DESIGN AND ANALYSIS OF COMBINATORIAL PROTEIN LIBRARIES CREATED BY SITE-DIRECTED RECOMBINATION

Thesis by

Jeffrey B. Endelman

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

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California
2005
(Defended May 12, 2005)

ABSTRACT

For many protein design problems, limited understanding of the relationship between sequence and function necessitates searching through a library of proteins to find the properties of interest. To accelerate this process, molecular models and optimization algorithms can be combined to design diverse libraries enriched in folded proteins. I apply this strategy to site-directed recombination, in which an alignment of p homologs is partitioned into f blocks, and the resulting gene fragments are combinatorially assembled to create a library with p^f chimeric sequences. To design the fragments, I present a dynamic programming algorithm that minimizes the average energy of the library, subject to constraints on fragment length. This algorithm works for any pairwise residue potential, several of which are compared for their ability to predict which chimeras retain the parental function and/or fold. The alignments of folded and unfolded chimeras are used to generate sequence-function relationships via logistic regression, a technique for fitting models to binary data. Compared to methods developed for alignments of naturally occurring proteins, logistic regression more readily distinguishes true interactions from correlations between strongly stabilizing but non-interacting residues.

ACKNOWLEDGMENTS

I would like to begin by thanking my thesis advisor Frances Arnold. Her ability to maintain an open door policy and quickly provide feedback on manuscripts is remarkable for a professor managing 20 people. She has challenged me to think critically and communicate effectively. From proposals to progress reports, I have learned a lot from Frances about how research is funded. I appreciate the freedom she has given me to dictate the direction of my thesis research.

I am also extremely thankful for the many wonderful students Frances has attracted to her research group. Chris Otey and Michelle Meyer have been great collaborators and kind enough to share their data with me. Without them this thesis would not have been possible! Joff Silberg was a terrific mentor during my first year at Caltech, and he was even patient enough to teach me some molecular biology. Jesse Bloom and Allan Drummond have been my sounding board for ideas and have suggested several of the approaches used in this thesis. Working with Marco Landwehr has been a pleasure. The Arnold group has been much more than a professional home for me. They have been my friends and support group through the ups and downs of graduate school.

There are several other people at Caltech I would like to thank. Steve Mayo, Niles Pierce, and Zhen-Gang Wang served on my candidacy and thesis committees, and Zhen-Gang has been a great resource for resolving theoretical issues. I need to thank Eric Zollars for our collaboration and Tom Treynor for our thoughtful conversations. I had the pleasure of working with Richard Murray, Christina Smolke, and Michael Elowitz

during the Synthetic Biology Competition of summer 2004. It was a lot of fun, and I appreciate the respect and responsibility I was given. To my fellow bioengineering students: what a trying first year we had! Thanks to Kevin McHale, Gwyneth Card, and many others for being there to commiserate and celebrate.

I came to Caltech after two years of grad school at UCSB, and I have many people to thank from that time, including Mark Henle, Eric Dunham, Matt Hansen, Grace Brannigan, Lawrence Lin, Jesse Epstein, Vanessa Hayward, and Rich Steinberg. I am grateful for the support of my advisor Jean Carlson during my transformation from physicist to bioengineer. With her help I traveled to Budapest and Aspen to learn about biological complexity. I'd like to thank John Doyle for his contagious enthusiasm and for inspiration.

I was fortunate to have several great mentors as an undergraduate student at Northwestern University. Thanks to Randy Snurr and John Torkelson in particular for launching me on a research career.

I'd like to acknowledge a few special individuals in southern California who helped me balance academia with life. I'll never forget wild buckwheat biscuits along the Arroyo Seco with Christopher Nyerges. To the Dervaes family: thanks for showing me the path to freedom. The folks at Tierra Miguel Farm have nourished my mind and body, and I'm grateful to Brian Brown for sharing with me his enthusiasm and vision for sustainability. Most of all, I want to thank Sarah Spivack for her love and for keeping me afloat during the tough times, especially my transition from UCSB to Caltech.

I am blessed to have such a loving family. It is comforting to know that any time of day I can pick up the phone and call my mom toll-free. Over the years I have turned to my dad for advice on so many issues. Respect and admiration from my brother are more precious to me than from anyone else. I am thankful for their unconditional support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I

The tradeoff between folding and diversity with recombination

Protein design seeks the amino acid sequence that encodes a protein with a desired set of properties (DeGrado, 2001). It is the inverse of the protein folding problem, in which one seeks the fold and function of a given amino acid sequence. One class of strategies for protein design involves searching through a library of proteins for the properties of interest (Arnold, 2000). These libraries are often created by altering the sequence of a parental protein whose features make it a good starting point for the fitness search. When the parental protein is mutated randomly, the fraction of functional mutants declines exponentially with the number of mutations (Daugherty *et al.*, 2000; Guo *et al.*, 2004). The fraction functional after one amino acid mutation, or neutrality, can vary from 0.35 to 0.55 depending on the protein structure (Bloom *et al.*, 2005).

When many amino acid changes are desired, random mutagenesis is too deleterious to be effective. For these design problems, recombination of homologous sequences is often appropriate because the mutations involve amino acids previously selected by nature to be compatible with the protein fold, albeit in a different genetic background (Stemmer, 1994; Crameri *et al.*, 1998; Ostermeier *et al.*, 1999; Stevenson & Benkovic, 2002). The conservative nature of recombination relative to random mutation was demonstrated clearly by Drummond *et al.* (Drummond *et al.*, 2005) with the beta-lactamase proteins TEM-1 (Jelsch *et al.*, 1993) and PSE-4 (Lim *et al.*, 2001). Using error-prone PCR to randomly mutate PSE-4, they estimated its neutrality to be 0.54. The

red line in Figure I-1 shows the fraction of functional proteins (F_f) expected with random mutation under the exponential relationship $F_f = 0.54^m$, where m is the number of amino acid changes relative to PSE-4. The recombination data of Drummond *et al.*, generated by selecting 32 unique functional chimeras (including the parents) from a library of 16,384 sequences, follow a very different trend (black squares, Figure I-1). The fraction functional (calculated by assuming chimeras not isolated by the selection are not functional) is roughly symmetric about the midpoint $m = D/2$, where $D = 151$ is the number of amino acid differences between TEM-1 and PSE-4. At low levels of mutation relative to either parent, the fraction functional (F_f) decreases exponentially, but by $m = D/2$ the slope is zero. Functional beta-lactamases at this midpoint, which are maximally different from TEM-1 and PSE-4, are at least 16 orders of magnitude more common in the recombination library than in the one created by random mutation.

Drummond *et al.* have proposed a simple model consistent with these data for the probability $p(F|m)$ that a chimera with m mutations will fold (Drummond *et al.*, 2005). For two parents that differ at D residues, each of the m amino acids from one parent makes $D - m$ pairs with residues from the other parent. Because these "novel" pairs are untested by nature within the context of the two parents, each one is assumed to have some probability $1 - q$ of disrupting folding. All other pairs are assumed to be nondisruptive. If each novel pair acts independently, the probability of retaining the parental fold is

$$p(F|m) = q^{m(D-m)} \equiv \rho^{m(D-m)/(D-1)}, \qquad \text{[I-1]}$$

which defines the recombinational tolerance ρ as the probability of folding for chimeras with one mutation. When $m = 0$ and $m = D$, the parents are recovered and the probability of folding is unity as required. The dotted blue line in Figure I-1, which is based on a maximum likelihood estimate of $\rho = 0.82$, shows that Equation I-1 describes the data fairly well. The main limitation of this model is the assumption that all novel residue pairs act equally. For a particular set of parents, it is acceptable to treat q as an average over the structural context of the residue pairs, but this probability will be different for other parents. To use the TEM-1/PSE-4 recombination data to predict folding with other beta-lactamase parents, a different model is needed.

Borrowing heavily from Voigt *et al.* (Voigt *et al.*, 2002) and Drummond (D.A. Drummond, personal communication), I propose using novel residue-residue contacts, which are more likely to be transferable across parents. The total number of novel residue-residue contacts, called "SCHEMA disruption" by Voigt *et al.*, can be written as a pairwise sum over residues (Silberg *et al.*, 2004):

$$E = \sum_{i} \sum_{j>i} C_{ij} \Delta_{ij} \ . \qquad \text{[I-2]}$$

The contact matrix C_{ij} depends solely on structural information, while Δ_{ij} uses only the parental sequence alignment. Voigt *et al*. defined $C_{ij} = 1$ if residues i and j are within 4.5 Å in the parental structure; otherwise $C_{ij} = 0$. The delta function $\Delta_{ij} = 0$ if the amino acids found in the chimera at positions i and j are also found together in any single parent at homologous positions. Otherwise, the i-j interaction is considered broken and $\Delta_{ij} = 1$. Although Equation I-2 is hardly an energy function in the thermodynamic sense, it can be

used to score the compatibility of a sequence with a target fold. For this reason I refer to Equation I-2 as the "SCHEMA energy," which also helps to distinguish it from the SCHEMA algorithm discussed in Chapter II.

If all non-contacting residues are assumed compatible, and if each novel residue contact acts independently with probability $1 - \pi$ to disrupt folding, the probability of folding decreases exponentially with SCHEMA energy (Meyer *et al.*, 2003):

$$p(F|E) = \pi^E. \qquad \text{[I-3]}$$

Figure I-2 is a plot of the fraction functional for TEM-1/PSE-4 chimeras vs. their SCHEMA energy. Some bins contain no functional chimeras, but the 90% confidence limits still appear on the graph. The dotted line, which is based on a maximum likelihood estimate of $\pi = 0.89$, shows that Equation I-3 describes the data reasonably well.

