not only less likely to confer resistance to ampicillin, but when they do they are impaired.

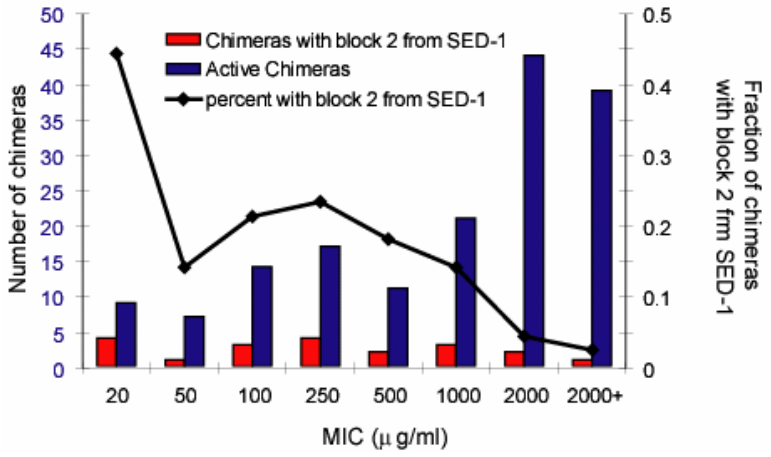


Figure V-5. Characterized functional chimeras with SED-1 at block 2 show less ampicillin resistance compared to the library as a whole.

We also identified sets of characterized functional chimeras that differ by only the parent at block 2 (Table V-5). Examination of all seven sets identified shows that the chimeras with block 2 from SED-1 always have a lower MIC than the same chimera inheriting block 2 from either TEM-1 or PSE-4. In some cases it is not a large difference (only 2-fold), but in many cases the effect is >10-fold (Table V-5).

Table V-5. Characterized Sets of Chimeras Differing Only by Block 2.

Chimera									MIC	T_m (°C)	K_m (μM)	k_{cat} (s^{-1})
1	S	T	T	P	T	T	S	S	4000	52	340 ± 20	450±160
2	S	P	T	P	T	T	S	S	4000	49	16 ± 1.5	280±60
3	S	S	T	P	T	T	S	S	2000	50	25 ± 4	160±39
4	T	T	T	P	T	P	T	T	4000	55	81 ± 1	1400±285
5	T	P	T	P	T	P	T	T	4000	50	12 ± 2.5	14±3
6	T	S	T	P	T	P	T	T	1000	45	25 ± 1.5	40±11
7	T	T	T	T	S	S	T	T	4000	55	168 ± 9	2900 ± 316
8	T	P	T	T	S	S	T	T	2250	49	5.5 ± 1	90 ± 22
9	T	S	T	T	S	S	T	T	1000	48	24 ± 4	60 ± 20
10	T	T	T	P	T	T	T	T	4000			
11	T	P	T	P	T	T	T	T	4000			
12	T	S	T	P	T	T	T	T	212			
13	T	T	T	P	S	T	T	T	4000			
14	T	P	T	P	S	T	T	T	4000			
15	T	S	T	P	S	T	T	T	20			
16	T	T	T	P	P	P	T	T	4000			
17	T	P	T	P	P	P	T	T	4000			
18	T	S	T	P	P	P	T	T	40			
19	T	P	T	T	T	T	T	T	4000			
20	T	S	T	T	T	T	T	T	200			
TEM-1	T	T	T	T	T	T	T	T	4000	55	268 + 49	700 ± 20
SED-1	S	S	S	S	S	S	S	S	4000	55	42 + 4.5	1050 ± 110
PSE-4	P	P	P	P	P	P	P	P	4000	55*		

Chimera sequences are represented by the parents each block is inherited from: P for PSE-4, S for SED-1, and T for TEM-1. Ampicillin MICs (μg/mL) were redetermined to increase fidelity and are not directly comparable with previously reported MICs. For those chimeras for which they were determined the T_m (°C), K_m (μM) and k_{cat} (s^{-1}) have been listed. *For PSE-4 the thermostability is as reported in the literature (Savoie et al. 2000).

The expression level for each of the 21 chimeric proteins was optimized to allow for purification. Analysis of periplasmic extracts shows that chimeras with SED-1 at block 2 have significantly less protein present in the periplasm (Figure V-6) than chimeras with a different parent at block 2. This indicates that a large part of the depressed MIC associated with SED-1 at block 2 may be due to low stability, poor

expression or inadequate transport to the periplasm. While these experiments were performed under high expression conditions, experiments under the screening conditions gave similar results based upon an activity assay performed with cell lysate.

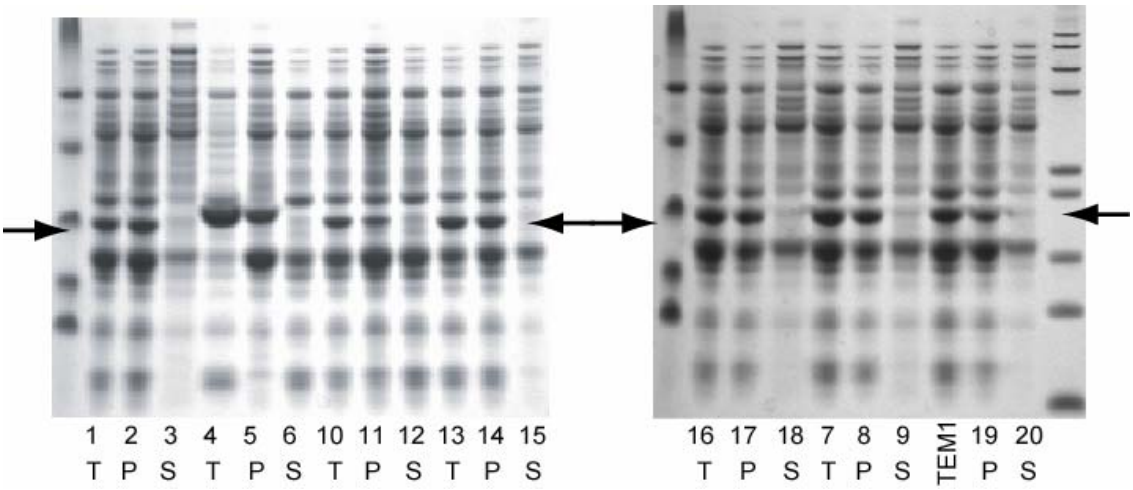


Figure V-6. Periplasmic extracts from chimeras grown under high protein expression conditions. The lactamase is ~30 KD and a band corresponding to that size (marked) is present in all chimeras that do not have SED-1 at block 2. The numbers correspond to chimeras listed on Table V-5, and the letter beneath indicates the identity of the parent at block 2: P for PSE-4, S for SED-1 and T for TEM-1.

The native signal sequence for each parent is included as part of the N-terminal block, and all of the parents are exported correctly to the periplasm. In the past it has been observed that a mutant lactamase can fail to reach the periplasm and become trapped in the cytoplasm when it is partially unfolded (Sideraki et al. 2001). Analysis of cell lysates and periplasmic extracts indicates that there is not a significant difference in activity between the whole cell lysates and the periplasmic extracts under normal expression conditions. However, there may be misfolded or inactive protein present in the cytoplasm. Western blots of whole cell lysates from cells grown under normal expression conditions using an antibody to TEM-1 show a similar pattern to the periplasmic extracts. Only six of the seven sets of chimeras can be examined in this way because the antibody

does not cross react with SED-1 sufficiently to detect chimeras (1, 2, and 3 from Table V-5) where blocks 1 and 8 are from TEM-1.

For three of the sets of chimeras, the members were purified to >95% purity. Kinetic studies were performed with the purified enzymes (Table V-5). The values of K_m and k_{cat} obtained for TEM-1 hydrolysis of ampicillin are consistent values reported in the literature (Schroeder et al. 2002). The chimeras are distinct from one another, but show no clear trend with regard to the identity of block 2. Circular dichroism spectra for TEM-1 and SED-1 and the chimeras were similar to that of PSE-4 (Savoie et al. 2000). T_m apparent was determined for each protein by observing ellipticity at 222 nm during a thermal melt from 1 to 99 °C. The transitions showed cooperative unfolding, and the T_ms agreed well with those determined from activity assays on cell lysates. These studies show that while there is variation in the thermostabilities of the enzymes, none of them are lower than ~45 °C (Table V-5). Thus, these proteins are probably not sufficiently unstable to cause the effect we observe at 37 °C. This is consistent with the fact that lowering the growth temperature to 20 from 37 °C does not have a large impact on the protein expression level. Additionally, adding the well-characterized M182T thermostabilizing mutation (Huang and Palzkill 1997) to four of the chimeras with SED-1 at block 2 (chimeras 3, 12, 15 and 20 from Table V-5) had no effect on the protein expression level, further indicating that the proteins are likely sufficiently stable to be expressed at 37 °C, although they do not accumulate in the cell.

SED-1 is strongly disfavored at position 2, and this effect appears consistent throughout characterized chimeras, even those that are similar to SED-1. The effect is most likely due to decreased expression of these chimeras compared to the parent

proteins. All chimeras with block 2 from SED-1 examined show very low levels of periplasmic protein, and the MICs of the remaining chimeras are consistent with low to no expression. While the expression levels are depressed, the thermostabilities of the purified proteins are not sufficiently low to cause the observed lack of expression. This result ties back to the GFP folding screen conducted in Chapter IV where several chimeras showed ampicillin resistance, but no significant GFP signal. Chimeras that confer resistance to ampicillin, and are stable when purified, do not necessarily accumulate in the cell. It does not take very much of an active lactamase to confer resistance to ampicillin, especially in our low stringency screen. The chimeric proteins which confer resistance but do not accumulate may aggregate or be broken down by the cell for some other reason besides low thermostability.

Despite the effect of block 2 on chimeric proteins, SED-1 itself is well expressed and the codons present in this block are not particularly rare in *E. coli*. It is possible the bias against block 2 has its origins at the mRNA level. However, because the mRNAs are large there is unlikely to be a specific change in RNA folding due to alteration of just four codons. It is also possible that there is a deleterious interaction between block 2 and some other region of the protein that causes the effect. We have not isolated any such interaction with these analyses, but the data are limited. Very few functional chimeras contain block 2, making identification of such an interaction more challenging.

Conclusions

Analysis of chimeric β-lactamases has allowed us to narrow the possible regions of sequence responsible for CTX-M lactamase altered substrate specificity. Nonconserved residues within the ω-loop are not likely contributing to altered substrate specificity because they can be inherited from proteins which do not share this property. Altered substrate specificity is also not likely tied solely to specific amino acid interactions with the substrate because a sequence block distant from the active site is necessary to confer altered substrate specificity.

We have also used the functional and nonfunctional chimeric β-lactamases to inform us about which regions of sequence space might be populated with additional lactamases. All of the clusters observed in Chapter IV (Figure V-1) fall within the areas of sequence space that are compatible with the one-body and two-body terms identified as favorable for producing a functional chimera by LRA. They all have TEM-1 at block 3, and the N- and C-termini originate either from the same parent, or PSE-4 is at the N-terminus and TEM-1 at the C-terminus. However, there are other smaller clusters of functional chimeras that were not originally detected that are likely underrepresented only because more chimeras were not characterized, not because those areas are incompatible with functional chimeras. Examining the critical interactions found in the chimeric β-lactamases shows how the regions of sequence space that functional chimeras populate are limited by specific pairwise interactions. Additionally, portions of sequence that do not appear to interact strongly with other parts of the protein can limit which chimeras function. Why exactly these portions of sequence are so deleterious or

advantageous is still not clear. However, these sequence portions are not necessarily thermostability limiting.

Methods

Cefotaxime Activity Screen

All unique sequences were inoculated from glycerol stocks and used to inoculate 96 deep-well plates which were grown to saturation as for the ampicillin resistance assays in Chapter IV. Aliquots were transferred onto LB agar plates containing various concentrations of cefotaxime (0.05 to 1 μg/mL) similarly to the ampicillin assay previously described (Chapter IV).

Logistic Regression Analysis

Logistic regression assumes that the probability of a chimera folding decreases with energy E according to the sigmoidal relationship

$$Pf = \frac{1}{1+e^{E}}, \tag{V-2}$$

We defined E as the sum of one- and two-body terms in Equation (VI-1). The significance of each term was calculated relative to a reference model that included only the one-body terms using the maximum likelihood test (Endelman et al. 2005). The individual one-body terms were removed from the model and the increase in deviance D measured.

$$D = -2\sum_{i=N_f+1}^{N} E_i + 2\sum_{i=1}^{N} \ln(1+e^{E}) \tag{V-3}$$

The magnitude of this increase follows the chi-square distribution with two degrees of freedom, which was used to calculate a p-value for each one-body term. The significance of two-body terms was determined by recording the decrease in the deviance and the p-value determined from the chi-square distribution with four degrees of freedom. The algorithm MINOS through the NEOS server was used for optimization. The GAMS input file necessary for this computation can be found in Appendix I.

Sequence Analysis

Evolutionary covariation between amino acids was examined using both Statistical Coupling Analysis and McLachlan Based Substitution Correlation. Java code for both of these algorithms was downloaded from http://www.afodor.net/ (Fodor and Aldrich 2004b, 2004a), and the full PFAM lactamase superfamily alignment used for calculation (Bateman et al. 2004).

Protein Purification

With the exception of A1A3, A1H6, and A2A4, proteins were expressed in 2x 1L cultures of TB +35 μg/mL Chl grown to saturation at 37 °C, 250 rpm. The remaining proteins were expressed in 6 x 500 ml TB cultures with the addition of 20 ng/mL anhydrotetracycline (inducer) to maintain a high expression level and grown 48 hours at 25 °C. The cells were pelleted (8000 xg, 8 min, 4C) and the periplasmic proteins isolated through osmotic shock. The cells were resuspended gently in 200 mL 30% sucrose, 30 mM Tris pH 8.0, 1 mM EDTA and allowed to incubate at room temperature for 10 minutes before repeating centrifugation as above. The supernatant was removed

completely and the cells were resuspended in 200 mL ice cold water with shaking and vortexing. Following a 10 minute incubation on ice the cells were centrifuged to pellet the cells (30 min, 15,000g, 4 °C). The supernatant was removed as the periplasmic extract and either dialyzed overnight at 4 °C against 20 mM Tris pH 8.0, or 1 M Tris pH 8.0 was added to a final concentration of 20 mM. SED-1 was dialyzed overnight to 20mM HEPES pH 7.0.

The buffered periplasmic extract was applied to a Q FF HP (Amersham) column and washed with 20 column volumes of 20 mM Tris pH 8.0. The protein was eluted in a gradient of 0-200 mM NaCL over 12 column volumes. Fractions were tested for lactamase activity using nitrocefin and purity assessed through SDS gel electrophoresis. The purest fractions were collected and concentrated using a Millipore Centriprep (YM10) to 0.7 mL and then applied to an S-100 gel filtration column (Amersham) run at 0.15 mL/min in 30 mM Tris pH 8.0. Fractions were tested for activity as above and purity verified through SDS gel electrophoresis.

SED-1 has an isoelectric point around 8 and therefore was purified using cation exchange chromatograhpy. SED-1 periplasmic extract was buffered in 20 mM HEPES pH 7.0, applied to 3 SP FF (Amersham) columns in series and washed with 20 column volumes of 20 mM HEPES pH 7.0. The protein was eluted in a gradient from 0 to 200 mM NaCl over 12 column volumes. The fractions were assayed for activity using nitrocefin and the purity of active was verified by gel electrophoresis. No gel filtration was necessary to obtain >95% purity.

MIC Determination

The MICs for Figure V-5 were determined in Chapter IV during the high-throughput screening. However, MIC determination for sets of chimeras in Table V-4 was repeated to ensure higher fidelity. These MIC's were determined through liquid culture dilutions rather than the plate assay. 500 uL cultures of LB with 35 ug/mL chloramphenicol were grown in deep-well 96-well plates at 37 °C for 18 hours and then diluted 1:1000. 10 uL was used to inoculate 96 well culture plates containing 90 uL of LB with varying concentrations of ampicillin (0, 2.5, 5, 10, 25, 50, 100, 250, 500, 1000, 2500, 5000, 10000 μg/mL). The cultures were grown for 18 hours at 37 °C and the OD_{590} measured. Cultures with an $OD_{590} > 0.1$ were considered grown.

Site-Directed Mutagenesis

The M182T mutation was introduced into chimeras 3, 12, 15 and 20 using quick-change mutagenesis with the following primer and its reverse complement: 5'-CGT GAC ACC ACG ACC CCT GTA GCA ATG G. The altered codon is underlined. The genes were sequenced in both directions to verify correct incorporation of the mutation and no additional mutations.

CD Spectroscopy

Purified proteins were diluted to 30 μM in KPO_4 buffer pH 7.0. Protein concentrations were determined by measuring absorbance at 280 nm. To verify the presence of a folded lactamase, circular dichroism wavelength scans from 200 to 250 nm at 1 nm increments with a 5 second averaging time were performed on a JASCO model J-

600 spectropolarimeter. To determine the apparent T_m, ellipticity was monitored at 222 nm during a thermal denaturation from 0 to 99 °C. The step size was 1 °C, the equilibration time 2 minutes, and the signal averaging time 30 seconds.

Catalytic Activity Assays

Enzyme kinetics were measured at 25 °C in 100 mM KPO_4 buffer pH 7.0. Degradation of ampicillin was measured by UV-Vis at 232 nm, ε_{232} for ampicillin is 912 $cm^{-1}M^{-1}$. Ampicillin concentrations between 10 and 500 μM were tested and protein concentrations ranged between 0.25 and 12 μM as appropriate to record a linear initial rate. Rate constants were fit using a Hanes-Woolf plot ([S]/v vs. [S]).

Rates were measured to compare whole cell lysates and periplasmic extracts using the chromagenic substrate nitrocefin at 25 °C in 100 mM KPO_4 buffer pH 7.0, 50 mg/mL of nitrocefin. Nitrocefin degradation to form a red product (ε_{486}=20,500) was measured by observing at 468 nm. A similar assay was performed at varying temperatures to estimate the T_m for comparison with CD measurements.

Chapter VI: Mutagenesis to Restore Chimera Function

Introduction

Identifying characteristics of functional and nonfunctional chimeras is one way to address the underlying reasons for why chimeric proteins are nonfunctional. Another approach to this question is to determine if nonfunctional chimeras can be rescued through mutagenesis. Given enough of the right mutations, any chimera can be rescued as it returns to a wild-type sequence. However, whether nonfunctional chimeras are only a few or many mutations away from functional sequences is unknown.

There are many low *E* chimeras that are nonfunctional. It is unknown whether these chimeras are nonfunctional due to specific deleterious interactions, general lack of stability, or some other unknown factors. The specific mutations responsible for rescuing chimera function can indicate whether particular broken contacts are critical for function or chimeras are generally destabilized. Specific mutations that only rescue one or a few chimeras likely are responsible for correcting specific broken contacts. Mutations that rescue many chimeras and seem independent of specific sequences are likely global stabilizers (Poteete et al. 1997).

It is also unclear if all nonfunctional chimeras are equally distant in sequence space from functional sequences or if some chimeras are more likely to be rescued through mutagenesis. Chimeras that have low *E* are much more likely to function than chimeras with high *E*. It is possible that nonfunctional low *E* chimeras are closer in sequence space to functional sequences and may be easily rescued using random mutagenesis. To address these questions we randomly mutated nonfunctional chimeras to

examine which ones regain function, and what mutations are responsible for restoring function. The known TEM-1 stablizing mutation M182T was identified in half of the rescued chimeras. More thermostable proteins are more robust to random mutagenesis (Bloom et al. 2005b). To investigate whether this was true for mutations introduced by recombination, we introduced M182T into randomly selected chimeras and estimated the proportion of the library that might retain function if a more thermostable parent had been used.

Results

Random Mutagenesis Rescues Lactamase Chimera Function

To determine if chimeras could be rescued by random mutagenesis, DNA from all of the nonfunctional chimeras identified in Chapter IV and listed in Appendix II was combined. To ensure that no DNA from active chimeras was present, the collected DNA was transformed into *E. coli* and the transformants selected on ampicillin to verify that no colonies were produced. Following this verification, the DNA was PCR amplified under error-prone conditions as a single pool. The PCR products were cloned back into the expression vector and selected on ampicillin. Many ampicillin resistant clones were identified. However, sequencing these clones revealed that they were either known functional sequences, or functional sequences that had not been previously characterized. Apparently, during the mutagenic PCR, recombination similar to DNA shuffling occurred and scrambled the chimeras, making this strategy of mutating the whole set of nonfunctional at once unusable.

As an alternative, 10 inactive chimeras were subjected to error-prone PCR individually. The chimeras were chosen based on their hypothesized likelihood of being rescued: they all have low *E,* and the N- and C-termini originate from the same parents (Table VI-1A). Of the 10 chimeras, 8 were rescued by at least one single mutation (Table VI-2). There are 132 chimeras with *E* less than 30, of which 46 are nonfunctional. The ease with which the selected chimeras were rescued indicates that it is likely that many low *E* chimeras can be rescued similarly. For one of the two chimeras not rescued, none of the mutations identified in other chimeras can be incorporated because they are not found in the chimera due to differing parent blocks. For the other chimera only the M182T mutation is possible.

Table VI-1. Randomly Mutated Chimeras

A									E	m
S	T	S	P	T	S	S	S	*	15	12
T	S	T	S	S	T	T	T	*	15	13
T	T	T	S	S	T	P	T	*	20	29
P	T	T	T	P	T	T	P		20	71
S	T	T	T	T	T	S	S	*	20	64
S	P	T	T	P	S	T	S	*	22	81
T	P	S	P	T	T	T	T		22	55
S	P	P	T	T	S	T	S	*	28	76
P	S	T	T	S	T	T	P	*	29	71
S	S	T	T	T	P	T	S	*	30	80

B								E	m
P	T	S	T	P	S	T	T	45	86
P	S	T	T	S	S	T	S	50	89
T	S	T	P	S	T	T	S	53	62
T	P	T	T	T	T	T	S	53	52
P	P	T	T	T	S	T	S	54	85
T	T	T	P	T	T	T	S	55	56
P	T	P	T	T	T	T	S	59	87
P	P	T	S	S	T	P	S	59	103
T	T	T	P	T	P	S	S	61	78
P	S	P	S	T	P	T	S	62	84
T	P	P	P	P	S	P	S	67	84
T	P	S	P	T	T	T	S	71	65

Chimeras rescued by random mutagenesis are marked by an *. The sequence of a chimera is represented by the parents it inherits its blocks from: P for PSE-4, T for TEM-1, and S for SED-1. A: The initial set of chimeras chosen for their low *E* values that were randomly mutated to see if chimera function could be restored. B: The second set of chimeras chosen with higher *E* values to examine whether chimeras of any *E* could be rescued by random mutagenesis.

To explore whether the ease with which chimeras were rescued is a general property of all chimeras or due to the optimized population chosen, an additional 12

chimeras were chosen at high levels of *E* for mutagenesis (Table VI-1B). None of these chimeras were rescued. To ensure that this result was not due to sparse sampling of the possible mutants, the libraries were over sampled by ~10-fold.

Table VI-2. Mutations that Rescue Nonfunctional Chimera

Block			**Signal Seq**			**1**	**2**	**3**					**4**	**5**		**6**	**7**		**8**	
amino acid residue			8	22	27	63	72	99	100	114	120	147	153	171	174	182	191	193	224	261
Sequence	E	M																		
PSE-4	0	0				N	F	K	A	G	D	G	R	E	L	T	N	F	V	V
SED-1	0	0	Q		H	E	S	K	A	G	A	N	R	T	P	S	R	L	G	L
TEM-1	0	0		F		D	F	Q	N	T	R	E	H	E	P	M	R	L	A	V
STSPTSSS	15	12					**S**													
			L				**S**													
TSTSSTTT	15	13														***T***				
														S		***T***				
TTTSSTPT	20	29														***T***				
								R								***T***			T	
												G	**G**				Y			
																		L		
				L						S								**L**		
PTTTPTTP	20	71														***T***				
SPTTPSTS	22	81													**P**					
															P					**A**
						G									**P**					
SPPTTSTS	28	76	***L***		L								R							
PSTTSTTP	29	71														***T***				
SSTTTPTS	30	80	***L***																	
			L									**G**								
									S											**A**

Only unique sequences are shown, and mutations appearing alone in a chimera or in more than one chimera are shown in bold.

Several Mutations Can Rescue Function

Table VI-2 shows a list of the mutations that rescue each chimera; only unique sequences are shown. About half of the mutations mutate a single amino acid to an amino acid found in one or more 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 be correcting an interaction that was mismatched in the chimera.

The mutations that change the amino acid present to the amino acid present in a different parent sequence are: F72S (block 2 from TEM-1), E147G (block 3 from TEM-1), H153R (block 4 from TEM-1), L174P (block 5 from PSE-4) and M182T (block 6 frm TEM-1). Some of these positions have been previously characterized. H153R and M182T in TEM-1 not only revert to the amino acid found in both PSE-4 and SED-1, but also are known stabilizing mutations frequently identified in extended-spectrum TEM-1 variants (Knox 1995). The remaining residues have not been explicitly characterized, but all of them were found to be variable in a site-saturation study of TEM-1 (Huang et al. 1996). Examining the specific contacts that may be restored by a reversion shows that F72S and F193L both decrease the *E* of the chimeras by 1 or 2 contacts, respectively, and that L174P increases the *E* by 1 contact. From these limited studies it is not clear whether these mutations are likely to rescue many chimeras or are limited to specific sequences. Many of the mutations were isolated in only one chimera. While they are usually possible in at least one other chimera tested, it is unknown whether they rescue function in other chimeras.

There are two mutations which rescue several different chimeras. TEM-1 M182T rescues 4 of the 8 chimeras, and SED-1 Q8L rescues 3 of 8 chimeras. Of the rescued chimeras, M182T is identified in every one that has block 6 from TEM-1. The only chimera tested for which this mutation was possible and not identified was not rescued by any mutation. M182T was also identified in all four rescued chimeras as a single mutation. TEM-1 M182T is a well characterized mutation commonly found in extended-

spectrum TEM-1 variants (Knox 1995). It has been shown to mediate the effects of other deleterious mutations by increasing the stability of the protein by 2.7 kcal/mol (Wang et al. 2002). This mutation most likely has a similar effect on the chimeric proteins, providing them with enough additional stability to fold correctly. While PSE-4 already has a methionine at this position, previous studies and the widespread appearance of M182T here indicates that it is likely a global stabilizing mutation rather than correcting specific broken interactions.

The second mutation isolated from several different chimeras is the SED-1 Q8L mutation. This mutation appears alone in one of the three chimeras. It is accompanied by one additional mutation in one chimera and two additional mutations in the third chimera. Interestingly this mutation is in the signal sequence of the protein and not part of the mature protein. SED-1 is much less well characterized than PSE-4 or TEM-1 and there is no protein sequencing data or crystal structure currently available to give a definitive starting residue for the mature protein. However, the hypothesized start of the mature protein based on multiple sequence alignments is significantly further into the protein sequence than Q8. Why exactly this mutation rescues activity is not currently known. SED-1 was originally cloned from *Citrobacter sedalaki,* and it is possible that the signal sequence presumably optimized for this organism may be less efficient at transporting the protein to the periplasm in *E. coli*. However, transport of wild-type SED-1 to the periplasm appears normal (see Chapter V). A mutation in the signal sequence rescuing activity has not been previously observed in lactamases, and this brings our attention to the potentially key role of intracellular transport in an *in vivo* viability based screen or selection.

TEM-1 M182T Can Increase the Fraction of Folded Chimeras in the Library

Because M182T effectively rescued a high percentage of chimeras, it was introduced into 31 randomly chosen nonfunctional chimeras that have TEM-1 at block 6. Of the 31 randomly chosen chimeras, four were rescued by this single mutation (Table VI-3). Chimeras with low *E* are more likely to be rescued (Figure VI-1). All of the chimeras rescued have $E<35$, and they also all have the N- and C-termini from the same parent.

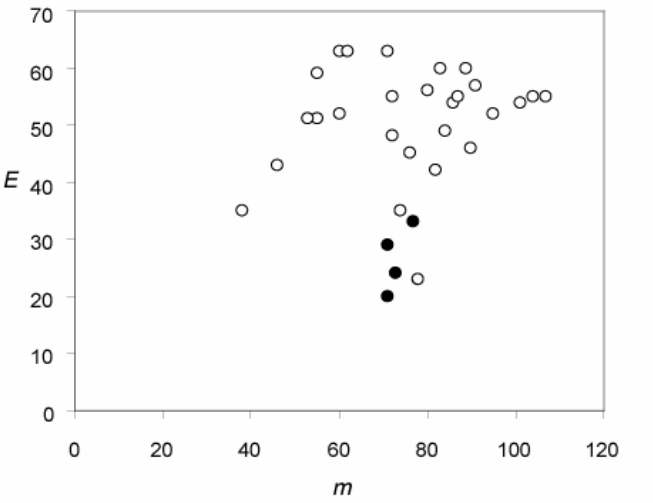


Figure VI-1. Rescued chimeras (solid points) are more likely to have low *E* than chimeras not rescued (open points) by M182T.

To estimate to effect of adding the M182T mutation to all chimeras in with TEM-1 we first calculated the probability of rescuing function P_{rescue} with respect to *E* for the small test set. This was done by fitting the 31 data points to a function of the form

$$P_{rescue} = \frac{1}{c + e^{bE+a}}, \qquad \text{(VI-1)}$$

where a, b, and c are parameters fit with the following constraints: $b \geq 0$, $0 \leq c \leq 1$. This function allows sigmoidal (c=1), exponential (c=0; a=0) and intermediate forms. For the

test set a = -4.8, b = 0.156, c = 1.0, and this fit corresponds well with P_{rescue} calculated for binned data (Figure VI-2). Using this function we calculated the probability of rescuing the remaining nonfunctional chimeras inheriting block 6 from TEM-1. Summing these probabilities shows that approximately 27 of the 442 nonfunctional chimeras (184 with block 6 from TEM-1) are likely to regain function if M182T was present in every chimera in the library. At low *E* nearly all chimeras should fold if the M182T mutation had been incorporated into the library (Figure VI-3A). However, the potentially rescued chimeras are spread over a wide range of *m* levels. Examining the extrapolated effect on fraction functional with respect to *m* shows that there are chimeras with high *m* that would likely function if TEM-1 M182T had been used rather than the wild-type TEM-1 (Figure VI-3B).

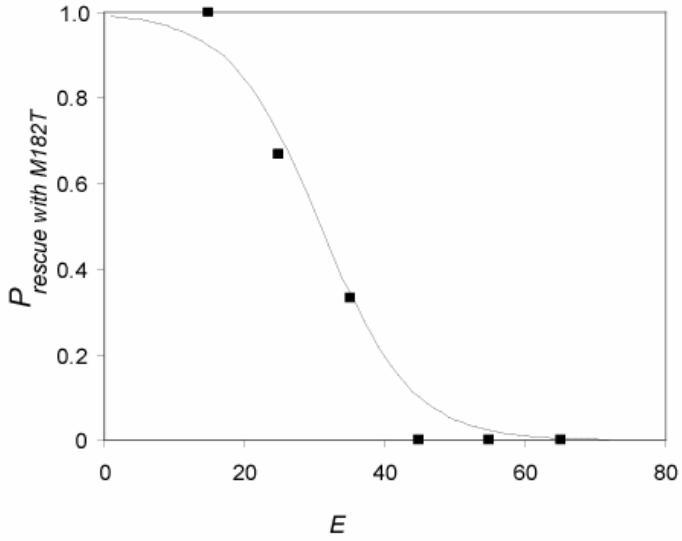


Figure VI-2. Probability of M182T rescuing function with respect to *E*. The points represent the fraction of chimeras rescued by M182T in a bin of 10 *E*. The curve is the fit of the individual data points (Figure VI-1) to Equation (VI-1).

Table VI-3. Randomly Chosen Chimeras M182T was Introduced Into

									E	m
P	T	T	T	P	T	T	P	*	20	71
S	S	T	T	T	T	T	S		23	78
P	T	T	P	P	T	T	P	*	24	73
P	S	T	T	S	T	T	P	*	29	71
P	S	T	P	S	T	T	P	*	33	77
P	T	P	T	S	T	T	T		35	74
P	S	T	T	P	T	T	T		35	38
P	T	S	T	P	T	T	T		42	82
P	T	T	T	T	T	S	T		43	46
P	S	P	S	T	T	T	T		45	76
P	P	T	T	S	T	S	S		46	90
T	P	T	T	S	T	S	S		48	72
P	P	T	T	S	T	T	S		49	84
P	S	T	P	T	T	S	T		51	55
T	S	T	T	T	T	T	S		51	53
P	P	P	P	T	T	P	S		52	60
P	P	T	P	P	T	S	S		52	95
P	S	T	T	T	T	P	S		54	101
P	P	T	S	T	T	S	S		54	86
T	T	T	T	S	T	P	S		55	72
P	P	T	P	T	T	T	S		55	87
P	S	T	P	T	T	P	S		55	107
P	T	T	P	T	T	P	S		55	104
S	T	S	S	P	T	T	T		56	80
P	P	T	S	S	T	T	S		57	91
T	S	S	S	T	T	T	S		59	55
P	S	P	P	T	T	T	S		60	83
P	S	P	T	T	T	T	S		60	89
P	P	S	T	S	T	T	S		63	60
P	S	P	T	P	T	S	T		63	71
P	T	S	T	T	T	T	S		63	62

Randomly chosen chimeras M182T was introduced into, they all inherit block 6 from TEM-1. The sequence of a chimera is represented by the parents it inherits its blocks from: P for PSE-4, T for TEM -1 and S for SED-1. Rescued chimeras are marked with an *.

