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.

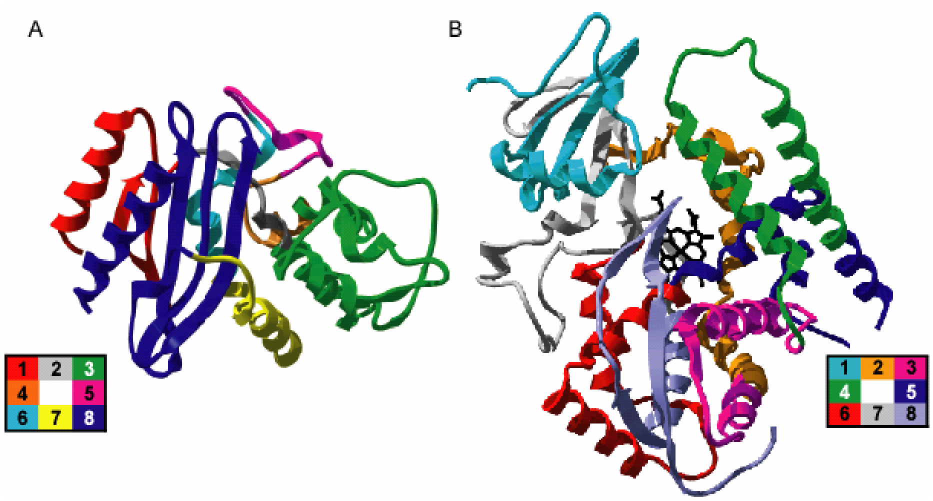

Figure VII-1. The three dimensional structures of A: β-lactamase chimera parent proteins (TEM-1, 1BTL) and B: cytochrome P450 parent proteins (CYP102A1, 1JPZ) with the independently exchangeable sequence blocks mapped to the structures. For lactamases the crossovers are after the following TEM-1 residues: 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

$$P_f = \frac{1}{c + e^{bE+a}}, \tag{VII-1}$$

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).

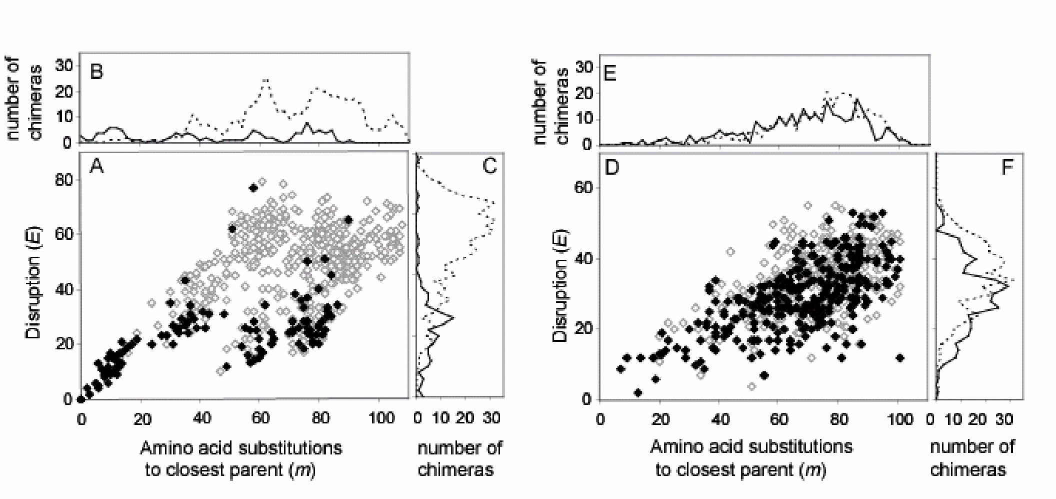

Figure VII-2. The *E* and *m* distributions for β-lactamase (A-C) and cytochrome P450 (D-F) chimeras. A, D: *E* vs. *m*, for unfolded chimeras (open points) and folded chimeras (solid points). B, E: Distribution of folded (solid line) and unfolded (dashed line) chimeras with respect to *E*. C, F: Distribution 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).

