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 x 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.

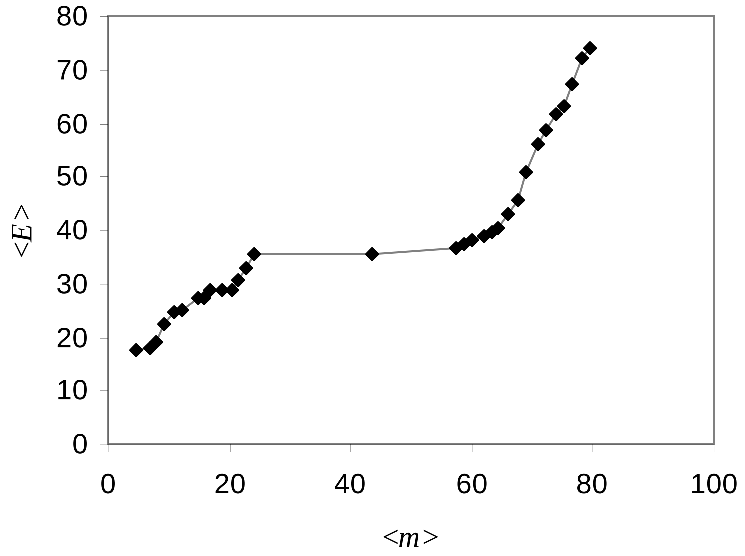

Figure III-1. A RASPP curve for nine crossover libraries made by recombining PSE-4, TEM-1 and SED-1. The minimum length constraint L=1, and the bin size set to 1 *m.*

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.

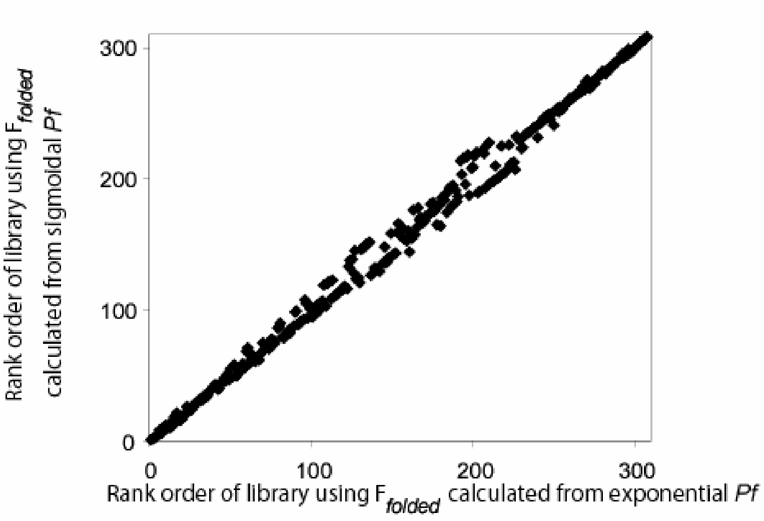

Figure III-2. The rank order of test libraries by F_{folded} using two different P_f functions (see methods) is strongly correlated indicating that rank ordering by F_{folded} is not sensitive to P_f.

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.

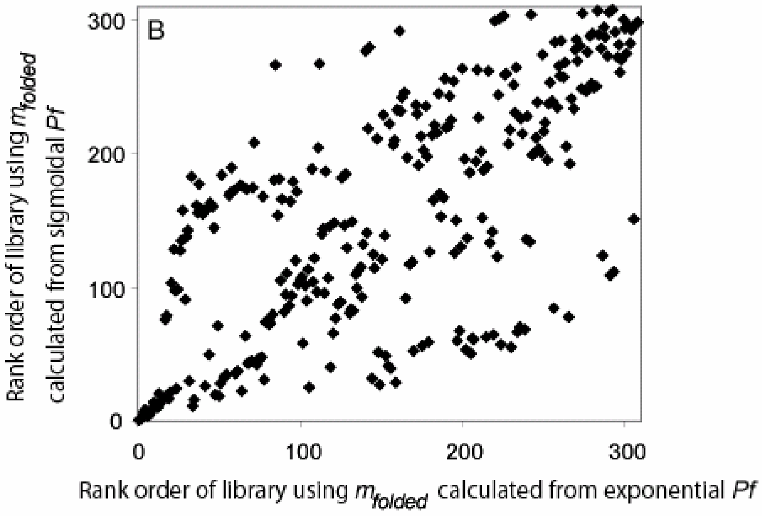

Figure III-3. The rank order of test libraries by m_{folded} using an exponential and a sigmoidal P_f function (see methods) shows that m_{folded} is very sensitive to $P_{f.}$ This indicates that m_{folded} is not likely a good measure to use for evaluating libraries if P_f is not known.