The probability of folding with respect to SCHEMA can be used to calculate the probability of folding with respect to mutation by conditioning:

$$p(F|m) = \sum_E p(F|E)p(E|m). \qquad \text{[I-4]}$$

$p(E|m)$ is the probability that a chimera with m mutations has E novel residue contacts. Rather than evaluate Equation I-4 directly, I use Jensen's inequality (Boyd & Vandenberghe, 2004) to derive a more readily calculated lower bound on $p(F|m)$, which holds as long as $p(F|E)$ is convex:

$$p(F|m) \geq p\left(F\left|\sum_E p(E|m)E\right.\right) = p\left(F\left|\langle E\rangle_m\right.\right). \qquad \text{[I-5]}$$

For the exponential model in Equation I-3, Equation I-5 becomes

$$p(F|m) \geq \pi^{\langle E \rangle_m}. \qquad \text{[I-6]}$$

$\langle E \rangle_m$, which is the average SCHEMA energy among chimeras with mutation level *m*, can be calculated by multiplying the number of novel residue pairs by the probability one makes contact, denoted χ. For two parents, there are $D(D-1)$ residue pairs between the D residues at which the parents differ, and $m(D-m)$ of these are novel for a chimera with m mutations. If the total number of contacts between the $D(D-1)$ pairs is C, then $\chi = C/(D(D-1))$, $\langle E \rangle_m = \chi(D-m)m$, and

$$p(F|m) \geq \left(\pi^{\chi}\right)^{m(D-m)} = \left(\pi^{C/D}\right)^{m(D-m)/(D-1)}. \qquad \text{[I-7]}$$

The ratio C/D is the average number of contacts per mutation, and thus $\pi^{C/D}$ is the probability that a mutation with this many contacts does not disrupt folding. Equations I-1 and I-7 together imply the recombinational tolerance $\rho \geq \pi^{C/D}$.

How conservative is this lower bound? There are C = 322 SCHEMA contacts among the D = 151 residues at which TEM-1 and PSE-4 differ, which yields $\pi^{C/D}$ = 0.78. This value is indeed lower than the maximum likelihood estimate of 0.82 for the recombinational tolerance. The solid blue curve in Figure I-1 shows that the lower bound in Equation I-6 is a fairly good approximation to Equation I-1 (dashed blue). The largest gap between the two models occurs at the midpoint between TEM-1 and PSE-4, where the SCHEMA bound underestimates the fraction folded by about one order of magnitude.

The data of Drummond *et al.* can be used with Equation I-6 to compare the tradeoff between folding and diversity for different parents. In principle any number of parents can be simulated, but counting the number of novel residue pairs quickly

becomes prohibitive. With three parents the combinatorics are tedious but manageable (see Methods). Instead of a single mutational distance, five degrees of freedom are needed to determine the average SCHEMA energy. These five variables can be averaged to give a three-dimensional label (m_1, m_2, m_3) describing the number of mutations relative to each parent. Figure I-3, which shows the probability of folding for four slices through (m_1, m_2, m_3)-space, is the analog of Figure I-1 when TEM-1 is partnered with PSE-4 and SED-1 (Petrella *et al.*, 2001). The top panel shows the plane containing all three parents, which lie at the corners of the triangle. These three beta-lactamases have roughly the same pairwise sequence identities (~40%), giving rise to the approximate threefold symmetry. The edges of the ternary diagram represent the least deleterious paths from one parent to another, along which the fraction folded remains above 10^{-4}. Chimeras at the center of the triangle have 107 mutations to the closest parent. The next three panels show slices through (m_1, m_2, m_3)-space at successively higher values of mutation, and hence the probability of folding decreases.

To more clearly visualize the tradeoff between folding and diversity for these three parents, Figure I-4 plots the probability of folding vs. mutation *to the closest parent,* rather than to a fixed parent. The solid curve represents TEM-1/PSE-4/SED-1, while the dashed curve is a reprint of the TEM-1/PSE-4 curve shown in solid blue in Figure I-1. The difference in how mutation is measured compared to Figure I-1 explains why the two-parent curve is only a half-parabola, terminating at the maximum mutation level of $D/2$. At low levels of mutation the tradeoff with three parents is similar to the two-parent case. As the mutation level increases, the probability of folding with three parents falls

below that for two, and much higher levels of mutation are possible with three parents. There is an interesting kink in the three-parent curve around 110 mutations, for which there is no analog with two parents. To understand this phenomenon, consider the group of residues at which all three parents have different amino acids (group *D* in Methods). At mutation levels below the kink, most of the chimeras inherit these maximally unconserved residues predominantly from two of the three parents. To reach mutation levels above the kink, however, chimeras must inherit the maximally unconserved residues from all three parents. It is precisely because these residues are different in all three parents that they create the most novel residue pairs. As the number of parents increases, I expect more of these transitions in the folding-diversity tradeoff curve.

Despite the conservative nature of recombination, highly mutated and folded chimeras may still be too rare to find when created randomly. According to Figure I-4, at least 10,000 TEM-1/PSE-4/SED-1 chimeras with 50 mutations must be checked before even one folded protein is expected. The odds of success can be improved by using computational methods to choose specific crossovers less likely to disrupt folding—a strategy called site-directed recombination (SDR). As illustrated in Figure I-5, with SDR an alignment of p homologs is partitioned into f blocks, and the resulting gene fragments are combinatorially assembled to create a library with p^f chimeric sequences (Hiraga & Arnold, 2003). This partitioning is equivalent to choosing $f - 1$ crossovers in the parental sequence alignment. Folded, site-directed chimeras with 50 mutations can be orders of magnitude more common than 1 in 10^4 for well-designed libraries, as I show in Chapter II.

Methods

To compute the average SCHEMA energy for chimeras derived from three parents, one must consider four different groups of residues. At the positions in group *A*, parents 2 and 3 have the same amino acid but parent 1 is different. At the positions in group *B*, parents 1 and 3 are the same but parent 2 is different, and in group *C* parent 3 is different from parents 1 and 2. Group *D* includes residues where all three parents are different. Let *a*, *b*, and *c* denote the number of residues in groups *A*, *B*, and *C*, respectively, that belong to the unique parent. The labels d_1, d_2, and d_3 indicate the number of residues from parents 1, 2, and 3, respectively, in group *D*.

Table I-1, which shows the number of novel residue pairs between each group, was constructed using the same logic as the two-parent case. For example, each of the *a* residues inherited from parent 1 in group *A* makes $A - a$ novel residue pairs with the other residues in that group, for a total of $a(A - a)$. To simplify the expressions in Table I-1, I have employed the complement overbar, e.g., $\bar{a} = A - a$. The below-diagonal entries in Table I-1 were left blank to prevent double counting. Whereas with two parents a single variable *m* is sufficient to specify the number of novel residue pairs, five variables (a, b, c, d_1, d_2) are needed with three parents. The number of mutations relative to parent 1 is $m_1 = \bar{a} + b + c + \bar{d}_1$, and the corresponding expressions for parents 2 and 3 are $m_2 = a + \bar{b} + c + \bar{d}_2$ and $m_3 = a + b + \bar{c} + \bar{d}_3$. If $n(a, b, c, d_1, d_2)$ denotes the number of novel residue pairs (the sum of the entries in Table I-1), then the average SCHEMA energy among chimeras at a particular point (m_1, m_2, m_3) is

$$\langle E \rangle_{(m_1,m_2,m_3)} = \frac{\sum_{a}\sum_{b}\sum_{c}\sum_{d_1}\sum_{\substack{d_2 \\ s.t.\ m_1=\bar{a}+b+c+\bar{d}_1 \\ m_2=a+\bar{b}+c+\bar{d}_2 \\ m_3=a+b+\bar{c}+\bar{d}_3}} \frac{A!}{a!\bar{a}!}\frac{B!}{b!\bar{b}!}\frac{C!}{c!\bar{c}!}\frac{D!}{d_1!d_2!d_3!}\,\chi\, n(a,b,c,d_1,d_2)}{\sum_{a}\sum_{b}\sum_{c}\sum_{d_1}\sum_{\substack{d_2 \\ s.t.\ m_1=\bar{a}+b+c+\bar{d}_1 \\ m_2=a+\bar{b}+c+\bar{d}_2 \\ m_3=a+b+\bar{c}+\bar{d}_3}} \frac{A!}{a!\bar{a}!}\frac{B!}{b!\bar{b}!}\frac{C!}{c!\bar{c}!}\frac{D!}{d_1!d_2!d_3!}} \quad .[I\text{-}8]$$

One could use a different contact probability for each entry in Table I-1, but I lumped all residue pairs together with a single χ. The beta-lactamases PSE-4 (parent 1), SED-1 (parent 2), and TEM-1 (parent 3) are characterized by A = 33, B = 46, C = 30, and D = 87, for a total of 196 amino acids, between which there are 538 contacts. The probability that a novel residue pair is in contact is thus χ = 538/(196×195) = 0.014. In practice Equation I-8 was evaluated without the constraints, and each term was simply added to the appropriate (m_1, m_2, m_3) bin. A similar procedure was used to evaluate the average SCHEMA energy at a fixed value of mutation relative to the closest parent, as in Figure I-4.

To generate Figure I-3, points in the (m_1, m_2, m_3) basis were first transformed to a new basis. In the old basis, parent 1 is located at (0, $A + B + D$, $A + C + D$), while in the new basis its position is (1, 0, 0). Parent 2 was transformed from ($A + B + D$, 0, $B + C + D$) to (0, 1, 0) and parent 3 from ($A + C + D$, $B + C + D$, 0) to (0, 0, 1). The slices shown in Figure I-3 are thus perpendicular to the (1, 1, 1) direction, at increasing distances from the origin.

Table I-1. Counting the number of novel residue pairs with three parents.

	Group *A*	Group *B*	Group *C*	Group *D*
Group *A*	$a\bar{a}$	ab	ac	$a\bar{d}_1 + \bar{a}d_1$
Group *B*		$b\bar{b}$	bc	$b\bar{d}_2 + \bar{b}d_2$
Group *C*			$c\bar{c}$	$c\bar{d}_3 + \bar{c}d_3$
Group *D*				$d_1\bar{d}_1 + d_2d_3$

Figure I-1. Comparing the fraction of functional proteins (F_f) by recombination vs. random mutation of PSE-4. Based on a structural alignment generated with Swiss-Pdb Viewer (Guex & Peitsch, 1997), PSE-4 and TEM-1 differ at 151 of the 263 amino acids in TEM-1. The red line, which was fit to error-prone PCR data by Drummond *et al.* (Drummond *et al.*, 2005), shows the exponential decline with random mutation ($F_f = 0.54^m$, where m is the number of amino acid mutations to PSE-4). The solid squares are the fraction of functional chimeras by recombination in ten evenly spaced bins (x-error bars). Each y-error bar is a 90% confidence interval based on the binomial distribution. The dashed blue line is a maximum likelihood fit of Equation I-1 to the recombination data (ρ = 0.82), and the solid blue line is the lower bound on folding derived with SCHEMA (Equation I-7, π = 0.89).