Overall, the effect of adding M182T to the entire library is significant but not enormous. For the characterized library the increase in overall fraction folded is about 5%. However, because TEM-1 is found more frequently at block 6 than the other parents (Chapter IV) (>60% of characterized chimeras inherit block 6 from TEM-1) this effect is magnified compared to a random population of protein chimeras.

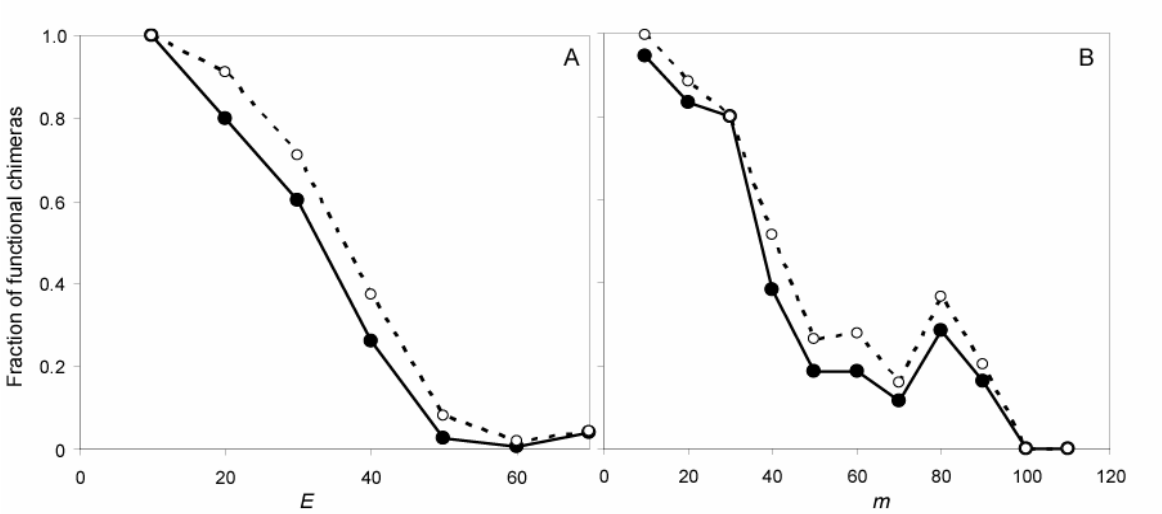


Figure VI-3. A: The fraction of functional chimeras in the library (solid points) and the fraction of the library folded if all possible library members contained M182T (open points) with respect to *E*. B: The fraction of functional chimeras in the library (solid points) and the fraction of the library folded if all possible library members contained M182T (open points) with respect to *m*. See methods for details.

Discussion

Many chimeras can be rescued with a single point mutation, either a global stabilizing mutation or a mutation that may correct specific broken contacts. Chimeras with low *E* are much more likely to be rescued than chimeras at high *E*, even if the chimera has many mutations to the closest parent (high *m*). This indicates that chimeras with low *E* are in an area of sequence space that is densely populated with folded proteins. Chimeras with low *E* are more likely to retain function and fold than chimeras at

higher *E*, and nonfunctional, low *E* chimeras are much closer to sequences that do encode folded functional proteins.

We have recently shown that more thermostable proteins are more tolerant to random mutations (Bloom et al. 2005b) and therefore can have a greater capacity to evolve (Bloom et al. 2006). Most mutations, beneficial or not, are destabilizing. More stable proteins are more likely to withstand a destabilizing mutation to fold correctly so that the phenotypic effects of that mutation are revealed. We have shown here that more thermostable proteins are likely more robust to mutations introduced through recombination as well as to randomly introduced mutations. Identifying mutations that increase thermostability indicates that starting with stabilized parents should increase the fraction of folded chimeras identified. This suggests that another way to increase the final fraction of folded variants in a recombination library is to begin with stabilized parent sequences.

Methods

Random Mutagenesis

DNA for inactive chimeras was sequenced prior to mutagenesis to confirm that no mutations were present in the chimera. Error-prone PCR was performed on each chimera in the following 100 uL reaction: 3 ng template, 1 μM forward and reverse primers listed on Table AII-3, 7mM $MgCl_2$, 75 μM $MnCl_2$, 200 μM dATP and dGTP, 50 μM dTTP and dCTP, 1x Applied Biosystems PCR buffer without $MgCl_2$ and 5 U of Applied Biosystems *taq* polymerase. Reactions were heated to 95 °C for 5 minutes then 14 cycles of 30 seconds at 95 °C, 30 seconds at 55 °C and 1 minute at 72 °C. PCR products were

digested with KpnI and PstI , cloned into pProTet (Clontech) cut with the same enzymes and transformed into XL-1 Blue (Stratagene).

Transformed *E. coli* were plated onto selective medium (35 ug/mL chloramphenocol and 10 ug/mL ampicillin) to identify sequences conferring resistance to ampicillin. A control aliquot was plated onto nonselective medium (35 ug/mL chloramphenicol) in order to assess how many chimeras were present in the selected sample. Colonies present on selective plates after 18 hours of growth at 37 °C were picked and the DNA extracted. The DNA was sequenced to identify mutations and retransformed into *E. coli* to verify plasmid conferred resistance. If no colonies were present on selective plates, 10 colonies were picked from nonselective plates to determine if the insert incorporation frequency, and typically 5 were sequenced to verify successful mutagenesis. A minimum of ~200,000 colonies were examined for each chimera not rescued. For rescued chimeras, typically many positive colonies were identified in much smaller libraries (~20,000 colonies).

Site-directed Mutagenesis

DNA for inactive chimeras was sequenced prior to mutagenesis to confirm that no mutations were present in the chimera. The TEM-1 M182T mutation was introduced using quick-change mutagenesis 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 XL-1 Blue (Stratagene) and plated onto selective and nonselective media as described above. Colonies growing on selective media after 18 hours at 37 °C were picked and the DNA extracted for

sequencing. For chimeras for which no colonies appeared on selective plates, 2 colonies were picked from the nonselective plates and the DNA extracted for sequencing to determine if the mutation was properly incorporated.

Extrapolation of Test Set to Library

The probability of M182T rescuing chimera function P_{rescue} was calculated by fitting the 31 data points to a function described by Equation (VI-1). This probability was applied to nonfunctional chimeras that inherited block 6 from TEM-1 (M182T is possible). To construct the figures, functional and nonfunctional chimeras in the naïve library (Appendix III) were counted for bins of 10 *E,* or 10 *m.* The point plotted for the naïve corresponds to the $N_{functional}/N_{total}$ for each bin. The point plotted for M182T added to the naïve library is $(N_{functional} + \Sigma P_{rescue})/N_{total}$, where ΣP_{rescue} is the sum of the probability of rescue for all chimeras within the bin.

Chapter VII: The Accuracy of SCHEMA Predictions of Chimera Folding on Different Protein Scaffolds

Introduction

The challenge of computationally predicting chimeric protein folding and function has produced several different energy functions specifically designed to score a chimeric protein's likelihood of folding (Voigt et al. 2002; Moore and Maranas 2003; Saraf and Maranas 2003; Saraf et al. 2004; Hernandez and LeMaster 2005) of function (Saraf et al. 2004). However, these energy functions are typically tested with only a few protein chimeras, or using chimeras derived from directed evolution experiments (Voigt et al. 2002). For chimeras derived from directed evolution experiments, the lack of characterization of the naïve populations makes it difficult to determine if the trends observed in identified chimeras are a result of the functional selection, or trends within the naïve population. The larger and better characterized populations of chimeras used to test energy functions tend to only include chimeras with a single crossover (Moore and Maranas 2003; Saraf et al. 2004). This results in a very limited pool of test cases that are all somewhat similar to one another. Additionally when a single crossover is allowed, the chimeras generated are a very specific type of chimera where the N- and C-termini always originate from different proteins and crossovers are generally more disruptive of folding as they move closer to the center of the protein. Using such chimeras it is difficult or impossible to assess the effects of noncontiguous protein portions inherited from the same parent, and thus the energy function's ability to predict folding for chimeras inheriting noncontiguous pieces from the same parent is questionable.

We have used the SCHEMA energy function (*E*) proposed by Voigt et al (Voigt et al. 2002) to design site-directed recombination libraries that have a large fraction of folded variants, using distantly related parental proteins (Chapter IV and Appendix III) (Otey et al. 2006). This energy function takes into account the three-dimensional structure of the protein and the sequence identity between the parents. By characterizing large numbers of both functional and nonfunctional chimeras in these libraries we have created large data sets that can be used to evaluate energy functions for predicting chimera folding. Additionally the chimeras produced from these data sets typically have several recombination sites, allowing noncontiguous portions of the sequence to occur from the same parent.

Two libraries have been characterized, one made with β-lactamases (Chapter IV), and one with cytochromes P450 (Otey et al. 2006) (Figure VII-1). The lactamase library was made by recombining eight sequence blocks from three β-lactamases (TEM-1, PSE-4 and SED-1) for a maximum size of 3^8 or 6,561 chimeras. The parental proteins are approximately 260 amino acids long and share ~40% sequence identity. From this library, 553 chimeras were characterized, 20% (111) of which confer resistance to ampicillin in a low stringency screen. On average the functional chimeras contain 46 amino acids substitutions relative to the closest parent sequence (see Table VII-1). The cytochrome P450 library recombined three proteins sharing approximately 65% sequence identity (CYP102A1, CPY102A2, and CYP103A3 known as A1, A2, and A3) to create a library the same size as the lactamase library (6,561 sequences). The cytochrome P450 heme domains are larger than the lactamases, with ~460 amino acids. Of the 628 characterized cytochromes P450, 45% (285) of the cytochrome P450 chimeras

incorporate the heme cofactor and thus fold correctly. The folded cytochrome P450 chimeras contain on average 67 amino acid substitutions to the closest parental sequence (see Table VII-1).

Table VII-1. Comparison of cytochrome P450 and lactamase library chimera properties

	Lactamase	Cytochrome P450
Number of chimeras	553	628
$\langle m \rangle$	66 ± 24	70 ± 18
$\langle E \rangle$	44 ± 17	32 ± 10
Number of folded chimeras	111	285
$\langle m \rangle_{\text{folded}}$	46 ± 28	67 ± 9
$\langle E \rangle_{\text{folded}}$	23 ± 12	29 ± 10
$\langle m \rangle_{\text{unfolded}}$	71 ± 20	72 ± 6
$\langle E \rangle_{\text{unfolded}}$	47 ± 12	34 ± 9

More than 73% of folded cytochromes P450 are catalytically active peroxygenases, indicating that the majority of sequences that fold correctly are active enzymes (Otey et al. 2006). Due to the sensitivity of the ampicillin resistance screen and the evidence that folded proteins are likely to have catalytic activity, it is likely that the majority of folded lactamase chimeras confer resistance to ampicillin. In this study we will consider the lactamase chimeras conferring ampicillin resistance as folded, and those that do not as not folded.

While the two libraries share many characteristics, they were constructed with proteins that have very different properties, including size, sequence identity and scaffold shape (Figure VII-1). In this work we ask the following questions of each data set: 1) How well does SCHEMA predict chimera folding? 2) How sensitive are predictions to the structural information incorporated? Asking these questions of multiple protein scaffolds with chimeras containing multiple crossovers allows us to determine whether

the energy function and its parameters apply to one specific protein or library choice or if they are likely to be generally applicable to protein chimeras.

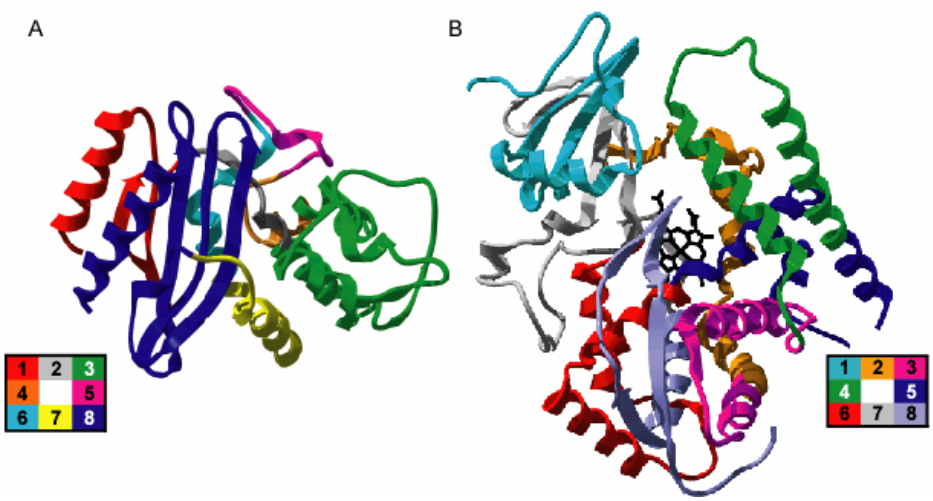


Figure VII-1. The three dimensional structures of A: β-lactamase chimera parent proteins (TEM-1, 1BTL) and B: cytochrome P450 parent proteins (CYP102A1, 1JPZ) with the independently exchangeable sequence blocks mapped to the structures. For lactamases the crossovers are after the following TEM-1 residues: Arg65, Lys73, Thr149, Arg161, Asp176, Leu190 and Gly218. For cytochromes P450 the crossovers are after CYP102A1 residues Glue64, Ile122, Tyr166, Val216, Thr268, Ala328, and Glu404.

Results and Discussion

Comparison of Cytochrome P450 and Lactamase Chimeras

In both the lactamase and cytochrome P450 libraries, chimeras with lower *E* are more likely to retain their fold. The <*E*> of all chimeras in the lactamase library 44 ± 17, and the <*E*> of folded chimeras is 23 ± 12. For cytochromes P450 the same is true, although the effect less pronounced. The <*E*> of all library chimeras is 32 ± 10, while the <*E*> for folded chimeras is 29 ± 10. Examining the spread of folded and unfolded chimeras over the <*m*> vs. <*E*> plot shows that, for both libraries, folded chimeras are spread over a large range of *m* levels. Although for lactamases, there is a significant trend

toward low *m* in folded chimeras (Figure VII-2). Examining the *E* vs. *m* distributions for lactamases and cytochromes P450s shows that the populations of folded and unfolded chimeras are better separated with respect to *E* for the lactamases (Figure VII-2).

The differences between lactamase and cytochrome P450 chimera folding with respect to *E* can be observed more clearly by calculating the probability of retaining fold (P_f) as a function of *E*. To accommodate both exponential and sigmoidal behaviors (Voigt et al. 2002; Meyer et al. 2003) we fit the folding data using maximum likelihood to a function of the form

$$P_f = \frac{1}{c + e^{bE+a}}, \qquad \text{(VII-1)}$$

subject to the constraints $b \geq 0$, $0 \leq c \leq 1$ which allows exponential (c=0), sigmoidal, (c=1, a=1) and intermediate behaviors (Figure VII-3).

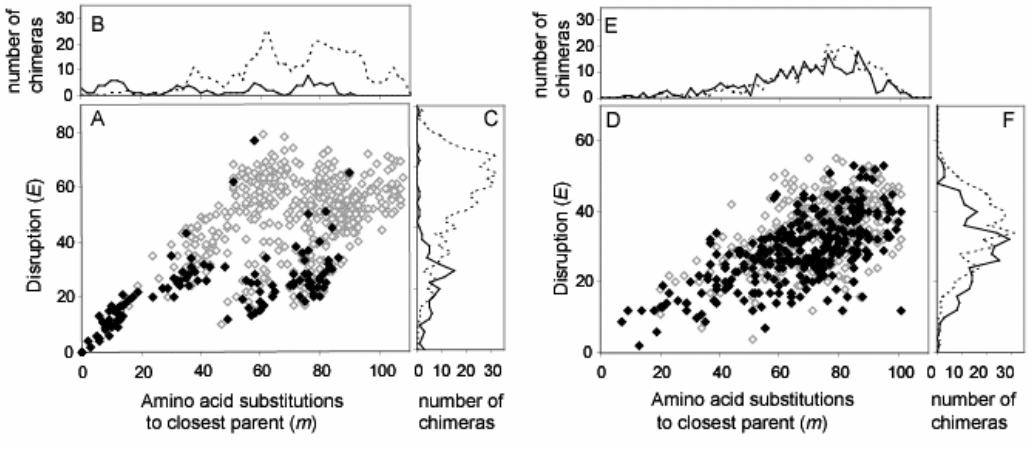


Figure VII-2. The *E* and *m* distributions for β-lactamase (A-C) and cytochrome P450 (D-F) chimeras. A, D: *E* vs. *m*, for unfolded chimeras (open points) and folded chimeras (solid points). B, E: Distribution of folded (solid line) and unfolded (dashed line) chimeras with respect to *E*. C, F: Distribution of folded and unfolded chimeras with respect to *m*. β-lactamase chimeras show a good separation between folded and unfolded chimeras. The naïve data sets of both cytochromes P450 (Appendix III) and β-lactamases were used for this analysis (Appendix III).

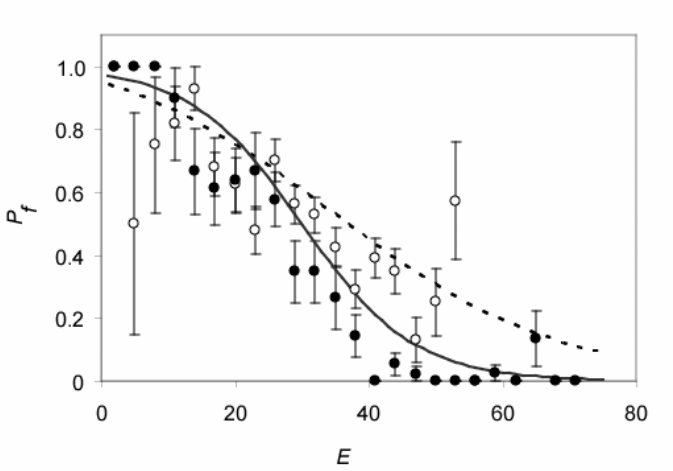


Figure VII-3. P_f (E) for lactamase chimeras (solid points, solid line) and cytochrome P450 chimeras (open points, dashed line). The points represent the fraction of folded chimeras in bins of 3 *E*. Curves represent the best fit of chimera folding data to Equation (VII-1). For lactamases a= 3.6, b = 0.12, c =1.0. For cytochromes P450 a= -2.1, b= 0.059, c= 0.93.

A significantly larger proportion of the cytochrome P450 chimeras retain their fold (45%) compared to the lactamase library (20%) (Figure VII-2). This is due to two factors. First, the P450 library has a lower $\langle E \rangle$ and a larger percentage of chimeras with low *E*. Second, in this experiment and in previous experiments, cytochromes P450 are universally more tolerant of *E* than lactamases (Voigt et al. 2002; Meyer et al. 2003; Otey et al. 2004; Otey et al. 2006). Examining the P_f determined using different sets of chimeras for both lactamases and cytochromes P450, it appears that cytochromes P450 are more tolerant to disruption (Figure VII-4). The curves forRASPP:PST, and the cytochrome P450 library are identical to the curves in Figure VII-3, and the curve for the 17 cytochrome P450s described by Otey et al. (2004) was determined by fitting the folding data for chimeras to Equation (VII-1) as described above (a = 5.8, b = 0.18, c = 1.0). The curves for the lactamase library described by Meyer et al. (2003) and the 12 lactamase chimeras described by Voigt et al. (2002) are reproduced from those works.

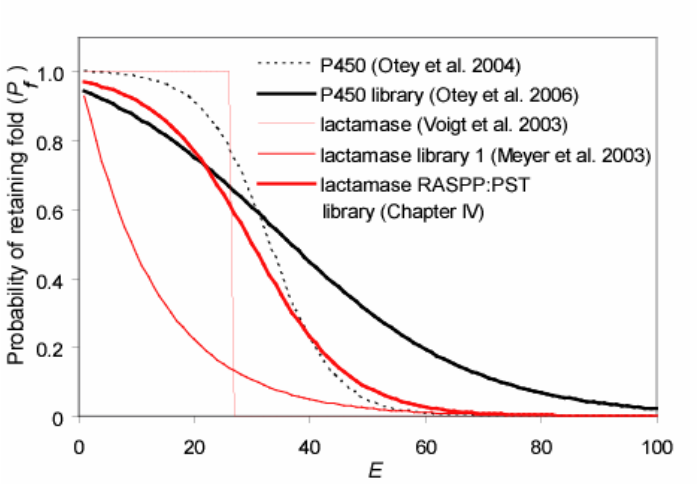


Figure VII-4. Different P_f functions can be calculated using various chimera data sets. All curves except those from (Meyer et al. 2003) and (Voigt et al. 2002) were fit to P_f in the form described in Equation (VII-1). The curves from Meyer et al. (2003) and Voigt et al. (2002) were included as described in the literature.

The extra tolerance of cytochromes P450 to *E* may stem from the higher degree of similarity between the parental cytochromes P450, and their larger size. There are 1814 amino acid contacts in the cytochrome P450 structure; in contrast there are only 1040 contacts in the lactamase structure. At the same number of contacts disrupted (*E*) a greater percentage of contacts in the lactamase are disrupted than in cytochrome P450s. Alternatively there may be other scaffold or sequence dependent affects.

Quantitative Comparison of SCHEMA Predictive Power

To quantitatively compare the predictive ability of SCHEMA on both data sets we used information theory to analyze the binary folding data (1 = folded, 0 = not folded). Given a set of chimeras, we cannot predict with 100% certainty whether a randomly chosen chimera is folded. If sequences with higher energies are less likely to be folded, this uncertainty or entropy can be reduced by knowing the energy for each sequence. The decrease in entropy is the mutual information between folding and the energy. An energy function with higher mutual information is better able to predict folding.

The uncertainty of chimera folding can be quantified by the Shannon entropy

$$H(F) = -\left[p \log_2 p + (1-p)\log_2 (1-p)\right], \qquad \text{(VII-2)}$$

where p is the fraction of chimeras folded (Adami 2004). The uncertainty, or entropy, can be reduced by knowing some predictive variable for each sequence. The conditional entropy H (F|E) measures the uncertainty when chimera energies are known and is an average over all energy values.

$$H(F \mid E) = \sum_{E_k} p(E_k) H(F \mid E_k), \qquad \text{(VII-3)}$$

where $p(E_k)$ is the fraction of chimeras with energy E_k, and H ($F \mid E_k$) is the conditional entropy associated with knowing whether a chimera has an energy E_k (Endelman 2005). The decrease in uncertainty associated with this knowledge, H (F) – H (F | E), is the mutual information.

The mutual information between folding and energy ranges from zero to the uncertainty of folding. The uncertainty of folding is determined by Equation (VII-2) and fraction of folded sequences in the data set p. When half the sequences in a population are folded, the uncertainty of folding is 1; as the fraction folded deviates from 0.5 it becomes easier to predict the folding status of a randomly chosen chimera and thus the uncertainty of folding decreases. The lactamase data set with 553 chimeras, 20% of which are folded, has a maximum mutual information of 0.72. The cytochrome P450 data set with 628 chimeras, 47% of which are folded, has a maximum mutual information of 0.96 (Figure VII-5).

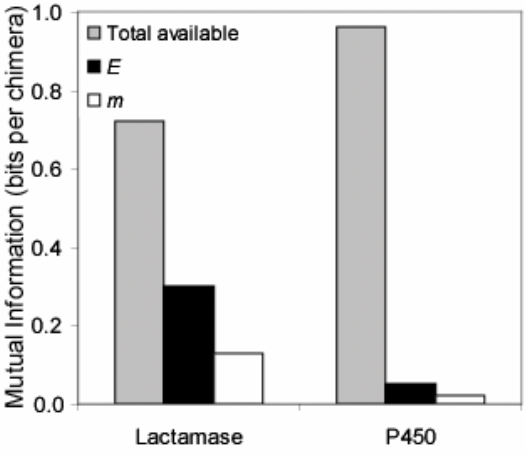


Figure VII-5. The total available mutual information (bits/chimera) for lactamase and cytochrome P450 data sets and the mutual information between folding and *E* and folding and *m* for each data set.

For both cytochromes P450 and lactamases, chimeras with lower SCHEMA disruption are more likely to fold correctly (Figure VII-3). However, SCHEMA predicts folded lactamase chimeras much better than it does folded cytochrome P450 chimeras (Figure VII-5). For lactamases nearly half of the available information is captured by *E;* while for cytochromes P450 less than 10% is captured. Calculating the mutual information between the number of mutations to the closest parent (*m*) and chimera folding shows that *m* has predictive power. However, *E* is a better predictor of chimera folding than *m* for both lactamases and cytochromes P450.

There are several potential reasons why *E* predicts lactamase folding better than cytochrome P450 folding. First, *E* is calculated using a static structure. The cytochrome P450 undergoes a conformational change on substrate binding that is not captured well by a single crystal structure (Arnold and Ornstein 1997). Second, it is unknown how well SCHEMA calculations derived from the structure of A1 reflect the contacts in A2 and A3 (whose structures are not available). For lactamases, calculation of *E* with structures from two of the three parents reveals few alterations when utilizing the different structures. Third, the parental cytochromes P450 share greater sequence identity than the parental

lactamases. The mutations introduced during lactamase recombination are likely less conservative and more deeply buried in the protein core, making a greater percentage of the disruptions counted by SCHEMA deleterious. Finally, it is possible that the differences are not due to specific structural or sequence properties, but that different scaffolds have different properties. Lattice protein studies indicate that while *E* is a good general predictor of chimeric protein folding, there is a great deal of variation in how well it performs that appears scaffold dependent (D. A. Drummond, personal communication).

The Effect of Imperfect Structural Information

The predictive ability of SCHEMA differs between the lactamase and cytochrome P450 libraries. Two of the possible reasons for this difference are tied to the unknown quality of the cytochrome P450 structural information and how well it applies to different protein conformations or to differences in the parent proteins. Often no structural information is available for a protein of interest, making the use of SCHEMA or any structure-based energy function difficult or impossible. In such situations there is frequently a structure available for one or more homologous proteins. To assess the effect of altering the structural information used by SCHEMA on its predictive abilities, we computed *E* for both lactamase and cytochrome P450 chimeras using structures of homologous proteins rather than the actual proteins recombined.

A search of the protein data bank identified many lactamase structures at varying levels of sequence identity to the parental proteins (Table VII-2). The cytochrome P450s were somewhat more difficult to analyze because no structures were available for proteins sharing 30-60% identity with the parents (Table VII-2). Using the structures

listed on Table VII-2, we calculated *E* for chimeras in the libraries and determined the mutual information between the new *E* values and the folding data. The parent protein sequences were aligned with the sequences from the structures using CLUSTALW (Chenna et al. 2003) to simulate a situation where no structural information is available.

The structure of any lactamase sharing more than 30% sequence identity on average with the parents predicts protein folding approximately as well as the structure of the protein of interest (Figure VII-6). The mutual information between chimera folding and *E* does not decrease very much until very distantly related (sharing <20% sequence identity on average with the parents) proteins are used for structural information. However the different structures sharing between 4 and 20% sequence identity to the parental sequence show a great deal of variation in the mutual information between *E* and chimera folding (Figure VII-6A). For the cytochromes P450 no definite conclusions can be drawn because there is not enough spread between the available structures on the sequence identity axis and because the mutual information between cytochrome P450 and folding is low. Since the structure of the proteins recombined does not yield particularly good predictions it is difficult to determine if the decreases associated with using alternative structures are significant. Some structures perform significantly worse than the A1 structure, others marginally better.

Table VII-2. Homologous Structures Used to Calculate *E*

	<Sequence	CE			DALI		
pdb ID	Identity>	Algn. Res.	RMSD	Z-score	Algn. Res.	RMSD	Z-score
Lactamases							
1BLS	6.00	204	3.4	4.4			
1DY6	39.00	256	1.7	7.4			
1FOF	7.00	222	3.2	6.3			
1KGE	32.00	253	2.2	7.3			
1MFO	37.33	257	1.8	7.4			
1QME	10.00	238	3.6	5.6			
1SKF	11.67	214	2.3	6.2			
4BLM	36.33	248	1.6	7.4			
1CI8	7.33				209	2.5	17
1EI5	11.67				204	3.0	14
1H8Y	7.33				220	3.4	19
1HZO	46.67				259	1.7	39
1IYO	49.00				259	1.7	39
1M6K	4.00				229	3.3	20
1MKI	2.67				212	3.4	15
1NRF	7.00				222	3.4	18
1RP5	9.33				231	3.5	19
1TVF	8.00				290	2.7	21
1XKZ	7.00				216	3.4	17
P450s							
1DT6	17.33	420	2.9	7.3			
1OXA	15.67	376	3.1	7.0			
1ROM	11.00	356	2.7	7.0			
1F4T	16.33	330	2.8	6.8			
1NR6	17.33				376	2.3	40
1SUO	14.67				430	3.1	36
1PQ2	14.67				433	3.4	35
1OG2	16.67				433	0*	35
1JIO	15.33				427	3.4	34
1ODO	17.33				370	3.0	33
1E9X	17.67				361	0*	33
1GWI	15.33				412	3.5	32
1LGF	15.67				368	3.0	32
1UED	13.67				359	3.3	31
1Q5E	15.00				363	3.1	31
1T88	10.67				368	3.3	30
1CPT	15.67				371	3.4	30

Structures used to determine whether *E* predictions are robust to altered structural information. *These values are as reported by the database, although I do not believe that they are correct.

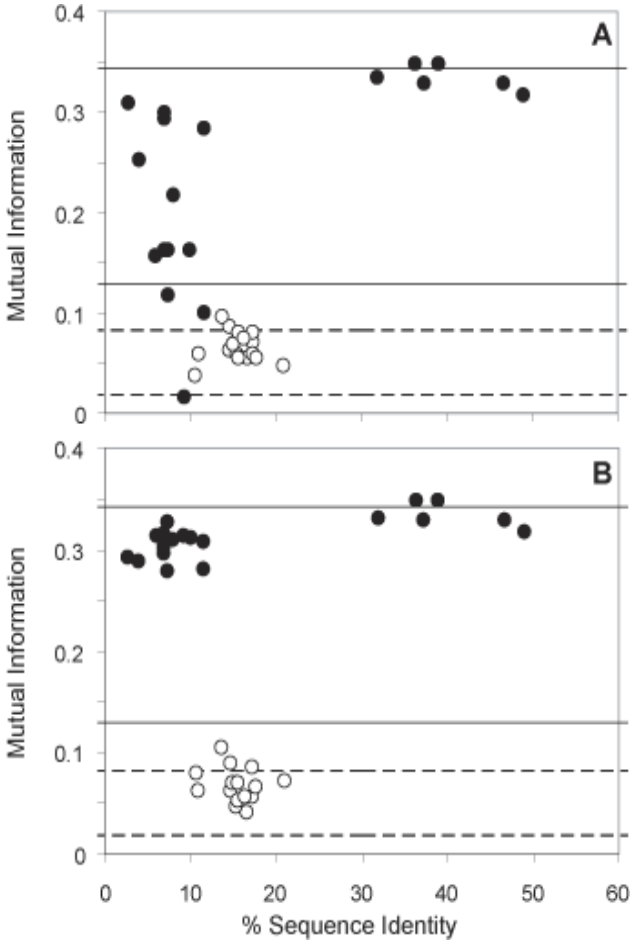


Figure VII-6. The mutual information (bits per chimera) between protein folding and *E* calculated using structural information from proteins homologous to the proteins recombined (Table VII-2). Solid points represent β-lactamases and open points cytochromes P450, the solid lines represent the mutual information of *m* (bottom line) or *E* calculated using the structure of PSE-4 (top line) for lactamases chimeras. Dashed lines indicate the same for cytochromes P450, *E* was calculated using the structure of CYP102A1. A: Sequences aligned using CLUSTALW. B: Sequences aligned using CE structural alignment tool (Shindyalov and Bourne 1998).

To further examine the relationship between mutual information and sequence identity for the lactamase structures, the mutual information was plotted vs. the length difference between the recombined proteins and the structurally characterized protein (Figure VII-7). The mutual information decreases as the length difference increases, suggesting that the alignment between the proteins may be affecting the performance of SCHEMA (Figure VII-7). This is not surprising because the reliability of CLUSTALW at low sequence identities is typically quite poor, especially if the sequences differ significantly in length (Thompson et al. 1999). Using structural alignments generated with Combinatorial Extension (Shindyalov and Bourne 1998) rather than CLUSTALW alignments to determine the SCHEMA disruption shows that structures of very distantly

related proteins can give good predictions, provided the proteins are aligned correctly (Figure VII-6B). Using structural rather than sequence alignments also improved the performance of the cytochrome P450 structures slightly, but due to the lack of diverse structures it is difficult to make strong conclusions.

