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.

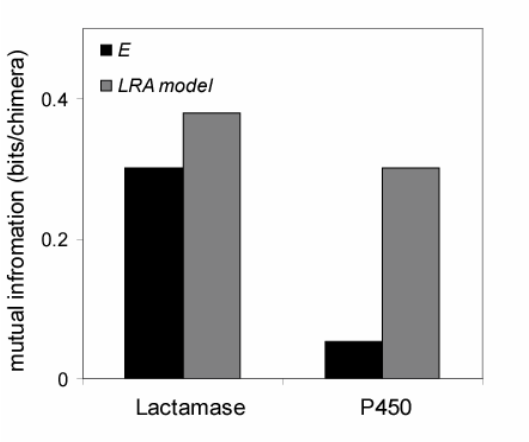

Figure VIII-1. The mutual information for chimera folding and different energy functions. *E* is the standard SCHEMA disruption that considers pairwise disruption. The LRA models are derived from the data and have both two-body (pairwise) and one-body contributions.

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).

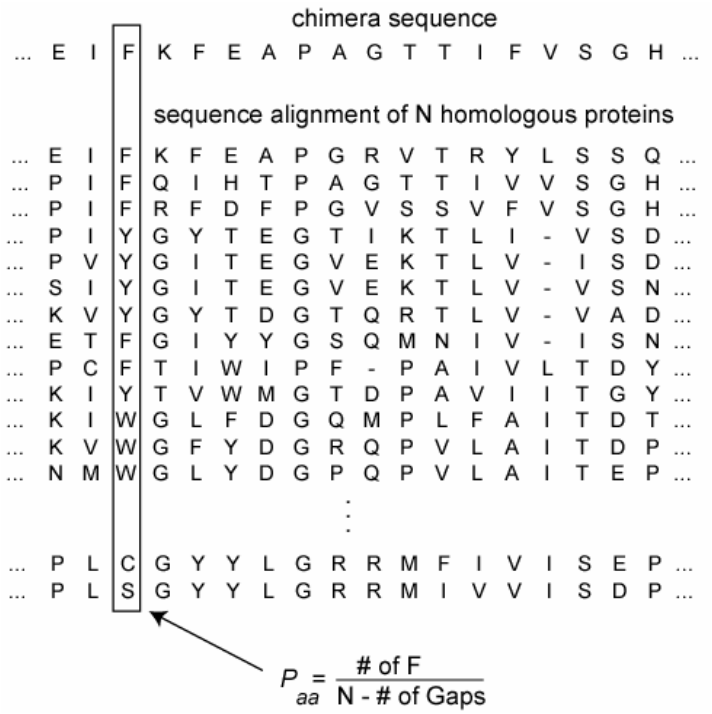

Figure VIII-2. The probability of finding the amino acid in the chimera in the multiple sequence alignment (P_{aa}) is determined by determining the frequency of the amino acid at the position and dividing by the total number of sequences not including sequences with gaps at the position.

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).

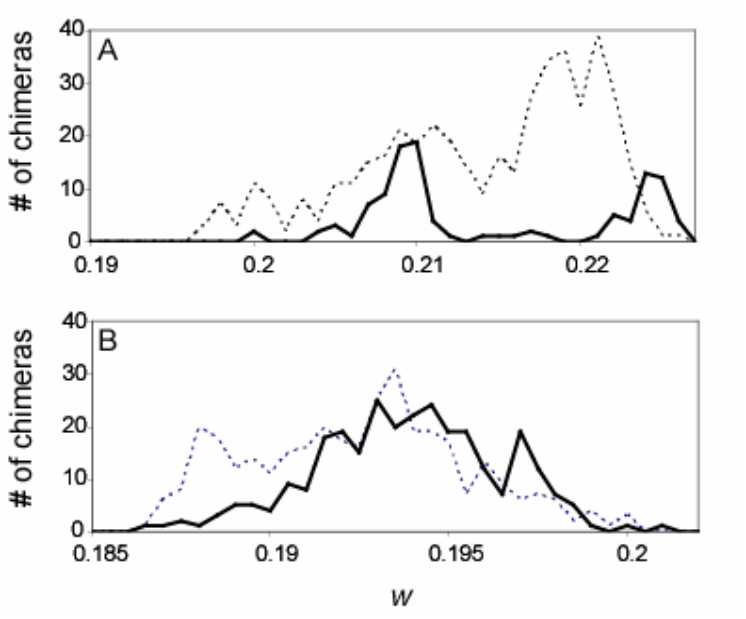

Figure VIII-3. Distribution of folded (solid line) and unfolded (dashed line) chimeras with respect to *w* shows that chimeras with low *w* are less likely to function in both A: β-lactamase chimeras and B: cytochrome P450 chimeras.