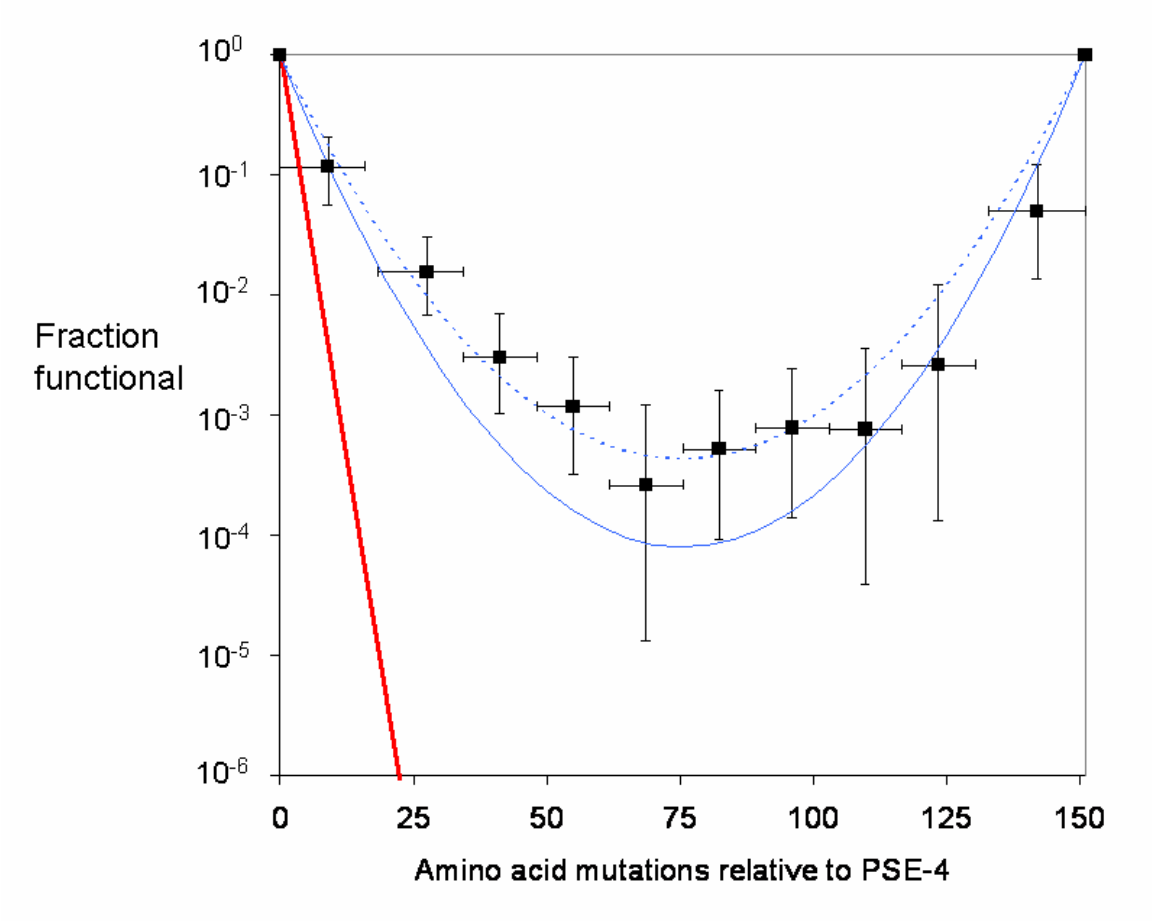

Fraction functional
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
0
25
50
75
100
125
150
Amino acid mutations relative to PSE-4

Figure I-2. The fraction of functional TEM-1/PSE-4 chimeras vs. SCHEMA energy. SCHEMA energies were calculated using the TEM-1 structure based on a structural alignment of the two proteins (Silberg *et al.*, 2004). The experimental data (black squares) are divided into ten evenly spaced bins (x-error bars). The y-error bar for each bin is the 90% confidence interval based on the binomial distribution. The dashed line is a maximum likelihood fit of the exponential model in Equation I-3 ($\pi = 0.89$).

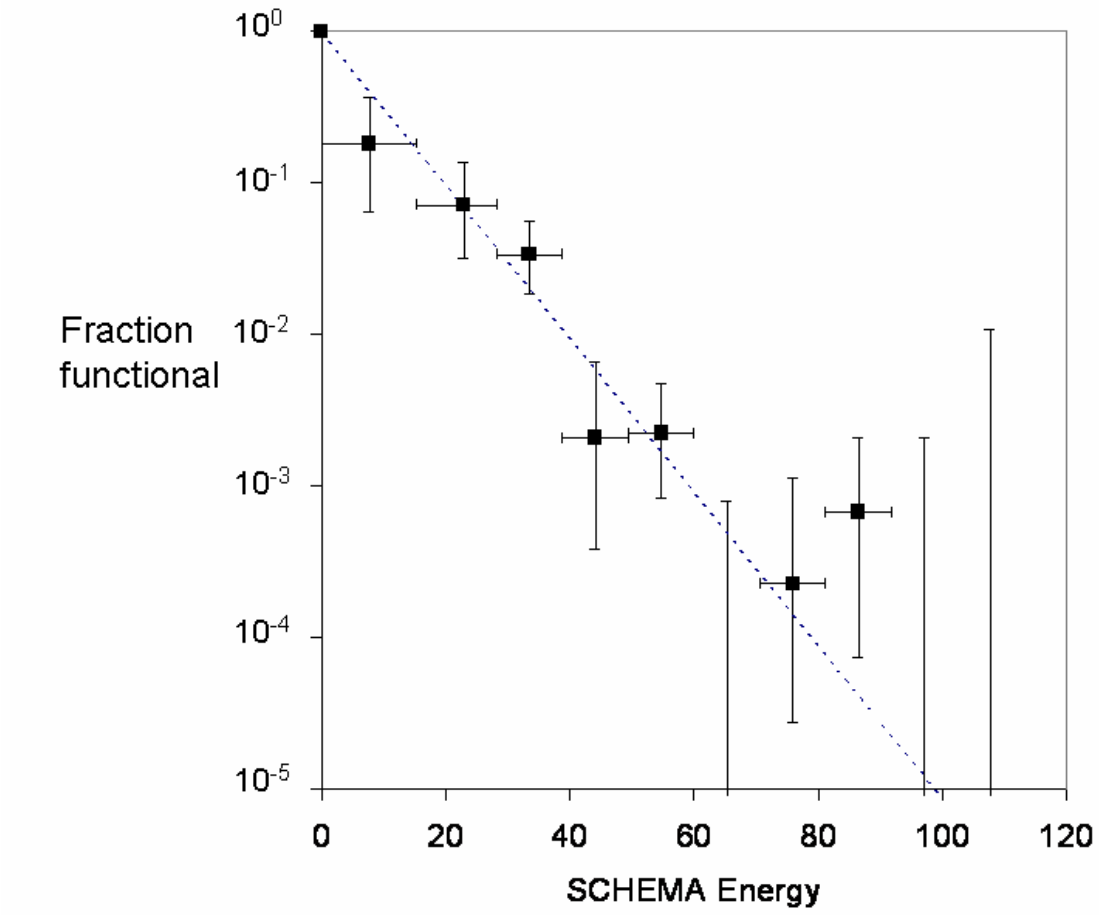

Fraction functional
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
0
20
40
60
80
100
120
SCHEMA Energy

Figure I-3. Ternary diagrams for TEM-1/PSE-4/SED-1. The probability of folding is shown for four slices through a three-dimensional space defined by the mutational distances to each parent (see Methods). The top slice is the plane containing all three parents, which lie at the corners of the triangle. Chimeras at the center of this triangle contain 107 mutations to the closest parent. The bottom three slices are parallel to the top one but at progressively higher levels of mutation (114, 121, and 128 are the maximum mutation values, relative to the closest parent). The white triangles show the outline of the top panel as a reference.

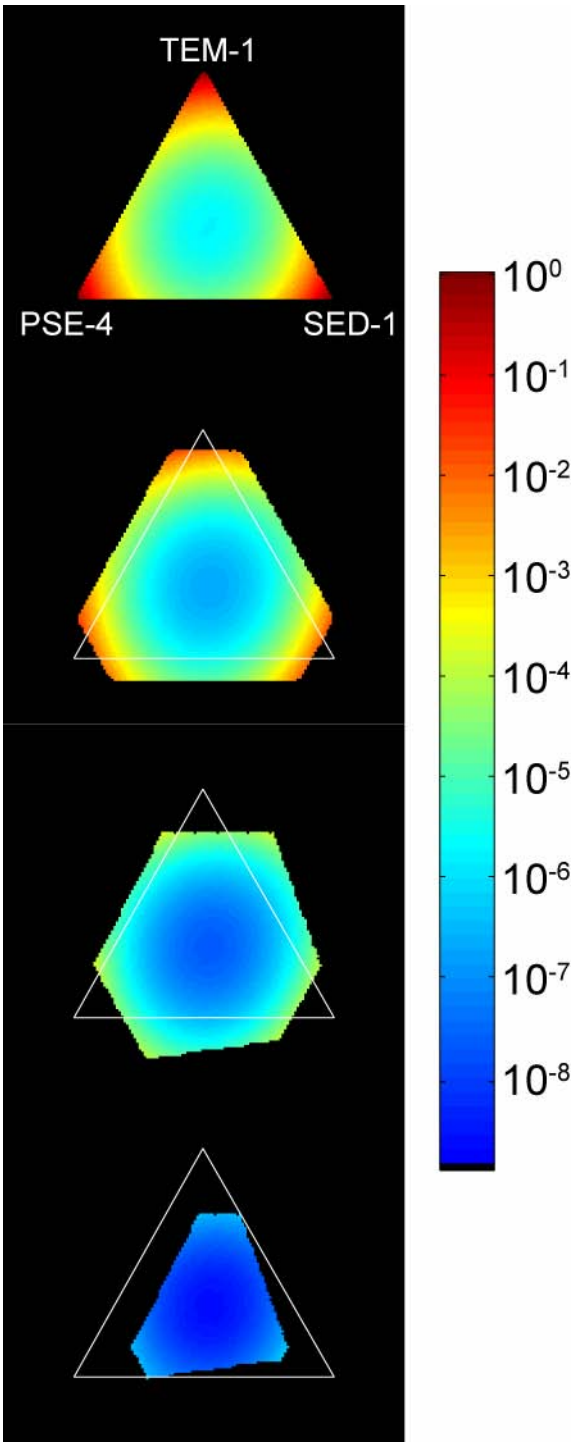

TEM-1
PSE-4
SED-1
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}

Figure I-4. Comparing the folding-diversity tradeoff for two vs. three beta-lactamase parents. The probability of folding (Equation I-6) is plotted against the number of amino acid mutations to the closest parent. The dashed curve for two parents is a half-parabola that terminates at the midpoint between the parents. The solid curve shows that many more mutations are possible with three parents, and it reveals an interesting transition at around 110 mutations (see text).