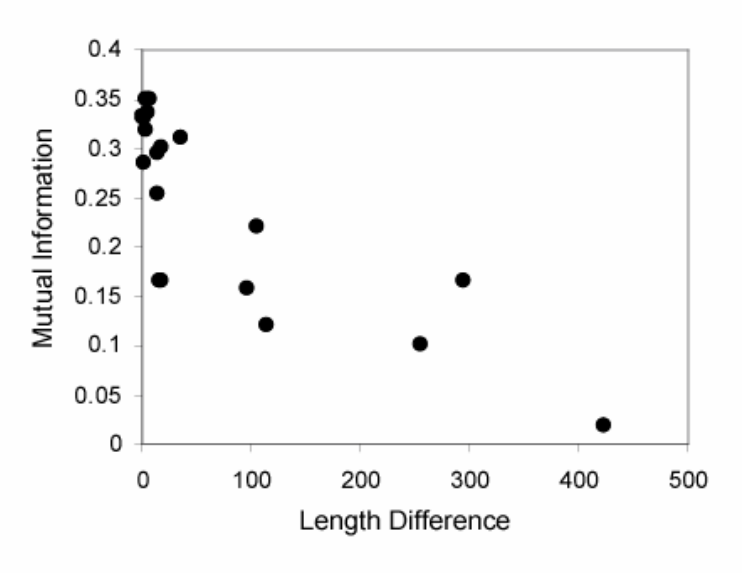


Figure VII-7. The mutual information between lactamase chimera folding and *E* calculated using a range of homologous structures. The sequences were aligned using CLUSTALW. As the length difference between the proteins recombined and the protein in the structure increases, there is a decline in the mutual information.

The topology of the lactamase fold is well conserved. However, many of the structures used for the analysis are of proteins that are very diverged from the proteins of interest, and all of them give relatively good predictions. Many of these proteins have very similar topologies to the lactamases recombined, but the structures themselves are not easily aligned as a whole. A good alignment among the proteins is essential to good results. SCHEMA not only takes into account structural contacts, but also the sequence of the parental proteins. The contacts broken in a chimera are mediated by the sequence identity between the proteins. If the alignment between the parental proteins and the structural contacts is incorrect, then the contacts are not treated appropriately. This results in a decrease in the mutual information between *E* and chimera folding. CLUSTALW alignments are not sufficient when there are large length differences between the proteins

corresponding to inserted or deleted domains. Protein family multiple sequence alignments could be used to identify large gap regions, overcoming this potential limitation. However, the success of this approach is dependent on quality of the multiple sequence alignment.

Minimal Structural Information Required to Calculate Accurate *E*

The robustness of *E* as a predictive measure using structures from distantly related proteins indicates that it is unlikely that *E* determined for cytochrome P450 chimeras is significantly affected by minor perturbations due to dynamics or slightly altered structures among the parent proteins. However, it also raises questions regarding how much structural information is required to accurately predict chimera folding. Computing *E* using only a Cα contact map (Cα distance <8 Å) shows a small decline in predictive ability compared to *E* calculated using the standard contact map (Figure VII-8). These results indicate that *E* captures overall structural topology, not necessarily specific side chain interactions. However, incorporation of sequence identity to remove contacts where the amino acid identities remain the same in the chimera compared with the parent sequence is an essential component for accurate predictions (Endelman 2005). This indicates that the amino acid side chain interactions, whether identified through proximity of any heavy atom or just Cα, are important for accurate predictions.

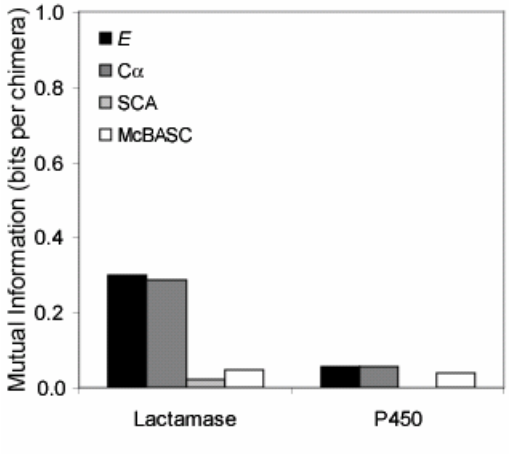


Figure VII-8. The mutual information for chimera folding and *E* calculated using standard contact maps where residues with any heavy atoms are within 4.5 Å (Voigt et al. 2002), Cα maps where residues with Cα within 8Å are contacting, and contact maps determined from the highest scoring 1% of covarying amino acids. Covarying amino acids were detected using statistical coupling analysis (SCA) (Lockless and Ranganathan 1999) and McLachlan Based Substitution Correlation (McBASC) (Gobel et al. 1994; Olmea et al. 1999).

Covarying Amino Acid Pairs Substituting for Structural Contacts

Given that *E* values calculated with only α-carbons are nearly as good as *E* values calculated using all the heavy atoms, we were curious whether structural information is necessary at all. Evolutionary amino acid covariation has been used in the past to predict Cβ-Cβ distances as well as to infer energetic coupling (Gobel et al. 1994; Lockless and Ranganathan 1999). To examine how well using amino acids with significant covariation scores might serve to replace structural contacts in calculating *E*, we used two covariation algorithms to score both lactamases and cytochrome P450s, Statistical Coupling Analysis (SCA) (Lockless and Ranganathan 1999) and McLachlan Based Substitution Correlation (MCBASC) (Gobel et al. 1994; Olmea et al. 1999). For each covariation algorithm the most significant 1% of covarying amino acids were used as contacts for calculating *E*. Figure VII-8 shows that amino acid covariation does not provide information that is useful for identifying chimeras that are likely to fold. The mutual information between *E* calculated using covarying amino acids determined using SCA or McBASC is very small

for lactamases and significantly decreased for cytochromes P450. This corresponds with the finding that amino acid evolutionary covariation has at best weak correlation with Cβ-Cβ distances (Fodor and Aldrich 2004b). The weak correlation that is present does not provide sufficient structural information for predictive values of *E.*

Conclusions

SCHEMA disruption *E* is a predictor of chimera folding. However, its accuracy for the two libraries examined here is different. In the case of the lactamases SCHEMA predictions are relatively accurate, capturing nearly ½ of the available information. For cytochrome P450s they are much poorer. The accuracy may depend on the protein scaffold and parental proteins chosen for recombination. The lactamase parents share much less sequence identity than the cytochrome P450 parents (~40% vs. ~60%). However the lactamase parents have approximately the same thermostability (Chapter V). While cytochrome P450 parents share more sequence identity, their thermostabilities differ by 11 °C (Otey et al. 2006). It is possible that these stability differences between the parental proteins contribute to the decreased accuracy of SCHEMA predictions. If different parents confer differing starting amounts of stability, then chimeras inheriting some blocks from a particular parent may be more likely to fold, mediating the effect of pairwise interactions that are measured by SCHEMA.

The accuracy of SCHEMA is not strongly influenced by the structure used to calculate the contact map so long as it has a similar topology to the protein of interest and the sequences are aligned correctly. This should allow many researchers that do not have structural information to take advantage of this approach toward library design.

However, structural information is necessary for accurate predictions. Trying to infer structural interactions from amino acid covariation is not an effective strategy. Whatever correlation there is between amino acid evolutionary covariation and distance in the three-dimensional structure is not sufficient to correctly identify a sufficient percentage of contacting residues.

Most of the other energy functions for predicting chimera folding use structural contacts to identify important residues pairs (Voigt et al. 2002; Moore and Maranas 2003; Saraf and Maranas 2003). However, one algorithm, FAMCLASH instead uses the conservation of pairwise charge, volume and hydrophobicity information (CVH) in the family of proteins as an indication of interacting residues rather than structure (Saraf et al. 2004). This metric penalizes interacting pairs of amino acids in the chimera where the chimeric amino acids result in a pairwise CHV outside the conserved range (clashes). Based on our results it is unlikely that these specific amino acid pairs are contributing greatly to chimera properties. This energy function was tested against 13 single-crossover DHFR chimeras and the number of clashes found to correlate well with chimera activity. However, only functional hybrids were characterized, and as with most single crossover chimera sets, there is a very simple curve displayed: low activity corresponds with a large number of clashes when the chimeras inherit roughly half protein from one parent and half from another. This effect is due the accumulation of deleterious pairwise interactions (Drummond et al. 2005), however this particular quantification of such interactions does not likely reflect the deleterious pairwise interactions any better than a structure based metric.

Methods

E Calculations

The structure of PSE-4 (1G68) was used with a CLUSTALW alignment of the lactamases TEM-1, SED-1 and PSE-4 to calculate SCHEMA disruption *E* for lactamase chimeras. The structure of CYP102A1 (1JPZ) was with a CLUSTALW alignment of cytochromes P450 CYP102A1, CYP102A2, and CYP102A3 to calculate *E* for cytochrome P450 chimeras. SCHEMA disruption is

$$E = \sum_i \sum_{j>i} C_{ij} \Delta_{ij} \ , \qquad \text{(VII-4)}$$

where C_{ij} =1 if any side-chain heavy atoms or main-chain carbons in residues *i* and *j* are within 4.5 Å. The Δ_{ij} function is based on the sequences of the parental proteins. $\Delta_{ij} = 0$ if amino acids i and j in the chimera are found together at the same positions in any parental protein sequence, otherwise $\Delta_{ij} = 1$. All of the code used to perform these calculations can be found as python scripts on the Arnold lab website http://www.che.caltech.edu/groups/fha/.

Mutual Information

The mutual information was calculated as described by Endelman (2005) and Matlab m-files to perform the computations are available on the Arnold lab website http://www.che.caltech.edu/groups/fha/.

For all comparisons the naïve data sets of chimeras were utilized for calculations. For lactamases this set consists of 553 chimeras, of which 111 confer resistance to ampicillin (Appendix III). For cytochrome P450s this set consists of 628 chimeras, of which 285 correctly bind the heme cofactor (Appendix III) (Otey et al. 2006).

Alternative Structures

Structural neighbors of the proteins were identified using both CE (http://cl.sdsc.edu/) (Shindyalov and Bourne 1998) searches of the protein data bank and the DALI database (http://ekhidna.biocenter.helsinki.fi/dali/start) (Holm and Sander 1996); representative structures were used wherever possible. Sequence alignments were performed using CLUSTALW (Chenna et al. 2003), and structural alignments using the CE pairwise alignment tool. SCHEMA calculations were performed as described above using the tools available on the Arnold lab website. A list of the structures used for this analysis and their average sequence identity to the sequences used and a measure of their structural identity can be found on Table VII-2.

Covariation Analysis

Evolutionary covariation between amino acids was examined using both Statistical Coupling Analysis and McLachlan Based Substitution Correlation. Java code for both of these algorithms was downloaded from http://www.afodor.net/ (Fodor and Aldrich 2004b, 2004a), and the full PFAM lactamase superfamily alignment used for calculation (Bateman et al. 2004). Alignments used for examination of consensus stabilization were the PFAM seed alignment, and a class A nonredundant alignment published by Axe (2004). The most significant 1% of amino acid correlations were used as the contacting residues for computing SCHEMA disruptions.

Chapter VIII: Improving Predictions of Chimera Folding Using Multiple Sequence Alignments

Introduction

There are many different energy functions for predicting chimera folding (Voigt et al. 2002; Moore and Maranas 2003; Saraf and Maranas 2003; Saraf et al. 2004). They take into account a variety of factors including three-dimensional structure, amino acid biophysical characteristics, and family multiple sequence alignment information, but all consider only pairwise terms. The development of pairwise energy functions is reflective of the properties of chimeric proteins. Unfavorable pairwise interactions are the largest contributors to chimera misfolding (Drummond et al. 2005).

We have used the energy function SCHEMA (*E*) to calculate the number of potentially unfavorable pairwise interactions that are generated by recombination in a chimera. This energy term is very simple and requires only a three-dimensional structure and the sequence of the proteins to be recombined. Using this energy function we have designed two libraries of chimeric proteins, one recombining class A β-lactamases and the other recombining cytochromes P450. We have characterized a large number of chimeras from each library, including 555 lactamases chimeras, 20% (111) of which are folded (Appendix III), and 628 cytochrome P450 chimeras, 45% (285) of which are folded (Appendix III) (Otey et al. 2006). The proteins recombined to make the two libraries have very different topologies, sizes, and sequence identity shared by the parents. Interestingly, while chimeras with lower *E* are more likely to fold for both lactamases and cytochromes P450, how well *E* predicts the folded chimera differs greatly between the two proteins (Chapter VII). Calculating the mutual information between

chimera folding and *E*, as described in Chapter VII, shows that lactamase chimera folding is predicted much more accurately than cytochrome P450 folding (Figure VIII-1).

Previous analyses of both chimera libraries included generating energy models using logistic regression analysis (LRA) to identify significant contributions to folding (Chapter V) (Otey et al. 2006). These models assign energies to interactions between sequence blocks (two-body terms) as well as to individual blocks (one-body terms). For lactamases a two-body term is the most significant (block 1-8 interaction) contributor to chimera folding, but there are also significant one-body terms (blocks 2 and 3). For cytochromes P450 one-body terms dominate whether a chimera folds (blocks 1, 5 and 7), but there is also a significant two-body term (block 1-7 interaction). In the process of creating these models, an energy value is assigned to each chimera corresponding to the sum of the one-body and two-body terms. This energy is predictive of chimera folding. Determining the mutual information between the LRA energies and chimera shows that, as expected, the LRA models more accurately predict chimera folding than *E* does because they are derived directly from the data (Figure VIII-1). For cytochromes P450 the LRA model is significantly better than *E*, capturing nearly seven times more information. For lactamases the LRA model predicts chimera folding better than *E*, but does not have the same large increase in mutual information.

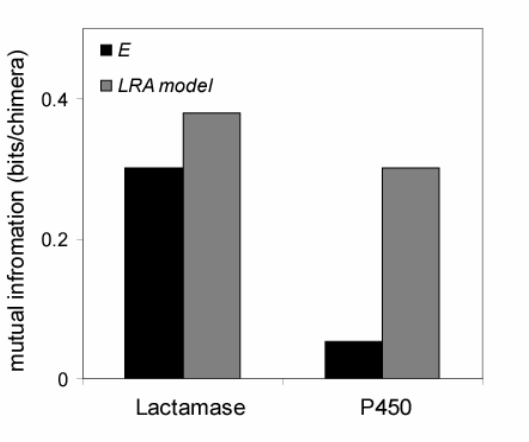


Figure VIII-1. The mutual information for chimera folding and different energy functions. *E* is the standard SCHEMA disruption that considers pairwise disruption. The LRA models are derived from the data and have both two-body (pairwise) and one-body contributions.

Previous attempts to improve SCHEMA have focused on altering the pairwise energy function (Endelman 2005; Saraf and Maranas 2003). However, the LRA models incorporate not only two-body (pairwise) terms, but also one-body terms. The one-body terms represent how an individual block inherited from a specific parent contributes to whether a chimera folds. This includes effects due to interactions between residues within the block as well as interactions between the residues and the solvent. The LRA models for both lactamases and cytochromes P450 show that one-body terms are important in determining whether a chimera will fold (Chapter V) (Otey et al. 2006). However, none of the current predictive energy functions for chimera folding, including SCHEMA, explicitly take into account any one-body information.

In this work we estimate the one-body terms that appear significant for predicting chimera folding from the LRA models. The strength of SCHEMA *E* is that it can be calculated *a priori* using relatively little information. In order to retain an energy function which can be easily calculated *a priori* we used only information that is readily available for most proteins, family multiple sequence alignments, to estimate one-body contributions to chimera folding. Finally we ask whether the estimates calculated can provide information that is useful for predicting chimera folding, and how this information can be combined with the existing pairwise energy function.

Consensus Sequence Stabilization Theory to Estimate One-Body Contributions

The contributions of individual amino acids to protein stability have been estimated in a variety of different ways (Mendes et al. 2002). However, such potentials incorporate pairwise terms, and are usually complex and computationally intensive. The

pairwise interactions in protein chimeras are fairly well predicted by the SCHEMA energy *E*. One of the strengths of SCHEMA is its simplicity. It does not require very much information and is even robust to imperfect structural information (Chapter VII). Ideally if a one-body term is added to the existing SCHEMA energy function it should not require more information than SCHEMA already incorporates, and should not be computationally intensive.

A potential approach to approximating individual amino acid contributions to protein stability is to calculate the probability of the amino acids found in a chimera at each position in a multiple sequence alignment (Figure VIII-2). This idea has its basis in the theory of consensus stabilization (Steipe et al. 1994). Consensus stabilization asserts that the amino acid with the highest frequency at a given position in a multiple sequence alignment of homologous proteins likely contributes the most stability to the protein. This idea is based on the theory that evolved populations of proteins share some canonical or prototype sequence which is the most mutationally robust (Bornberg-Bauer and Chan 1999) and stable sequence for a particular fold (Xia and Levitt 2004). This sequence accumulates mutations which are usually destabilizing, but selectively neutral so long as the protein continues to fold and function. In a population of proteins with marginal stability, where stability is the only selective property, amino acid frequencies are fixed with probabilities related to their effects on stability (Steipe et al. 1994; Dokholyan and Shakhnovich 2001).

Using consensus stabilization to approximate single amino acid contributions to protein stability is based on several assumptions which often may not apply to real proteins. First, that most mutations have independent contributions to stability, and

second that the set of homologous proteins analyzed reflects the stability of the protein and not some other selected property. Despite these potential limitations, the general concept of consensus stabilization has been implemented in several different proteins to increase thermostability. While not all consensus mutations increase thermostability, most appear to have stabilizing or neutral effects (Steipe et al. 1994; Nikolova et al. 1998; Wang et al. 1999; Lehmann et al. 2000; Lehmann et al. 2002).

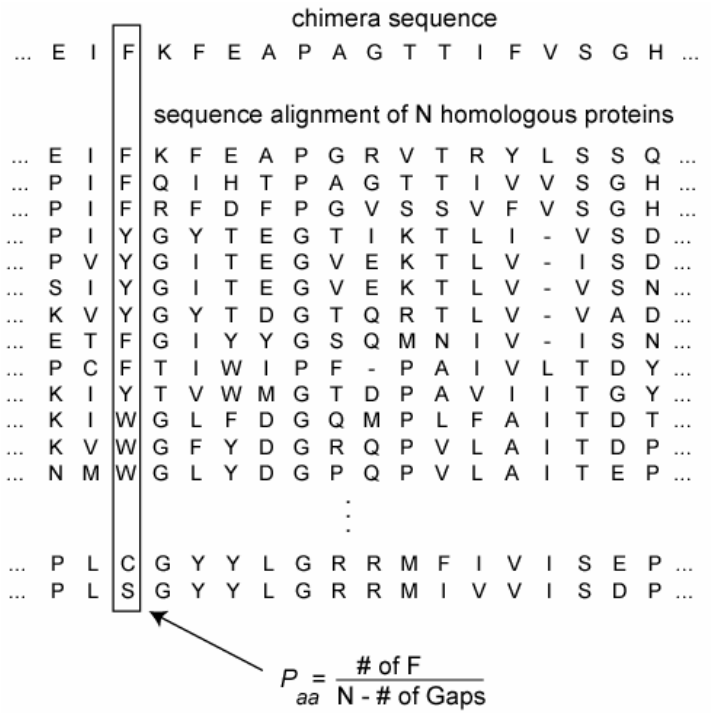


Figure VIII-2. The probability of finding the amino acid in the chimera in the multiple sequence alignment (P_{aa}) is determined by determining the frequency of the amino acid at the position and dividing by the total number of sequences not including sequences with gaps at the position.

One-Body Energy Term: *w*

To implement the consensus stabilization theory into a scoring function for chimeras we first obtained sequence alignments for both the cytochromes P450 and the lactamases. The choice of a high-quality sequence alignment is essential to determining a good representation of the amino acid probabilities. An alignment that is inaccurate or contains many sequences similar to one of the parents could potentially lead to a flawed analysis (Ewart et al. 2003). For cytochromes P450 the alignment used was a manually

corrected alignment of 238 cytochrome P450 sequences (Nelson 2005). The sequences in this alignment share on average 18% sequence identity, and <0.1% of the sequence pairs shared greater than 90% identity. For the lactamases, the PFAM seed alignment for lactamases was utilized (Bateman et al. 2004). This alignment contains 130 sequences that share on average 17% identity. No sequences sharing >80% identity are in the alignment. The full PFAM alignment for lactamases (1485 sequences) contains many variants of TEM-1, making its use for this type of application limited.

Once an alignment was obtained, we calculated the frequency of each amino acid at each position in the alignment. Many consensus stabilization experiments with proteins have identified the consensus amino acid for each position and mutated the residue existing in the protein of interest to this residue (Lehmann et al. 2000; Lehmann et al. 2002). The term consensus amino acid can indicate the amino acid that appears most frequently, or can indicate the amino acid occurring at a probability greater than some threshold. Rather than determine if the chimera matches the consensus sequence exactly, we calculated the probability of each parental amino acid (P_{aa}) at all positions in the alignment (Figure VIII-2). For cytochromes P450 the P_{aa} varies between the maximum of 1.00 and 0.00425. The average P_{aa} was 0.19. For lactamases the P_{aa} varies between 0.992 and 0.00752. The average P_{aa} was 0.21. Some positions are highly conserved (all or nearly all sequences have the same amino acid). For other positions, the amino acid present in the parent only appears in the parent. The variation in P_{aa} over all positions is not the same as the variation in P_{aa} of different parents at the same position. For cytochrome P450s where the parental amino acids are not conserved the

ΔP_{aa} = | (P_{aa}(*parent 1*) – P_{aa} (*parent 2*)) | varies between 0.68 and 0.00425 with an average of 0.104, for lactamases the ΔP_{aa} varies between 0.80 and 0.075, with an average of 0.145.

To compute a one-body score (*w)* for a chimera, the P_{aa} for each amino acid in the chimera is averaged over the sequence,

$$w=<P_{aa}>. \qquad \text{(VIII-1)}$$

A higher *w* indicates a chimeric sequence closer to the prototype sequence, and more likely to fold. For both lactamases and cytochromes P450, sequences with lower *w* are less likely to fold (Figure VIII-3). For lactamases there appears to be a bimodal distribution among folded chimeras. Examining the *m* vs. *w* distribution of folded and unfolded chimeras shows that the lower *w* lactamase chimeras are usually chimeras with few mutations, while the higher *w* lactamase chimeras that are likely to fold are chimeras with more mutations (Figure VIII-4). For cytochromes P450 both folded and unfolded chimeras are distributed over a range of *w* values, but chimeras with lower *w* are less likely to fold (Figure VIII-3).

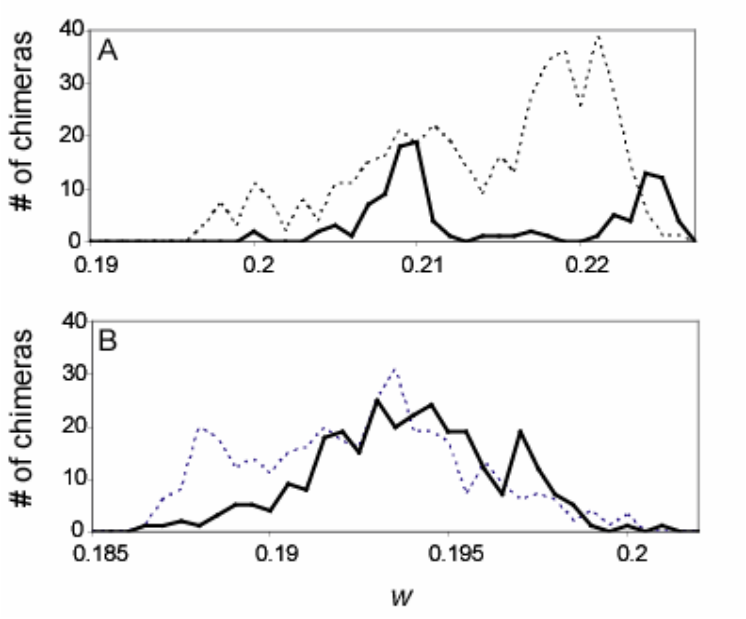


Figure VIII-3. Distribution of folded (solid line) and unfolded (dashed line) chimeras with respect to *w* shows that chimeras with low *w* are less likely to function in both A: β-lactamase chimeras and B: cytochrome P450 chimeras.

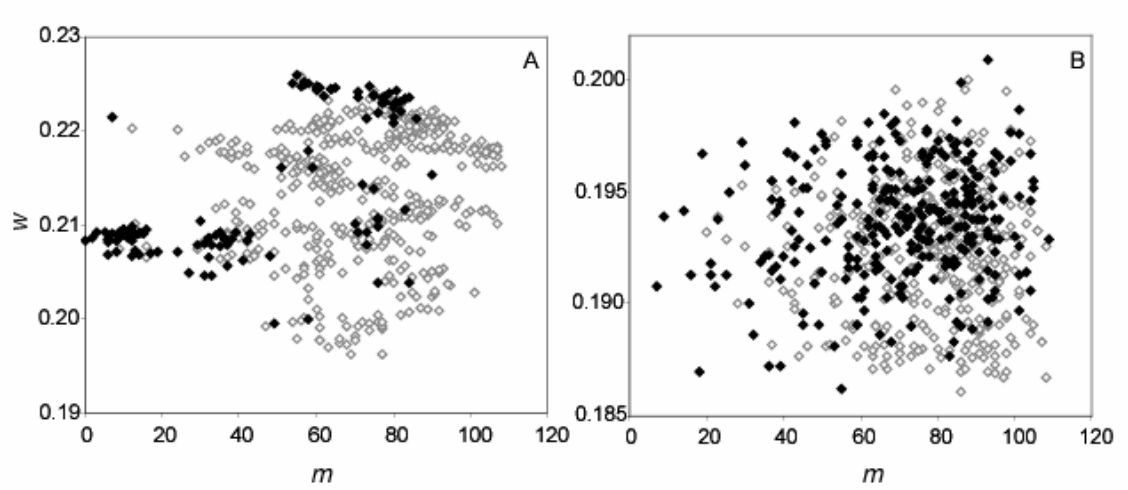


Figure VIII-4. *w* vs. *m* of A: lactamase chimeras and B: cytochrome P450 chimeras. Open points represent unfolded chimeras and closed points represent folded chimeras. For lactamases the parent *w* values are: PSE-4 *w* = 0.201, SED-1 *w* = 0.227, TEM-1 *w* = 0.213. For cytochromes P450 the parent *w* values are: A1 *w* = 0.191, A2 *w* = 0.192, A3 *w* = 0.195.

Our mutual information calculation relies on a fit of the energy to a probability function (P_f) that assumes increased energy leads to increased misfolding (Equation (VIII-2), Chapter VII).

$$P_f = \frac{1}{c + e^{bE_g + a}}, \qquad \text{(VIII-2)}$$

Where *a*, *b*, and *c* are fit parameters and E_g is a generic energy term that is substituted by the energy of interest. Therefore we inverted *w* to calculate the mutual information between folding and the one-body weight. Calculating the mutual information between 1/*w* and chimera folding shows that 1/*w* is a better predictor of cytochrome P450 folding than *E* (Figure VIII-5), but contributes almost no information toward lactamase folding when fit to the definition of P_f shown above. This is not surprising considering the distribution of folded lactamase chimera with respect to *w*, and indicates that additional variables may need to be incorporated to predict chimera folding for proteins generally.

Calculating the mutual information between folding and $E = -w$ gives a similar result (Figure VIII-5).

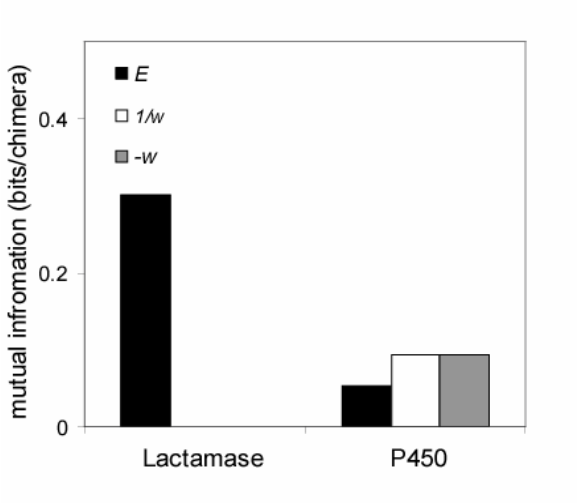


Figure VIII-5. Mutual information between chimera folding and E, $1/w$ and $-w$. Both $-w$ and $1/w$ are more predictive of chimera folding than E for cytochromes P450. However, for lactamases neither has significant predictive power.

Individual Block Contributions to *w*

To visualize the contribution of each library sequence block to w, it can be broken down into the individual components for each block $w_{block,}$

$$w_{block} = \frac{\sum_{i=blockstart}^{i=blockend} P_{aa}(i)}{N}, \qquad \text{(VIII-3)}$$

where $P_{aa}(i)$ is P_{aa} for the amino acid at position i, *blockstart* and *blockend* are the starting and ending residues of the sequence block, and N is the total number of amino acids present in the protein. The sum of the w_{block} terms corresponding to a chimeric sequence is the same as its w. The w_{block} for cytochrome P450 sequence blocks does not differ greatly among the parents in most cases (Figure VIII-6A). However, where significant differences do exist (standard deviation >5%), they correspond well with chimeric protein folding data. Blocks 1 and 7 are significant one-body terms important for determining cytochrome P450 folding (Otey et al. 2006), and they show the greatest variability between the parental sequences. Additionally, the parents that are favored in

folded chimeras (A2 for block 1 and A3 for block 7) display higher $<P_{aa}>$. However, calculating the *w* for each parent shows that it does not correspond directly to the parent's thermostability. A1 is more thermostable than both A2 and A3, but it has a lower *w* (0.191 as opposed to 0.192 and 0.195). The w_{block} values for lactamase blocks are more variable between blocks as well as between different parents at the same block. This is due to the larger differences in block size in the lactamase library as well as the decreased sequence identity shared by the parents. The biggest contributor to lactamase *w* is block 3, and the parent favored at block 3 in folded chimeras (TEM-1) is also the parent with the highest *w* for this block (Figure VIII-6B). Additionally, the lactamase parents have approximately the same thermostability, but *w* differs (PSE-4 *w* = 0.201, SED-1 *w* = 0.227, TEM-1 *w* = 0.213).

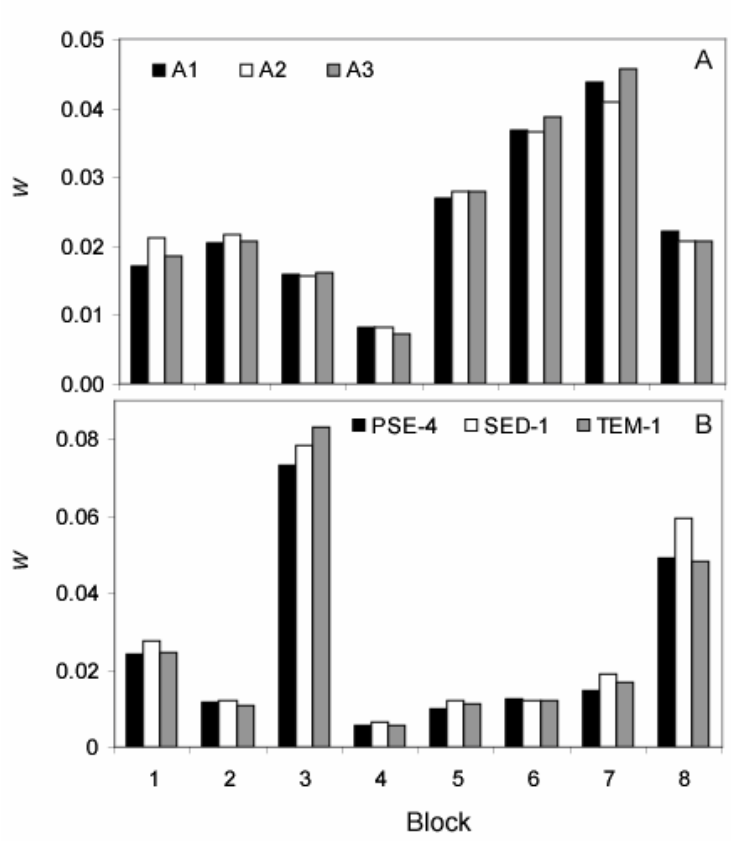


Figure VIII-6. The one-body weighting term w_{block} determined for each exchangeable block of sequence in the A: cytochrome P450 and B: β-lactamase libraries. Parents with higher w_{block} block values are more likely to be found in functional chimeras than parents with low w_{block} values.