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 $<m/E>$. This measure effectively identifies libraries with the most mutations per disrupted contact. Examining plots of $<m/E>$ vs. $<m>$ and $<E>$ vs. $<m>$ for the test libraries described above shows that $<m/E>$ 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).

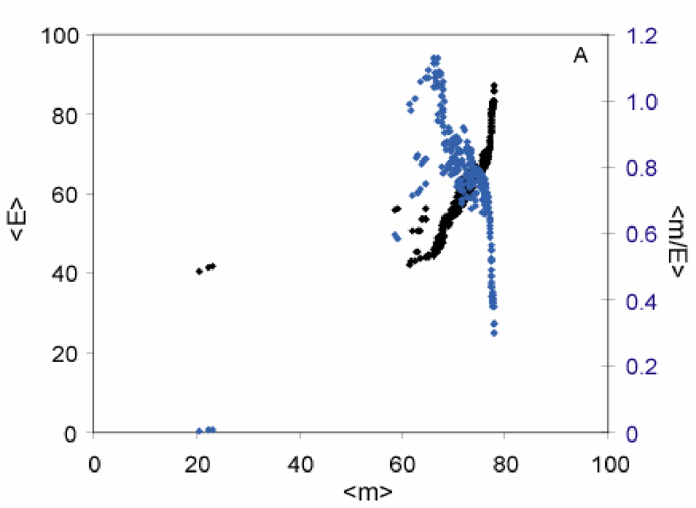

Figure III-4. Comparing <*E*> vs. <*m*> (black points) and <*m/E*> vs. <*m*> (blue points) shows that for RASPP test libraries there is a maximum <*m/E*> that corresponds to a plateau in the <*E*> vs. <*m*> curve.

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.

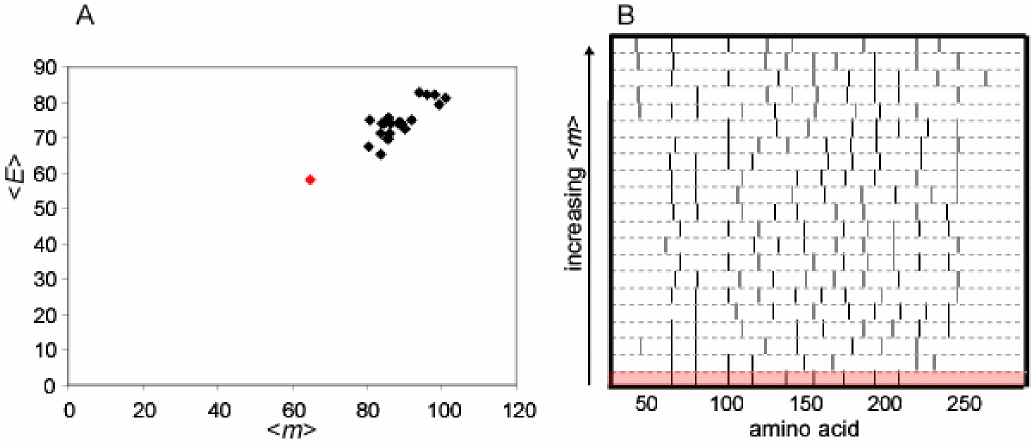

Figure III-5. <*E*> vs. <*m*> for the top 22 of 3,000 randomly enumerated libraries. The library shown in red was chosen for construction. B: The crossover locations for all the libraries shown to the right; the highlighted library was constructed. For all libraries the N- and C-termini together are 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= $(1- (f_dE/n))^n$, where n is total number of contacts (322), and f_d=0.075 (Meyer et al. 2003).