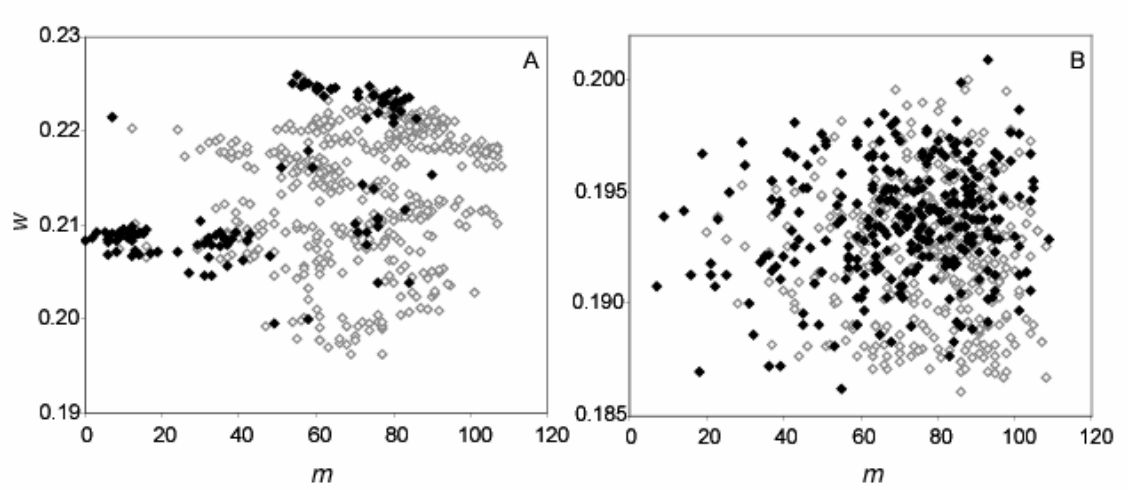

Figure VIII-4. *w* vs. *m* of A: lactamase chimeras and B: cytochrome P450 chimeras. Open points represent unfolded chimeras and closed points represent folded chimeras. For lactamases the parent *w* values are: PSE-4 *w* = 0.201, SED-1 *w* = 0.227, TEM-1 *w* = 0.213. For cytochromes P450 the parent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

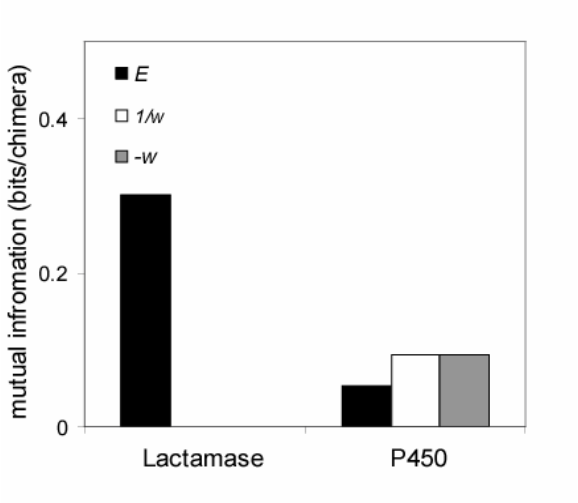

Figure VIII-5. Mutual information between chimera folding and E, $1/w$ and $-w$. Both $-w$ and $1/w$ are more predictive of chimera folding than E for cytochromes P450. However, for lactamases neither has significant predictive power.

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).