Combining *E* and 1/*w*

The *1/w* term alone predicts cytochrome P450 folding better than *E*, however for lactamases 1/*w* has little predictive power. It is not surprising that estimates of one-body terms alone are not enough to predict chimera folding because the potentially deleterious pairwise terms introduced by recombination are not explicitly being addressed. The LRA energies combine one-body and two-body terms in an additive manner to predict protein folding. To emulate these models we combined the *a priori* estimate of one-body energies (*w)* with the *a priori* estimate of two-body energies *E*.

To bring *w* together with the existing pairwise energy function *E*, *w* is first normalized by the variation in the population of all possible chimeras created from the parents to give the normalized weight *W* (Equation (VIII-4)). *W* for most chimeras should be between 0 and 1.

$$W = \frac{w - (<< P_{aa} >> -3\sigma_{aa})}{<< P_{aa} >> +3\sigma_{aa}}, \qquad \text{(VIII-4)}$$

where $<<P_{aa}>>$ is the mean $<P_{aa}>$ for all possible chimeras, and σ_{aa} is the standard deviation on $<<P_{aa}>>$ for the population of all possible chimeras. The combined energy function (E_w) is the sum of the SCHEMA disruption, *E*, and the reciprocal of the normalized weight, 1/*W* (Equation (VIII-5)).

$$E_W = E + \frac{c}{W}, \qquad \text{(VIII-5)}$$

where *c* is a constant parameter. The parameter *c* that determines the relative weighting of *E* and *1/W* was optimized independently for both lactamases and cytochromes P450. In both cases the optimal value was close to 1.0 (0.93 ± 0.07 for lactamases and 1.0 ± 0.2 for cytochromes P450). The value of *c* is sensitive to the normalization of *w*. Without

normalization, the optimal value of *c* is very different for lactamases and cytochromes P450 (32 and 165 respectively). This is likely due to the different levels of sequence identity among the parental proteins. The cytochromes P450 parents share higher sequence identity, therefore *w* for chimeras has a smaller range than for lactamases (0.143 vs. 0.297). The normalization allows the variation between parental sequences to be standardized into the same range for any potential sets of parents. Thus, the parameter *c* that amplifies this variation will vary less from protein to protein.

Based on the mutual information between E_w and chimera folding, E_w is a better predictor of chimera folding than either *1/w* or *E* alone for both lactamases and cytochromes P450 (Figure VIII-7). Tenfold cross-validation to compare *E* with E_w (c=1) shows that E_w is significantly better for predicting chimera folding for both lactamases and cytochromes P450. While E_w is a significantly better predictor of both lactamase and cytochrome P450 chimera folding, its increase compared to *E* is much larger for cytochromes P450 than for lactamases. This is anticipated because *E* captures nearly 85% of the information captured by the LRA model for lactamases, while for cytochromes P450 *E* performed poorly compared to the LRA model. There is more information that can be captured by adding a one-body term to a model of cytochrome P450 folding than for lactamases.

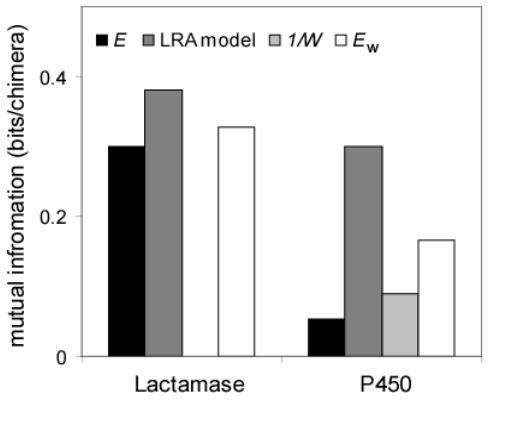


Figure VIII-7. The mutual information between folding and various predictive energy functions. *E*, *1/w* and E_w can be calculated *a priori* but the LGA energy was calculated directly from the chimera folding data and represents the best that a model incorporating one- and two-body terms can predict the data.

Comparison of the E vs. m and E_w vs. m plot for folded and unfolded lactamase chimeras shows that the plots looks very similar (Figure VIII-8A, B). The biggest difference at first glance is that the values are shifted ~18 higher. However, careful examination shows that many unfolded chimeras in the low E range are not in the low E_w range, and that the distribution of folded chimeras with respect to E_w is somewhat narrower. For cytochromes P450 the plot of E vs. m is very different than the plot of E_w vs. m (Figure VIII-8C, D). In the E_w vs. m plot chimeras are spread over a wider range than the E vs. m plot with high E_w chimeras more likely to be unfolded.

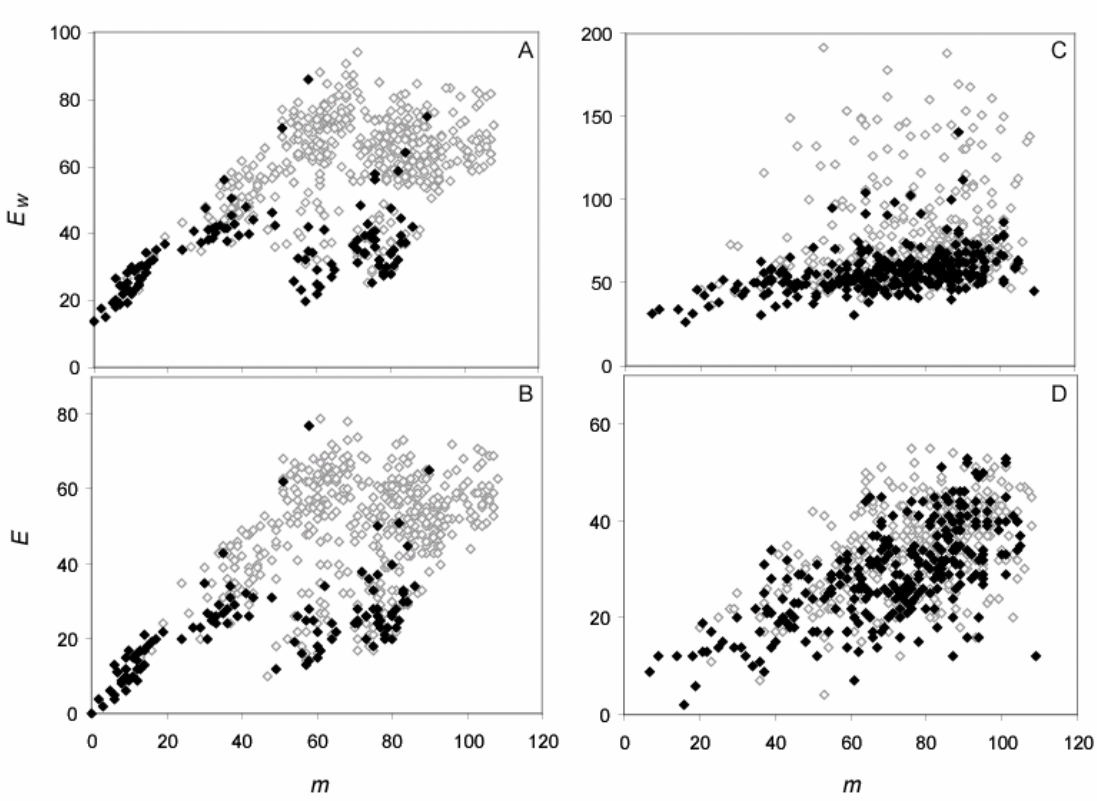


Figure VIII-8. E vs. m and E_w vs. m for lactamase (A, B) and cytochrome P450 (C, D) folded (solid point) and unfolded (open point) chimeras.

Discussion

For both lactamases and cytochromes P450 the energy models derived using LRA are significantly better at predicting chimera function than SCHEMA *E*. These models showed that the pairwise terms included in most energy functions are not the only important factors governing chimera folding, but that some blocks were inherited from particular parents more frequently in folded chimeras independent of pairwise interactions. We estimated the individual contributions of each amino acid position to chimera folding by calculating the average probability of finding the amino acid present in the chimera in a multiple sequence alignment of homologous proteins, *w*. This measure was effective for predicting cytochrome P450 folding without the addition of any pairwise contributions. When *w* was combined with the SCHEMA disruption (*E)* which estimates the pairwise contributions, the resulting function (E_w) showed significant improvement for predicting both lactamase and cytochrome P450 chimera folding.

There are undoubtedly many one-body effects that are not captured by this simple model, and it is also possible that one-body effects are not the only properties captured by *w*. However in both the lactamases and cytochromes P450, adding an estimation of the one-body term based on multiple sequence alignments increases the predictive power of the energy function. The strength of this prediction is variable depending on the protein, but represents a real improvement.

Other energy functions designed to predict chimera folding only take into account pairwise terms. Most energy functions use structural information to identify the interacting pairs of amino acids (Voigt et al. 2002; Moore and Maranas 2003; Saraf and Maranas 2003). However, one uses conservation of pairwise additive charge, volume and

hydrophobicity (CVH) properties in a family of proteins to identify interacting residues (Saraf et al. 2004). The pairwise interactions changed by recombination in chimeras are usually counted and the count then mediated by some additional information. In Voigt et al. (2002) the count is only mediated by the sequence identity between the parental sequences so that when the residue identities remain the same, the clash is not counted (Voigt et al. 2002). Moore and Maranas (2003) use mean-field calculations to approximate the complete set of residue-residue coupling compatible with a fold, and penalize chimeric residue pairs that fall outside this set. In both of his works Saraf mediated the interacting residue pairs using amino acid biophysical information. Residue pairs where the additive CVH was altered were considered clashing (Saraf and Maranas 2003; Saraf et al. 2004); counting a smaller subset of potential clashes compared to SCHEMA when structural information is used to identify the interacting residue pairs. Despite the use of multiple sequence alignments by Saraf et al. to identify interacting residues (2004), there are no one-body terms explicitly incorporated and the family sequence information is used in a very different way than it is used here.

The folding of chimeras for the two proteins used in this study is predicted differently by the two terms used to compose E_w. While both SCHEMA E and E_w are predictors of chimera folding for both proteins, the amount of information provided by the one-body and two-body terms is different. Lactamase chimera folding is better predicted by pair-wise interactions. The one-body weight w adds information, but alone it is not effective. The pairwise term is dominant in the final energy value. Cytochrome P450 chimera folding is predicted more evenly by the one-body and pairwise terms, and they are nearly additive when combined. There are many potential reasons why the two

proteins may behave differently. First, the cytochrome P450 parental proteins are larger, and have several subdomains while the lactamases are smaller and have two closely connected subdomains. It is possible that the structure of the cytochrome P450 is more modular and that pairwise interactions are less important. The LGA models identified pairs of interacting blocks for both lactamases and cytochromes P450. In both cases the interacting blocks each formed interacting β–stands. The β-sheet in the lactamases is a much larger percentage of the structure (16% vs. 7%) than the β–domain is in the cytochromes P450, and the pairwise disruptions larger in the lactamase because of the lower sequence identity between the parents. Finally, the cytochrome P450 parents have different thermostabilities and this may obscure the pairwise effects.

Studies with model proteins have suggested that evolved proteins sharing the same structure exist on neutral networks. On these neutral networks there is a prototype sequence that is the most mutationally robust sequence (Bornberg-Bauer and Chan 1999; Xia and Levitt 2004). It has also been shown that more thermostable proteins are more robust to random mutations (Poteete et al. 1997; Bloom et al. 2005a), and to mutations introduced by recombination (Chapter VI). With the one-body weights we are essentially estimating a chimera's similarity to the prototype sequence. Chimeras that are far away from the prototype sequence are likely less stable and less prone to fold correctly. Chimeras that are closer to the prototype sequence are more likely to fold. We have developed an energy function that combines an approximation of the effects due to deleterious interactions introduced in a chimera by recombination with an estimation of a chimera's inherent stability that is a significant improvement upon examining pairwise interactions alone.

Methods

LRA Energies

A chimera's LRA energy is the sum of its one-body and two-body energies. Tables VIII-1 and VIII-2 list the relevant energies for lactamases and cytochromes P450 respectively (Endelman 2005).

Table VIII-1. One- and Two-body Energy Terms Used to Calculate LRA Energies for Cytochromes P450.

Two-body Terms	Parent at Block 7		
Parent at Block 1	A1	A2	A3
CYP102A1	-0.9	1.3	-0.4
CYP102A2	0.1	-1.3	1.2
CYP102A3	0.8	0.0	-0.8
One-body Terms	Parent		
Block	A1	A2	A3
1	0.5	-1.0	0.5
5	1.4	-0.8	-0.6
7	0.3	1.0	-1.4

Table VIII-2. One- and Two-Body Energ Terms Used to Calculaute LRA Engies for Lactamases

Two-body Terms	Parent at Block 8		
Parent at Block 1	PSE-4	SED-1	TEM-1
PSE-4	-1.2	1.7	-0.5
SED-1	-0.1	-2.8	2.8
TEM-1	1.3	1	-2.3
One-body Terms	Parent		
Block	PSE-4	SED-1	TEM-1
2	-0.5	1.1	-0.5
3	-0.6	1.1	-1.7

Calculation of Chimera One-Body Weights

The probability of finding each parental amino acid in the multiple sequence alignment for all positions was determined for each parental protein sequence. Gaps were excluded from the calculations. For cytochromes P450 the alignment used was

obtained from Dave Nelson (Nelson 2005), for lactamases the PFAM seed alignment was used (Bateman et al. 2004). C++ code to derive the parental probabilities from a multiple sequence alignment can be found in Appendix I. To determine *w* for each chimera, the probability of identifying the amino acid at each position was summed over all positions and divided by the total number of residues. Positions not in the multiple sequence alignment were not included in the average. *w* was normalized to the population of all possible chimeras according to Equation (VIII-4) to determine *W*. *W* will vary between 0 and 1 for most chimeras. The standard deviation on the population of all possible chimeras was estimated by calculating the standard deviation of *w* in a population of 100,000 randomly determined chimeras. Enumerating the entire population of possible chimeras is computationally intractable because there are 3^N possible chimeras, with three possible parent sequences and *N* amino acid positions.

The parameter *c* was optimized between 0 and 10 to give the largest mutual information between E_w and folding. The error on the measurement is determined by splitting the data into ten equal partitions and independently optimizing *c*, to obtain an average and standard deviation. To verify significant improvement of predictions with E_w compared with *E*, tenfold cross-validation was performed with c=1.0. For the tenfold cross-validation the data were split into 10 equally sized partitions. For each partition the data were fit using the other 90% of the data. The energy function was scored by its ability to predict the remaining 10% of the data by the change in mutual information (M). For lactamases the change in mutual information ($M(E_w)$-$M(E)$) was 0.0186 ± 0.010 and for cytochromes P450 the change in mutual information was 0.0938 ± 0.007. In both cases each partition displayed a positive change.

Appendix I: Computer Code

Crossovers.cc

```
#include <iostream.h>
#include <fstream.h>
#include <stdlib.h>
#include <math.h>
main()
{
const int libraryN=1000;    //number of libraries
const int crossN=5;     // number of crossovers
const int MINRES=24;  //minimum residue
const int MAXRES=290; //maximum residue
 int count;
int Crossovers[crossN];
 count=0;
int tog=0; //a toggle for loop generating random crossovers.
const int MINEND=15; //minimum distance to the ends of the protein
const int MINDIST=15; //minimum distance between crossovers.

 int temp; // temporary variable for sorting algorithm.
 ofstream crossoverfile ("5Crossovers.txt", ios::out);
 for (int i=0; i<libraryN; i++){
            tog=0;
            while (tog<=0){
              for (int j=0;j<crossN;j++){
                Crossovers[j]=(rand()%(MAXRES-MINRES))+MINRES;
              }
              for (int pass=0; pass<crossN-1; pass++){
                for (int k=0; k < crossN-1; k++){
                  if (Crossovers[k]> Crossovers[k+1]){
                        temp=Crossovers[k];
                        Crossovers[k]=Crossovers[k+1];
                        Crossovers[k+1]=temp;
                  }}}
                        //generates the list of random crossovers in array Crossovers
                        //the crossovers appear in chronological order
              count=count+1;
              int togl=1;
              if (Crossovers[0]<=MINRES+MINEND)
                togl=0;

              for (int k=0; k<crossN-1; k++){
                if (Crossovers[k+1]<((Crossovers[k])+MINDIST))
                  togl=0;
              }
              if (Crossovers[crossN-1]>MAXRES-MINEND)
                togl=0;
                        tog=togl;
            }
            for (int j=0; j<crossN; j++){
              crossoverfile<<Crossovers[j]<<"\t";
            }
            crossoverfile<<endl;
 }
 cout<<count<<endl;
return 0; }
```

Testlibraries.cc

```
#include <iostream.h>
#include <fstream.h>
#include <stdlib.h>
#include <math.h>

int minimum(int array[], int number){
  int M=1000;
  for (int i=1; i<=number; i++){
    if (array[i]<=M){
      M=array[i];
    }}
  return M;
}

float average(int array[], int number){
  float sum=0;
  float A=0;
  if (number!=0){
  for (int i=0;i<number; i++){
      sum=sum+array[i];
    }
  A=sum/number;
  }
return A;
}

float standarddev (int array[], int number, float  average){
float deviation=0, sum=0;
float S=0;
 if (number !=0){
  for (int i=0; i<number; i++){
    deviation=array[i]-average;
    deviation=deviation*deviation;
    sum=sum+deviation;
  }
    S=sqrt(sum/number);
 }
return S;
}

main()
{
  //contants that we need througout the whole program
    const int MAXN = 1506;
  //This number is dependent upon the pdbout file.
  const float DC = 4.5;
  // temporary variables for input before putting into structure
  char TATOM[5], TAATYPE[5];
  int  TATOMN, TRESN;
  float TX_CORR, TY_CORR, TZ_CORR;
  // define loop variable
  int i;
 //define a structure PDBin that contains all the info we want
      static struct PDBin{
        int   ATOMN;   //the atom number, unique identifier
```

```
    int   RESN;    // the residue number which the atom belongs to
    float X_CORR;       // x coordinate
    float Y_CORR;       // y coordinate
    float Z_CORR;       // z coordinate
  }protein[MAXN];    //the name of the structure type protein with maxn atoms
  // designed to take the data
  ifstream pdb("atomfile.txt");
  //opens file test.txt and calls it pdb
  //if the file cannot be opened return an error message
  if (pdb.bad()){
   cerr<<"error couldn not open test.txt";
   exit (8);
  }
  for (i=1; i<MAXN; ++i){
   pdb >> TATOMN>>TRESN>>TX_CORR>>TY_CORR>>TZ_CORR;
   protein[i].ATOMN = TATOMN;
   protein[i].RESN =TRESN;
   protein[i].X_CORR = TX_CORR;
   protein[i].Y_CORR = TY_CORR;
   protein[i].Z_CORR = TZ_CORR;
  }
   //This ends the section of code that is necessary for intaking the PDB file
  //The data is stored in structure protein
 // this next section is for determining a contact matrix between all the residues in the //protein.  It does not
//include any identity characteristics
  //First we are going to create a matrix in which to put the data
  //Then we are going determine the distance measurements and fill in the matrix.
    //These variables are required for the contact matrix formation.
     int j, k;
  int MAXRES, MINRES;
  float XX, YY, ZZ, Distance;
  MAXRES=protein[MAXN-1].RESN;
  MINRES=protein[1].RESN;
  int CMatrix[MAXRES+1][MAXRES+1];
  for (i=1;i<=MAXRES;i++){
   for (j=1;j<=MAXRES;j++)
         CMatrix[i][j]=0;
  }
   //We also require constants MAXN and DC for this code segment.
  int count=0;
  for (i=26;i<=MAXRES;i++){
   //this loop goes through each residue in the structure protein
   for (j=1; j<=MAXN; j++){
         // This loop goes through all atoms in the structure protein
         if (protein[j].RESN==i){
           //If atom (j) is in residue (i) then:
           for (k=j; k<MAXN; k++){
             //loop through all atoms again
             if(protein[k].RESN!=i){
               //check and make sure that the residue is not the one being examined
               XX=(protein[k].X_CORR-protein[j].X_CORR)*(protein[k].X_CORR-protein[j].X_CORR);
               YY=(protein[k].Y_CORR-protein[j].Y_CORR)*(protein[k].Y_CORR-protein[j].Y_CORR);
               ZZ=(protein[k].Z_CORR-protein[j].Z_CORR)*(protein[k].Z_CORR-protein[j].Z_CORR);
               Distance=sqrt(XX+YY+ZZ);
               //Determine distance between atoms j and k
               if (Distance<DC){
```

```
                    CMatrix[i][protein[k].RESN]=1;
                    CMatrix[protein[k].RESN][i]=1;
                    //makes both halves of the contact matrix;
                }}}}}}
    count=0;
    for (i=1;i<MAXRES;i++){
      for (j=1;j<MAXRES;j++)
            if (CMatrix[i][j]==1)
              count=count+1;
    }
    // This is the end of the code for generating the contact matrix.
    // This contact matrix is used for the remainder of the program.
// the next job is to correct for identitiy in the contact matrix.
            ifstream align1and2("TEM1PSE4.txt");
            //opens alignment between parent 1 and parent 2
            ifstream align1and3("PSE4SED1.txt");
            //opens alignment between parent 1 and parent 3
            ifstream align2and3("TEM1SED1.txt");
            //opens alignment between parent 2 and parent 3
            ifstream align1and4("PSE4AST.txt");
            //opens alignment between parent 1 and parent 4
            ifstream align2and4("TEM1AST.txt");
            //opens alignment between parent 2 and parent 4
            ifstream align3and4("AST1SED1.txt");
            //opens alignment between parent 3 and 4
            if (align1and2.bad()){
      cerr<<"error couldn not open TEM1PSE4.txt";
      exit (8);
    }
            if (align1and3.bad()){
      cerr<<"error couldn not open TEM1SHV1.txt";
      exit (8);
    }
            if (align2and3.bad()){
      cerr<<"error couldn not open PSE4SHV1.txt";
      exit (8);
    }
            if (align1and4.bad()){
              cerr<<"error could not open 1to4";
              exit (8);
            }
            if (align2and4.bad()){
              cerr<<"error could not open 2to4";
              exit(8);
            }
            if (align3and4.bad()){
              cerr<<"error could not open 3to4";
              exit (8);
                }
            //error messages for bad file inputs
            int index, value, l;
            static int alignmaster[5][5][350];
            static int mastercontact[5][5][350][350];
            int parentN=4;
            for (i=1; i<=parentN; i++){
                    for (j=1; j<=parentN; j++){
```

```
                        for (k=1; k<350; k++){
                                alignmaster[i][j][k]=1;
                                for (l=1;l< 350; l++){
                                mastercontact[i][j][k][l]=0;
                                }}}}
            //initialize matrices with correct ones or zeros
            //declare input variables and the matrix into which they are put
            for (i=MINRES; i<=MAXRES;i++){
              align1and2>>index>>value;
              alignmaster[1][2][index]=value;
              alignmaster[2][1][index]=value;
              align2and3>>index>>value;
              alignmaster[2][3][index]=value;
              alignmaster[3][2][index]=value;
              align1and3>>index>>value;
              alignmaster[1][3][index]=value;
              alignmaster[3][1][index]=value;
              align1and4>>index>>value;
              alignmaster[1][4][index]=value;
              alignmaster[4][1][index]=value;
              align2and4>>index>>value;
              alignmaster[2][4][index]=value;
              alignmaster[4][2][index]=value;
              align3and4>>index>>value;
              alignmaster[3][4][index]=value;
              alignmaster[4][3][index]=value;
                        }
            for (i=1; i<=parentN; i++){
              for (j=i;j<=parentN; j++){
                for (k=MINRES; k<=MAXRES;k++){
                  for (l=MINRES; l<=MAXRES; l++){
            mastercontact[i][j][k][l]=alignmaster[i][j][k]*alignmaster[i][j][l]*CMatrix[k][l];
            mastercontact[j][i][k][l]=alignmaster[i][j][k]*alignmaster[i][j][l]*CMatrix[k][l];
                 }}}}
            //this finished the contact array entering and identity correction
            //next we need to open files for outputting data and inputing crossover points.
            ifstream cross("crossovers.txt");
            ofstream output("9Clibsnew2.txt", ios::out);
            //we also start doing each library one at a time now
            //first read in the crossovers, then generate the chimeras
            //and finally evaluate each chimera in the library
            //compile the data and write to the output file.
            int Maxdis=55;
            int counter=0;
            int libraryN=29; //number of libraries to analyze
            int crossN=9;    //number of crossovers
            int total=262144; // possible number of chimeras in each library
            static int Chimeras[262144][300];
            for (i=0;i<total; i++){
              for (k=1;k<=MAXRES; k++){
                Chimeras[i][k]=0;
              }}
// this part must be modified for a greater number of crossovers.
int C1, C2, C3, C4, C5, C6, C7, C8, C9;  // these are the fragments created by 8 //crossovers.
int Crossovers[crossN];
 int sum,c, count1;
```

```
 float  AvD, AvM, AvDG, AvMG, AvDD, AvMD;
 float StD, StM, StDG, StMG;
int mut[parentN+1];
for(j=1;j<=parentN; j++)
  mut[j]=0;
static int gdisruption[262144];
static int gmutation [262144];
static int disruption[262144];
static int mutation[262144];
double  fractionH, fractionG, probabilityH, probabilityG, MutationP;
for (j=0;j<total; j++){
  disruption[j]=-1;
  mutation[j]=0;
  gmutation[j]=0;
  gdisruption[j]=-1;
 }
  for (j=0; j<libraryN; j++){ //this goes through each library
   counter=0;
   c=0;
   count1=0;
   for (k=0; k<crossN; k++){
     cross>>Crossovers[k];
   }
  for (C1=1; C1<=parentN; C1++){
     for (C2=1;C2<=parentN;C2++){
       for (C3=1;C3<=parentN;C3++){
              for (C4=1;C4<=parentN;C4++){
                for (C5=1;C5<=parentN;C5++){
                  for(C6=1;C6<=parentN;C6++){
                    for(C7=1;C7<=parentN; C7++){
                          for(C8=1;C8<=parentN; C8++){
                            for(C9=1;C9<=parentN; C9++){
                              for (i=1; i<=Crossovers[0]; i++)
                                Chimeras [counter][i]=C1;
                              for (i=(Crossovers[0]+1); i<=Crossovers[1]; i++)
                          Chimeras [counter][i]=C2;
                    for (i=(Crossovers[1]+1); i<=Crossovers[2]; i++)
                          Chimeras [counter][i]=C3;
                    for (i=(Crossovers[2]+1); i<=Crossovers[3]; i++)
                          Chimeras [counter][i]=C4;
                    for (i=(Crossovers[3]+1); i<=Crossovers[4]; i++)
                          Chimeras [counter][i]=C5;
                    for (i=(Crossovers[4]+1); i<=Crossovers[5]; i++)
                          Chimeras [counter][i]=C6;
                    for(i=(Crossovers[5]+1); i<=Crossovers[6]; i++)
                          Chimeras[counter][i]=C7;
                    for(i=(Crossovers[6]+1); i<=Crossovers[7]; i++)
                          Chimeras[counter][i]=C8;
                    for(i=(Crossovers[7]+1); i<=Crossovers[8]; i++)
                          Chimeras[counter][i]=C9;
                    for(i=(Crossovers[8]+1); i<=MAXRES; i++)
                          Chimeras[counter][i]=C1;
                          counter=counter+1;
              } } } } }}}}}
  fractionH=0;
  fractionG=0;
```

```
  MutationP=0;
  for (i=0; i<total; i++){
   sum=0;
   for (k=MINRES; k<=MAXRES; k++){
    for (l=k+1; l<=MAXRES; l++){
          if (Chimeras[i][k]!=Chimeras[i][l])
           sum=sum+mastercontact[(Chimeras[i][k])][(Chimeras[i][l])][k][l];
    }
     for (l=1; l<=parentN; l++){
          if (Chimeras[i][k]!=l){
          if ((alignmaster[(Chimeras[i][k])][l][k])==1)
           mut[l]=mut[l]+1;
          }}}
   disruption[i]=sum;
   probabilityH=pow((1-((.1*sum)/322)),322);
   probabilityG=pow((1-((.04*sum)/322)),322);
   fractionG=fractionG +probabilityG;
   fractionH=fractionH+probabilityH;
   mutation[i]=minimum (mut, parentN);
   MutationP=MutationP+(probabilityG*mutation[i]);
   for (l=1; l<=parentN; l++){
    mut[l]=0;
   }
   if (disruption[i]<Maxdis){
    gdisruption[c]=disruption[i];
    gmutation[c]=mutation[i];
    c=c+1;
   } }
  AvD=average(disruption, total);
  AvM=average(mutation, total);
  AvDG=average(gdisruption, c);
  AvMG=average(gmutation, c);
  StM=standarddev(mutation, total, AvM);
  StD=standarddev(disruption, total, AvD);
  StDG=standarddev(gdisruption, c, AvDG);
  StMG=standarddev(gmutation, c, AvMG);
                output<<AvD<<"\t"<<AvM<<"\t"<<StD<<"\t"<<StM<<"\t"
                      <<c<<"\t"<<AvDG<<"\t" <<AvMG<<"\t"<<StDG<<"\t"
                      <<StMG<<"\t"<<fractionG<<"\t"<<fractionH<<"\t"<<MutationP<<"\t";
  for (i=0; i<crossN; i++){
   output<<Crossovers[i]<<"\t";
  }
  output<<endl;
 }
return 0;
}
```