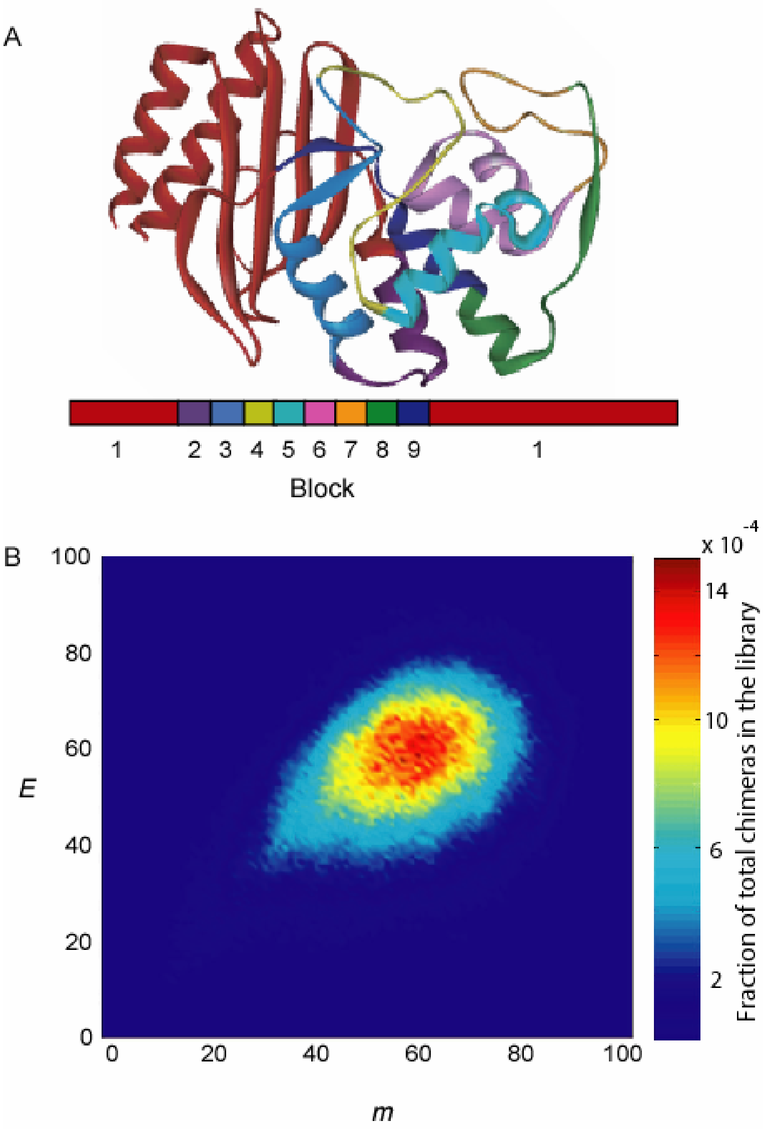

Figure III-6. An overview of the RandE:APST library. A: The differently colored sequence blocks are mapped to the structure of TEM-1. B: The *E* vs. *m* density plot of chimeras in the library shows a single large peak in the population with a slight tail toward low *E* and *m*.

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.

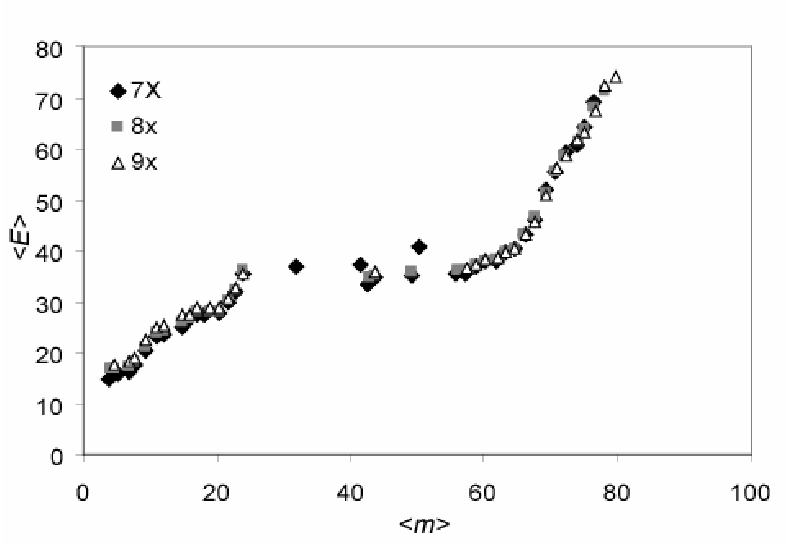

Figure III-7. RASPP curves generated for TEM-1, PSE-4, and SED-1 using 7, 8, or 9 crossovers show that there is no gain in fraction folded at a given level of diversity associated with 8 or 9 crossovers vs. 7.

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.