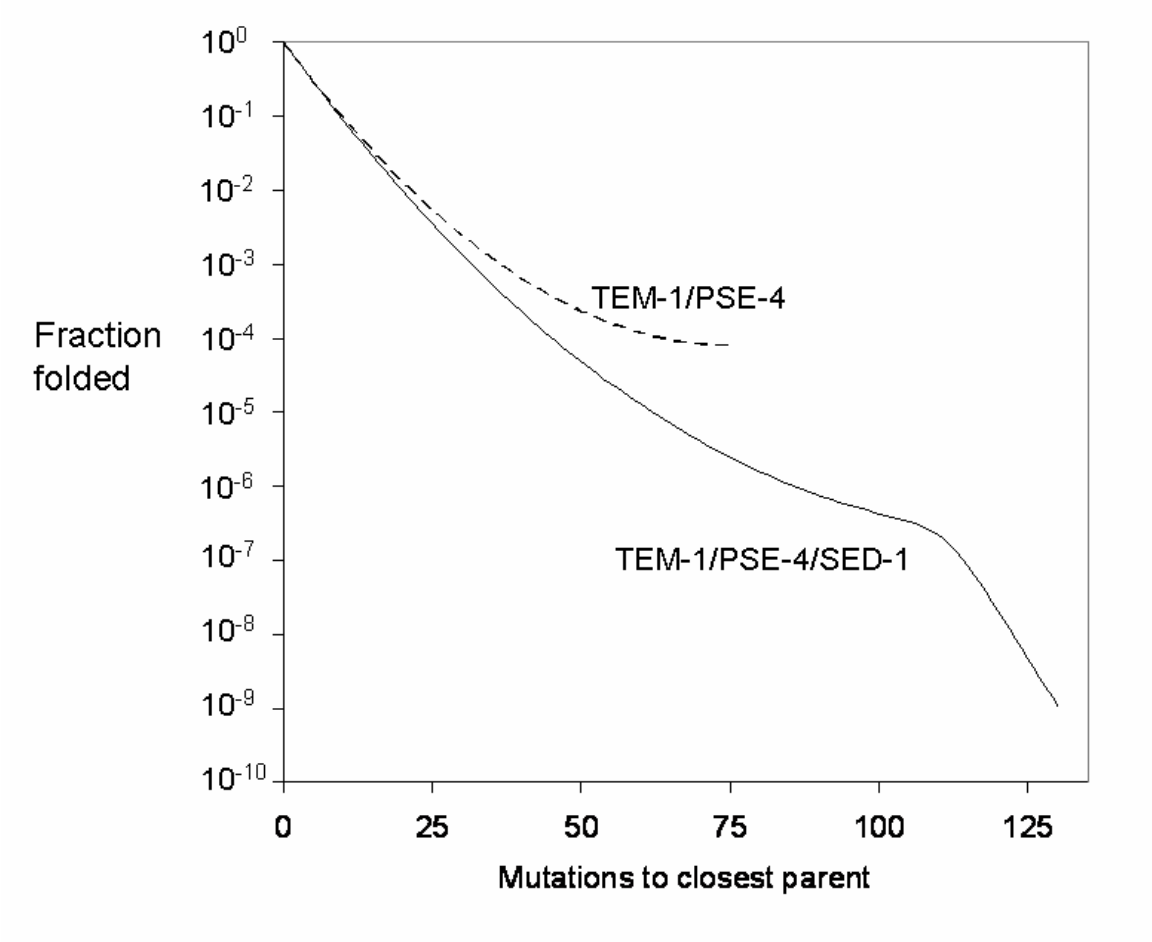
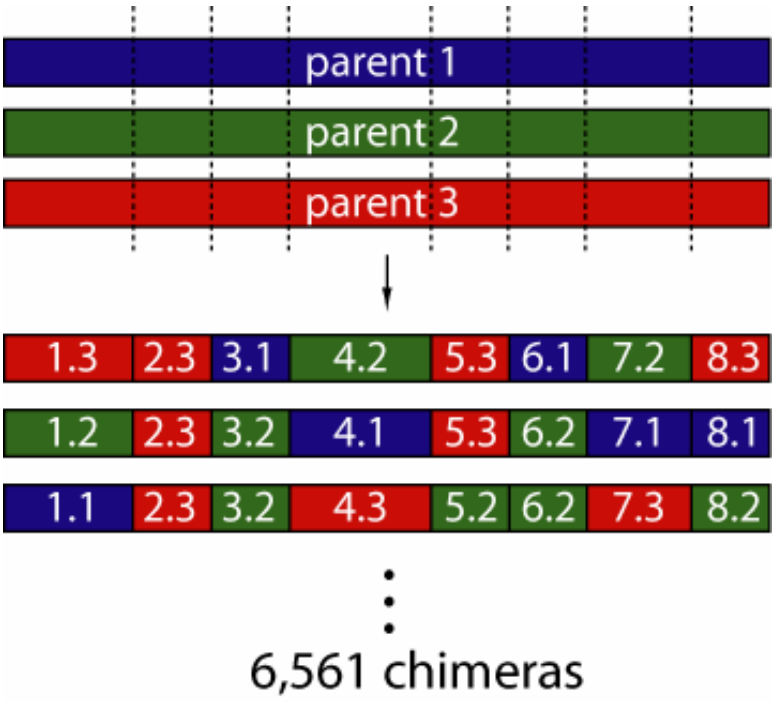

Fraction folded
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
TEM-1/PSE-4
TEM-1/PSE-4/SED-1
0
25
50
75
100
125
Mutations to closest parent

Figure I-5. Site-directed recombination with three parents and eight blocks. After the parents are aligned and split into blocks, the gene fragments are combinatorially assembled to create a library with 3^8 = 6,561 chimeras (including the parents). Fragment *i.x* is the peptide inherited from parent *x* at block *i*.


parent 1
parent 2
parent 3
1.3 2.3 3.1 4.2 5.3 6.1 7.2 8.3
1.2 2.3 3.2 4.1 5.3 6.2 7.1 8.1
1.1 2.3 3.2 4.3 5.2 6.2 7.3 8.2
6,561 chimeras

Chapter II

Site-directed recombination as a shortest-path problem

As with other kinds of computational protein design (Hellinga & Richards, 1991; Dahiyat & Mayo, 1996; Voigt *et al.*, 2001b; Hayes *et al.*, 2002; Kuhlman *et al.*, 2003), site-directed recombination (SDR) has two key ingredients: an energy function and optimization algorithm. Depending on the desired level of molecular detail and acceptable kinds of prior information (e.g., evolutionary vs. physicochemical), a wide variety of energy functions are available to predict how compatible chimeric sequences are with the target fold (Gordon *et al.*, 1999; Lazaridis & Karplus, 2000; Mendes *et al.*, 2002). My focus is on energy functions that involve one- and two-body interactions between residues. The SCHEMA energy belongs to this class of pairwise residue potentials.

The optimization component of SDR involves algorithms that minimize the chimeric energies, and hence improve the probability of folding, by judiciously choosing crossovers. The SCHEMA algorithm proposed by Voigt *et al.* uses the SCHEMA energy to calculate the local disruption caused by crossovers (Voigt *et al.*, 2002). When plotted against residue number, this map of local disruption was called the SCHEMA profile. Originally, minima in the profile were recommended as crossovers for site-directed recombination, but subsequent work by Voigt (Voigt, 2002) and the experimental results of Meyer *et al.* (Meyer *et al.*, 2003) showed that profile minima are not necessarily ideal for optimizing libraries with many crossovers.

Algorithms for library design must consider not only the fraction folded but also the diversity of the proteins. Searching through a library of folded proteins that closely resemble each other or the parents is not much better than searching through a mostly unfolded library. The design goals and library creation method should dictate what kinds of diversity and how much are needed, but our understanding of this subject is limited. In several studies, libraries with more mutations were better for adaptive evolution of enzyme function (Crameri *et al.*, 1998; Zaccolo & Gherardi, 1999; Daugherty *et al.*, 2000). Because our capacity for searching through libraries is finite, more diversity cannot always be better (Voigt *et al.*, 2001a). In theory there should be an optimal diversity (Ostermeier, 2003), but this concept is of limited use when little is known about the fitness landscape.

These two design criteria, diversity and folding, are at odds because most mutations are neutral or deleterious to protein structure. In Chapter I the tradeoff between folding and mutation was illustrated for chimeras with any number of crossovers. Site-directed recombination can ameliorate this tradeoff, yielding libraries closer to the optimal tradeoff surface, i.e., the highest fraction folded at different levels of diversity. Among all libraries of fixed size at a desired level of diversity, those on the optimal tradeoff surface provide the most attempts at the protein design goal.

In general, optimizing the tradeoff between folding and diversity for site-directed recombination is hard, but one formulation of the problem can be solved efficiently. I will show that finding the crossovers (X_1, X_2, …, X_n) that minimize the average energy $\langle E \rangle$ of the library, subject to constraints on the length L of each peptide fragment,

$$\min_{(x_1,x_2,\ldots,x_n)} \quad \langle E \rangle \qquad \text{[II-1]}$$

$$\text{subject to} \quad L_{min} \leqslant L \leqslant L_{max},$$

is equivalent to finding the shortest path between nodes in a network. This is a well-studied combinatorial optimization problem for which the global minimum can be found efficiently by dynamic programming (Korte & Vygen, 2002). Just as the energy of one protein sequence measures how likely it is to fold, the average energy of all sequences in a library is an aggregate measure of "foldability." Minimizing $\langle E \rangle$ is thus one way to enrich a library in folded proteins, although the effectiveness of this strategy depends on the relationship between $\langle E \rangle$ and the fraction folded, which in turn depends on the energy function. The constraints on fragment length [L_{min}, L_{max}] provide an indirect but computationally tractable way to control library diversity.