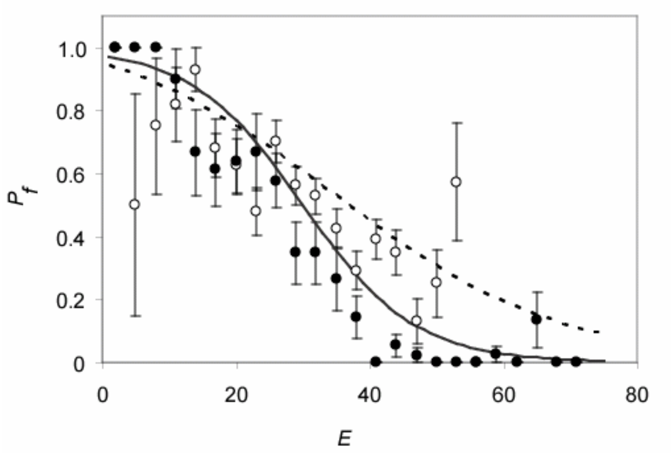

Figure VII-3. P_f *(E)* for lactamase chimeras (solid points, solid line) and cytochrome P450 chimeras (open points, dashed line). The points represent the fraction of folded chimeras in bins of 3 *E*. Curves represent the best fit of chimera folding data to Equation (VII-1). For lactamases *a*= 3.6,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

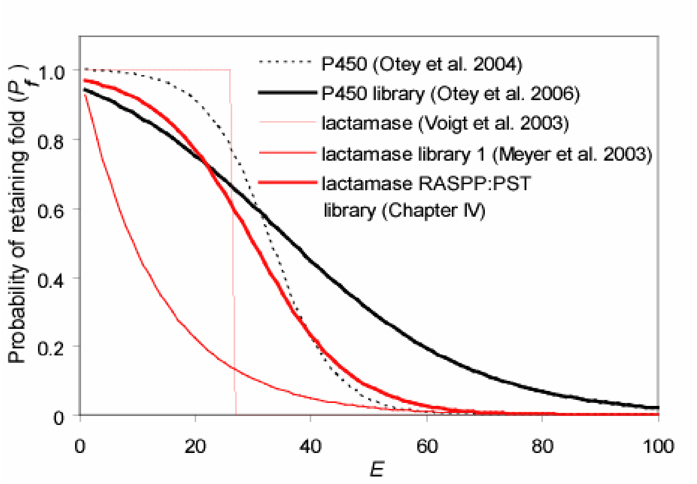

Figure VII-4. Different P_f functions can be calculated using various chimera data sets. All curves except those from (Meyer et al. 2003) and (Voigt et al. 2002) were fit to P_f in the form described in Equation (VII-1). The curves from Meyer et al. (2003) and Voigt et al. (2002) were included as described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

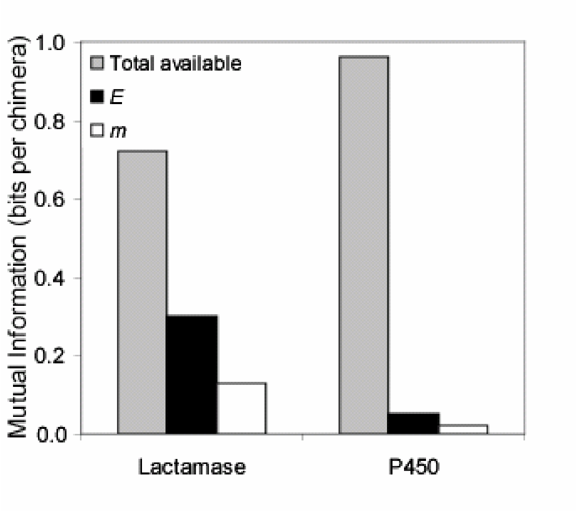

Figure VII-5. The total available mutual information (bits/chimera) for lactamase and cytochrome P450 data sets and the mutual information between folding and *E* and folding and *m* for each data set.

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.