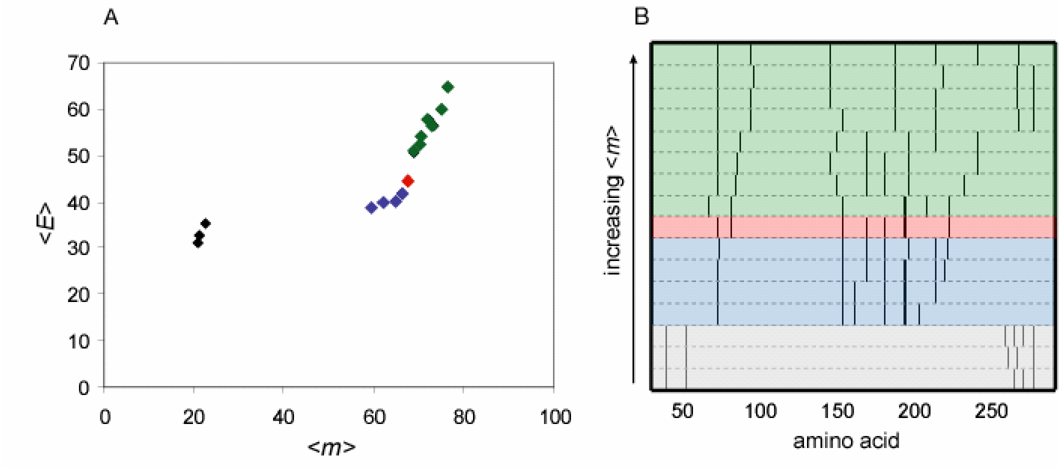

Figure III-8. A: The <*E*> vs. <*m*> RASPP curve generated for TEM-1, PSE-4 and SED-1 using seven crossovers. The libraries break into three regions that are colored black, blue and green. The red point represents the library chosen for construction. B: The crossover locations for the RASPP libraries 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.

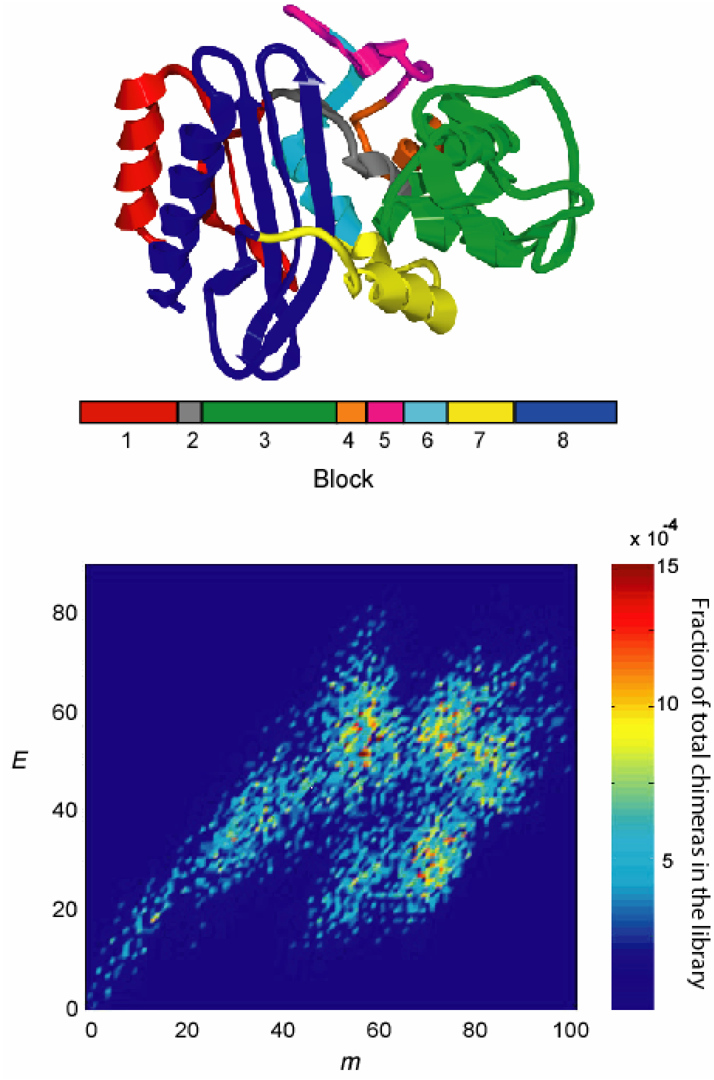

Figure III-9. RASPP:PST library chosen for construction. A: The differently colored sequence blocks are mapped to the structure of TEM-1. B: The *E* vs. *m* plot of chimeras in the library shows a relatively diffuse population which has bimodal properties in both the *E* and the *m* dimensions.

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 <*E*>, a higher F_{folded}, and a higher <*m*> 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 <*E*>, which RASPP uses as its minimization criterion is a good surrogate for the fraction of folded variants. The binning of RASPP libraries by <*m*> 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. $\Delta_{\mathrm{ij}} = 0$ if amino acids i and j in the chimera are found together at the same positions in any parental protein sequence, otherwise $\Delta_{\mathrm{ij}} = 1$. 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 $(1- (f_d E/n))^n$, 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/.