Methods

Protein sequences

Two different protein families were used. The cytochrome P450 homologs CYP102A1, CYP102A2, and CYP102A3 (Nelson, 2005) were aligned with ClustalW (Thompson *et al.*, 1994), and SCHEMA calculations were done with the CYP102A1 structure 1JPZ (Haines *et al.*, 2001). P450 residues were numbered from the N-terminus of CYP102A1. The beta-lactamase homologs TEM-1 (Jelsch *et al.*, 1993), PSE-4 (Lim *et al.*, 2001), and SED-1 (Petrella *et al.*, 2001) were also aligned with ClustalW, and

SCHEMA calculations were done with TEM-1 structure 1BTL (Jelsch *et al.*, 1993). Beta-lactamase residues were numbered from the N-terminus of TEM-1.

From SCHEMA energy to fraction folded

The probability of folding was assumed to decrease exponentially with SCHEMA energy, using the same maximum likelihood fit presented in Chapter I. This model was used to estimate the diversity and fraction folded for SDR libraries. If each chimera folds independently, the library can be modeled as a binomial experiment (Silberg *et al.*, 2004), for which the fraction expected to fold (F_f) is

$$F_f = \frac{1}{T}\sum_{i=1}^{T} p\left(F|E_i\right). \qquad \text{[II-2]}$$

Excluding the parents (they are known to be folded), the number of binomial trials T for a library with p parents and n crossovers is $p^{n+1} - p$. The diversity of each library was defined as the number of amino acid mutations expected for a folded protein:

$$D = \frac{\sum_k p\left(F|E_k\right) m_k}{\sum_k p\left(F|E_k\right)}. \qquad \text{[II-3]}$$

For a particular chimera, m is the number of amino acid mutations to the closest parent.

SCHEMA algorithm

The SCHEMA algorithm calculates the local disruption Γ_i, within a window of length w along the primary sequence, caused by a crossover immediately after residue i (Voigt *et al.*, 2002):

$$\Gamma_i = \frac{1}{\sqrt{w}} \sum_{j=i-w+1}^{i} \sum_{k=j}^{j+w-2} \sum_{l=k+1}^{j+w-1} C_{kl} \Pi_{kl} \, . \qquad \text{[II-4]}$$

The contact matrix C_{kl} is the same as defined in Equation I-2, and Π_{kl} is the probability that the residue pair *k*-*l* is "novel," and hence potentially disruptive, when each residue is inherited randomly. (This is a different normalization than the one used by Voigt *et al.* but makes no qualitative difference to the profile.) For recombination of *p* parents, there are p^2 amino acid combinations for every pair of positions. Because the parental residue combinations are not novel (by definition), the maximum value for Π_{kl} is $1 - 1/p$. I used a window size of 14 residues.

Results

Proof of the shortest-path problem

For any energy function with one- and two-body interactions between residues σ_k,

$$E = \sum_i e(\sigma_i) + \sum_i \sum_{j>i} e(\sigma_i, \sigma_j) \, , \qquad \text{[II-5]}$$

Equation II-1 is equivalent to finding the shortest path in the directed graph of Figure II-1. To prove this equivalence, first I establish a one-to-one correspondence between feasible *n*-crossover libraries (those that satisfy the constraints in Equation II-1) and *n*-paths, which are paths that connect node 0 to any node in column *n* of Figure II-1. Then I show that the total length of each *n*-path equals the average energy of the corresponding library, which means the shortest *n*-path is the minimum of Equation II-1.

To create the one-to-one correspondence, nodes in Figure II-1 are selectively connected. If the node X_k visited in column $k \leqslant n$ represents the position of the kth crossover, then a path that visits node X_1 in the first column defines the first peptide fragment as $[1, X_1]$ (amino acid residues 1 to X_1, inclusive). To satisfy the constraints on fragment length, node 0 should be connected to a node in the first column if and only if (iff) $L_{min} \leqslant X_1 \leqslant L_{max}$. Similarly, an arc from node X_1 in the first column to node X_2 in the second column defines the second peptide fragment as $[X_1 + 1, X_2]$. Thus node X_1 is connected to node X_2 iff $L_{min} \leqslant X_2 - X_1 \leqslant L_{max}$. This process is continued until the last column, where an arc from X_{n-1} to X_n defines two peptide fragments: $[X_{n-1} + 1, X_n]$ and $[X_n + 1, N]$ for a protein of length N. Thus node X_{n-1} is connected to node X_n iff $L_{min} \leqslant X_n - X_{n-1} \leqslant L_{max}$ and $L_{min} \leqslant N - X_n \leqslant L_{max}$.

Arc lengths are assigned so that the total length of each n-path equals the average energy of the corresponding library:

$$\sum_{k=1}^{n} A(X_{k-1}, X_k) = \langle E \rangle_{(X_1, X_2, \ldots, X_n)}, \qquad \text{[II-6]}$$

where $A(X_{k-1},X_k)$ is the arc length from node X_{k-1} to node X_k ($X_0 = 0$), and the subscript on $\langle E \rangle$ explicitly denotes the crossovers. To satisfy Equation II-6, arc lengths from node 0 are assigned the average energy of a library with one crossover immediately following residue X_1:

$$A(0, X_1) = \langle E \rangle_{(X_1)}. \qquad \text{[II-7]}$$

Arc lengths between columns are assigned the incremental change in energy associated with the next crossover:

$$A(\mathrm{X}_{k-1},\mathrm{X}_k)=\langle E\rangle_{(\mathrm{X}_1,\mathrm{X}_2,\ldots,\mathrm{X}_{k-1},\mathrm{X}_k)}-\langle E\rangle_{(\mathrm{X}_1,\mathrm{X}_2,\ldots,\mathrm{X}_{k-1})}\ ,\ k\geqslant 2. \qquad \text{[II-8]}$$

To be consistent, the right-hand side of Equation II-8 must be independent of all crossovers except X_{k-1} and X_k. This holds true for any pairwise residue potential (Equation II-5). First note that for a library with p parents, the average energy can be written as a sum over inheritance patterns:

$$\langle E\rangle_{(\mathrm{X}_1,\mathrm{X}_2,\ldots,\mathrm{X}_n)}=\frac{1}{p^{n+1}}\sum_{S_{0,\mathrm{X}_1}}\sum_{S_{\mathrm{X}_1,\mathrm{X}_2}}\cdots\sum_{S_{\mathrm{X}_n,N}}E, \qquad \text{[II-9]}$$

where $S_{i,j}$ denotes the parent from whom the peptide fragment [i + 1, j] (residues i + 1 to j, inclusive) is inherited. When a fragment contains only one residue, I use the shorthand $S_{i-1,i}\equiv S_i$. Combining Equations II-5 and II-9, the arc length in Equation II-8 can be rewritten as

$$\frac{1}{p^{k+1}}\sum_{S_{0,\mathrm{X}_1}}\cdots\sum_{S_{\mathrm{X}_{k-2},\mathrm{X}_{k-1}}}\left[\sum_{S_{\mathrm{X}_{k-1},\mathrm{X}_k}}\sum_{S_{\mathrm{X}_k,N}}-p\sum_{S_{\mathrm{X}_{k-1},N}}\right]\sum_{r=1}^{N}\sum_{t=r+1}^{N}e(\sigma_r,\sigma_t)$$

$$=\frac{1}{p^{k+1}}\sum_{S_{0,\mathrm{X}_1}}\cdots\sum_{S_{\mathrm{X}_{k-2},\mathrm{X}_{k-1}}}\left[\sum_{S_{\mathrm{X}_{k-1},\mathrm{X}_k}}\sum_{S_{\mathrm{X}_k,N}}-p\sum_{S_{\mathrm{X}_{k-1},N}}\right]\sum_{r=\mathrm{X}_{k-1}+1}^{\mathrm{X}_k}\sum_{t=\mathrm{X}_k+1}^{N}e(\sigma_r,\sigma_t) \qquad \text{[II-10]}$$

$$=\frac{1}{p^2}\left[\sum_{S_{\mathrm{X}_{k-1},\mathrm{X}_k}}\sum_{S_{\mathrm{X}_k,N}}-p\sum_{S_{\mathrm{X}_{k-1},N}}\right]\sum_{r=\mathrm{X}_{k-1}+1}^{\mathrm{X}_k}\sum_{t=\mathrm{X}_k+1}^{N}e(\sigma_r,\sigma_t) \qquad \text{[II-11]}$$

$$=\sum_{r=\mathrm{X}_{k-1}+1}^{\mathrm{X}_k}\sum_{t=\mathrm{X}_k+1}^{N}\frac{1}{p^2}\left[\sum_{S_r}\sum_{S_t}-p\sum_{S_r=S_t}\right]e(\sigma_r,\sigma_t). \qquad \text{[II-12]}$$

Equation II-10 follows because the operator in brackets, which is the difference of two inheritance sums, is only nonzero for interactions between the fragments [X_{k-1} + 1, X_k]

and $[X_k + 1, N]$. Trivial evaluation of the $k - 1$ inheritance sums outside the brackets yields Equation II-11. In Equation II-12 the order of the sums is swapped to simplify the notation: first sum over parents and then sum over residues.

Algorithmic complexity

Shortest-path problems can be solved efficiently because of their recursive structure (Lawler, 1976; Korte & Vygen, 2002). In the case of Figure II-1, the length of the shortest path U_j^k from node 0 to node j in column k can be computed using the shortest paths from node 0 to all nodes in column $k - 1$:

$$U_j^k = \min_i \left(U_i^{k-1} + A(i, j) \right). \qquad \text{[II-13]}$$

No information from other columns is needed. This property is the basis for dynamic programming. Using forward induction, RASPP finds the shortest path to every node in the first column, then the shortest path to every node in the second column, etc. Each evaluation of Equation II-13 requires O(N) operations. This is repeated for all O(N) nodes in a column and for each of the n columns, yielding a running time of O(N^2n).

In the process of finding the shortest n-path, RASPP also finds the shortest path to every column $k \leqslant n$. These path lengths do not quite correspond to the solution of Equation II-1 with k crossovers because the set of arc connections to the "last" column must satisfy a different set of constraints, as discussed above. To find optimal libraries with any fixed number of crossovers $k \leqslant n$, the arc connections between column $k - 1$ and column k are updated, and Equation II-13 is solved O(N^2) times as before. This can

be repeated for all values $k \leqslant n$ with a total running time of $O(N^2 n)$, the same as a single iteration of RASPP.

The complexity analysis must also include the time needed to calculate the arc lengths. The first evaluation of Equation II-12 requires $O(N^2 p^2)$ pairwise energy calculations, but only $O(Np^2)$ are needed to compute each subsequent arc length. This means all $O(N^2)$ distinct arc lengths can be constructed with complexity $O(N^3 p^2)$. When combined with the running time for dynamic programming, the overall complexity is $O(N^3 p^2 + N^2 n)$.