Librarytest.cc

```
#include <iostream.h>
#include <fstream.h>
#include <stdlib.h>
#include <math.h>
int minimum(int array[], int number){
  int M=1000;
  for (int i=1; i<=number; i++){
    if (array[i]<=M){
      M=array[i];
    } }
  return M;
}
float average(int array[], int number){
  float sum=0;
  float A=0;
  if (number!=0){
  for (int i=0;i<number; i++){
      sum=sum+array[i];
    }
  A=sum/number;
  }
return A;
}
float standarddev (int array[], int number, float  average){
float deviation=0, sum=0;
float S=0;
 if (number !=0){
  for (int i=0; i<number; i++){
    deviation=array[i]-average;
    deviation=deviation*deviation;
    sum=sum+deviation;
  }
    S=sqrt(sum/number);
 }
return S;
}

main()
{
  //contants that we need througout the whole program
  const int MAXN = 2057;
  const float DC = 4.5;
    // temporary variables for input before putting into structure
  char junk[4], junk2[2], junk5[2];
  float junk3, junk4;
  char TATOM[5], TAATYPE[5];
  int  TATOMN, TRESN;
  float TX_CORR, TY_CORR, TZ_CORR;
  // define loop variable
  int i;
  //define a structure PDBin that contains all the info we want
   static struct PDBin{
        int   ATOMN;   //the atom number, unique identifier
        int   RESN;    // the residue number which the atom belongs to
        float X_CORR;       // x coordinate
```

```
    float Y_CORR;       // y coordinate
    float Z_CORR;       // z coordinate
  }protein[MAXN];    //the name of the structure type protein with maxn atoms
  // designed to take the data
  ifstream pdb("Atomfile.txt");
  //opens file test.txt and calls it pdb
  //if the file cannot be opened return an error message
  if (pdb.bad()){
   cerr<<"error couldn not open test.txt";
   exit (8);
  }
  for (i=1; i<MAXN; ++i){
   pdb >>TATOMN>>TRESN>>TX_CORR>>TY_CORR>>TZ_CORR;
   protein[i].ATOMN = TATOMN;
   protein[i].RESN =TRESN;
   protein[i].X_CORR = TX_CORR;
   protein[i].Y_CORR = TY_CORR;
   protein[i].Z_CORR = TZ_CORR;
  }
   //This ends the section of code that is necessary for intaking the PDB file
  //The data is stored in structure protein
  // this next section is for determining a contact matrix between all the residues in the protein.  It does not
include any identity characteristics
  //First we are going to create a matrix in which to put the data
  //Then we are going determine the distance measurements and fill in the matrix.
     //These variables are required for the contact matrix formation.
    int j, k;
  int MAXRES, MINRES;
  float XX, YY, ZZ, Distance;
  MAXRES=protein[MAXN-1].RESN;
  MINRES=protein[1].RESN;
 int CMatrix[MAXRES+1][MAXRES+1];
  for (i=1;i<=MAXRES;i++){
   for (j=1;j<=MAXRES;j++)
        CMatrix[i][j]=0;
  }
     //We also require constants MAXN and DC for this code segment.
  int count=0;
  for (i=26;i<=MAXRES;i++){
   //this loop goes through each residue in the structure protein
 for (j=1; j<=MAXN; j++){
        // This loop goes through all atoms in the structure protein
        if (protein[j].RESN==i){
          //If atom (j) is in residue (i) then:
          for (k=j; k<MAXN; k++){
           //loop through all atoms again
                  if(protein[k].RESN!=i){
            //check and make sure that the residue is not the one being examined
            XX=(protein[k].X_CORR-protein[j].X_CORR)*(protein[k].X_CORR-protein[j].X_CORR);
           YY=(protein[k].Y_CORR-protein[j].Y_CORR)*(protein[k].Y_CORR-protein[j].Y_CORR);
            ZZ=(protein[k].Z_CORR-protein[j].Z_CORR)*(protein[k].Z_CORR-protein[j].Z_CORR);
             Distance=sqrt(XX+YY+ZZ);
            //Determine distance between atoms j and k
            if (Distance<DC){
                  CMatrix[i][protein[k].RESN]=1;
                  CMatrix[protein[k].RESN][i]=1;
```

```
                    //makes both halves of the contact matrix;
                }} }}}}
    count=0;
    for (i=1;i<MAXRES;i++){
      for (j=1;j<MAXRES;j++)
            if (CMatrix[i][j]==1)
              count=count+1;
    }
       // This is the end of the code for generating the contact matrix.
    // This contact matrix is used for the remainder of the program.
// the next job is to correct for identitiy in the contact matrix.
            ifstream align1and2("TEM1PSE4.txt");
            //opens alignment between parent 1 and parent 2
            ifstream align1and3("PSE4SED1.txt");
            //opens alignment between parent 1 and parent 3
            ifstream align2and3("TEM1SED1.txt");
            //opens alignment between parent 2 and parent 3
            ifstream align1and4("PSE4AST.txt");
            //opens alignment between parent 1 and parent 4
            ifstream align2and4("TEM1AST.txt");
            //opens alignment between parent 2 and parent 4
            ifstream align3and4("AST1SED1.txt");
            //opens alignment between parent 3 and 4
            if (align1and2.bad()){
      cerr<<"error couldn not open TEM1PSE4.txt";
      exit (8);
    }
            if (align1and3.bad()){
      cerr<<"error couldn not open TEM1SHV1.txt";
      exit (8);
    }
            if (align2and3.bad()){
      cerr<<"error couldn not open PSE4SHV1.txt";
      exit (8);
    }
            if (align1and4.bad()){
              cerr<<"error could not open 1to4";
              exit (8);
            }
            if (align2and4.bad()){
              cerr<<"error could not open 2to4";
              exit(8);
            }
            if (align3and4.bad()){
              cerr<<"error could not open 3to4";
              exit (8);
                }
            //error messages for bad file inputs
            int index, value, l;
            static int alignmaster[5][5][350];
            static int mastercontact[5][5][350][350];
            int parentN=4;

            for (i=1; i<=parentN; i++){
                        for (j=1; j<=parentN; j++){
                                    for (k=1; k<350; k++){
```

```
                                        alignmaster[i][j][k]=1;
                                        for (l=1;l< 350; l++){
                                        mastercontact[i][j][k][l]=0;
                                        }}}}
                    //initialize matrices with correct ones or zeros
                    //declare input variables and the matrix into which they are put
                    for (i=MINRES; i<=MAXRES;i++){
                      align1and2>>index>>value;
                      alignmaster[1][2][index]=value;
                      alignmaster[2][1][index]=value;
                      align2and3>>index>>value;
                      alignmaster[2][3][index]=value;
                      alignmaster[3][2][index]=value;
                      align1and3>>index>>value;
                      alignmaster[1][3][index]=value;
                      alignmaster[3][1][index]=value;
                      align1and4>>index>>value;
                      alignmaster[1][4][index]=value;
                      alignmaster[4][1][index]=value;
                      align2and4>>index>>value;
                      alignmaster[2][4][index]=value;
                      alignmaster[4][2][index]=value;
                      align3and4>>index>>value;
                      alignmaster[3][4][index]=value;
                      alignmaster[4][3][index]=value;
                                }
                    for (i=1; i<=parentN; i++){
                      for (j=i;j<=parentN; j++){
                        for (k=MINRES; k<=MAXRES;k++){
                          for (l=MINRES; l<=MAXRES; l++){
                    mastercontact[i][j][k][l]=alignmaster[i][j][k]*alignmaster[i][j][l]*CMatrix[k][l];
                    mastercontact[j][i][k][l]=alignmaster[i][j][k]*alignmaster[i][j][l]*CMatrix[k][l];
                         }}}}
                    //this finished the contact array entering and identity correction
                    //next we need to open files for outputting data and inputing crossover points.
                    ifstream cross("libraryX.txt");
                    ofstream output("Libraryxs.txt", ios::out);
                    //we also start doing each library one at a time now
                    //first read in the crossovers, then generate the chimeras
                    //and finally evaluate each chimera in the library
                    //compile the data and write to the output file.
                    int Maxdis=55;
                    int counter=0;
                    int crossN=9;    //number of crossovers
                    int total=262144; // possible number of chimeras in each library
                    static int Chimeras[262144][300];
                    for (i=0;i<total; i++){
                      for (k=1;k<=MAXRES; k++){
                        Chimeras[i][k]=0;
                      }}
// this part must be modified for a greater number of crossovers.
int C1, C2, C3, C4, C5, C6, C7, C8, C9;  // these are the fragments created by 9 crossovers.
int Crossovers[crossN];
 int sum,c, count1;
int mut[parentN+1];
for(j=1;j<=parentN; j++)
```

```
  mut[j]=0;
static int disruption[262144];
static int mutation[262144];
 for (j=0;j<total; j++){
  disruption[j]=-1;
  mutation[j]=0;
 }
cout<<"made it here"<<endl;
   counter=0;
   c=0;
   count1=0;
   for (k=0; k<crossN; k++){
     cross>>Crossovers[k];
   }

   for (C1=1; C1<=parentN; C1++){
     for (C2=1;C2<=parentN;C2++){
       for (C3=1;C3<=parentN;C3++){
              for (C4=1;C4<=parentN;C4++){
                for (C5=1;C5<=parentN;C5++){
                  for(C6=1;C6<=parentN;C6++){
                    for(C7=1;C7<=parentN; C7++){
                          for(C8=1;C8<=parentN; C8++){
                            for(C9=1; C9<=parentN; C9++){
                    for (i=1; i<=Crossovers[0]; i++)
                          Chimeras [counter][i]=C1;
                    for (i=(Crossovers[0]+1); i<=Crossovers[1]; i++)
                          Chimeras [counter][i]=C2;
                    for (i=(Crossovers[1]+1); i<=Crossovers[2]; i++)
                          Chimeras [counter][i]=C3;
                    for (i=(Crossovers[2]+1); i<=Crossovers[3]; i++)
                          Chimeras [counter][i]=C4;
                    for (i=(Crossovers[3]+1); i<=Crossovers[4]; i++)
                          Chimeras [counter][i]=C5;
                    for (i=(Crossovers[4]+1); i<=Crossovers[5]; i++)
                          Chimeras [counter][i]=C6;
                    for(i=(Crossovers[5]+1); i<=Crossovers[6]; i++)
                          Chimeras[counter][i]=C7;
                    for(i=(Crossovers[6]+1); i<=Crossovers[7]; i++)
                          Chimeras[counter][i]=C8;
                    for(i=(Crossovers[7]+1); i<=Crossovers [8]; i++)
                          Chimeras[counter][i]=C9;
                    for(i=(Crossovers[8]+1); i<=MAXRES; i++)
                          Chimeras[counter][i]=C1;
                    counter=counter+1;
                          }}}}}}}}}
   for (i=0; i<total; i++){
     sum=0;
     for (k=MINRES; k<=MAXRES; k++){
       for (l=k+1; l<=MAXRES; l++){
              if (Chimeras[i][k]!=Chimeras[i][l])
                sum=sum+mastercontact[(Chimeras[i][k])][(Chimeras[i][l])][k][l];
       }
             for (l=1; l<=parentN; l++){
              if (Chimeras[i][k]!=l){
              if ((alignmaster[(Chimeras[i][k])][l][k])==1)
```

```
            mut[l]=mut[l]+1;
           }
      }
    }
    disruption[i]=sum;
    mutation[i]=minimum (mut, parentN);
    for (l=1; l<=parentN; l++){
      mut[l]=0;
    }
    output<<disruption[i]<<"\t"<<mutation[i]<<"\t";
    for (l=0; l<crossN;l++)
      output<<Chimeras[i][Crossovers[l]]<<"\t";
    output<<Chimeras[i][291]<<endl;

  }

return 0;
}
```

Libraryparse.cc

```
#include <stdio.h>
#include <iostream.h>
#include <fstream.h>
#include <cstdlib>
#include <cmath>

float average(float array[], int number){
  float sum=0;
  float A=0;
  if (number!=0){
  for (int i=0;i<number; i++){
     sum=sum+array[i];
    }
  A=sum/number;
  }
return A;
}

float Pfunction(float E){
  float Pf=0;
  Pf=1/(1+exp((0.138*E)-3.44));
  return Pf;
 }

float Pfunctionold(float E){
  float Pf=0;
  Pf=pow(1-0.0734*(E/322), 322);
  return Pf;
}

main ()
{
 ifstream datafile("RASPPdataX7.txt");
   if (datafile.bad()){
      cerr<<"error could not open file.txt";
      exit (8);
    }
  ofstream output("RASPPalllibrarydataME.txt", ios::out);
  //for each library loop through and calculate all the Pf and Pfold
  float libEmean, libMmean, libMPfinal, libMPoldfinal, Ffold, Ffoldold, libPf, libPfold, libmf, libmfold,
AvgME;
  int value1, value2, value3, i, j;
  float chimM[6561];
  float chimPf[6561], chimPfold[6561], chimE[6561], ME[6561];
  for (j=1; j<=1450; j++){
    libEmean=0;
    libMmean=0;
    libMPfinal=0;
    libMPoldfinal=0;
    Ffold=0;
    Ffoldold=0;
    libPf=0;
    libPfold=0;
    libmf=0;
    libmfold=0;
```

```
   for (i=0; i<6561; i++){
     chimPfold[i]=0;
     chimPf[i]=0;
     chimE[i]=0;
     chimM[i]=0;
   }
   for (i=0;i<6561; i++){
     datafile>>value1>>value2>>value3;
     if(i==0){
          cout<<j<<"\t"<<value1<<"\t"<<value2<<"\t"<<value3<<endl;
     }
     chimPf[i]=Pfunction(value2);
     chimPfold[i]=Pfunctionold(value2);
     chimE[i]=value2;
     chimM[i]=value3;
     libPf=libPf+chimPf[i];
     libPfold=libPfold+chimPfold[i];
     libmf=libmf+(chimPf[i]*chimM[i]);
     libmfold=libmfold+(chimPfold[i]*chimM[i]);
          if(value2==0)
                    ME[i]=0;
          else
                    ME[i]=value3/value2;
   }
   libEmean=average(chimE, 6561);
   libMmean=average(chimM, 6561);
   libMPfinal=libmf/libPf;
   libMPoldfinal=libmfold/libPfold;
   Ffold=libPf/6561;
   Ffoldold=libPfold/6561;
   AvgME=average(ME, 6561);

output<<j<<"\t"<<libEmean<<"\t"<<libMmean<<"\t"<<AvgME<<"\t"<<Ffold<<"\t"<<libMPfinal<<"\t"<
<Ffoldold<<"\t"<<libMPoldfinal<<endl;
  }

 return 0;
}
```

Librarysimulates.py
(needs other python tools from http://www.che.caltech.edu/~groups/fha)

```
#! /usr/bin/env python

import sys, os, math, string, random
import pdb, schema

def main ():
	#Read the parents
	parent_list = schema.readMultipleSequenceAlignmentFile (file('lac-msa.txt','r'))
	parents = [p for (key,p) in parent_list]
	pdb_alignment_list = schema.readMultipleSequenceAlignmentFile(file('PSE4-1G68.txt','r'))
	pdb_alignment= [p for (key, p) in pdb_alignment_list]

	# Read in the contact map
	pdb_residues = pdb.File().read(file('1G68.pdb','r'))
	residues = schema.alignPDBResidues(pdb_residues, pdb_alignment[1], pdb_alignment[0],
parents[0], ['A',' '])
	pdb_contacts = schema.getPDBContacts (residues, 4.5)
	contacts =schema.getSCHEMAContacts(pdb_contacts, parents)


	filename=file('7XRASPPdata.txt', 'w')
	filename.write("# E	m\n")

	lines = file('7Xraspplib.txt', 'r').readlines()

	for line in lines:
		if line[0] == '#':
			continue
		flds = line.split()
		crossovers = [int(x) for x in flds]
		filtered_contacts = schema.getSCHEMAContactsWithCrossovers(contacts, parents,
crossovers)
		fragments = schema.getFragments(crossovers, parents[0])
		p=len(parents)
		n=len(fragments)
		for i in range (p**n):
			#make chimeras into block pattersn
			n2c = schema.base(i,p)
			chimera_blocks = ''.join(['1']*(n-len(n2c))+['%d'%(int(x)+1,) for x in n2c])
			E = schema.getChimeraDisruption(chimera_blocks, filtered_contacts,
fragments, parents)
			m = schema.getChimeraShortestDistance(chimera_blocks, fragments, parents)
			filename.write("%d\t%d\t%d\n" % (i , E, m))



main ()
```

GAMS file for LRA

```
Sets

i chimera number /1*163/
j1 block /1*8/
p1 parent1   /1*3/
;

alias (j1, j2);
alias (j1,j3);
alias(j1,j4);
alias (p1, p2);

Parameters

chimeras(i,j1,p1)
/
 1.1.1 = 0,  1.1.2 = 0,  1.1.3 = 1,  1.2.1 = 1,  1.2.2 = 0,  1.2.3 = 0,  1.3.1 = 0,  1.3.2 = 0,  1.3.3 = 1,
1.4.1 = 1,  1.4.2 = 0,  1.4.3 = 0,  1.5.1 = 0,  1.5.2 = 1,  1.5.3 = 0,  1.6.1 = 0,  1.6.2 = 0,  1.6.3 = 1,  1.7.1
= 0,  1.7.2 = 0,  1.7.3 = 1,  1.8.1 = 0,  1.8.2 = 0,  1.8.3 = 1,
```

All chimeras represented in this particular format where 1 (31312333) and 163 (13313332) are each a chimera

```
 163.1.1 = 0,  163.1.2 = 1,  163.1.3 = 0,  163.2.1 = 0,  163.2.2 = 0,  163.2.3 = 1,  163.3.1 = 0,  163.3.2
= 0,  163.3.3 = 1,  163.4.1 = 1,  163.4.2 = 0,  163.4.3 = 0,  163.5.1 = 0,  163.5.2 = 0,  163.5.3 = 1,
163.6.1 = 0,  163.6.2 = 0,  163.6.3 = 1,  163.7.1 = 0,  163.7.2 = 0,  163.7.3 = 1,  163.8.1 = 0,  163.8.2 =
1,  163.8.3 = 0

/
*0 = unfolded , 1 = functional
fold(i)
/
1          0
163        0
/;

Variables
E(i)
Eo
Es(j1,p1)
Ep(j1,p1,j2,p2)
D;
Equations
energy_defn(i)
deviance
dof_single(j1)
dof_pair_row(j1,j2,p1)
dof_pair_col(j1,j2,p2);
energy_defn(i) .. E(i) =e= Eo + sum(j1,sum(p1$chimeras(i,j1,p1),Es(j1,p1))) +
sum(j1,sum(p1$chimeras(i,j1,p1),sum(j2$(ord(j2) > ord(j1)),sum(p2$chimeras(i,j2,p2),Ep(j1,p1,j2,p2)))));
dof_single(j1) .. 0 =e= sum(p1,Es(j1,p1));
dof_pair_row(j1,j2,p1) .. 0 =e= sum(p2,Ep(j1,p1,j2,p2));
dof_pair_col(j1,j2,p2) .. 0 =e= sum(p1,Ep(j1,p1,j2,p2));
deviance .. D =e= -2*sum(i$(fold(i)=0),E(i)) + 2*sum(i,log(1+exp(E(i))));
```

```
Model
logistic /all/;
logistic.optfile = 1;
Scalar enrg_limit /40/;
*Solve reference model the one bodies and that's it.
Es.up(j1,p1) = enrg_limit;
Es.lo(j1,p1) = -enrg_limit;
Ep.fx(j1, p1, j2, p2) = 0;
Solve logistic using nlp minimizing D;
*these are paramters to store the differences in logistic function D
*dbase is the base value from the last model (assigned D.1 to last model)
*delta pair is for pair of fragments added
*delta single of for removing a single fragment
Parameter
        Dbase,
        deltaD_singlefragment(j3),
        deltaD_pairoffragments(j3,j4);

        Dbase = D.l;
*This loop goes through all single fragments and removes/adds them to calculate the change in model from
removing these fragments
        loop(j3$(ord(j3) > 0),
                Es.fx(j3,p1) = 0;
        Solve logistic using nlp minimizing D;
                deltaD_singlefragment(j3) = Dbase - D.l;
*reset parameter bounds on energies
                Es.up(j3,p1) = enrg_limit;
                Es.lo(j3,p1) = -enrg_limit;
        );
*now we go through each fragment and try adding each pair of fragments to calculate the change in the
model from adding this pair
        loop(j4$(ord(j4)>0),
                Ep.up('1',p1,j4,p2) = enrg_limit;
                Ep.lo('1',p1,j4,p2) = -enrg_limit;
        Solve logistic using nlp minimizing D;
                deltaD_pairoffragments('1',j4) = Dbase - D.l;
                Ep.fx('1', p1,j4,p2) = 0
        );
        loop(j4$(ord(j4)>0),
                Ep.up('2',p1,j4,p2) = enrg_limit;
                Ep.lo('2',p1,j4,p2) = -enrg_limit;
        Solve logistic using nlp minimizing D;
                deltaD_pairoffragments('2',j4) = Dbase - D.l;
                Ep.fx('2', p1,j4,p2) = 0
        );
        loop(j4$(ord(j4)>0),
                Ep.up('3',p1,j4,p2) = enrg_limit;
                Ep.lo('3',p1,j4,p2) = -enrg_limit;

        Solve logistic using nlp minimizing D;
                deltaD_pairoffragments('3',j4) = Dbase - D.l;
                Ep.fx('3', p1,j4,p2) = 0
        );

        loop(j4$(ord(j4)>0),
                Ep.up('4',p1,j4,p2) = enrg_limit;
```

```
            Ep.lo('4',p1,j4,p2) = -enrg_limit;

        Solve logistic using nlp minimizing D;
            deltaD_pairoffragments('4',j4) = Dbase - D.l;
            Ep.fx('4', p1,j4,p2) = 0
        );


        loop(j4$(ord(j4)>0),
            Ep.up('5',p1,j4,p2) = enrg_limit;
            Ep.lo('5',p1,j4,p2) = -enrg_limit;

        Solve logistic using nlp minimizing D;
            deltaD_pairoffragments('5',j4) = Dbase - D.l;
            Ep.fx('5', p1,j4,p2) = 0
        );

        loop(j4$(ord(j4)>0),
            Ep.up('6',p1,j4,p2) = enrg_limit;
            Ep.lo('6',p1,j4,p2) = -enrg_limit;

        Solve logistic using nlp minimizing D;
            deltaD_pairoffragments('6',j4) = Dbase - D.l;
            Ep.fx('6', p1,j4,p2) = 0
        );


        loop(j4$(ord(j4)>0),
            Ep.up('7',p1,j4,p2) = enrg_limit;
            Ep.lo('7',p1,j4,p2) = -enrg_limit;

        Solve logistic using nlp minimizing D;
            deltaD_pairoffragments('7',j4) = Dbase - D.l;
            Ep.fx('7', p1,j4,p2) = 0
        );

        loop(j4$(ord(j4)>0),
            Ep.up('8',p1,j4,p2) = enrg_limit;
            Ep.lo('8',p1,j4,p2) = -enrg_limit;

        Solve logistic using nlp minimizing D;
            deltaD_pairoffragments('8',j4) = Dbase - D.l;
            Ep.fx('8', p1,j4,p2) = 0
        );


Display Dbase;
Display deltaD_singlefragment;
Display deltaD_pairoffragments;
```

FAM.cc

```
#include <stdio.h>
#include <iostream.h>
#include <fstream.h>
#include <cstdlib>
#include <cmath>
#include <ctime>
main()
{
 int i,j; //loop variables
 //open the alignment file
 ifstream alignment("lactamasePfam.txt");
 if(alignment.bad()){
  cerr<<"error could not open alignment";
  exit(8);
 }

 ofstream distributions("distributionlac.txt", ios::out);
 ofstream scores ("scoreslac.txt", ios::out);

 const int size=703;// (j)
 const int numbersequences=133; //(i)
 static int sequencestore[numbersequences][size];
 char aasequence[size];
 char name;

 //extract each line(i) and assign each letter(j) the appropriate info

 for (i=0; i<numbersequences; i++){
  alignment>>aasequence;
  //cout<<aasequence<<endl<<endl;
  for(j=0; j<size; j++){
   if(aasequence[j]=='-'|| aasequence[j]=='*'|| aasequence[j]=='.'){
         sequencestore[i][j]=0;     }
   else if(aasequence[j]=='A'|| aasequence[j]=='a'){
         sequencestore[i][j]=1;     }
   else if (aasequence[j]=='C'|| aasequence[j]=='c'){
         sequencestore[i][j]=2;     }
   else if (aasequence[j]=='D'|| aasequence[j]=='d'){
         sequencestore[i][j]=3;     }
   else if (aasequence[j]=='E'|| aasequence[j]=='e'){
         sequencestore[i][j]=4;     }
   else if (aasequence[j]=='F'|| aasequence[j]=='f'){
         sequencestore[i][j]=5;      }
   else if (aasequence[j]=='G'|| aasequence[j]=='g'){
         sequencestore[i][j]=6;     }
   else if (aasequence[j]=='H'|| aasequence[j]=='h'){
         sequencestore[i][j]=7;     }
   else if (aasequence[j]=='I'|| aasequence[j]=='i'){
         sequencestore[i][j]=8;     }
   else if (aasequence[j]=='K'|| aasequence[j]=='k'){
         sequencestore[i][j]=9;     }
   else if (aasequence[j]=='L'|| aasequence[j]=='l'){
         sequencestore[i][j]=10;     }
   else if (aasequence[j]=='M'|| aasequence[j]=='m'){
         sequencestore[i][j]=11;     }
```

```
    else if (aasequence[j]=='N'|| aasequence[j]=='n'){
            sequencestore[i][j]=12;      }
    else if (aasequence[j]=='P'|| aasequence[j]=='p'){
            sequencestore[i][j]=13;      }
     else if (aasequence[j]=='Q'|| aasequence[j]=='q'){
            sequencestore[i][j]=14;       }
    else if (aasequence[j]=='R'|| aasequence[j]=='r'){
            sequencestore[i][j]=15;       }
     else if (aasequence[j]=='S'|| aasequence[j]=='s'){
             sequencestore[i][j]=16;       }
     else if (aasequence[j]=='T'|| aasequence[j]=='t'){
             sequencestore[i][j]=17;      }
     else if (aasequence[j]=='V'|| aasequence[j]=='v'){
             sequencestore[i][j]=18;      }
     else if (aasequence[j]=='W'|| aasequence[j]=='w'){
             sequencestore[i][j]=19;       }
            else if  (aasequence[j]=='Y'|| aasequence[j]=='y'){
              sequencestore[i][j]=20;    }
    else {
            sequencestore[i][j]=21;
                }    } }

  int bins[size][22];
  int a;
  for (j=0; j<size; j++){
    for (a=0; a<22; a++){
      bins[j][a]=0;
    }
    for (i=0; i<numbersequences; i++){
      bins[j][sequencestore[i][j]]=bins[j][sequencestore[i][j]]+1;
    }
    if(sequencestore[0][j]!=0){
      distributions<<j<<"\t"<<sequencestore[0][j]<<"\t";
      for (a=0; a<22; a++){
            distributions<<bins[j][a]<<"\t";
      }
      distributions<<endl;
    } }

  float weight[3][j];
  float  sum[j];
  for (j=0; j<size; j++){
    sum[j]=0;
    if(sequencestore[0][j]!=0 && sequencestore[0][j]!=21){

      for (a=1; a<21; a++){
            sum[j]=sum[j]+bins[j][a];
      }
      for (a=0; a<3; a++){
            weight[a][j]=bins[j][sequencestore[a][j]]/sum[j];
      }
scores<<j<<"\t"<<sum[j]<<"\t"<<sequencestore[0][j]<<"\t"<<weight[0][j]<<"\t"<<weight[1][j]<<"\t"<<w
eight[2][j]<<endl;
    }    }}
```

Appendix II: Primers and Oligonucleotides Used for Construction and Analysis of Recombination Libraries

Table AII-1. Primers used in the construction of RandE:APST and RandEPST libraries for the parent proteins: P, PSE-4; S, SED-1; A, AST-1; T, TEM-1. All primers are named for the parent P2F, the block P2F, and whether they are forward (coding strand) or reverse (noncoding strand) P2F. Underlined regions are 5' overhangs used for construction. Letters shown in bold are single base mutations from the native sequence to either make the overhangs match or remove restriction sites.

Primers that are annealed to form internal fragments all 5' to 3'

P2F	Cqcttcccgttaacaagtacttttaaaacaatagcttgcgctaaatta
P2R	**CAA**TAATTTAGCGCAAGCTATTGTTTTAAAAGTACTTGTTAACGGGAA
S2F	Cq**C**tttgcgatgtgcagcaccagtaaggtcatgaccgccgccgcggta
S2R	**C**AATACCGCGGCGGCGGTCATGACCTTACTGGTGCTGCACATCGCAAA
A2F	Cq**C**ttcccgatggcgtccacgttcaagggcctggcgtgcggggcgctg
A2R	**CAA**CAGCGCCCCGCACGCCAGGCCCTTGAACGTGGACGCCATCGGGAA
T2F	CG**c**TTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTA
T2R	**C**AATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAA
P3F	**TTG**tatgatgctgagcaaggaaaagttaatcccaatagtacagtcgagattaagaaagc
P3R	TC**A**GCTTTCTTAATCTCGACTGTACTATTGGGATTAACTTTTCCTTGCTCAGCATCATA
S3F	Tt**G**aaacagagtgaaacccatgacggtattttgcagcaaaaaatgaccattaaaaaagc
S3R	TC**A**GCTTTTTTAATGGTCATTTTTTGCTGCAAAATACCGTCATGGGTTTCACTCTGTTT
A3F	**TTG**cgcgagcatcccctgtcgacgGgctacttcgatcaggtgatccactactccgccgc
A3R	TC**A**GCGGCGGAGTAGTGGATCACCTGATCGAAGTAGCCCGTCGACAGGGGATGCTCGCG
T3F	TT**g**TCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAA
T3R	TCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGA
P4F	**T**gatcttgtgacctattcccctgtaatagaaaagcaagtagggcaggcaatc
P4R	**C**GTGATTGCCTGCCCTACTTGCTTTTCTATTACAGGGGAATAGGTCACAAGA
S4F	**T**gatctgaccaactggaatcccgtaacagagaaatatgtgggtaatacgatg
S4R	CGTCATCGTATTACCCACATATTTCTCTGTTACGGGATTCCAGTTGGTCAGA
A4F	**T**gagctggtcgagtattcgccggtgaccgagacccgggtcgagaccggcatg
A4R	**C**GTCATGCCGGTCTCGACCCGGGTCTCGGTCACCGGCGAATACTCGACCAGC
T4F	TGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATG
T4R	**C**GTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAG
P5F	Ac**g**ctcgatgatgcgtgcttcgcaactatgactacaagtgataatactgcggcaaatatcatc
P5R	TAGGATGATATTTGCCGCAGTATTATCACTTGTAGTCATAGTTGCGAAGCACGCATCATCGAG
S5F	Ac**g**ttagctgagctaagcgcagcgacgttacagtacagcgataataccgccatgaataaactg
S5R	**T**AGCAGTTTATTCATGGCGGTATTATCGCTGTACTGTAACGTCGCTGCGCTTAGCTCAGCTAA
A5F	Ac**g**gtccgggaactgtgcgacgccgcgatcacggtttccgacaacacggcgggcaatcagttg
A5R	**T**AGCAACTGATTGCCCGCCGTGTTGTCGGAAACCGTGATCGCGGCGTCGCACAGTTCCCGGAC
T5F	AC**g**GTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTT
T5R	**T**AGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTAC

P6F	Ctaagtgctgtaggtggccccaaaggcgttactgatttttaagacaaatt
P6R	CCCAATTTGTCTTAAAAAATCAGTAACGCCTTTGGGGCCACCTACAGCACT
S6F	Ct**A**gcgcatcttggcggccccggcaacgtcacggcgtttgcacgttccatt
S6R	**C**CCAATGGAACGTGCAAACGCCGTGACGTTGCCGGGGCCGCCAAGATGCGC
A6F	Ct**A**aaactgctcggtggaccggagggattcaccgcgtccctgcgttccctc
A6R	**C**CCGAGGGAACGCAGGGACGCGGTGAATCCCTCCGGTCCACCGAGCAGTTT
T6F	CT**a**ACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATG
T6R	CCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGT
P7F	Gqqgacaaagagactcgtctagaccgtattgagcctgatttaaatgaaggtaagctcggt
P7R	ATCACCGAGCTTACCTTCATTTAAATCAGGCTCAATACGGTCTAGACGAGTCTCTTTGTC
S7F	Gq**G**gacacgacgtttcgtctcgatcgcaaagagccggaattaaacaccgccattcccggc
S7R	ATCGCCGGGAATGGCGGTGTTTAATTCCGGCTCTTTGCGATCGAGACGAAACGTCGTGTC
A7F	Gq**G**gacgccacgtcgcggctggaccgctgggagaccgacctgaacaccgcgattcccggg
A7R	ATCCCCGGGAATCGCGGTGTTCAGGTCGGTCTCCCAGCGGTCCAGCCGCGACGTGGCGTC
T7F	GGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAAC
T7R	**A**TCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATC
P8F	Gatttgagggatacgacaactcctaaggcaatagccagtactttgaataaatttttattt
P8R	GCCAAATAAAAATTTATTCAAAGTACTGGCTATTGCCTTAGGAGTTGTCGTATCCCTCAA
S8F	Gatgagcgcgacacaacatcgccgctggc**g**atggccaaaagtctgcgtaaactcacgctg
S8R	GCCCAGCGTGAGTTTACGCAGACTTTTGGCCAT**c**GCCAGCGGCGATGTTGTGTCGCGCTC
A8F	Gatgagcgcgataccaccaccccggccgcgctcgccgccgactaccgcgcgctcgtcgtc
A8R	GCCGACGACGAGCGCGCGGTAGTCGGCGGCGAGCGCGGCCGGGGTGGTGGTATCGCGCTC
T8F	GA**t**GAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACT
T8R	GCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTC
P9F	Gg**c**tccgcgctatctgaaatgaaccagaaaaaattagagtct
P9R	CCAAGACTCTAATTTTTTCTGGTTCATTTCAGATAGCGCGGA
S9F	Ggcgacgcgctggcagggccccagcgcgcgcagcttgtcgac
S9R	CCAGTCGACAAGCTGCGCGCGCTGGGGCCCTGCCAGCGCGTC
A9F	Ggcgatgtcctcggcgcacccgaacgcgaccagct**t**aaggca
A9R	CCATGCCTT**a**AGCTGGTCGCGTTCGGGTGCGCCGAGGACATC
T9F	GGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGAC
T9R	CCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTC
	Primers to construct the plasmids containing the stop sequence
P1R	**GCATGC**TCAGCTACTTA**GCTCTTC**AGCGCTGATTGCCATTGTAATCCCAAT
S1R	**GCATGC**TCAGCTACTTA**GCTCTTC**AGCGCTCGTCTGCGCGGTACAG
A1R	**GCATGC**TCAGCTACTTA**GCTCTTC**AGCGTTCGTCGGCGCGGTGGGCGACG
T10R	**GCATGC**TCAGCTACTTA**GCTCTTC**AGCGTTCTTCGGGGCGAAAACTCTC
P10F	**Gaagagc**taagtagctga**gcatgcGCTCTTC**ttggatggtgaacaatcaagtcac
S10F	**Agc**taagtagctga**gcatgcGCTCTTC**c**Tgg**ctgaaaggcaacaccaccg
A10F	**c**taagtagctga**gcatgcGCTCTTC**a**tgg**ctcgtcgccaacaccaccgg
T10F	**gagc**taagtagctga**gcatgcGCTCTTC**CTGGATGGAGGCGGATAAAG

Table AII-2. Other primers involved with synthesis of RandE:APST and RandE:PST libraries. Including the primers necessary to construct the cassette between blocks 1 and 10 (first set of primers). The primers to PCR amplify from the 5' and 3' end to put blocks 1 and 10 into the plasmid. The primers to PCR amplify just blocks 2-9 with Sap1 sites on the ends so that overhangs can be regenerated.