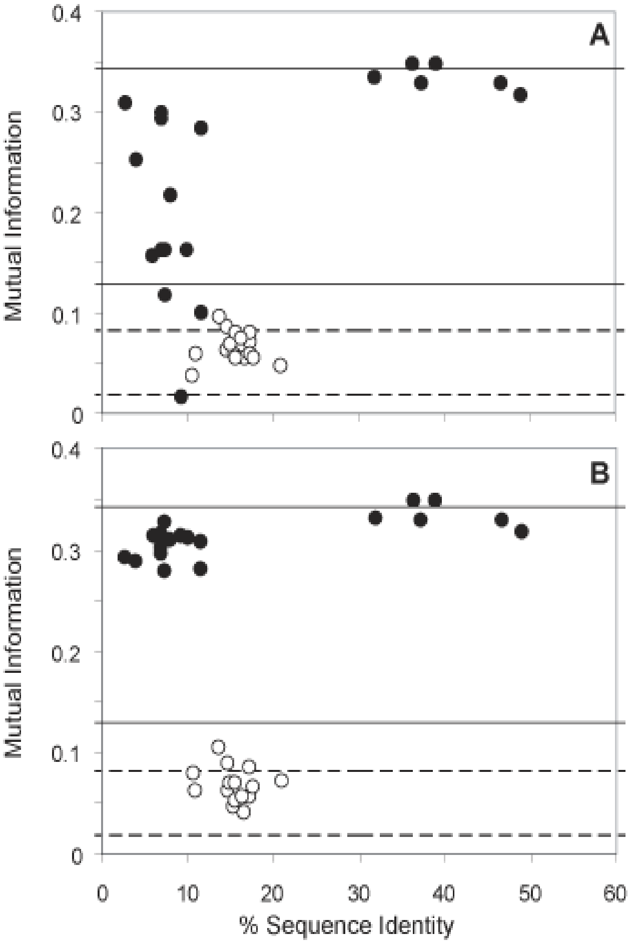

Figure VII-6. The mutual information (bits per chimera) between protein folding and *E* calculated using structural information from proteins homologous to the proteins recombined (Table VII-2). Solid points represent β-lactamases and open points cytochromes P450, the solid lines represent the mutual 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.

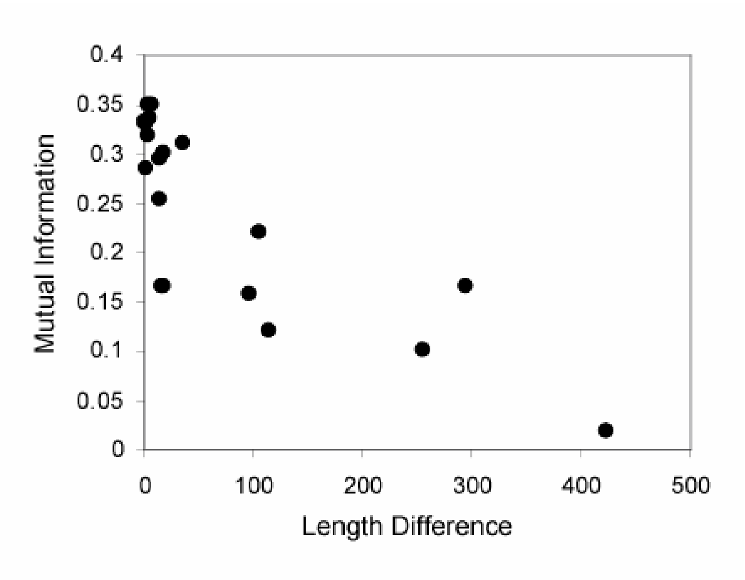

Figure VII-7. The mutual information between lactamase chimera folding and *E* calculated using a range of homologous structures. The sequences were aligned using CLUSTALW. As the length difference between the proteins recombined and the protein in the structure increases, there is a decline in the mutual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

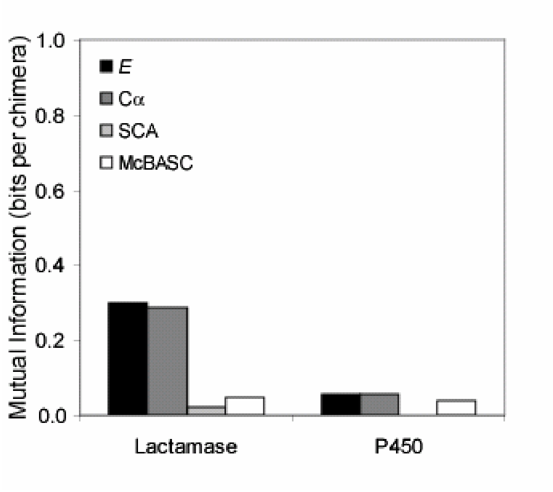

Figure VII-8. The mutual information for chimera folding and *E* calculated using standard contact maps where residues with any heavy atoms are within 4.5 Å (Voigt et al. 2002), Cα maps where residues with Cα within 8Å are contacting, and contact maps determined from the highest scoring 1% of covarying 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.