To generate libraries with different diversities, the length constraints $[L_{min}, L_{max}]$ are adjusted over a range of values. In the absence of experimental constraints, L_{min} can vary from 1 to $N/(n + 1)$ and L_{max} from $N/(n + 1)$ to $N - nL_{min}$. This requires $O(N^2/n)$ iterations of RASPP, but since the arc lengths are not recalculated each time, the total running time is only $O(N^4 + N^3 p^2)$.

Case studies using the SCHEMA energy

The theoretical development thus far has been valid for any potential with pairwise interactions between residues (Equation II-5). To present computational results, I now specialize to the SCHEMA energy, which counts the number of structural contacts disrupted when portions of the sequence are inherited from different parents (Equation I-2). In this case the formula for the arc lengths (Equation II-12) has a particularly simple form:

$$A(\mathrm{X}_{k-1}, \mathrm{X}_k) = \sum_{r=\mathrm{X}_{k-1}+1}^{\mathrm{X}_k} \sum_{t=\mathrm{X}_k+1}^{N} C_{rt}\Pi_{rt} \; . \qquad \text{[II-15]}$$

Π_{rt}, which is the probability of breaking the *r*-*t* interaction when residues *r* and *t* are inherited at random, also appears in the definition of the SCHEMA profile (Equation II-4).

Consider making a seven crossover library using the heme domains of cytochrome P450 homologs CYP102A1, CYP102A2, and CYP102A3, which share roughly 65% of their 456 amino acids (Nelson, 2005). By varying the length constraints for $n = 7$ crossovers, 2,052 libraries were generated, of which 391 are distinct. These are plotted in gray in Figure II-2. As L_{min} increases and L_{max} decreases, the crossovers become more evenly spaced, resulting in libraries with higher $\langle E \rangle$ and higher diversity, as measured by the average number of amino acid mutations to the closest parent $\langle m \rangle$. Designing libraries with more crossovers increases the levels of diversity accessible by SDR, but adding fragments also complicates construction of the library. In this example, the choice of $n = 7$ provides enough mutants for screening ($3^8 = 6{,}561$ chimeras) and sufficiently high levels of mutation for laboratory evolution based on data from previous experiments (Otey *et al.*, 2004).

The lowest-energy RASPP libraries at increasing values of $\langle m \rangle$ define the solid "RASPP-curve" shown in Figure II-2. To determine how well RASPP-curves approximate the optimal energy-diversity tradeoff surface, I enumerated all four-crossover libraries for CYP102A1/CYP102A2 and for TEM-1/ PSE-4, the two beta-

lactamases introduced in Chapter I. These two sets of parents have 25 and 20 million possible libraries, respectively, as shown in Figure II-3. At most levels of mutation, the RASPP-curve provides a good estimate of the lowest energy possible. Exceptions occur in mutation ranges where RASPP does not produce any libraries, e.g., $30 < \langle m \rangle < 35$ for the cytochromes P450. A similar mutation gap can be seen in Figure II-2 at $40 < \langle m \rangle < 50$. Such gaps are to be expected when using constraints on fragment length as a surrogate for $\langle m \rangle$. Changing the parents or the number of crossovers can shift the location of a gap, as seen by comparing Figures II-2 and II-3, which differ in both respects.

The pattern of optimal crossovers varies dramatically along a RASPP-curve. Figure II-4 shows the elements of secondary structure for CYP102A1 (Ravichandran *et al.*, 1993) corresponding to crossovers along the RASPP-curve of Figure II-2. At low values of $\langle m \rangle$, RASPP favors the ends of the protein to minimize structural disruption. The resulting chimeras inherit a single, large fragment from one parent, and most of the remaining fragments contain only a few residues. To create libraries with higher $\langle m \rangle$, RASPP must spread out the crossovers and penetrate the middle of the polypeptide chain, and the algorithm often cuts through secondary structure motifs. For example, the most commonly chosen crossover (after residue 214, which shows up as a long horizontal black line in Figure II-4) lies in the middle of a long α helix covering the substrate binding pocket. Two other consistently good regions for recombination (residues 248–255 and 256–276) are also helical.

Before choosing optimal crossover locations, one must decide upon a set of parents for recombination. RASPP-curves provide a rapid and reliable way of determining which parents yield the lowest-energy libraries in a desired diversity range. To illustrate, consider choosing among three combinations of cytochrome P450 homologs: A1/A2, A1/A3, or A1/A2/A3. Even though a library with three parents has more chimeras than one with two parents, the comparison is fair because any random, experimental sample will on average have the same $\langle E \rangle$ as the entire library.

The RASPP-curves for these alternative designs, shown in Figure II-5, reveal significant differences at mutation levels $\langle m \rangle > 40$. For $40 < \langle m \rangle < 60$, the combination A1/A2 is better than A1/A3 because the former has lower energy. This would be difficult to ascertain by other means, since A1/A2 and A1/A3 both have 65% sequence identity, and on average their nonconserved residues make the same number of contacts. For $40 < \langle m \rangle < 50$, A1/A2 also has lower energy than A1/A2/A3. For $50 < \langle m \rangle < 60$, A1/A2 and A1/A2/A3 have comparable energy, but A1/A2 is still preferable because adding a third parent increases the cost and complexity of library construction. All three parents are needed to build libraries with $\langle m \rangle > 60$.

Discussion

Equation II-12 is the key theoretical result that shows dynamic programming can be used for SDR library design. In this respect, Equation II-12 is analogous to the dead-end elimination (DEE) theorem (Goldstein, 1994; Pierce *et al.*, 2000), which has enabled

many successes in protein sequence design (Dahiyat & Mayo, 1996; Looger *et al.*, 2003). However, Equation II-12 and the DEE theorem have very different consequences for computational protein design. RASPP finds the global energy minimum for Equation II-1 in $O(N^3p^2 + N^2n)$ operations for a protein of length N, making it efficient in theory and practice (Papadimitriou & Steiglitz, 1998). In contrast, DEE requires an exponential number of operations $O(a^N)$ in the worst case. This is unavoidable (unless P = NP) because finding the amino acid sequence with minimum energy is NP-hard (Pierce & Winfree, 2002). Averaging over the library transforms protein design from a hard problem to an easy one (see also Appendix A).

Figure II-6 shows how RASPP (open squares, solid curve) compares with the SCHEMA algorithm (closed triangle) using the beta-lactamase homologs TEM-1, PSE-4, and SED-1. The SCHEMA profile for these three parents, shown in Figure II-7, has seven pronounced minima and thus encodes one seven-crossover library. Assuming an exponential folding model, I expect 0.1% of the profile library to be folded and each folded protein to have 46 mutations (see Methods). The RASPP library with comparable diversity is nearly 3% folded, which is 20 times more than the library chosen by SCHEMA and several hundred times more than the probability of folding among all TEM-1/PSE-4/SED-1 chimeras with 46 mutations (dashed curve, reproduced from Figure I-4). When the crossovers for the RASPP-library are plotted on the profile of Figure II-7 (closed squares), no preference for minima over maxima is observed. Although the crossovers selected by SCHEMA are locally optimal on an individual basis, when combined into a multi-crossover library their performance is variable (Voigt, 2002). In

my experience, when all profile minima are included as crossovers, the library is unlikely to be competitive with RASPP solutions. However, if only a subset of the profile minima is needed, e.g., making a four-crossover library for TEM-1/PSE-4/SED-1, then some combinations may be quite good.

The main limitation with RASPP is its assumption that the parents are inherited with equal probability at every block. Of the p^2 fragment-fragment combinations for each block pair, many have never been tested before by nature and hence are likely to be deleterious. Some of these disruptive interactions can be avoided by optimizing the crossover locations, but not all. Yet another way to minimize the library energy is by omitting fragments that do not pair well with other fragments. This modification to the SDR paradigm is easily implemented in the laboratory, and the algorithm OPTCOMB has recently been developed for designing these kinds of combinatorial libraries (Saraf *et al.*, 2005). The solutions found by RASPP are a subset of those possible with OPTCOMB, but this added flexibility comes at a cost. Unlike RASPP, OPTCOMB is not guaranteed to be efficient for large proteins. In practice this may not be a concern, in which case OPTCOMB would be preferred for its ability to find libraries closer to the optimal folding-diversity tradeoff surface.

Portions of this chapter were reproduced with the permission of Oxford University Press from
Endelman, J.B., Silberg, J.J., Wang, Z.G. & Arnold, F.H. (2004) *Protein Eng. Des. Sel.*, **17**, 589–594.

Figure II-1. Site-directed recombination as a shortest-path problem. Every feasible *n*-crossover library can be represented as an *n*-path from node 0 to column *n*. The node visited in column *k* corresponds to the position of the *k*th crossover, shown here for a protein of length *N*. To constrain the length of each peptide fragment, nodes in adjacent columns are selectively connected. Arc lengths are assigned so that the total length of each *n*-path equals the average energy of the corresponding library.