	Primers to construct the plasmids containing the stop sequence 3' end of block 1 and 5' end of block 10
P1R	**GCATGC**TCAGCTACTTA**GCTCTTC**AGCGCTGATTGCCATTGTAATCCCAAT
S1R	**GCATGC**TCAGCTACTTA**GCTCTTC**AGCGCTCGTCTGCGCGGTACAG
A1R	**GCATGC**TCAGCTACTTA**GCTCTTC**AGCGTTCGTCGGCGCGGTGGGCGACG
T10R	**GCATGC**TCAGCTACTTA**GCTCTTC**AGCGTTCTTCGGGGCGAAAACTCTC
P10F	**Gaagagc**taagtagctga**gcatgcGCTCTTC**ttggatggtgaacaatcaagtcac
S10F	**Agc**taagtagctga**gcatgcGCTCTTC**c**Tgg**ctgaaaggcaacaccaccg
A10F	**c**taagtagctga**gcatgcGCTCTTC**a**tgg**ctcgtcgccaacaccaccgg
T10F	**gagc**taagtagctga**gcatgcGCTCTTC**CTGGATGGAGGCGGATAAAG

	Primers for KpnF at 5' end of block 1
PkpnF	CAAGCTTGGTACCCatgcttttatataaaatgtgtgacaa
SkpnF	CAAGCTTGGTACCCatgcttaaggaacggtttcgccag
AkpnF	CAAGCTTGGTACCCgtgactttctccgctctccccttcc
TkpnF	CAAGCTTGGTACCCatgagtattcaacatttccgtgtc
	Primers for Pst1 Rev at 3' end of block 10
PSE4PstRMk	CAACCTGCAGCCATGGGtcagcgcgactgtgatgtataa
SED1PstRMk	CAACCTGCAGGAATTCGTTACTTTCCTTCCGTCACAATTTTCGC
AST1PstRMk	CAACCTGCAGactagtGCTATCCGAGCGCGTCGACCACC
TEMPstRmk	CAACCTGCAGCAGCTGGttaccaatgcttaatcagtgagg
	Primers to PCR amplify insert and add Sap1 site to 5'
PSE4PCRSAPF	CggcGacta**gctcttc**g**cgc**ttcccgttaacaagtactt
SED1PCRSAPF	cggcGacta**gctcttc**g**cgC**tttgcgatgTgcagcaccagt
TEM1PCRSAPF	cggcGacta**gctcttc**A**CGc**TTTCCAATGATG**AGC**ACTTTT
	Primers to PCR amplify insert and add Sap1 site to 3'
PSE4PCRSAPR	CggcGacta**gctcttc**TCCAAGACTCTAATTTTTTCTGGTTC
TEM1PCRSAPR	cggcGacta**gctcttc**TCCAGTCTATTAATTGTTGCCGGGAA
SED1PCRSAPR	**cggcGactagctcttcGCCAGTCGACAAGCTGCGCGCGC**

Table AII-3. Primers for construction of RASPP:PST using SISDC (Hiraga and Arnold 2003). First, the tag sequences for each recombination site are shown. The overhangs are in italics, the BsaX1 site is in bold and the Nde1 site is underlined. The primers used for all PCR reactions are shown below. They are named for the parent P64F, the recombination site amino acid P64F, and whether they are for the coding sequence (F), or noncoding sequence (R). For the primers, capital letters are part of the tags, lower-case letters match the gene.

Tag 64:	--- TCT GGC AGA **AC** GGACT **CTCC** ATATGGC *CGC*
	GCG AGA CCG TCT **TG** CCTGA **GAGG** TATACCG ---
Tag 73:	*aaa* ACC CTT GAG **AC** GTTGC **CTCC** ATATGCT *AAA*
	TTT TGG GAA CTC **TG** CAACG **GAGG** TATACGA ---
Tag 148:	Acc GGC AAC CGT **AC** CGGTA **CTCC** ATATGAT *ACC*
	TGG CCG TTG GCA **TG** GCCAT **GAGG** TATACTA ---
Tag 176:	--- TCG TTA GCC **AC** AAGGC **CTCC** ATATGCG *GAT*
	CTA AGC AAT CGG **TG** TTCCG **GAGG** TATACGC ---
Tag 190:	--- CAA TGC GTG **AC** ATTCG **CTCC** ATATGTC *TTG*
	AAC GTT ACG CAC **TG** TAAGC **GAGG** TATACAG ---
Tag 218:	--- CGC CTT GAC **AC** TGCCA **CTCC** ATATGTA *GGC*
	CCG GCG GAA CTG **TG** ACGGT **GAGG** TATACAT ---

P1F	ccgCTCGAGGGTACCCatgcttttatataaaatgtgtgaca
T1F	ccgCTCGAGGGTACCCatgagtattcaacatttccgtgt
S1F	ccgCTCGAGGGTACCCatgcttaaggaacggtttcgcc
P64F	GGCAGAACGGACTCTCCATATGGCCGCttcccgttaacaagta
P64R	GGAGAGTCCGTTCTGCCAGAGCGctgattgccattgtaatccc
T64F	GGCAGAACGGACTCTCCATATGGCCGCtttccaatgatgagca
T64R	GGAGAGTCCGTTCTGCCAGAGCGttcttcggggcgaaaac
S64F	GGCAGAACGGACTCTCCATATGGCCGCtttgcgatgtgcagca
S64R	GGAGAGTCCGTTCTGCCAGAGCGctcgtctgcgcggtacagc
P73F	CTTGAGACGTTGCCTCCATATGCTAAAacaatagcttgcgctaaat
P73R	GGAGGCAACGTCTCAAGGGTTTTaaaagtacttgttaacgg
T73F	CTTGAGACGTTGCCTCCATATGCTAAAgttctgctatgtggcgcgg
T73R	GGAGGCAACGTCTCAAGGGTTTTaaaagtgctcatcattgg
S73F	CTTGAGACGTTGCCTCCATATGCTAAAgtcatgaccgccgccgcgg
S73R	GGAGGCAACGTCTCAAGGGTTTTactggtgctgcacatcgc
P149F	AACCGTACCGGTACTCCATATGATACCgatttttaagacaaattgggga
P149R	GGAGTACCGGTACGGTTGCCGGTaacgcctttggggccacct
T149F	AACCGTACCGGTACTCCATATGATACCgctttttttgcacaacatgggga
T149R	GGAGTACCGGTACGGTTGCCGGTtagctccttcggtcctccga
S149F	AACCGTACCGGTACTCCATATGATACCgcgtttgcacgttccattggcg
S149R	GGAGTACCGGTACGGTTGCCGGTgacgttgccggggccgc

P161F	CCGCCTCGAGGCTCTTCCtcgtctagaccgtattgagcctga
P161R	AAAACTGCAGGCTCTTCAcgagtctctttgtccccaatttg
T161F	CCGCCTCGAGGCTCTTCCtcgccttgatcgttgggaaccgg
T161R	AAAACTGCAGGCTCTTCGcgagttacatgatcccccatgttg
S161F	CCGCCTCGAGGCTCTTCTtcgtctcgatcgcaaagagccgg
S161R	AAAACTGCAGGCTCTTCAcgaaacgtcgtgtcgccaatggaacg
P176F	TTAGCCACAAGGCCTCCATATGCGGATttgagggatacgacaacccc
P176R	GGAGGCCTTGTGGCTAACGAATCaccgagcttaccttcatttaa
T176F	TTAGCCACAAGGCCTCCATATGCGGATgagcgtgacaccacgatgcc
T176R	GGAGGCCTTGTGGCTAACGAATCgtttggtatggcttcattcag
S176F	TTAGCCACAAGGCCTCCATATGCGGATgagcgcgacacaacatcgcc
S176R	GGAGGCCTTGTGGCTAACGAATCgccgggaatggcggtgt
P190F	TGCGTGACATTCGCTCCATATGTCTTGaataaatttttatttggttccgc
P190R	GGAGCGAATGTCACGCATTGCAAagtactggctattgccttagg
T190F	TGCGTGACATTCGCTCCATATGTCTTGcgcaaactattaactggcgaacta
T190R	GGAGCGAATGTCACGCATTGCAAcgttgttgccattgctacag
S190F	TGCGTGACATTCGCTCCATATGTCTTGcgtaaactcacgctgggcgacgc
S190R	GGAGCGAATGTCACGCATTGCAAacttttggccatggccagcgg
P218F	CTTGACACTGCCACTCCATATGTAGGCaatttactacgttcagtattgcc
P218R	GGAGTGGCAGTGTCAAGGCGGCCagtgacttgattgttcaccatc
T218F	CTTGACACTGCCACTCCATATGTAGGCccacttctgcgctcggccc
T218R	GGAGTGGCAGTGTCAAGGCGGCCtgcaactttatccgcctccat
S218F	CTTGACACTGCCACTCCATATGTAGGCcagagcattcgtgccggcct
S218R	GGAGTGGCAGTGTCAAGGCGGCCtccggtggtgttgcctttc
PendR	AAAACTGCAGAAGCTTtcagcgcgactgtgatgtat
TendR	AAAACTGCAGAAGCTTttaccaatgcttaatcagtgagg
SendR	AAAACTGCAGAAGCTTttactttccttccgtcacaattttc

Table AII-4. PCR amplification of each half library of the small library during construction was done with the following primer sets. Primers sequences can be found on Table AII-3.

Front Half-Library	
P1F	P161R
P1F	S161R
P1F	T161R
S1F	P161R
S1F	S161R
S1F	T161R
T1F	P161R
T1F	S161R
T1F	T161R

Back Half-Library	
P161F	PendR
P161F	SendR
P161F	TendR
S161F	PendR
S161F	SendR
S161F	TendR
T161F	PendR
T161F	SendR
T161F	TendR

Table AII-5. Probes for DNA Hybridization to sequence chimeras in the smaller lactamase library. Stringency wash conditions: all washes contain 0.5% SDS and the indicated concentration of SSC.

	sequence	Stringency wash
PSEprobe1	GTTGAACAAGACGTTAAGGCAATTGAAG	2x
TEMprobe1	CGCTGGTGAAAGTAAAAGATGCTGAAG	1x
SEDprob1	GTTCAGAAAAAGCTGGCGGCG	0.5x
PSEprobe2	CGCTTCCCGTTAACAAGTACTTTT	2x
TEMprobe2	CGcTTTCCAATGATGAGCACTTTT	2x
SEDprobe2	CGCTTTGCGATGTGCAGCACCAGT	1x
PSEprobe3	GAAAAGTTAATCCCAATAGTACAGTCGAGATTAAG	2x
TEMprobe3	GCAACTCGGTCGCCGC	1x
SEDprobe3	GGTATTTTGCAGCAAAAAATGACCATTAAAAAAG	2x
PSEprobe4	GATTTTTTAAGACAAATTGGGGACAAAGAGAC	2x
TEMprobe4	CTTTTTTGCACAACATGGGGGATC	2x
SEDprobe4	GCACGTTCCATTGGCGACAC	2x
PSEprobe5	GCCTGATTTAAATGAAGGTAAGCTCGG	2x
TEMprobe5	CGGAGCTGAATGAAGCCATACC	1x
SEDprobe5	GGAATTAAACACCGCCATTCCCG	1x
PSEprobe6	CAACCCCTAAGGCAATAGCCAGTAC	2x
TEMprobe6	GCCTGTAGCAATGGCAACAacg	2x
SEDprobe6	CGCTGGCCATGGCCAAAAG	2x
PSEprobe7	GTTCCGCGCTATCTGAAATGAACC	2x
TEMprobe7	GAACTACTTACTCTAGCTTCCCGGC	2x
SEDprobe7	CTGAAAGGCAACACCACCGGA	0.5x
PSEprobe8	GGAGAGCATCAAGCCCCAATTATTG	2x
TEMprobe8	GGGGCCAGATGGTAAGCCC	2x
SEDprobe8	GATGCGAAATGGCGTAAAGATGTCC	0.5x

Appendix III: Characterized Chimeras

Table AIII-1. Functional chimeras (third column in MIC in μg/mL of ampicillin for each chimera) from naïve RASPP:PST library, 111 total sequences. The sequences are designated by their block pattern: 1 represents PSE-4, 2 SED-1, and 3 TEM-1. A chimera's freezer stock location is in the first column and the sequence is determined by the sequences of the parental genes in the following blocks: (Ambler standard numbering) 1-65, 66-73, 74-149, 150-161, 162-176, 177-190, 191-218, 219-290.

11E10	1	1	2	1	3	3	3	3	10		12F6	2	3	3	2	2	3	3	2	2000
10E4	1	1	3	1	1	1	3	3	10		4E4	2	3	3	2	3	2	1	2	25
18C9	1	1	3	1	2	3	3	1	10		9B9	2	3	3	2	3	3	2	2	2000
4A4	1	1	3	1	3	1	3	3	1000		2B3	2	3	3	3	1	2	3	2	100
17F11	1	1	3	1	3	3	3	1	25		9E10	2	3	3	3	2	2	2	2	2000
10F8	1	1	3	1	3	3	3	3	50		10C8	2	3	3	3	2	3	3	2	500
20D2	1	1	3	3	1	3	3	1	25		11C8	2	3	3	3	3	1	1	2	1000
4D7	1	1	3	3	2	1	3	3	1000		2H6	2	3	3	3	3	1	3	2	100
20F8	1	1	3	3	2	3	3	1	100		12B3	2	3	3	3	3	2	1	2	250
11F6	1	1	3	3	2	3	3	3	1000		2C9	2	3	3	3	3	2	2	2	1000
10D12	1	1	3	3	3	1	3	3	500		11E6	2	3	3	3	3	3	2	2	2000
20B12	1	1	3	3	3	2	3	1	100		10F1	2	3	3	3	3	3	3	2	500
20D7	1	1	3	3	3	3	3	3	100		4H9	3	1	1	1	2	3	3	3	1000
20C4	1	2	3	1	2	1	3	1	10		2B9	3	1	1	3	3	3	3	3	1000
1F2	1	2	3	2	3	3	2	2	10		12E3	3	1	2	2	2	1	3	2	100
18G12	1	2	3	3	2	1	3	1	10		12H8	3	1	3	1	1	1	3	3	2000
20C12	1	2	3	3	3	1	2	1	10		1E2	3	1	3	1	1	3	1	3	50
4E1	1	3	2	2	3	2	3	2	100		10D8	3	1	3	1	1	3	3	3	500
18E6	1	3	2	3	3	3	3	3	10		11A11	3	1	3	1	2	1	3	3	2000
20H7	1	3	3	1	1	3	3	3	25		1G4	3	1	3	1	2	2	3	3	500
17C4	1	3	3	1	2	3	3	3	50		10B12	3	1	3	1	2	3	3	3	2000
10H11	1	3	3	1	3	3	3	3	100		9C3	3	1	3	1	3	1	1	3	500
10D9	1	3	3	2	2	2	3	3	500		10F12	3	1	3	1	3	2	3	3	2000
20B11	1	3	3	3	2	2	3	3	2000		11B11	3	1	3	1	3	3	3	3	2000
10C3	1	3	3	3	2	3	3	3	250		2E8	3	1	3	2	2	3	3	3	1000
2F1	1	3	3	3	3	2	3	3	1000		3B5	3	1	3	2	3	2	2	3	1000
18A7	1	3	3	3	3	3	3	3	500		20G5	3	1	3	2	3	2	3	3	2000
4H4	2	1	3	1	3	1	3	2	1000		10G12	3	1	3	2	3	3	3	3	1000
2C8	2	1	3	1	3	2	2	2	1000		20H9	3	1	3	3	1	3	1	3	25
3B7	2	1	3	1	3	2	3	2	1000		18H3	3	1	3	3	1	3	3	3	1000
12F11	2	1	3	1	3	3	2	2	2000		18H4	3	1	3	3	2	2	3	3	25
3C11	2	1	3	2	1	1	3	2	1000		11G11	3	1	3	3	2	3	1	3	1000
9C4	2	1	3	2	2	1	3	2	2000		3H7	3	1	3	3	2	3	3	3	1000
1B9	2	1	3	2	2	3	2	2	1000		1D12	3	1	3	3	3	2	3	3	1000
4G3	2	1	3	2	3	2	2	2	1000		11D3	3	1	3	3	3	3	3	3	2000
1B8	2	1	3	2	3	2	3	2	1000		9F1	3	2	3	1	2	3	3	3	50

10B5	2	1	3	2	3	3	2	2	2000
9D4	2	1	3	2	3	3	3	2	2000
2F4	2	1	3	3	2	1	3	2	1000
17C10	2	1	3	3	2	2	3	2	1000
17B11	2	1	3	3	2	3	3	2	1000
11A7	2	1	3	3	3	1	3	2	1000
11B5	2	1	3	3	3	2	2	2	2000
4A3	2	1	3	3	3	2	3	2	1000
12E9	2	1	3	3	3	3	2	2	2000
11E11	2	1	3	3	3	3	3	2	1000
1E11	2	2	2	2	3	3	2	2	100
2A3	2	2	3	1	3	2	2	2	100
1G8	2	2	3	1	3	3	2	2	100
19B10	2	2	3	3	2	3	2	1	500
11E12	2	3	1	2	3	3	3	2	1000
12G4	2	3	1	3	2	3	2	2	250
1F10	2	3	1	3	3	2	3	2	1000
20D3	2	3	3	1	2	3	3	2	1000
9H11	2	3	3	1	3	3	3	2	100
20G6	2	3	3	2	1	1	2	2	2000

10C2	3	2	3	1	3	1	3	3	25
3F3	3	2	3	1	3	3	2	2	1000
12F5	3	2	3	1	3	3	3	3	250
19F8	3	2	3	3	2	2	3	3	500
9E9	3	3	3	1	1	1	3	3	2000
11C3	3	3	3	1	1	3	1	3	1000
19B4	3	3	3	1	1	3	3	3	2000
2D11	3	3	3	1	2	1	3	3	1000
17A7	3	3	3	1	2	3	3	3	2000
17B4	3	3	3	1	3	1	3	3	2000
2H10	3	3	3	1	3	3	3	3	1000
3E4	3	3	3	2	1	1	1	3	1000
4E2	3	3	3	2	1	1	2	3	50
3A8	3	3	3	2	1	3	3	3	1000
10B4	3	3	3	2	2	3	3	3	2000
18C3	3	3	3	3	1	3	3	3	2000
20F12	3	3	3	3	2	2	3	3	2000
12B7	3	3	3	3	2	3	3	3	250
20E2	3	3	3	3	3	3	3	3	2000

Table AIII-2. Nonfunctional chimeras from the naïve library described for Table AIII-1.

2C10	1	1	1	1	1	2	2	2	0
12F3	1	1	1	1	2	2	1	2	0
10A10	1	1	1	1	2	3	3	2	0
9D12	1	1	1	1	3	2	3	2	0
4A5	1	1	1	1	3	3	1	2	0
10F6	1	1	1	2	1	1	3	2	0
11A6	1	1	1	2	1	2	2	3	0
4C10	1	1	1	2	1	3	1	2	0
1H1	1	1	1	2	1	3	2	2	0
20A7	1	1	1	2	1	3	3	3	0
11B4	1	1	1	2	2	2	1	2	0
20B7	1	1	1	2	2	3	3	2	0
1E8	1	1	1	2	3	2	1	2	0
3B10	1	1	1	2	3	3	1	2	0
3D10	1	1	1	3	1	1	3	3	0
12C11	1	1	1	3	1	3	3	2	0
2B10	1	1	1	3	1	3	3	3	0
1D3	1	1	1	3	2	1	3	3	0
19H2	1	1	1	3	2	3	3	3	0
1F5	1	1	1	3	3	2	2	2	0
3E8	1	1	1	3	3	3	2	2	0
11G3	1	1	1	3	3	3	3	2	0
9B8	1	1	2	1	1	1	3	2	0
2G2	1	1	2	1	1	3	3	3	0
9F3	1	1	2	1	2	3	1	2	0
9G8	1	1	2	1	3	3	3	2	0
12D10	1	1	2	2	1	3	1	3	0
20B10	1	1	2	2	1	3	2	2	0
19E7	1	1	2	2	1	3	3	3	0
20E11	1	1	2	2	2	2	3	3	0
12F2	1	1	2	2	2	3	3	2	0
3F2	1	1	2	2	3	2	2	2	0
10H5	1	1	2	2	3	3	1	2	0
10B6	1	1	2	2	3	3	3	2	0
2A5	1	1	2	2	3	3	3	3	0
17H11	1	1	2	3	1	1	3	2	0
18E5	1	1	2	3	2	3	3	2	0
11C5	1	1	2	3	3	1	3	2	0
11H10	1	1	2	3	3	3	1	2	0
1H10	1	1	2	3	3	3	3	2	0
18G8	1	1	2	3	3	3	3	3	0
12F10	1	1	3	1	1	2	3	2	0
10A4	1	1	3	1	1	3	2	2	0
4D2	1	1	3	1	1	3	3	2	0
18G4	1	1	3	1	1	3	3	3	0
2D4	1	1	3	1	2	3	1	2	0
2E4	1	1	3	1	2	3	3	2	0
19H3	1	1	3	1	2	3	3	3	0
4A2	1	1	3	1	3	1	1	2	0
12A4	1	1	3	1	3	2	2	2	0

2E6	1	1	3	1	3	2	3	2	0
11C9	1	1	3	1	3	3	1	2	0
4G6	1	1	3	1	3	3	2	2	0
11C2	1	1	3	1	3	3	3	2	0
3H8	1	1	3	2	1	1	3	2	0
12C6	1	1	3	2	1	3	3	2	0
20F7	1	1	3	2	1	3	3	3	0
11F10	1	1	3	2	2	3	1	2	0
17E7	1	1	3	2	2	3	3	1	0
18F1	1	1	3	2	2	3	3	2	0
2E1	1	1	3	2	3	1	3	2	0
9G3	1	1	3	2	3	1	3	3	0
10F10	1	1	3	2	3	2	1	3	0
3E6	1	1	3	2	3	2	2	2	0
10H2	1	1	3	2	3	2	3	2	0
11D7	1	1	3	2	3	3	1	2	0
11H9	1	1	3	2	3	3	2	2	0
10F4	1	1	3	2	3	3	3	2	0
19H5	1	1	3	2	3	3	3	3	0
9F10	1	1	3	3	1	1	1	2	0
12F1	1	1	3	3	1	1	1	3	0
19H4	1	1	3	3	1	1	3	3	0
1G11	1	1	3	3	1	2	1	2	0
4C12	1	1	3	3	1	2	3	2	0
10D5	1	1	3	3	1	3	1	2	0
1F7	1	1	3	3	1	3	3	2	0
17C7	1	1	3	3	1	3	3	3	0
17D4	1	1	3	3	2	1	3	1	0
19E2	1	1	3	3	2	1	3	2	0
1C7	1	1	3	3	2	3	2	2	0
10A12	1	1	3	3	2	3	3	2	0
2G4	1	1	3	3	3	1	1	2	0
10A7	1	1	3	3	3	1	3	2	0
4H8	1	1	3	3	3	2	1	2	0
1F12	1	1	3	3	3	2	2	2	0
10G6	1	1	3	3	3	2	3	2	0
1E5	1	1	3	3	3	3	1	2	0
1D5	1	1	3	3	3	3	2	2	0
10A6	1	1	3	3	3	3	3	2	0
12C10	1	2	1	1	3	3	3	2	0
19C11	1	2	1	2	1	3	3	3	0
12B8	1	2	1	2	2	3	3	2	0
20F3	1	2	1	2	2	3	3	3	0
4A10	1	2	1	2	3	1	3	2	0
10F2	1	2	1	2	3	2	3	2	0
11G4	1	2	1	2	3	3	2	2	0
12E11	1	2	1	2	3	3	3	2	0
1B2	1	2	1	2	3	3	3	3	0
2F9	1	2	1	3	1	3	2	3	0
19C12	1	2	1	3	1	3	3	3	0

10H8	1	2	1	3	2	3	2	2	0
11A3	1	2	1	3	3	1	3	2	0
11F9	1	2	1	3	3	2	3	3	0
4G9	1	2	1	3	3	3	3	2	0
2C5	1	2	2	1	1	3	2	2	0
3E3	1	2	2	1	1	3	3	2	0
2D9	1	2	2	1	1	3	3	3	0
1H6	1	2	2	1	2	1	3	3	0
12E1	1	2	2	1	2	2	3	3	0
17F9	1	2	2	1	2	3	3	1	0
12C7	1	2	2	1	2	3	3	3	0
11H5	1	2	2	1	3	1	3	2	0
9H8	1	2	2	2	1	2	3	2	0
1D8	1	2	2	2	2	3	2	2	0
11B2	1	2	2	2	3	3	1	2	0
10H10	1	2	2	2	3	3	3	2	0
9E8	1	2	2	3	1	2	1	2	0
18D11	1	2	2	3	1	3	3	3	0
18D9	1	2	2	3	2	3	3	3	0
12D5	1	2	2	3	3	1	1	2	0
11C12	1	2	2	3	3	3	1	2	0
3D7	1	2	2	3	3	3	2	2	0
1H8	1	2	2	3	3	3	3	2	0
12H6	1	2	3	1	1	1	2	2	0
20E5	1	2	3	1	1	1	3	1	0
19B12	1	2	3	1	1	1	3	3	0
18E9	1	2	3	1	1	2	3	3	0
12B10	1	2	3	1	1	3	3	2	0
20B8	1	2	3	1	1	3	3	3	0
20F4	1	2	3	1	2	1	3	2	0
10E9	1	2	3	1	2	2	3	3	0
18F3	1	2	3	1	2	3	3	1	0
12A6	1	2	3	1	3	1	3	2	0
10G8	1	2	3	1	3	2	1	2	0
12A2	1	2	3	1	3	2	2	2	0
10C12	1	2	3	1	3	2	3	2	0
3C2	1	2	3	1	3	2	3	3	0
11F5	1	2	3	1	3	3	1	2	0
10A9	1	2	3	1	3	3	2	2	0
9C5	1	2	3	1	3	3	2	3	0
1D10	1	2	3	1	3	3	3	2	0
4G7	1	2	3	2	1	3	3	1	0
11H3	1	2	3	2	1	3	3	3	0
17H4	1	2	3	2	2	3	3	3	0
1F8	1	2	3	2	3	1	3	2	0
2F2	1	2	3	2	3	2	1	2	0
10B8	1	2	3	3	1	1	2	3	0
11B9	1	2	3	3	1	3	1	3	0
17D2	1	2	3	3	1	3	3	3	0
1E1	1	2	3	3	2	2	2	2	0

1H11	1	2	3	3	2	2	3	2	0
19A6	1	2	3	3	2	3	3	1	0
11E5	1	2	3	3	2	3	3	2	0
17F8	1	2	3	3	2	3	3	3	0
10C5	1	2	3	3	3	1	3	2	0
12E4	1	2	3	3	3	3	1	2	0
9C9	1	2	3	3	3	3	3	2	0
1B6	1	3	1	1	1	2	3	3	0
11D5	1	3	1	2	2	3	1	3	0
12G2	1	3	1	2	3	2	3	2	0
2G12	1	3	1	2	3	3	3	2	0
10E11	1	3	1	3	1	2	1	2	0
4F4	1	3	1	3	1	3	3	3	0
10C10	1	3	1	3	2	2	1	2	0
17H3	1	3	1	3	2	3	3	3	0
3A11	1	3	1	3	3	1	3	2	0
20G10	1	3	1	3	3	3	3	2	0
18G3	1	3	1	3	3	3	3	3	0
12H5	1	3	2	1	2	2	3	3	0
3D6	1	3	2	1	3	3	1	2	0
10G10	1	3	2	1	3	3	3	2	0
9E7	1	3	2	2	1	1	1	3	0
18A5	1	3	2	2	1	1	3	2	0
12E5	1	3	2	2	2	2	1	3	0
11G8	1	3	2	2	2	2	3	3	0
2E12	1	3	2	2	3	1	3	2	0
3B8	1	3	2	2	3	3	3	2	0
20B5	1	3	2	3	1	2	2	3	0
12D4	1	3	2	3	1	2	3	3	0
18G9	1	3	2	3	1	3	3	2	0
11D11	1	3	2	3	1	3	3	3	0
9H9	1	3	2	3	2	3	3	2	0
11D8	1	3	2	3	2	3	3	3	0
2D7	1	3	2	3	3	2	3	3	0
1C10	1	3	2	3	3	3	3	2	0
2E2	1	3	3	1	1	1	1	2	0
11C7	1	3	3	1	1	1	2	3	0
9H4	1	3	3	1	1	1	3	2	0
11H11	1	3	3	1	1	2	3	2	0
17F5	1	3	3	1	1	3	2	3	0
19C7	1	3	3	1	1	3	3	1	0
4G1	1	3	3	1	1	3	3	2	0
20C2	1	3	3	1	2	1	3	1	0
12C3	1	3	3	1	2	1	3	2	0
1G2	1	3	3	1	2	2	2	3	0
19F11	1	3	3	1	2	3	1	2	0
9B3	1	3	3	1	3	1	1	2	0
9D10	1	3	3	1	3	1	2	2	0
2B2	1	3	3	1	3	2	1	2	0
9A3	1	3	3	1	3	3	1	2	0

11F4	1	3	3	1	3	3	2	2	0
2A2	1	3	3	1	3	3	2	3	0
1H7	1	3	3	1	3	3	3	2	0
18E12	1	3	3	2	1	3	3	1	0
4G8	1	3	3	2	2	3	3	3	0
2A7	1	3	3	2	3	3	1	2	0
20E7	1	3	3	3	1	2	3	1	0
4F8	1	3	3	3	1	2	3	2	0
10G4	1	3	3	3	1	3	1	2	0
17E2	1	3	3	3	1	3	3	1	0
12G12	1	3	3	3	1	3	3	2	0
2H4	1	3	3	3	1	3	3	3	0
4C4	1	3	3	3	2	1	2	2	0
18E1	1	3	3	3	2	1	3	3	0
1B3	1	3	3	3	2	2	1	2	0
2D10	1	3	3	3	2	2	2	2	0
10F7	1	3	3	3	2	2	3	2	0
3H9	1	3	3	3	2	3	1	2	0
1D2	1	3	3	3	2	3	2	3	0
2B8	1	3	3	3	3	1	1	2	0
9G10	1	3	3	3	3	1	3	2	0
1B5	1	3	3	3	3	2	1	2	0
2A9	1	3	3	3	3	2	3	2	0
1G7	1	3	3	3	3	3	2	2	0
10F5	1	3	3	3	3	3	2	3	0
1D9	1	3	3	3	3	3	3	2	0
17A2	2	1	1	3	1	3	3	1	0
11G12	2	1	1	3	3	2	3	2	0
2F7	2	1	1	3	3	3	3	2	0
1F1	2	1	2	1	1	3	3	2	0
4A8	2	1	2	1	2	2	3	3	0
2E10	2	1	2	1	3	3	3	2	0
17G8	2	1	2	3	1	3	3	1	0
19B6	2	1	2	3	1	3	3	2	0
19G12	2	1	2	3	2	3	3	3	0
1C6	2	1	2	3	3	1	1	2	0
11F7	2	1	2	3	3	3	1	2	0
10A2	2	1	2	3	3	3	3	2	0
17C9	2	1	3	1	1	3	1	1	0
10D3	2	1	3	1	1	3	3	3	0
4E3	2	1	3	1	2	2	1	3	0
18B2	2	1	3	1	2	3	3	1	0
9G7	2	1	3	1	3	3	1	2	0
9B2	2	1	3	1	3	3	3	3	0
12D3	2	1	3	3	1	1	3	2	0
11G1	2	1	3	3	1	2	3	2	0
12H3	2	1	3	3	1	3	1	2	0
17D8	2	1	3	3	1	3	2	1	0
19A10	2	1	3	3	1	3	3	1	0
19G11	2	1	3	3	2	1	3	3	0

20D8	2	1	3	3	2	3	1	3	0
17G10	2	1	3	3	2	3	3	3	0
18F12	2	2	1	1	3	3	3	2	0
20C7	2	2	1	2	2	3	3	2	0
20B6	2	2	1	3	1	3	3	3	0
10D7	2	2	1	3	3	1	3	2	0
19D11	2	2	2	2	2	3	3	3	0
20B3	2	2	2	3	2	3	3	3	0
4C2	2	2	2	3	3	1	1	2	0
1E3	2	2	3	1	1	3	2	2	0
10F3	2	2	3	1	1	3	3	2	0
19B3	2	2	3	1	1	3	3	3	0
20E1	2	2	3	1	2	3	3	3	0
3E9	2	2	3	1	3	3	3	2	0
4D3	2	2	3	2	3	1	2	2	0
10B10	2	2	3	3	1	3	3	3	0
12F9	2	2	3	3	2	3	1	2	0
3G8	2	2	3	3	2	3	2	3	0
20E9	2	2	3	3	2	3	3	1	0
20C3	2	2	3	3	2	3	3	2	0
10B9	2	2	3	3	3	1	3	2	0
4B2	2	2	3	3	3	2	1	3	0
3D9	2	2	3	3	3	2	3	2	0
17D7	2	2	3	3	3	3	2	3	0
1D4	2	2	3	3	3	3	3	2	0
3F10	2	2	3	3	3	3	3	3	0
2B4	2	3	1	3	2	3	3	2	0
9D5	2	3	1	3	3	3	2	1	0
11E1	2	3	1	3	3	3	2	2	0
4H6	2	3	1	3	3	3	2	3	0
4A7	2	3	2	1	2	3	1	3	0
10B3	2	3	2	1	3	1	1	2	0
2G8	2	3	2	1	3	2	2	2	0
3C12	2	3	2	2	1	3	3	2	0
17H10	2	3	2	2	1	3	3	3	0
3B4	2	3	2	2	2	3	3	2	0
17D5	2	3	2	2	2	3	3	3	0
9G5	2	3	2	2	3	2	1	2	0
9A10	2	3	2	2	3	3	3	3	0
2B12	2	3	3	1	1	3	3	2	0
20B9	2	3	3	3	2	1	3	1	0
1E6	2	3	3	3	3	3	1	2	0
3A12	3	1	1	1	1	2	1	2	0
4A12	3	1	1	1	1	3	3	3	0
4F1	3	1	1	1	3	2	1	2	0
11D12	3	1	1	1	3	2	3	2	0
3C7	3	1	1	1	3	3	1	2	0
2E9	3	1	1	1	3	3	2	2	0
17F3	3	1	1	2	1	2	1	3	0
17A12	3	1	1	2	1	3	3	2	0