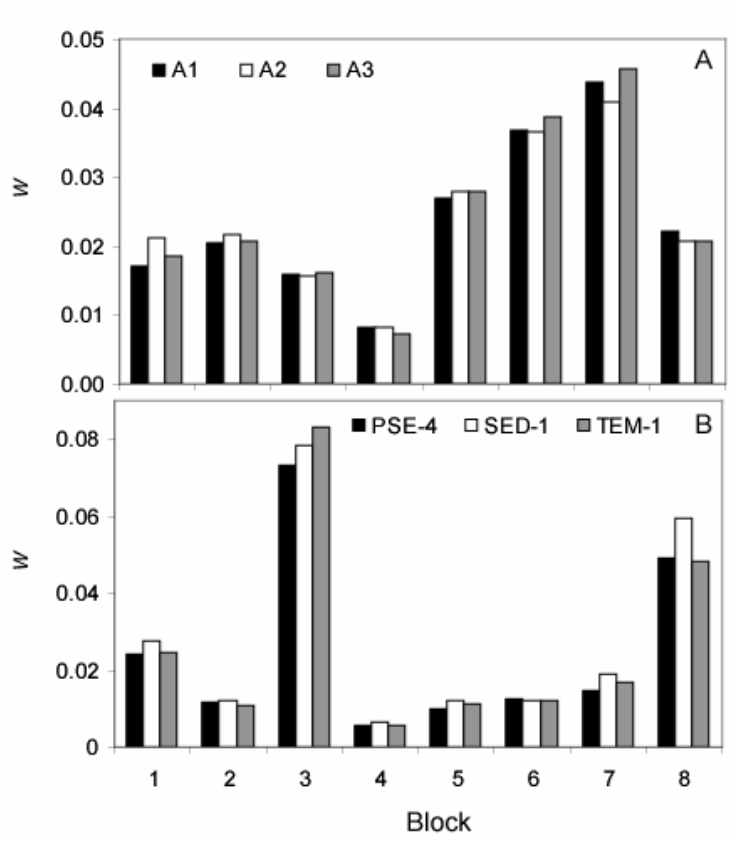

Figure VIII-6. The one-body weighting term w_{block} determined for each exchangeable block of sequence in the A: cytochrome P450 and B: β-lactamase libraries. Parents with higher w_{block} block values are more likely to be found in functional chimeras than parents with low w_{block} values.

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.

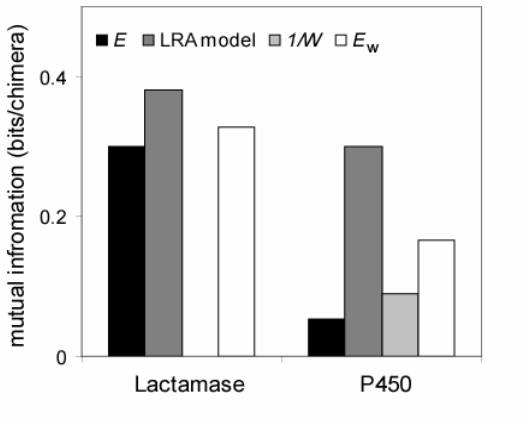

Figure VIII-7. The mutual information between folding and various predictive energy functions. *E*, *1/w* and E_w can be calculated *a priori* but the LGA energy was calculated directly from the chimera folding data and represents the best that a model incorporating one- and two-body terms can predict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

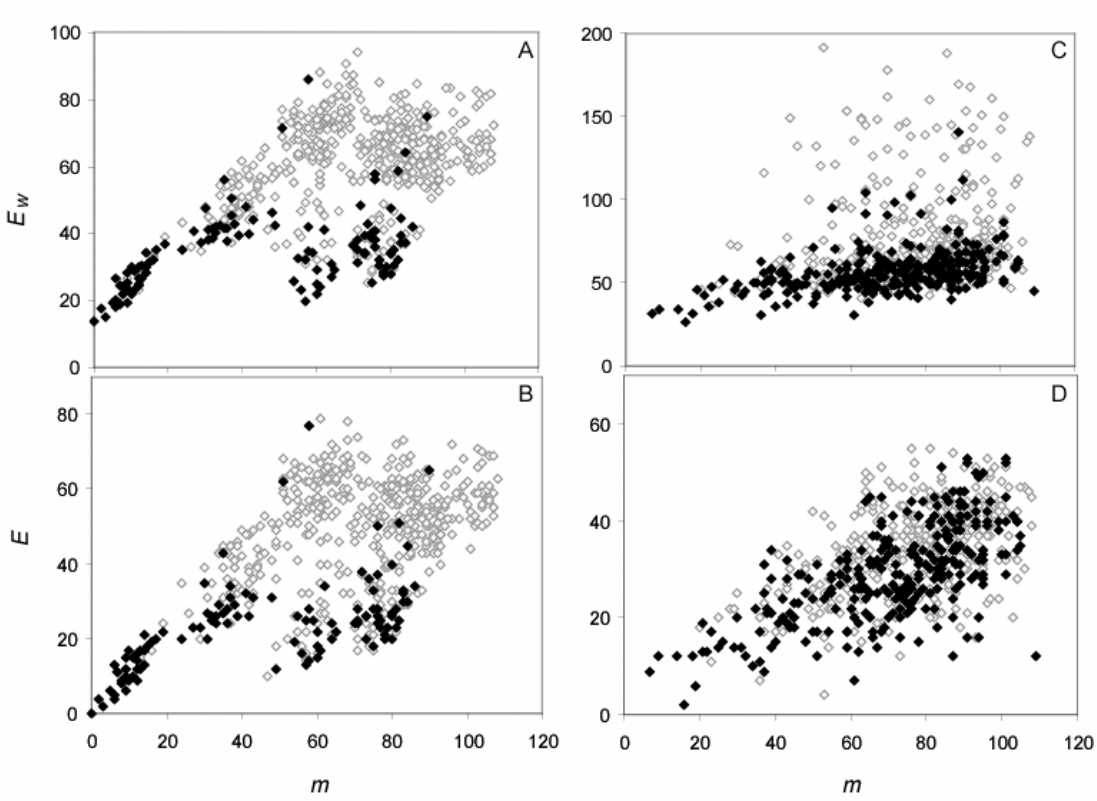

Figure VIII-8. E vs. m and E_w vs. m for lactamase (A, B) and cytochrome P450 (C, D) folded (solid point) and unfolded (open point) chimeras.

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.