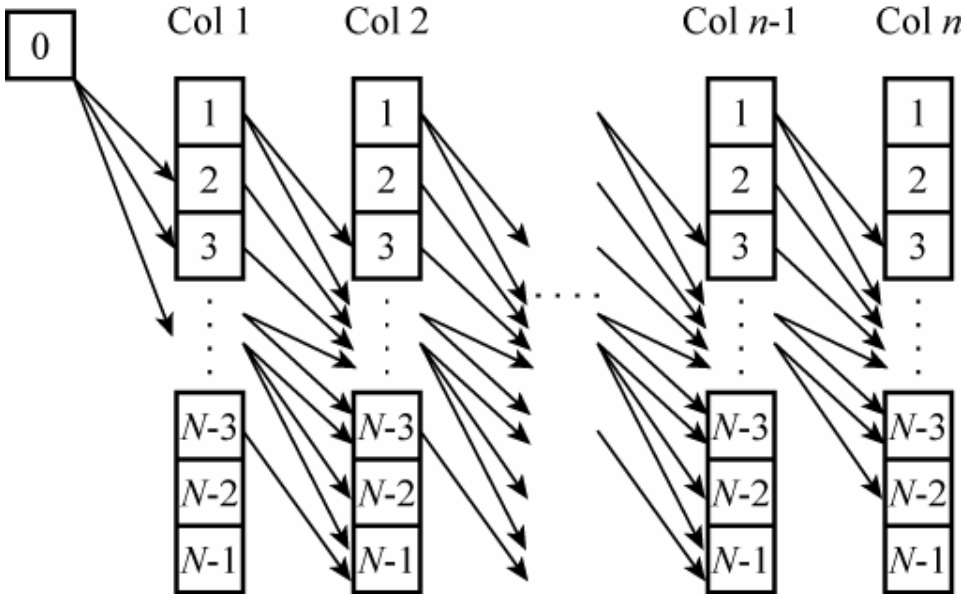

0
Col 1
Col 2
Col n-1
Col n
1
2
3
N-3
N-2
N-1

Figure II-2. SCHEMA energy vs. diversity for seven-crossover cytochrome P450 libraries. Equation II-1 was solved by RASPP for length constraints L_{min} = 1 to $N/(n + 1)$ and $L_{max} = N/(n + 1)$ to $N - nL_{min}$, where N = 197 nonconserved residues. A plot of $\langle E \rangle$ vs. $\langle m \rangle$ for the 391 distinct libraries in this set (gray squares) reveals that no RASPP libraries fall in the range $40 < \langle m \rangle < 50$. This is a consequence of using constraints on fragment length as a surrogate for $\langle m \rangle$. The RASPP-curve (black line) was generated by dividing the $\langle m \rangle$-axis into bins of 1.5 mutations and keeping the lowest-energy library within each bin (black squares).

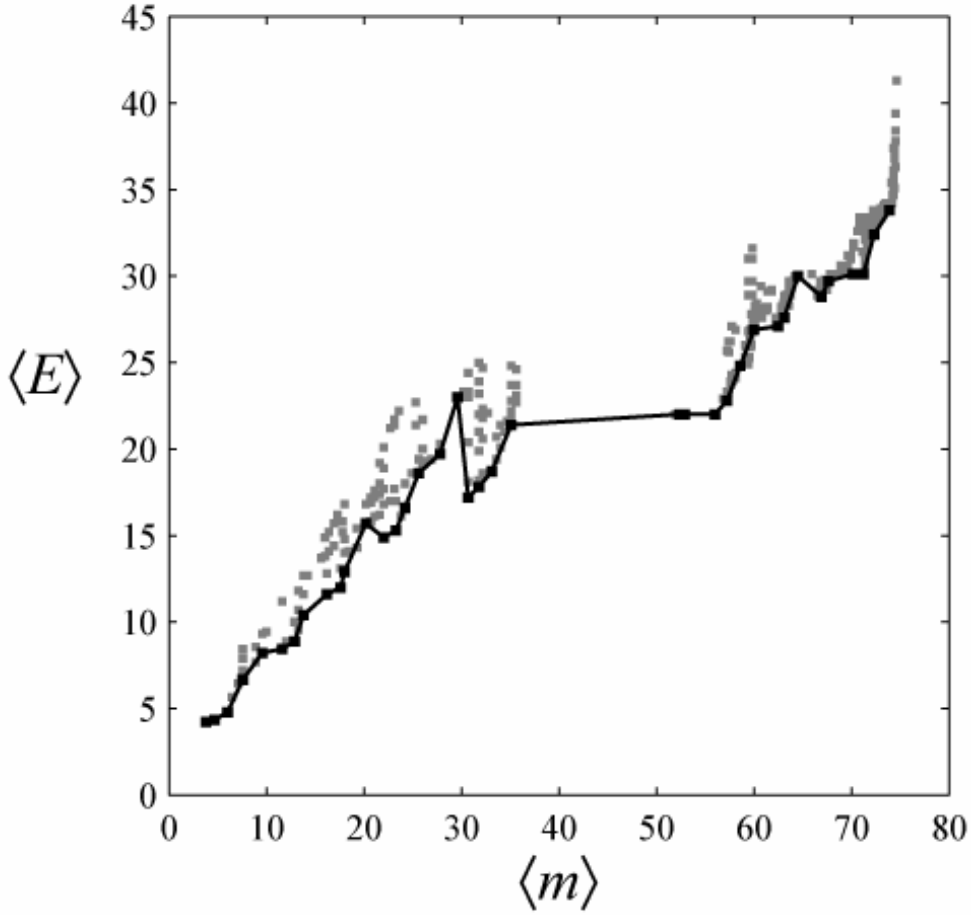

45
40
35
30
25
20
15
10
5
0
⟨E⟩
0
10
20
30
40
50
60
70
80
⟨m⟩

Figure II-3. RASPP-curves approximate the optimal tradeoff between SCHEMA energy and diversity. All four-crossover libraries (gray) were enumerated for cytochromes P450 CYP102A1/A2 and β-lactamases TEM-1/PSE-4 (25 and 20 million libraries, respectively). In both cases, the RASPP-curve (black line) closely approximates the optimal energy-diversity tradeoff surface at most values of mutation. One glaring exception is the range $30 < \langle m \rangle < 35$ for the cytochromes P450, in which the RASPP-curve substantially underestimates the minimum energy. This happens because there are no RASPP libraries in this range (as was true for $40 < \langle m \rangle < 50$ in Figure II-2).

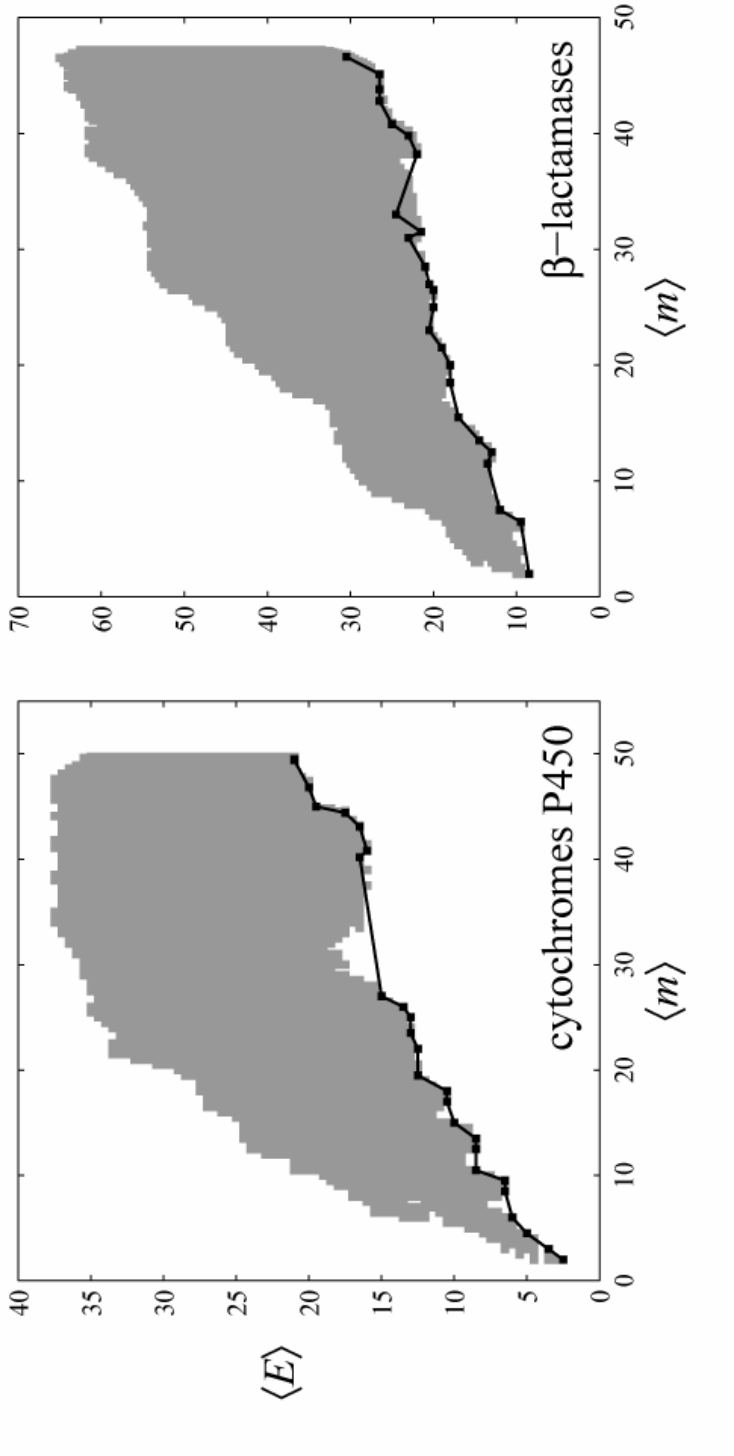

β–lactamases
cytochromes P450
⟨E⟩
⟨m⟩
⟨m⟩

Figure II-4. Crossovers along the RASPP-curve. Crossover locations (dark horizontal bars) are shown for every library along the RASPP-curve of Figure II-2 (CYP102A1/A2/A3, $n = 7$ crossovers). When crossovers fall in a contiguous region of conserved residues, they are depicted at the position closest to the N-terminus. Long horizontal black lines indicate regions consistently chosen by RASPP. There are two vertical axes. On the far right are the residue numbers for CYP102A1. The second axis, labeled as 2^o, depicts secondary structure motifs along the polypeptide chain of CYP102A1 (Ravichandran *et al.*, 1993). Boxes filled solid gray represent β strands; boxes filled solid white represent 3_{10} helices; boxes filled with a black and white gradient represent α helices. Many of the crossovers chosen by RASPP lie within secondary structure motifs.