2F3	3	1	1	2	2	1	1	3	0
3H1	3	1	1	2	3	1	3	3	0
11A10	3	1	1	2	3	3	3	2	0
18A4	3	1	1	2	3	3	3	3	0
11B12	3	1	1	3	3	1	3	2	0
9H7	3	1	1	3	3	3	3	2	0
9D3	3	1	2	1	1	2	1	3	0
11G5	3	1	2	1	2	2	3	2	0
18F11	3	1	2	1	2	3	3	2	0
3G4	3	1	2	1	3	3	2	2	0
3H3	3	1	2	1	3	3	3	2	0
10G3	3	1	2	1	3	3	3	3	0
4E7	3	1	2	2	2	3	2	2	0
12C4	3	1	2	2	3	1	3	2	0
9E1	3	1	2	2	3	1	3	3	0
10D6	3	1	2	2	3	3	2	2	0
12B12	3	1	2	2	3	3	3	2	0
17G3	3	1	2	3	1	3	1	2	0
18B7	3	1	2	3	2	3	3	3	0
10H6	3	1	2	3	3	1	2	2	0
12B11	3	1	2	3	3	1	3	2	0
9F6	3	1	2	3	3	3	1	2	0
11E2	3	1	2	3	3	3	2	2	0
2C7	3	1	2	3	3	3	3	3	0
11B6	3	1	3	1	1	1	2	2	0
9F4	3	1	3	1	1	2	2	2	0
18C2	3	1	3	1	1	3	1	2	0
10C7	3	1	3	1	1	3	2	3	0
12A11	3	1	3	1	1	3	3	2	0
11H6	3	1	3	1	2	1	3	2	0
18C10	3	1	3	1	2	3	3	2	0
11H8	3	1	3	1	3	1	2	2	0
11B8	3	1	3	1	3	1	3	2	0
2B11	3	1	3	1	3	2	3	2	0
12F8	3	1	3	1	3	3	1	2	0
12D8	3	1	3	1	3	3	3	2	0
18D8	3	1	3	2	1	3	3	2	0
12E10	3	1	3	2	2	2	1	2	0
2C3	3	1	3	2	2	3	1	2	0
1A11	3	1	3	2	3	1	1	2	0
11G9	3	1	3	2	3	1	3	2	0
12A8	3	1	3	2	3	2	3	2	0
2G7	3	1	3	2	3	3	3	2	0
3F4	3	1	3	3	1	1	2	2	0
20G1	3	1	3	3	1	1	3	2	0
4G5	3	1	3	3	1	3	3	2	0
4A6	3	1	3	3	2	1	1	2	0
12B5	3	1	3	3	2	3	2	2	0
11D10	3	1	3	3	3	3	2	2	0
11F1	3	1	3	3	3	3	3	2	0

12H4	3	2	1	1	1	1	1	3	0
18C7	3	2	1	2	2	3	3	2	0
4B9	3	2	1	2	2	3	3	3	0
12F12	3	2	1	2	3	2	1	2	0
3F6	3	2	1	2	3	2	3	2	0
1F11	3	2	1	2	3	3	3	3	0
10G7	3	2	1	3	3	2	3	2	0
4F7	3	2	1	3	3	2	3	3	0
1C3	3	2	1	3	3	3	3	2	0
9F5	3	2	2	1	3	3	3	2	0
12B9	3	2	2	2	1	2	3	2	0
18D2	3	2	2	2	1	3	3	2	0
12B6	3	2	2	2	2	2	3	3	0
4E11	3	2	2	2	2	3	1	2	0
10E12	3	2	2	2	2	3	3	2	0
12A3	3	2	2	2	2	3	3	3	0
12D6	3	2	2	2	3	3	2	2	0
12G10	3	2	2	2	3	3	3	2	0
2C11	3	2	2	2	3	3	3	3	0
10H3	3	2	2	3	1	3	3	2	0
11D4	3	2	3	1	1	1	2	2	0
12C8	3	2	3	1	1	1	2	3	0
9E2	3	2	3	1	1	2	3	2	0
3A10	3	2	3	1	1	2	3	3	0
10H4	3	2	3	1	2	3	1	2	0
11A2	3	2	3	1	2	3	2	2	0
18G11	3	2	3	1	2	3	3	2	0
2C4	3	2	3	1	3	1	2	2	0
12F7	3	2	3	1	3	1	3	2	0
3D2	3	2	3	1	3	2	2	2	0
19D12	3	2	3	1	3	2	3	2	0
12A10	3	2	3	1	3	3	1	2	0
10A11	3	2	3	1	3	3	3	2	0
3B9	3	2	3	2	1	1	3	2	0
1H9	3	2	3	2	1	2	2	2	0
11F2	3	2	3	2	1	2	2	3	0
1H3	3	2	3	2	1	3	3	3	0
9D7	3	2	3	2	2	1	3	2	0
10A3	3	2	3	2	2	3	2	2	0
20G8	3	2	3	2	2	3	3	1	0
12E12	3	2	3	2	2	3	3	2	0
11C11	3	2	3	2	2	3	3	3	0
10E7	3	2	3	2	3	1	2	2	0
1B10	3	2	3	2	3	2	1	2	0
10E10	3	2	3	2	3	2	2	2	0
1E12	3	2	3	2	3	3	2	2	0
11E7	3	2	3	2	3	3	3	2	0
11D2	3	2	3	3	1	3	1	2	0
12E2	3	2	3	3	1	3	3	3	0
4C5	3	2	3	3	3	3	2	2	0

10F11	3	2	3	3	3	3	3	2	0
2B6	3	3	1	1	3	3	2	2	0
19D10	3	3	1	2	1	3	2	3	0
4F5	3	3	1	2	2	3	1	2	0
9H2	3	3	1	2	3	3	2	2	0
4D4	3	3	1	2	3	3	3	2	0
19A12	3	3	1	3	1	3	3	1	0
3D11	3	3	1	3	3	3	3	3	0
17E11	3	3	2	1	1	2	1	3	0
3D3	3	3	2	1	3	3	3	2	0
19C4	3	3	2	2	2	1	3	3	0
11C6	3	3	2	2	2	2	3	2	0
19D4	3	3	2	2	2	3	3	2	0
1F4	3	3	2	2	2	3	3	3	0
10D10	3	3	2	2	3	3	1	2	0
12H2	3	3	2	2	3	3	2	3	0
3A3	3	3	2	2	3	3	3	2	0
19C9	3	3	2	3	1	1	3	3	0
1B4	3	3	3	1	1	1	3	2	0
4C3	3	3	3	1	1	2	3	3	0
9G11	3	3	3	1	1	3	1	2	0
12A5	3	3	3	1	1	3	3	2	0
11D6	3	3	3	1	3	1	2	2	0
4E6	3	3	3	1	3	2	3	2	0
1E7	3	3	3	1	3	3	1	2	0
11G2	3	3	3	1	3	3	2	2	0
3C10	3	3	3	1	3	3	3	2	0
3F8	3	3	3	2	1	1	3	2	0
3F11	3	3	3	2	1	3	1	2	0
12C12	3	3	3	2	2	3	1	3	0
19B9	3	3	3	2	2	3	3	2	0
12H9	3	3	3	2	3	1	1	2	0
2D2	3	3	3	2	3	2	1	2	0
2F10	3	3	3	2	3	2	3	2	0
12F4	3	3	3	2	3	3	1	2	0
10G9	3	3	3	2	3	3	2	2	0
12G7	3	3	3	3	2	3	1	2	0
18E4	3	3	3	3	2	3	3	1	0
1F3	3	3	3	3	3	1	2	2	0
11C4	3	3	3	3	3	2	1	2	0
9H5	3	3	3	3	3	2	3	2	0
9A9	3	3	3	3	3	3	3	2	0

Table AIII-3. Functional lactamase chimeras selected on ampicillin prior to probe hybridization. Not part of the naïve library, but from the RASPP:PST library. The third column in the MIC for ampicillin in μg/mL.

23A11	1	1	3	1	2	1	3	3	100
24D11	1	1	3	3	1	2	3	1	50
23D12	1	1	3	3	2	2	3	1	50
21F6	1	1	3	3	2	2	3	3	50
21A5	1	1	3	3	3	2	3	3	100
24A10	1	1	3	3	3	3	3	1	2000
23A6	1	3	3	1	1	1	3	3	50
24G4	1	3	3	1	2	1	3	3	500
23B7	1	3	3	1	3	1	3	3	500
21C2	1	3	3	2	1	2	3	3	100
22F12	1	3	3	2	2	3	3	2	100
22A10	2	1	3	1	2	1	3	2	500
22B2	2	1	3	1	2	3	3	2	1000
21A4	2	1	3	1	3	3	3	2	250
21G11	2	1	3	3	1	3	3	3	250
24A6	2	1	3	3	2	2	2	2	1000
24D8	2	1	3	3	3	1	2	2	50
22B8	2	2	2	2	2	2	2	2	2000
23B6	2	2	3	1	3	3	1	2	1000
23G12	2	3	1	3	2	3	3	3	1000
24C7	2	3	3	1	3	1	1	2	250
23E1	2	3	3	1	3	1	3	2	10
21C8	2	3	3	1	3	2	3	2	500
23E7	2	3	3	1	3	3	2	2	1000
21C3	2	3	3	2	1	2	3	2	250
22B7	2	3	3	2	2	2	2	2	2000

22C7	2	3	3	2	3	2	2	2	2000
21F5	2	3	3	2	3	2	3	2	1000
22A2	2	3	3	2	3	3	3	2	500
23H5	2	3	3	3	1	2	2	2	2000
24F4	2	3	3	3	2	2	1	2	50
22B5	2	3	3	3	2	2	3	2	500
22A6	2	3	3	3	3	1	2	2	2000
21G4	2	3	3	3	3	2	3	2	500
24A7	3	1	1	1	3	3	3	3	500
24D12	3	1	3	1	1	2	3	3	1000
24G10	3	1	3	1	1	3	3	1	2000
23F5	3	1	3	1	3	1	1	2	500
21D6	3	1	3	1	3	1	3	3	2000
24D3	3	2	2	1	3	3	1	2	500
22E9	3	2	3	1	1	1	3	3	50
21A3	3	2	3	2	1	2	3	3	50
21D12	3	2	3	2	2	2	3	3	250
22E1	3	2	3	3	3	3	3	3	100
21H4	3	3	3	1	2	2	1	3	500
22A11	3	3	3	1	2	2	3	3	2000
22C6	3	3	3	1	2	3	1	2	50
24A8	3	3	3	1	3	2	3	3	2000
22D8	3	3	3	2	2	2	3	3	2000
22F3	3	3	3	2	3	3	3	3	200
21C9	3	3	3	3	1	2	3	3	2000

Table AIII-4. Cytochrome P450 naïve chimeras (third column, 1 for functional, 0 for nonfunctional, 2 for P420 peak counted as nonfunctional) from naïve library, 628 total sequences. The sequences are designated by their block pattern: 1 represents CYP102A1, 2, CYP102A2, and 3, CYP102A3. A chimeras sequence is determined by the sequences of the parental genes in the following blocks: 1-64, 65-122, 123-166, 166-216, 216-268, 269-328, 329-404, 405-460 (Otey et al. 2006).

1	1	1	1	2	1	2	3	0
1	1	1	1	2	2	1	2	1
1	1	1	1	3	2	2	3	0
1	1	1	1	3	2	3	3	1
1	1	1	3	1	3	1	3	1
1	1	1	3	2	2	2	3	0
1	1	1	3	2	2	3	2	0
1	1	1	3	2	3	2	3	0
1	1	2	1	3	1	3	3	1
1	1	2	3	1	2	3	2	0
1	1	2	3	2	1	1	1	0
1	1	2	3	2	3	2	3	0
1	1	2	3	2	3	3	3	1
1	1	3	1	2	2	3	3	1
1	1	3	1	3	2	2	3	2
1	1	3	3	1	1	2	3	0
1	1	3	3	2	2	2	1	0
1	1	3	3	2	3	3	3	1
1	1	3	3	3	1	2	2	0
1	1	3	3	3	2	1	2	1
1	1	3	3	3	3	2	3	2
1	2	1	3	3	2	2	3	0
1	2	2	1	1	2	2	2	0
1	2	2	1	2	1	1	2	1
1	2	2	1	2	2	1	1	0
1	2	2	1	2	2	2	3	2
1	2	2	3	1	2	3	1	0
1	2	2	3	2	1	1	1	2
1	2	2	3	3	1	1	2	0
1	2	2	3	3	3	2	3	0
1	2	3	1	1	3	3	3	1
1	2	3	3	1	1	2	3	1
1	2	3	3	1	2	1	1	0
1	2	3	3	1	2	2	1	2
1	2	3	3	2	1	2	3	2
1	2	3	3	2	2	2	3	1
1	2	3	3	2	3	3	3	1
1	2	3	3	3	3	3	1	1
1	3	1	3	2	2	2	3	0

2	3	3	1	3	2	3	2	1
2	3	3	1	3	3	2	2	2
2	3	3	3	1	1	1	2	0
2	3	3	3	1	2	1	2	0
2	3	3	3	1	2	3	2	0
2	3	3	3	1	3	2	3	1
2	3	3	3	2	2	2	2	1
2	3	3	3	2	3	1	1	1
2	3	3	3	2	3	2	2	0
2	3	3	3	2	3	2	3	1
2	3	3	3	3	1	1	1	1
2	3	3	3	3	1	2	2	2
2	3	3	3	3	1	2	3	1
2	3	3	3	3	2	1	1	1
2	3	3	3	3	2	2	2	1
2	3	3	3	3	2	2	3	0
2	3	3	3	3	2	3	2	1
2	3	3	3	3	2	3	3	1
2	3	3	3	3	3	2	3	1
3	1	1	1	1	2	3	3	1
3	1	1	1	2	1	2	1	0
3	1	1	1	3	1	3	2	1
3	1	1	1	3	3	2	1	0
3	1	1	1	3	3	2	3	1
3	1	1	1	3	3	3	2	1
3	1	1	3	1	2	3	3	1
3	1	1	3	1	3	1	2	0
3	1	1	3	1	3	2	3	2
3	1	1	3	2	2	2	1	0
3	1	1	3	2	2	2	3	0
3	1	1	3	2	2	3	2	1
3	1	1	3	2	3	1	1	0
3	1	1	3	2	3	1	2	0
3	1	1	3	2	3	3	3	1
3	1	1	3	3	1	1	2	0
3	1	1	3	3	1	2	3	0
3	1	1	3	3	2	1	2	0
3	1	1	3	3	2	3	3	1
3	1	1	3	3	3	3	1	1

1	3	1	3	2	3	2	2	0
1	3	1	3	2	3	3	3	2
1	3	1	3	3	3	2	3	0
1	3	2	1	2	1	2	2	0
1	3	2	1	2	3	2	1	0
1	3	2	1	3	1	3	1	1
1	3	2	3	1	3	3	2	0
1	3	2	3	2	1	2	3	0
1	3	2	3	2	3	1	1	0
1	3	2	3	2	3	2	3	0
1	3	2	3	3	1	3	3	0
1	3	2	3	3	2	3	3	0
1	3	2	3	3	3	2	2	0
1	3	3	1	1	3	1	1	2
1	3	3	3	1	1	2	3	2
1	3	3	3	1	3	3	3	2
1	3	3	3	2	2	2	3	0
1	3	3	3	2	3	3	2	2
1	3	3	3	3	1	2	2	1
1	3	3	3	3	1	2	3	2
1	3	3	3	3	1	3	1	1
1	3	3	3	3	2	2	2	0
1	3	3	3	3	2	2	3	2
1	3	3	3	3	2	3	3	2
1	3	3	3	3	3	2	3	2
2	1	1	1	1	1	1	2	0
2	1	1	1	1	2	1	2	1
2	1	1	1	1	3	1	2	0
2	1	1	1	1	3	2	2	0
2	1	1	1	2	2	1	2	1
2	1	1	1	2	2	2	2	1
2	1	1	1	2	2	3	2	1
2	1	1	1	2	3	1	2	1
2	1	1	1	3	1	1	1	1
2	1	1	1	3	1	1	2	1
2	1	1	1	3	2	1	2	1
2	1	1	1	3	2	2	1	1
2	1	1	1	3	2	2	2	2
2	1	1	1	3	3	2	2	1
2	1	1	3	1	1	1	1	0
2	1	1	3	1	2	1	2	0
2	1	1	3	1	3	2	1	0
2	1	1	3	2	1	1	2	1
2	1	1	3	2	1	1	3	1
2	1	1	3	2	1	2	1	0
2	1	1	3	2	3	1	1	1
2	1	1	3	2	3	1	3	1
2	1	1	3	3	1	2	3	2

3	1	2	1	1	1	2	2	0
3	1	2	1	1	1	3	2	0
3	1	2	1	1	2	1	1	0
3	1	2	1	1	3	1	2	0
3	1	2	1	2	1	1	2	1
3	1	2	1	2	1	1	3	0
3	1	2	1	2	1	3	2	2
3	1	2	1	2	2	1	1	0
3	1	2	1	2	3	2	1	1
3	1	2	1	3	1	2	2	0
3	1	2	1	3	2	2	3	0
3	1	2	1	3	2	3	2	1
3	1	2	1	3	3	2	3	1
3	1	2	3	1	2	1	1	0
3	1	2	3	1	3	1	1	0
3	1	2	3	1	3	2	3	0
3	1	2	3	2	2	3	1	1
3	1	2	3	2	3	1	2	1
3	1	2	3	2	3	2	2	0
3	1	2	3	2	3	3	2	1
3	1	2	3	3	1	1	1	2
3	1	2	3	3	1	2	2	0
3	1	2	3	3	1	3	3	0
3	1	2	3	3	2	2	1	1
3	1	2	3	3	2	2	2	0
3	1	2	3	3	2	3	3	1
3	1	2	3	3	3	3	3	1
3	1	3	1	1	1	1	2	0
3	1	3	1	1	1	2	2	0
3	1	3	1	1	2	1	2	0
3	1	3	1	1	3	1	2	0
3	1	3	1	2	2	1	2	1
3	1	3	1	2	2	2	1	1
3	1	3	1	2	2	2	2	1
3	1	3	1	2	2	3	1	1
3	1	3	1	2	2	3	3	1
3	1	3	1	3	1	1	1	1
3	1	3	1	3	1	2	3	0
3	1	3	1	3	1	3	2	1
3	1	3	1	3	1	3	3	1
3	1	3	1	3	2	2	2	0
3	1	3	1	3	2	2	3	1
3	1	3	1	3	2	3	3	1
3	1	3	1	3	3	2	1	0
3	1	3	3	1	2	2	1	0
3	1	3	3	1	2	2	2	0
3	1	3	3	1	2	2	3	0
3	1	3	3	1	3	3	2	0

2	1	1	3	3	1	3	1	1
2	1	1	3	3	2	1	2	1
2	1	1	3	3	2	2	3	1
2	1	1	3	3	3	2	1	1
2	1	1	3	3	3	2	2	1
2	1	1	3	3	3	3	1	1
2	1	1	3	3	3	3	2	1
2	1	2	1	1	1	1	3	0
2	1	2	1	1	1	2	2	0
2	1	2	1	1	2	1	1	0
2	1	2	1	1	2	2	2	0
2	1	2	1	1	3	2	1	1
2	1	2	1	2	1	1	2	0
2	1	2	1	2	1	2	2	1
2	1	2	1	2	1	2	3	1
2	1	2	1	2	2	1	2	2
2	1	2	1	2	3	3	3	1
2	1	2	1	3	1	2	1	1
2	1	2	1	3	2	1	2	1
2	1	2	2	2	1	1	2	1
2	1	2	3	2	1	1	2	1
2	1	2	3	2	1	2	2	1
2	1	2	3	2	1	3	2	1
2	1	2	3	2	2	1	2	0
2	1	2	3	2	2	3	1	1
2	1	2	3	3	1	1	2	0
2	1	2	3	3	1	3	2	1
2	1	2	3	3	2	1	2	1
2	1	2	3	3	2	2	1	1
2	1	2	3	3	3	1	2	1
2	1	2	3	3	3	2	2	2
2	1	3	1	1	1	1	1	0
2	1	3	1	1	3	1	1	0
2	1	3	1	1	3	3	1	0
2	1	3	1	1	3	3	3	0
2	1	3	1	2	1	1	1	1
2	1	3	1	2	1	1	2	1
2	1	3	1	2	1	2	1	2
2	1	3	1	2	1	2	3	1
2	1	3	1	2	2	1	2	0
2	1	3	1	2	3	2	1	0
2	1	3	1	3	1	1	2	1
2	1	3	1	3	3	1	2	1
2	1	3	1	3	3	2	2	1
2	1	3	3	1	1	1	1	0
2	1	3	3	1	1	1	2	2
2	1	3	3	1	1	3	1	0
2	1	3	3	1	3	1	2	0

3	1	3	3	2	1	1	2	0
3	1	3	3	2	1	3	1	1
3	1	3	3	2	1	3	2	0
3	1	3	3	2	1	3	3	1
3	1	3	3	2	2	2	1	0
3	1	3	3	2	2	3	2	1
3	1	3	3	2	2	3	3	1
3	1	3	3	2	3	1	2	1
3	1	3	3	2	3	2	2	1
3	1	3	3	2	3	2	3	1
3	1	3	3	3	1	1	2	0
3	1	3	3	3	2	2	2	0
3	1	3	3	3	2	2	3	0
3	1	3	3	3	2	3	2	0
3	1	3	3	3	2	3	3	1
3	1	3	3	3	3	1	1	0
3	1	3	3	3	3	2	2	1
3	1	3	3	3	3	3	2	1
3	1	3	3	3	3	3	3	1
3	2	1	1	1	1	1	2	0
3	2	1	1	1	1	2	1	0
3	2	1	1	1	1	2	3	0
3	2	1	1	1	2	1	1	2
3	2	1	1	1	3	1	1	0
3	2	1	1	2	2	1	2	1
3	2	1	1	2	2	3	2	0
3	2	1	1	2	3	1	1	0
3	2	1	1	2	3	2	1	1
3	2	1	1	3	1	1	2	0
3	2	1	3	1	1	3	3	1
3	2	1	3	1	2	1	2	0
3	2	1	3	1	3	1	1	0
3	2	1	3	2	2	1	1	0
3	2	1	3	2	2	1	2	0
3	2	1	3	2	2	2	1	0
3	2	1	3	3	1	1	1	1
3	2	1	3	3	1	1	3	0
3	2	1	3	3	1	2	2	0
3	2	1	3	3	2	1	2	0
3	2	1	3	3	2	2	3	0
3	2	1	3	3	2	3	2	1
3	2	1	3	3	2	3	3	1
3	2	1	3	3	3	1	1	0
3	2	1	3	3	3	1	2	0
3	2	1	3	3	3	2	1	0
3	2	1	3	3	3	2	3	0
3	2	1	3	3	3	3	1	1
3	2	2	1	1	1	1	1	0

2	1	3	3	1	3	1	3	0
2	1	3	3	2	1	1	3	1
2	1	3	3	2	1	3	1	1
2	1	3	3	2	2	1	2	1
2	1	3	3	2	3	2	3	1
2	1	3	3	2	3	3	3	1
2	1	3	3	3	1	2	1	0
2	1	3	3	3	1	3	2	1
2	1	3	3	3	1	3	3	0
2	1	3	3	3	2	1	2	1
2	1	3	3	3	3	1	2	1
2	1	3	3	3	3	3	1	0
2	1	3	3	3	3	3	2	0
2	2	1	1	1	2	3	1	0
2	2	1	1	2	1	1	1	1
2	2	1	1	2	3	2	1	1
2	2	1	1	3	1	3	2	0
2	2	1	1	3	3	1	3	1
2	2	1	2	1	3	3	1	0
2	2	1	3	1	1	1	1	0
2	2	1	3	1	1	1	2	0
2	2	1	3	1	1	3	2	0
2	2	1	3	1	1	3	3	0
2	2	1	3	1	2	2	2	0
2	2	1	3	1	3	2	1	0
2	2	1	3	2	3	1	2	1
2	2	1	3	2	3	3	1	1
2	2	1	3	3	1	1	2	1
2	2	1	3	3	2	1	1	1
2	2	1	3	3	2	1	2	1
2	2	1	3	3	3	1	2	1
2	2	2	1	1	1	3	2	0
2	2	2	1	1	2	2	2	0
2	2	2	1	1	3	3	1	0
2	2	2	1	1	3	3	2	0
2	2	2	1	2	2	3	2	1
2	2	2	1	2	3	1	2	1
2	2	2	1	3	1	1	1	1
2	2	2	1	3	1	1	2	1
2	2	2	1	3	2	2	2	1
2	2	2	1	3	3	1	2	1
2	2	2	1	3	3	2	1	1
2	2	2	3	1	1	1	3	0
2	2	2	3	1	1	2	2	0
2	2	2	3	1	2	1	1	0
2	2	2	3	1	2	1	2	0
2	2	2	3	1	2	2	3	1
2	2	2	3	1	2	3	2	0

3	2	2	1	1	1	1	2	1
3	2	2	1	1	1	3	3	0
3	2	2	1	1	2	1	2	0
3	2	2	1	2	1	1	1	0
3	2	2	1	2	1	2	2	2
3	2	2	1	2	1	3	3	1
3	2	2	1	2	2	3	1	1
3	2	2	1	2	3	1	1	0
3	2	2	1	2	3	2	3	1
3	2	2	1	3	1	2	3	1
3	2	2	1	3	1	3	2	1
3	2	2	1	3	3	1	1	0
3	2	2	1	3	3	1	2	0
3	2	2	1	3	3	3	3	1
3	2	2	3	1	1	2	2	0
3	2	2	3	1	2	2	2	0
3	2	2	3	1	3	3	2	0
3	2	2	3	2	1	1	1	0
3	2	2	3	2	1	3	3	0
3	2	2	3	2	2	1	2	0
3	2	2	3	2	2	1	3	0
3	2	2	3	2	2	2	1	0
3	2	2	3	2	2	2	2	2
3	2	2	3	2	3	3	1	1
3	2	2	3	2	3	3	3	1
3	2	2	3	3	1	1	2	0
3	2	2	3	3	1	2	2	0
3	2	2	3	3	1	2	3	0
3	2	2	3	3	2	2	2	1
3	2	2	3	3	2	3	2	0
3	2	2	3	3	2	3	3	0
3	2	2	3	3	3	3	2	1
3	2	3	1	1	1	3	1	1
3	2	3	1	1	1	3	2	0
3	2	3	1	1	2	1	2	0
3	2	3	1	1	2	2	1	0
3	2	3	1	1	3	2	2	1
3	2	3	1	2	2	1	2	1
3	2	3	1	2	2	3	3	1
3	2	3	1	2	3	1	1	1
3	2	3	1	2	3	2	3	1
3	2	3	1	2	3	3	1	1
3	2	3	1	2	3	3	2	1
3	2	3	1	3	1	2	2	0
3	2	3	1	3	1	3	3	1
3	2	3	1	3	2	1	2	0
3	2	3	1	3	2	2	2	0
3	2	3	1	3	2	3	2	1

2	2	2	3	1	3	1	1	0
2	2	2	3	1	3	1	2	0
2	2	2	3	1	3	3	3	0
2	2	2	3	2	1	1	2	1
2	2	2	3	2	1	2	2	1
2	2	2	3	2	2	1	2	1
2	2	2	3	2	2	1	3	2
2	2	2	3	2	2	2	2	1
2	2	2	3	2	3	1	2	1
2	2	2	3	2	3	2	3	1
2	2	2	3	3	1	1	2	1
2	2	2	3	3	2	2	1	1
2	2	2	3	3	2	2	2	1
2	2	2	3	3	2	2	3	1
2	2	2	3	3	3	3	1	0
2	2	2	3	3	3	3	3	0
2	2	3	1	1	1	2	1	0
2	2	3	1	1	2	1	2	0
2	2	3	1	1	2	3	1	0
2	2	3	1	1	3	3	2	0
2	2	3	1	2	1	2	3	1
2	2	3	1	2	1	3	2	0
2	2	3	1	2	2	1	1	1
2	2	3	1	2	2	2	1	1
2	2	3	1	2	2	2	2	1
2	2	3	1	2	2	3	1	1
2	2	3	1	2	3	1	2	1
2	2	3	1	2	3	2	2	1
2	2	3	1	3	2	2	1	1
2	2	3	1	3	2	2	2	1
2	2	3	1	3	2	3	2	1
2	2	3	1	3	2	3	3	1
2	2	3	1	3	3	3	1	1
2	2	3	1	3	3	3	3	0
2	2	3	3	1	1	2	1	0
2	2	3	3	1	1	3	3	0
2	2	3	3	1	2	1	1	0
2	2	3	3	1	2	1	2	0
2	2	3	3	1	2	2	1	1
2	2	3	3	1	2	2	2	0
2	2	3	3	1	3	2	1	0
2	2	3	3	1	3	3	2	0
2	2	3	3	2	1	1	2	1
2	2	3	3	2	1	1	3	1
2	2	3	3	2	1	2	3	1
2	2	3	3	2	1	3	2	1
2	2	3	3	2	2	1	1	1
2	2	3	3	2	2	2	2	1

3	2	3	1	3	3	1	3	1
3	2	3	1	3	3	3	2	1
3	2	3	3	1	1	1	1	0
3	2	3	3	1	1	1	2	0
3	2	3	3	1	1	2	2	0
3	2	3	3	1	1	3	2	0
3	2	3	3	1	2	1	2	0
3	2	3	3	1	2	2	1	0
3	2	3	3	1	2	2	2	0
3	2	3	3	1	3	1	1	0
3	2	3	3	1	3	1	3	0
3	2	3	3	2	1	1	1	0
3	2	3	3	2	1	1	2	0
3	2	3	3	2	1	2	3	0
3	2	3	3	2	1	3	3	1
3	2	3	3	2	2	1	1	0
3	2	3	3	2	2	1	2	0
3	2	3	3	2	2	2	2	0
3	2	3	3	2	2	2	3	1
3	2	3	3	2	3	2	3	1
3	2	3	3	3	1	2	2	0
3	2	3	3	3	2	1	2	0
3	2	3	3	3	2	2	3	1
3	2	3	3	3	2	3	2	1
3	2	3	3	3	3	1	1	0
3	2	3	3	3	3	1	2	1
3	2	3	3	3	3	2	2	2
3	2	3	3	3	3	2	3	1
3	2	3	3	3	3	3	3	1
3	3	1	1	1	1	2	2	0
3	3	1	1	1	2	1	2	0
3	3	1	1	1	2	2	2	0
3	3	1	1	1	3	1	2	0
3	3	1	1	2	1	1	2	2
3	3	1	1	3	1	1	1	1
3	3	1	1	3	1	2	1	0
3	3	1	1	3	2	1	2	1
3	3	1	1	3	2	2	3	0
3	3	1	3	1	1	2	2	0
3	3	1	3	1	1	2	3	0
3	3	1	3	1	3	3	2	0
3	3	1	3	1	3	3	3	1
3	3	1	3	2	2	2	2	0
3	3	1	3	2	2	2	3	0
3	3	1	3	2	3	2	2	0
3	3	1	3	3	1	2	1	0
3	3	1	3	3	1	3	1	1
3	3	1	3	3	2	2	3	0

2	2	3	3	2	2	3	2	1
2	2	3	3	2	3	1	2	1
2	2	3	3	2	3	2	1	1
2	2	3	3	2	3	2	2	1
2	2	3	3	3	1	1	1	0
2	2	3	3	3	1	3	2	1
2	2	3	3	3	1	3	3	1
2	2	3	3	3	2	1	2	1
2	2	3	3	3	2	2	1	1
2	2	3	3	3	2	2	2	1
2	2	3	3	3	2	3	1	1
2	2	3	3	3	3	1	3	1
2	2	3	3	3	3	2	3	1
2	3	1	1	1	1	1	2	0
2	3	1	1	1	2	1	2	0
2	3	1	1	2	1	2	3	0
2	3	1	1	2	2	1	3	0
2	3	1	1	2	2	2	2	0
2	3	1	1	3	1	1	1	1
2	3	1	1	3	2	1	2	1
2	3	1	1	3	3	1	1	1
2	3	1	1	3	3	1	2	1
2	3	1	2	2	2	1	2	1
2	3	1	3	1	3	2	3	1
2	3	1	3	1	3	3	2	0
2	3	1	3	2	1	1	1	2
2	3	1	3	2	1	2	1	1
2	3	1	3	2	2	1	2	1
2	3	1	3	2	2	2	1	1
2	3	1	3	2	2	3	1	1
2	3	1	3	2	3	2	2	0
2	3	1	3	2	3	2	3	1
2	3	1	3	3	1	1	2	1
2	3	1	3	3	1	2	1	1
2	3	1	3	3	3	1	1	1
2	3	1	3	3	3	1	2	0
2	3	1	3	3	3	2	1	1
2	3	1	3	3	3	3	3	1
2	3	2	1	1	1	2	1	0
2	3	2	1	1	1	3	1	0
2	3	2	1	1	1	3	2	1
2	3	2	1	1	2	2	2	0
2	3	2	1	1	3	1	1	0
2	3	2	1	1	3	3	2	0
2	3	2	1	2	1	1	2	1
2	3	2	1	2	2	1	2	1
2	3	2	1	2	2	3	1	1
2	3	2	1	2	3	1	2	0

3	3	1	3	3	2	3	3	0
3	3	1	3	3	3	2	1	0
3	3	1	3	3	3	2	3	2
3	3	1	3	3	3	3	2	0
3	3	1	3	3	3	3	3	1
3	3	2	1	1	1	1	2	0
3	3	2	1	1	2	1	1	0
3	3	2	1	1	3	1	2	0
3	3	2	1	1	3	2	1	0
3	3	2	1	2	2	2	2	0
3	3	2	1	2	3	1	1	1
3	3	2	1	2	3	1	2	0
3	3	2	1	2	3	1	3	0
3	3	2	1	3	1	1	2	0
3	3	2	1	3	2	1	1	1
3	3	2	1	3	2	3	2	1
3	3	2	3	1	2	1	2	0
3	3	2	3	1	2	2	1	0
3	3	2	3	1	3	1	2	0
3	3	2	3	1	3	3	3	0
3	3	2	3	2	1	1	2	0
3	3	2	3	2	1	2	2	0
3	3	2	3	2	1	2	3	0
3	3	2	3	2	2	2	2	0
3	3	2	3	2	2	2	3	0
3	3	2	3	2	2	3	3	1
3	3	2	3	2	3	1	2	1
3	3	2	3	2	3	2	2	2
3	3	2	3	2	3	2	3	0
3	3	2	3	2	3	3	3	1
3	3	2	3	3	1	1	2	0
3	3	2	3	3	1	3	1	1
3	3	2	3	3	2	2	1	0
3	3	2	3	3	2	2	2	0
3	3	2	3	3	2	2	3	0
3	3	2	3	3	2	3	3	1
3	3	2	3	3	3	2	3	0
3	3	2	3	3	3	3	3	1
3	3	3	1	1	1	2	2	0
3	3	3	1	1	2	2	3	0
3	3	3	1	1	2	3	1	1
3	3	3	1	1	3	1	1	0
3	3	3	1	1	3	1	2	0
3	3	3	1	1	3	2	2	0
3	3	3	1	1	3	3	2	0
3	3	3	1	2	2	3	3	0
3	3	3	1	2	3	1	2	0
3	3	3	1	2	3	2	2	1

2	3	2	1	2	3	3	2	1
2	3	2	1	3	1	2	3	1
2	3	2	1	3	2	2	3	1
2	3	2	1	3	2	3	1	0
2	3	2	1	3	2	3	2	1
2	3	2	1	3	3	1	1	1
2	3	2	1	3	3	2	2	1
2	3	2	3	1	1	2	1	2
2	3	2	3	1	2	1	2	0
2	3	2	3	1	3	2	3	0
2	3	2	3	2	2	1	1	1
2	3	2	3	2	2	1	2	1
2	3	2	3	2	2	2	1	2
2	3	2	3	2	2	2	3	0
2	3	2	3	2	2	3	3	1
2	3	2	3	2	3	2	3	1
2	3	2	3	3	2	3	2	1
2	3	2	3	3	3	2	2	0
2	3	2	3	3	3	3	3	1
2	3	3	1	1	1	1	2	0
2	3	3	1	1	2	2	1	0
2	3	3	1	1	2	2	2	0
2	3	3	1	1	2	3	3	1
2	3	3	1	1	3	1	3	0
2	3	3	1	2	1	2	2	1
2	3	3	1	2	1	3	1	1
2	3	3	1	2	2	2	3	1
2	3	3	1	2	3	1	1	1
2	3	3	1	2	3	1	2	1
2	3	3	1	2	3	2	3	1
2	3	3	1	3	1	2	1	0
2	3	3	1	3	1	3	3	1
2	3	3	1	3	2	1	2	1
2	3	3	1	3	2	2	2	1
2	3	3	1	3	2	3	1	0

3	3	3	1	2	3	2	3	0
3	3	3	1	2	3	3	3	1
3	3	3	1	3	1	2	2	0
3	3	3	1	3	2	2	3	0
3	3	3	1	3	2	3	3	0
3	3	3	1	3	3	1	1	0
3	3	3	1	3	3	2	3	1
3	3	3	3	1	1	2	2	0
3	3	3	3	1	1	3	3	0
3	3	3	3	1	2	2	3	2
3	3	3	3	1	2	3	2	1
3	3	3	3	1	2	3	3	1
3	3	3	3	1	3	1	1	0
3	3	3	3	1	3	2	1	2
3	3	3	3	1	3	3	1	0
3	3	3	3	1	3	3	3	1
3	3	3	3	2	1	1	2	0
3	3	3	3	2	1	2	1	0
3	3	3	3	2	1	2	3	0
3	3	3	3	2	1	3	3	1
3	3	3	3	2	2	1	1	0
3	3	3	3	2	2	2	1	2
3	3	3	3	2	2	2	3	0
3	3	3	3	2	2	3	2	1
3	3	3	3	2	2	3	3	1
3	3	3	3	2	3	2	3	1
3	3	3	3	3	1	3	3	0
3	3	3	3	3	2	2	2	0
3	3	3	3	3	2	2	3	0
3	3	3	3	3	2	3	1	1
3	3	3	3	3	2	3	2	1
3	3	3	3	3	3	1	1	0
3	3	3	3	3	3	2	1	2
3	3	3	3	3	3	2	3	1
3	3	3	3	3	3	3	2	0

References

Adami, C. 2004. Information theory in molecular biology. *Physics of Life Reviews* **1:** 3-22.

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. A standard numbering scheme for the class A beta-lactamases. *Biochemical Journal* **276:** 269-272.

Arnold, F.H. 1998. When blind is better: protein design by evolution. *Nature Biotechnology* **16:** 617-618.

Arnold, F.H. 2000. *Advances in Protein Chemistry*. Academic Press, San Diego.

Arnold, G.E., and Ornstein, R.L. 1997. Molecular dynamics study of time-correlated protein domain motions and molecular flexibility: cytochrome P450BM-3. *Biophysical Journal* **73:** 1147-1159.

Axe, D. 2004. Estimating the Prevalence of Protein Sequences Adopting Functional Enzyme Folds. *Journal of Molecular Biology* **341:** 1295-1315.

Back, K., and Chappell, J. 1996. Identifying functional domains within terpene cyclases using a domain-swapping strategy. *Proceedings of the National Academy of Science, USA* **93:** 6841-6845.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffths-Jones, S., Khanna, A., Marshal, M., Maxon, S., Sonnhammer, E.L.L., et al. 2004. The Pfam Protein Families Database. *Nucleic Acids Research* **32:** D138-D141.

Bernstein, S., and Bernstein, R. 1999. *Elements of Statistics II: Inferential Statistics*. McGraw-Hill.

Bloom, J.D., Labthavikul, S.T., Otey, C.R., and Arnold, F.H. 2006. Protein stability promotes evolvability. *Proceedings of the National Academy of Science, USA* **109:** 5869-5874.

Bloom, J.D., Meyer, M.M., Meinhold, P., Otey, C.R., MacMillan, D., and Arnold, F.H. 2005a. Evolving Strategies for Enzyme Engineering. *Current Opinion in Structural Biology* **15:** 447-452.

Bloom, J.D., Silberg, J.J., Wilke, C.O., Drummond, D.A., Adami, C., and Arnold, F.H. 2005b. Thermodynamic prediction of protein neutrality. *Proceedings of the National Academy of Science, USA* **102:** 606-611.

Bogarad, L.D., and Deem, M.W. 1999. A hierarchical approach to protein molecular evolution. *Proceedings of the National Academy of Science, USA* **96:** 2591-2595.

Bornberg-Bauer, E., and Chan, H.S. 1999. Modeling evolutionary landscapes: Mutational stability, topology and superfunnels in sequence space. *Proceedings of the National Academy of Sciences USA* **96:** 10689-10694.

Chen, Y., Delman, J., Sirot, J., Shoichet, B., and Bonnet, R. 2005. Atomic resolution structures of CTX-M beta-lactamases: Extended spectrum activities from increased mobility and decreased stability. *Journal of Molecular Biology* **348:** 349-362.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research* **31:** 3497-3500.

Crameri, A., Raillard, S.-A., Bermudez, E., and Stemmer, W.P.C. 1998. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391:** 288-291.

Cui, Y., Wong, W.H., Bornberg-Bauer, E., and Chan, H.S. 2002. Recombinatoric exploration of novel folded structures: A heteropolymer-based model of protein evolutionary landscapes. *Proceedings of the National Academy of Sciences USA* **99:** 809-814.

Daugherty, P.S., Chen, G., Iverson, B.I., and Georgiou, G. 2000. Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. *Proceedings of the National Academy of Sciences USA* **97:** 2029-2034.

Decousser, J.W., Poirel, L., and Nordmann, P. 2001. Characterization of a Chromosomally Encoded Extended-Spectrum Class A beta-lactamase from *Kluyvery cryocrescens*. *Antimicrobial Agents and Chemotherapy* **45:** 3595-3598.

Dideberg, O., Charlier, P., Wery, J.P., Dehottay, P., Dusart, J., Erpicum, T., Frere, J.-M., and Ghuysen, J.M. 1987. The crystal structure of the beta-lactamase of *Streptomyces albus G* at 0.3 nm resolution. *Biochem J.* **245:** 991-993.

Dokholyan, N.V., and Shakhnovich, E. 2001. Understanding Hierarchical Protein Evolution from First Principles. *Journal of Molecular Biology* **312:** 289-307.

Dooner, H.K., and Martinez-Ferez, I.M. 1997. Recombination Occurs Uniformly within the bronze Gene, a Meiotic Recombination Hotspot in the Maize Genome. *The Plant Cell* **9:** 1633-1646.

Drummond, D.A., Silberg, J.J., Meyer, M.M., Wilke, C.O., and Arnold, F.H. 2005. On the conservative nature of intragenic recombination. *Proceedings of the National Academy of Sciences USA* **102:** 5280-5385.

Endelman, J.B. 2005. Design and Analysis of Combinatorial Protein Libraries Created by Site-Directed Recombination. In *Bioengineering*, pp. 122. California Institute of Technology, Pasadena.

Endelman, J.B., Bloom, J.D., Otey, C.R., Landwehr, M., and Arnold, F.H. 2005. Inferring interactions from an alignment of folded and unfolded protein sequences. *arXiv:q-bio.BM/0505018.*

Endelman, J.B., Silberg, J.J., Wang, Z.-G., and Arnold, F.H. 2004. Site-directed protein recombination as a shortest-path problem. *Protein Engineering, Design & Selection* **17:** 589-594.

Ewart, S., Huber, T., Annemarie, H., and Pluckthun, A. 2003. Biophysical Properties of Human Antibody Variable Domains. *Journal of Molecular Biology* **325:** 531-553.

Feil, E.J., Holmes, E.C., Bessen, D.E., Chan, M.S., Day, N.P.J., Enright, M.C., Goldstein, R., Hood, D.W., Kalla, A., Moore, C.E., et al. 2001. Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences, USA* **98:** 182-187.

Feil, E.J., Maiden, M.C.J., Achtman, M., and Spratt, B.G. 1999. The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Molecular Biology and Evolution* **16:** 1496-1502.

Fodor, A.A., and Aldrich, R.W. 2004a. Influence of Conservation on Calculations of Amino Acid Covariance in Multiple Sequence Alignments. *Proteins* **56:** 211-221.

Fodor, A.A., and Aldrich, R.W. 2004b. On Evolutionary Conservation of Thermodynamic Coupling in Proteins. *Journal of Biological Chemistry* **279:** 19046-19050.

Fu, H., Zheng, Z., and Dooner, H.K. 2002. Recombination rates between adjacent genetic and retrotransposon regions in maize vary by 2 orders of magnitude. *Proceedings of the National Academy of Sciences, USA* **99:** 1082-1087.

Gibbs, M.D., Nevalainen, K.M., and Bergquist, P.L. 2001. Degenerate oligonucleotide gene shuffling (DOGS): a method for enhancing the frequency of recombination with family shuffling. *Gene* **271:** 13-20.

Gobel, U., Sander, C., Schneider, R., and Valencia, A. 1994. Correlated Mutations and Residue Contacts In Proteins. *Proteins-Structure Function And Genetics* **18:** 309-317.

Govindarajan, S., and Goldstein, R.A. 1996. Why are some protein structures so common? *Proceedings of the National Academy of Sciences USA* **93:** 2241-3345.

Guex, N., and Peitsch, M.C. 1997a. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **18:** 2714-2723.

Guo, H.H., Choe, J., and Loeb, L. 2004. Protein tolerance to random amino change. *Proceedings of the National Academy of Sciences, USA* **101:** 9205-9210.

Hernandez, G., and LeMaster, D.M. 2005. Hybrid Native Partitioning of Interactions Among Nonconserved Residues in Chimeric Proteins. *Proteins: Structure, Function and Bioinformatics* **60:** 723-731.

Herzberg, O. 1991. Refined crystal structure of beta-lactamase from *Staphlococcus aureus* PC1 at 2.0 A resolution. *Journal of Molecular Biology* **217:** 701-719.

Hiraga, K., and Arnold, F.H. 2003. A general method for sequence-independent site-directed chimeragenesis. *Journal of Molecular Biology* **330:** 287-296.

Holm, L., and Sander, C. 1996. Mapping the protein universe. *Science* **273:** 595-603.

Horton, R.M., Hunt, H.D., Ho, S.N., Pullen, J.K., and Pease, L.R. 1989. Engineering hybrid genes without the use of restriction enzymes: gene splicing by overlap extension. *Gene* **77:** 61-68.

Huang, W., Petrosino, J., Hirsch, M., Shenkin, P.S., and Palzkill, T. 1996. Amino Acid Sequence Determinants of b-lactamase Structure and Activity *Journal of Molecular Biology* **258:** 688-703.

Huang, W.Z., and Palzkill, T. 1997. A natural polymorphism in beta-lactamase is a global suppressor. *Proceedings of the National Academy of Sciences, USA* **94:** 8801-8806.

Ibuka, A., Taguchi, A., Ishiguro, M., Fushinobu, S., Ishii, Y., Kamitori, S., Okuyama, K., Yamaguchi, K., Konno, M., and Matsuzawa, H. 1999. Crystal structure of the E166A mutant of extended-spectrum beta-lactamase Toho-1 at 1.8 angstrom resolution. *Journal of Molecular Biology* **285:** 2079-2087.

Jacoby, G., and Bush, K. 2005. Lahey Clinic page on "Amino Acid Sequences for TEM, SHV and OXA Extended-Spectrum and Inhibitor Resistant beta-lactamases

Jelsch, C., Mourey, L., Masson, J.M., and Samama, J.P. 1993. Crystal structure of *Escherichia coli* TEM1 beta-lactamase at 1.8 A resolution. *Proteins* **16:** 364-383.

Joern, J.M., Meinhold, P., and Arnold, F.H. 2002. Analysis of shuffled gene libraries. *Journal of Molecular Biology* **316:** 643-656.

Kawarasaki, Y., Griswold, K.E., Stevenson, J.D., Selzer, T., and Benkovic, S.J. 2003. Enhanced crossover SCRATCHY: construction and high-throughput screening of a combinatorial library containing multiple non-homologous crossovers. *Nucleic Acids Research* **31:** E126.

Keefe, A.D., and Szostak, J.W. 2001. Functional proteins from a random-sequence library. *Nature* **410:** 715-718.

Knox, J.R. 1995. Extended-Spectrum and Inhibitor-Resistant Tem-Type Beta-Lactamases - Mutations, Specificity, And 3-Dimensional Structure. *Antimicrobial Agents and Chemotherapy* **39:** 2593-2601.

Knox, J.R., and Moews, P.C. 1991. Beta-lactamase of *Baccillus licheniformis* refinement at 2A resolution and analysis of hydration. *Journal of Molecular Biology* **220:** 435-455.

Kushiro, T., Shibuya, M., and Ebizuka, Y. 1999. Chimeric Triterpene Synthase: A Possible Model for Multifunctional Triterpene Synthase. *Journal of the American Chemical Society* **121:** 1208-1216.

Kuzin, A.P., Nukaga, M., Nukaga, Y., Hujer, A.M., Bonomo, R.A., and Knox, J.R. 1999. Structure of the SHV-1 β-lactamase. *Biochemistry* **38:** 5720-5727.

Laurent, F., Poirel, L., Naas, T., Chaibl, E.B., Labia, R., Boiron, P., and Nordmann, P. 1999. Biochemical-Genetic Analysis and Distribution of FAR-1, a Class A beta-lactamase from *Bocardia farcinica*. *Antimicrobial Agents and Chemotherapy* **43:** 1644-1650.

Lehmann, M., Kostrewa, D., Wyss, M., Brugger, R., D'Arcy, A., Pasamontes, L., and Van Loon, A. 2000. From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. *Protein Engineering* **13:** 49-57.

Lehmann, M., Loch, C., Middendorf, A., Studer, D., Lassen, S.F., Pasamontes, L., Van Loon, A., and Wyss, M. 2002. The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Engineering* **15:** 403-411.

Li, H., Helling, R., Tang, C., and Wingreen, N. 1996. Emergence of Preferred Structures in a Simple Model of Protein Folding. *Science* **273:** 666-669.

Lim, D., Sanschagrin, F., Passmore, L., De Castro, L., Levesque, R.C., and Strynadka, N.C. 2001a. Insights into the molecular basis for the carbenicillinase activity of PSE-4 beta-lactamase from crystallographic and kinetic studies. *Biochemistry* **40:** 395-402.

Lim, D., Sanschagrin, F., Passmore, L., De Castro, L., Levesque, R.C., and Strynadka, N.C.J. 2001b. Insights into the Molecular Basis for the Carbenicillinase Activity of PSE-4 β-lactamase from Crystallographic and Kinetic Studies. *Biochemistry* **40:** 395-402.

Lockless, S.W., and Ranganathan, R. 1999. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science* **286:** 295-299.

Lutz, S., Ostermeier, M., Moore, G.L., Maranas, C.D., and Benkovic, S.J. 2001. Creating multiple-crossover DNA libraries independent of sequence identity. *Proceedings of the National Academy of Sciences USA* **98:** 11248-11253.

Madinier, I., Fosse, T., Giudicelli, J., and Labia, R. 2001. Cloning and Biochemical Characterization of a Class A beta-lactamase from *Prevotella intermedia*. *Antimicrobial Agents and Chemotherapy* **45:** 2386-2389.

Markiewicz, P., Kleina, L.G., Cruz, C., Ehret, S., and Miller, J.H. 1994. Genetic-Studies Of The Lac Repressor: Analysis of 4000 Altered *E. coli* lac Repressors Reveals Essential and Nonessential Residues, As Well As Spacers Which Do Not Require A Specific Sequence. *Journal of Molecular Biology* **240:** 421-433.

Matagne, A., LaMotte-Brasseur, J., and Frere, J.-M. 1998. Catalytic properties of class A lactamases: efficiency and diversity. *Biochemical Journal* **330:** 581-598.

Maveyraud, L., Mourey, L., Pedalacq, J.-D., Guillet, V., Kotra, L.K., Mobashery, S., and Samama, J.P. 1998. Structural Basis for Clinical Longevity of Carbapenem Antibiotics in the Face of Challenge by the Common Class A beta-lactamases from Antibiotic-Resistant Bacteria. *Journal of the American Chemical Society* **120:** 9748-9752.

Maxwell, K.L., Mittermaier, A.K., Forman-Kay, J.D., and Davidson, A.R. 1999. A simple *in vivo* assay for increased protein solubility. *Protein Science* **8:** 1908-1911.

Meinhold, P., Joern, J.M., and Silberg, J.J. 2003. Analysis of Shuffled Libraries by Oligonucleotide Probe Hybridization. In *Directed Evolution Library Creation*. (eds. F.H. Arnold, and G. Georgiou), pp. 177-187. Humana Press, Totowa, New Jersey.

Mendes, J., Guerois, R., and Serrano, L. 2002. Energy estimation in protein design. *Current Opinion in Structural Biology* **12:** 441-446.

Meyer, M.M., Silberg, J.J., Voigt, C.A., Endelman, J.B., Mayo, S.L., Wang, Z.-G., and Arnold, F.H. 2003. Library analysis of SCHEMA-guided recombination. *Protein Science* **12:** 1686-1693.

Moore, G.L., and Maranas, C.D. 2003. Identifying residue-residue clashes in protein hybrids by using a second-order mean-field approach. *Proceedings of the National Academy of Sciences, USA* **100:** 5091-5096.

Morimotoa, S., and Tamura, A. 2004. Key Elements for Protein Foldability Revealed by a Combinatorial Approach among Similarly Folded but Distantly Related Proteins. *Biochemistry* **43:** 6596-6605.

Nelson, D. 2005. Cytochrome P450 Homepage http://drnelson.utmem.edu/CytochromeP450.html

Nicot, C., Relat, J., Woldegiorgis, Haro, D., and Marrerro, P.F. 2002. Pig Liver Carnitine Palmitoyltransferase. *Journal of Biological Chemistry* **277**: 10044-10049.

Nikolova, P.V., Henckel, J., Lane, D.P., and Fersht, A.R. 1998. Semirational design of active tumor suppressor p53 DNA binding domain with enhanced stability. *Proceedings of the National Academy of Sciences, USA* **95:** 14675-14680.

O'Maille, P., Bakhtina, M., and Tsai, M. 2002. Structure-based Combinatorial Protein Engineering (SCOPE). *Journal of Molecular Biology* **321:** 677.

Olmea, O., Rost, B., and Valencia, A. 1999. Effective use of sequence correlation and conservation in fold recognition. *Journal of Molecular Biology* **193:** 1221-1239.

Orencia, M.C., Yoon, J.S., Ness, J.E., Stemmer, W.P.C., and Stevens, R.C. 2001. Predicting the emergence of antibiotic resistance by directed evolution and structural analysis. *Nature* **8:** 238-242.

Ostermeier, M. 2003. Synthetic gene libraries: in search of the optimal diversity. *Trends in Biotechnology* **21:** 244-247.

Ostermeier, M., Nixon, A.E., and Benkovic, S.L. 1999a. Incremental Truncation as a Strategy in the Engineering of Novel Biocatalysts. *Bioorganic and Medicinal Chemistry* **7:** 2139-2144.

Ostermeier, M., Shim, J.H., and Benkovic, S.J. 1999b. A combinatorial approach to hybrid enzymes independent of DNA homology. *Nature Biotechnology* **17:** 1205-1209.

Osuna, J., Perez-Blancas, A., and Soberon, X. 2002. Improving a circularly permuted TEM-1 beta-lactamase by directed evolution. *Protein Engineering* **15:** 463-470.

Otey, C.R., Landwehr, M., Endelman, J.B., Hiraga, K., Bloom, J.D., and Arnold, F.H. 2006. Structure-Guided Recombination Creates an Artificial Family of Cytochromes P450. *Public Library of Science Biology* **4:** e112.

Otey, C.R., Silberg, J.J., Endelman, J.B., Bandara, G., and Arnold, F.H. 2004. Functional Evolution and Structural Conservation in Chimeric Cytochromes P450: Calibrating a Structure-Guided Approach. *Chemistry and Biology* **11:** 309-318.

Patrick, W.M., and Firth, A.E. 2005. Strategies and computational tools for improving randomized protein libraries. *Biomolecular Engineering* **22:** 105-112.

Petrella, S., Clermont, D., Casin, I., Jarlier, V., and Sougakoff, W. 2001. Novel Class A beta-lactamase Sed-1 from *Citrocater sedlakii*: Genetic Diversity of beta-lactamases within the *Citrobacter* Genus. *Antimicrobiall Agents and Chemotherapy* **45:** 2287-2298.

Petrella, S., Pernot, L., and Sougakoff, W. 2004. Crystallization and preliminary X-ray diffraction study of the class A beta-lactamase SED-1 and its mutant SED-G238C from *Citrobacter sedlakii*. *Acta Crystallographica Section D* **D60:** 125-128.

Petrosino, J.F., and Palzkill, T. 1996. Systematic Mutagenesis of the Active Site Omega Loop of TEM-1 β-lactamase. *Journal of Bacteriology* **178:** 1821-1828.

Philipps, B., Hennecke, J., and Glockshuber, R. 2003. FRET-based *in vivo* screening for protein folding and increased protein stability. *Journal of Molecular Biology* **327:** 239-249.

Poirel, L., Laurent, F., Naas, T., Labia, R., Boiron, P., and Nordmann, P. 2001. Molecular and Biochemical Analaysis of AST-1, a Class A beta-lactamase from *Norcardia asteroides* Sensu Stricto. *Antimicrobial Agents and Chemotherapy* **45:** 878-882.

Poteete, A.R., Rennell, D., Bouvier, S.E., and Hardy, L.W. 1997. Alteration of T4 lysozyme structure by second-site reversion of deleterious mutations. *Protein Science* **6:** 2418-2425.

Rennell, D., Bouvier, S.E., Hardy, L.W., and Poteete, A.R. 1991. Systematic mutation of bacteriphage-T4 lysozyme. *Journal of Molecular Biology* **222:** 67-87.

Sabbagh, Y., Theriault, E., Sanschagrin, F., Voyer, N., Palzkill, T., and Levesque, R.C. 1998. Characterization of a PSE-4 mutant with different properties in relation to penicillanic acid sulfones: importance of residues 216 to 218 in class A beta-lactamases. *Antimicrobial Agents Chemotherapy* **42:** 2319-2325.

Sanschagrin, F., Theriault, E., Sabbagh, Y., Voyer, N., and Levesque, R.C. 2000. Combinatorial biochemistry and shuffling of TEM, SHV and *Streptomyces albus* omega loops in PSE-4 class A β-lactamase. *Journal of Antimicrobial Chemotherapy* **45:** 517-519.

Saraf, M.C., Horswill, A.R., Benkovic, S.J., and Maranas, C.D. 2004. FamClash: A method for ranking the activity of engineered enzymes. *Proceedings of the National Academy of Science, USA* **101:** 4142-4147.

Saraf, M.C., and Maranas, C.D. 2003. Using a residue clash map to functionally characterize protein recombination hybrids. *Protein Engineering* **16:** 1025-1034.

Savoie, A., Sanschagrin, F., Palzkill, T., Voyer, N., and Levesque, R.C. 2000. Structure-function analysis of alpha-helix H4 using PSE-4 as a model enzyme representative of class A beta-lactamases. *Protein Engineering* **13:** 267-274.

Schroeder, W.A., Locke, T.R., and Jensen, S.E. 2002. Resistance to beta-lactamase Inhibitor Protein Does Not Paralell Resistance to Clavulanic Acid in TEM beta-lactamase Mutants. *Antimicrobial Agents and Chemotherapy* **46:** 3568-3573.

Shimamura, T., Ibuka, A., Fushinobu, S., Wakagi, T., Ishiguro, M., Ishii, Y., and Matsuzawa, H. 2002. Acyl-intermediate structures of the extended-spectrum class A beta-lactamase, TOHO-1, in complex with cefotaxime, cephalothin and benzylpenicillin. *Journal of Biological Chemistry* **277:** 46601-46608.

Shimizu-Ibuka, A., Matsuzawa, H., and Sakai, H. 2004. An engineered disulfide bond between residues 69 and 238 in extended-spectrum beta-lactamase Toho-1 reduces its activity toward third-generation cephalosporins. *Biochemistry* **43:** 15737-15745.

Shindyalov, I.N., and Bourne, P.E. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering* **11:** 739-747.

Sideraki, V., Huang, W.Z., Palzkill, T., and Gilbert, H.F. 2001. A secondary drug resistance mutation of TEM-1 beta-lactamase that suppresses misfolding and aggregation. *Proceedings of The National Academy of Sciences USA* **98:** 283-288.

Sieber, V., Martinez, C.A., and Arnold, F.H. 2001. Libraries of hybrid proteins from distantly related sequences. *Nature Biotechnology* **19:** 456-460.

Solaiman, F., Zink, M.A., Xu, G., Grunkemeyer, J., Cosgrove, D., Saenz, J., and Hodgson, C.P. 2000. Modular retro-vectors for transgenic and therapeutic use. *Mol Reprod Dev* **56:** 309-315.

Steipe, B., Schiller, B., Pluckthun, A., and Steinbacher, S. 1994. Sequence Statistics Reliably Predict Stabilizing Mutations In A Protein Domain. *Journal of Molecular Biology* **240:** 188-192.

Stemmer, W.P.C. 1994. Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370:** 389-391.

Swarent, P., Maveyraud, L., Raquet, X., Cabantous, S., Duez, C., Pedalacq, J.-D., Mariotte-Boyer, S., Mourey, L., Labia, R., Nicolas-Chanoine, M.-H., et al. 1998. X-ray analysis of the NMC-A b-lactamase at 1.64 A Resolution, a Class A Carbapenemase with Broad Substrate Specificity. *Journal of Biological Chemistry* **273:** 26714-26721.

Taverna, D.M., and Goldstein, R.A. 2002. Why Are Proteins So Robust to Site Mutations? *Journal of Molecular Biology* **315:** 479-484.

Teo, J.W.P., Suwanto, A., and Poh, C.L. 2000. Novel beta-lactamase from two environmental isolates of *Vibrio harveyi*. *Antimicrobial Agents and Chemotherapy* **44:** 1309-1314.

Therrien, C., Sanschagrin, F., Palzkill, T., and Levesque, R.C. 1998. Roles of Amino Acids 161 to 179 in the PSE-4 omega loop in Substrate specificity and in resistance to ceftazidime. *Antimicrobial Agents and Chemotherapy* **42:** 2576-2583.

Thompson, J.D., Plewniak, F., and Poch, O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research* **27:** 2682-2690.

Tranier, S., Bouthors, A.T., Maveyraud, L., Guillet, V., Sougakoff, W., and Samama, J.P. 2000. The High Resolution Crystal Structure for Class A beta-lactamase Per-1 Reveals the Bases for its Increase in Breadth of Activity. *Journal of Biological Chemistry* **275:** 28075-28082.

Voigt, C.A., Kauffman, S., and Wang, Z.G. 2001a. Rational evolutionary design: The theory of *in vitro* protein evolution. In *Advances in Protein Chemistry, vol 55*, pp. 79-160.

Voigt, C.A., Martinez, C., Wang, Z.-G., Mayo, S.L., and Arnold, F.H. 2002. Protein Building blocks preserved by recombination. *Nature Structural Biology* **9:** 553-558.

Voigt, C.A., Mayo, S.L., Arnold, F.H., and Wang, Z.-G. 2001b. Computational method to reduce the search space for directed protein evolution. *Proceedings of the National Academy of Sciences, USA* **98:** 3778-3783.

Waldo, G.S., Standish, B.M., Berendzen, J., and Terwilliger, T.C. 1999. Rapid protein-folding assay using green fluorescent protein. *Nature Biotechnology* **17:** 691-695.

Wang, Q., Buckle, A.M., Foster, N.W., Johnson, C.M., and Fersht, A.R. 1999. Design of highly stable functional GroEL minichaperones. *Protein Science* **8:** 2186-2193.

Wang, X., Misasov, G., and Shoichet, B. 2002. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *Journal of Molecular Biology* **320:** 85-95.

Xia, Y., and Levitt, M. 2002. Roles of mutation and recombination in the evolution of protein thermodynamics. *Proceedings of the National Academy of Sciences, USA* **99:** 10382-19387.

Xia, Y., and Levitt, M. 2004. Funnel-Like Organization in Sequence Space Determines the Distributions of Protein Stability and Folding Rate Preferred by Evolution. *Proteins: Structure, Function and Bioinformatics* **55:** 107-114.

Zaccolo, M., and Gherardi, E. 1999. The effect of high-frequency random mutagenesis on *in vitro* protein evolution: A study on TEM-1 beta-lactamase. *Journal of Molecular Biology* **285:** 775-783.