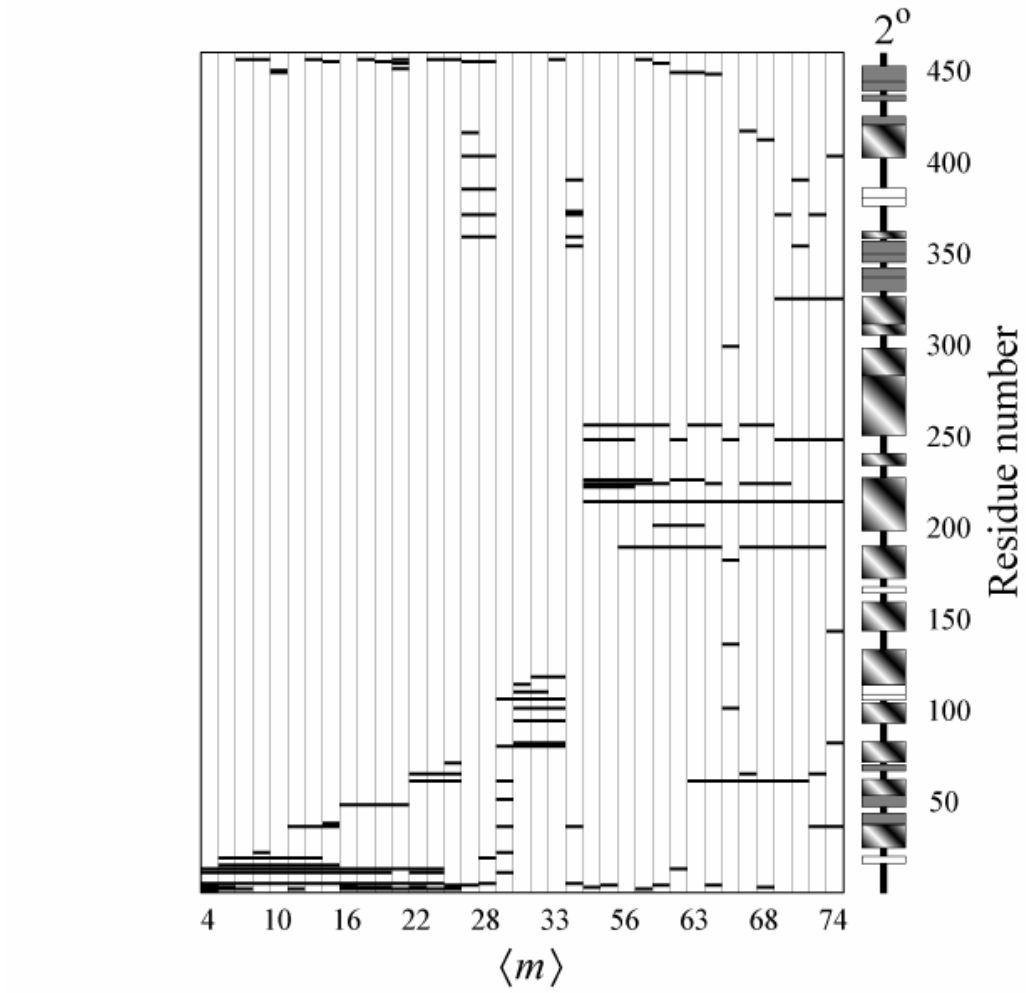

2°
450
400
350
300
250
200
150
100
50
Residue number
4
10
16
22
28
33
56
63
68
74
⟨m⟩

Figure II-5. RASPP-curves guide the choice of parents. Three alternative sets of cytochromes P450 are compared: CYP102A1/A2, A1/A3, and A1/A2/A3, each with nine crossovers. The RASPP-curves, computed as described in Figure II-2, represent the lowest-energy libraries possible for each set of parents. The optimal set of parents in a target range of mutation is the one with the lowest RASPP-curve. At low values of mutation, all sets of parents are comparable. For $40 < \langle m \rangle < 50$, A1/A2 is preferred because it has lower energy than A1/A3 or A1/A2/A3. All three parents are needed to create libraries with $\langle m \rangle > 60$.

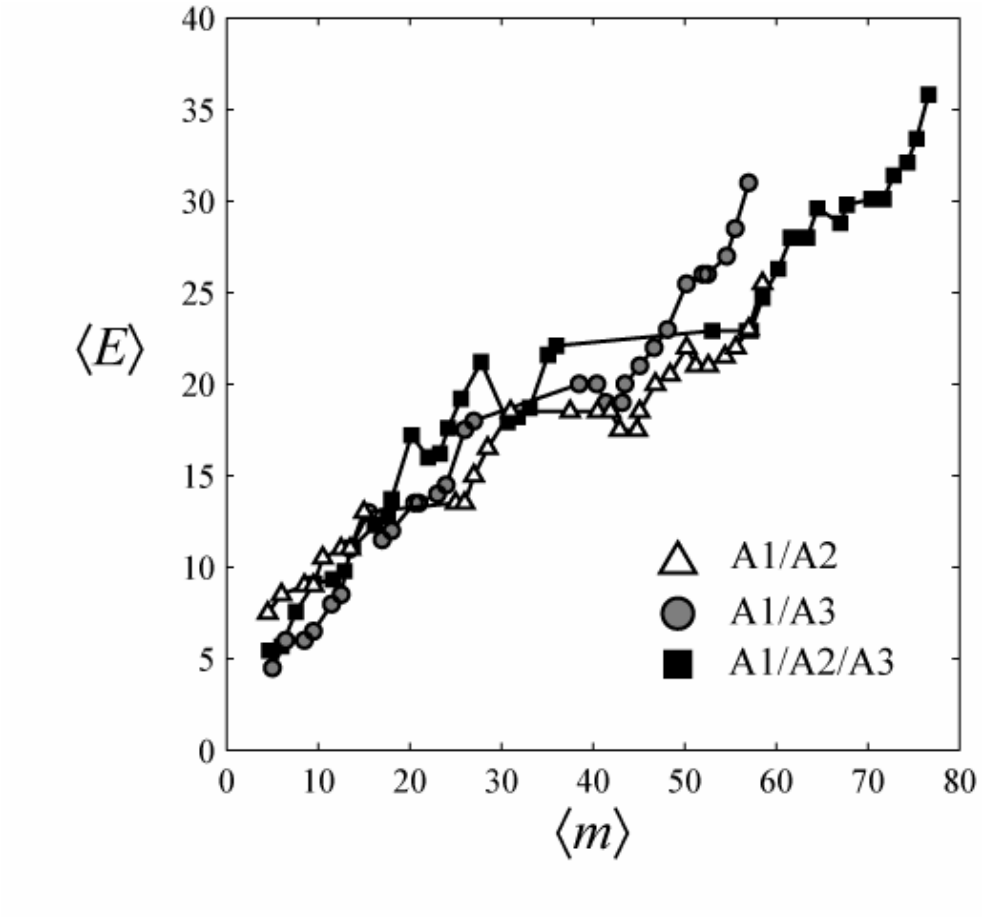

40
35
30
25
20
15
10
5
0
0
10
20
30
40
50
60
70
80
⟨E⟩
⟨m⟩
A1/A2
A1/A3
A1/A2/A3

Figure II-6. RASPP vs. the SCHEMA algorithm. The SCHEMA profile for beta-lactamase parents TEM-1/PSE-4/SED-1 (see Figure II-7) encodes one seven-crossover library, shown here as a closed triangle. The seven-crossover libraries along the RASPP-curve for these parents are shown as open squares connected by a solid line. For both algorithms, the fraction folded and number of amino acid mutations per folded protein were calculated by assuming the probability of folding decays exponentially with SCHEMA energy (see Methods). The RASPP library with diversity comparable to that chosen by the SCHEMA algorithm has twenty times more folded proteins. The dashed curve, which is reprinted from Figure I-4, shows the probability of folding among all TEM-1/PSE-4/SED-1 chimeras.

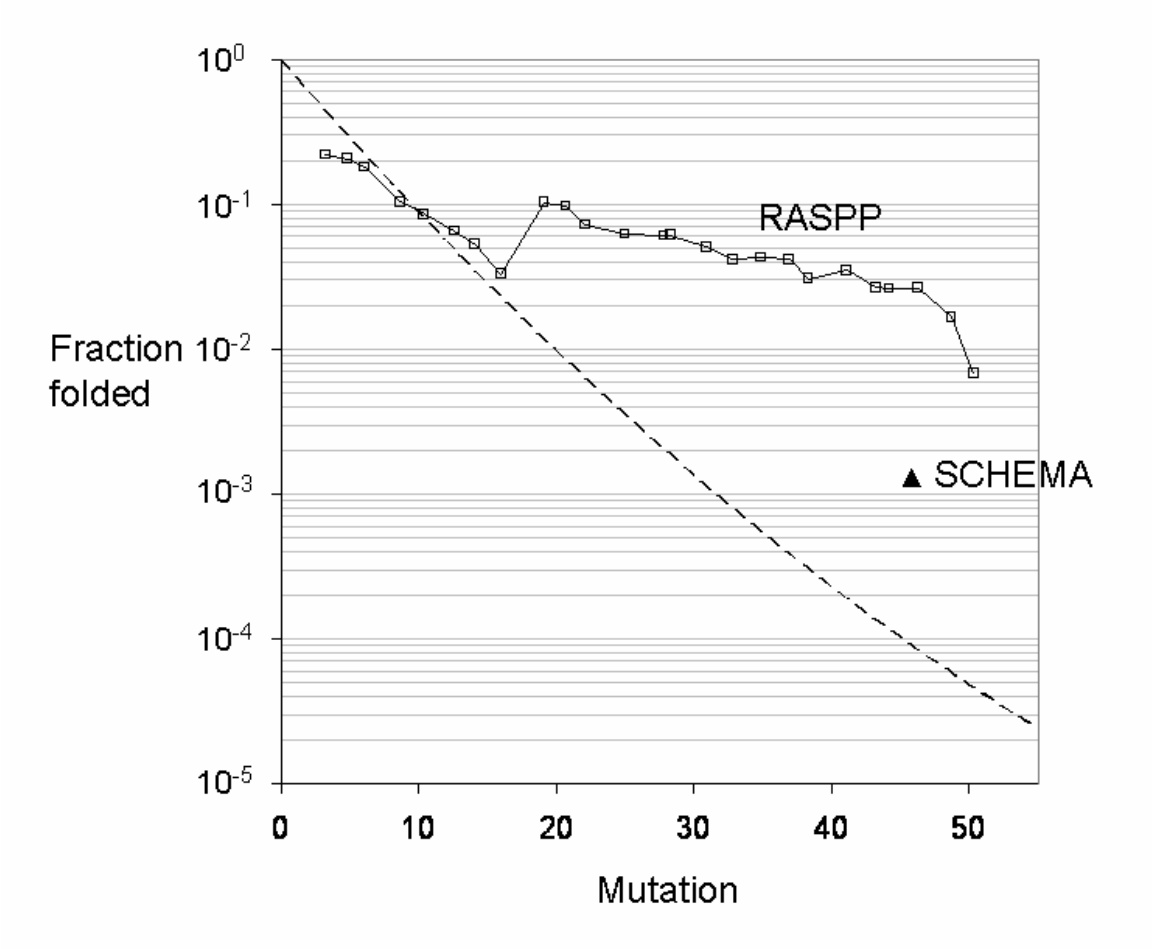
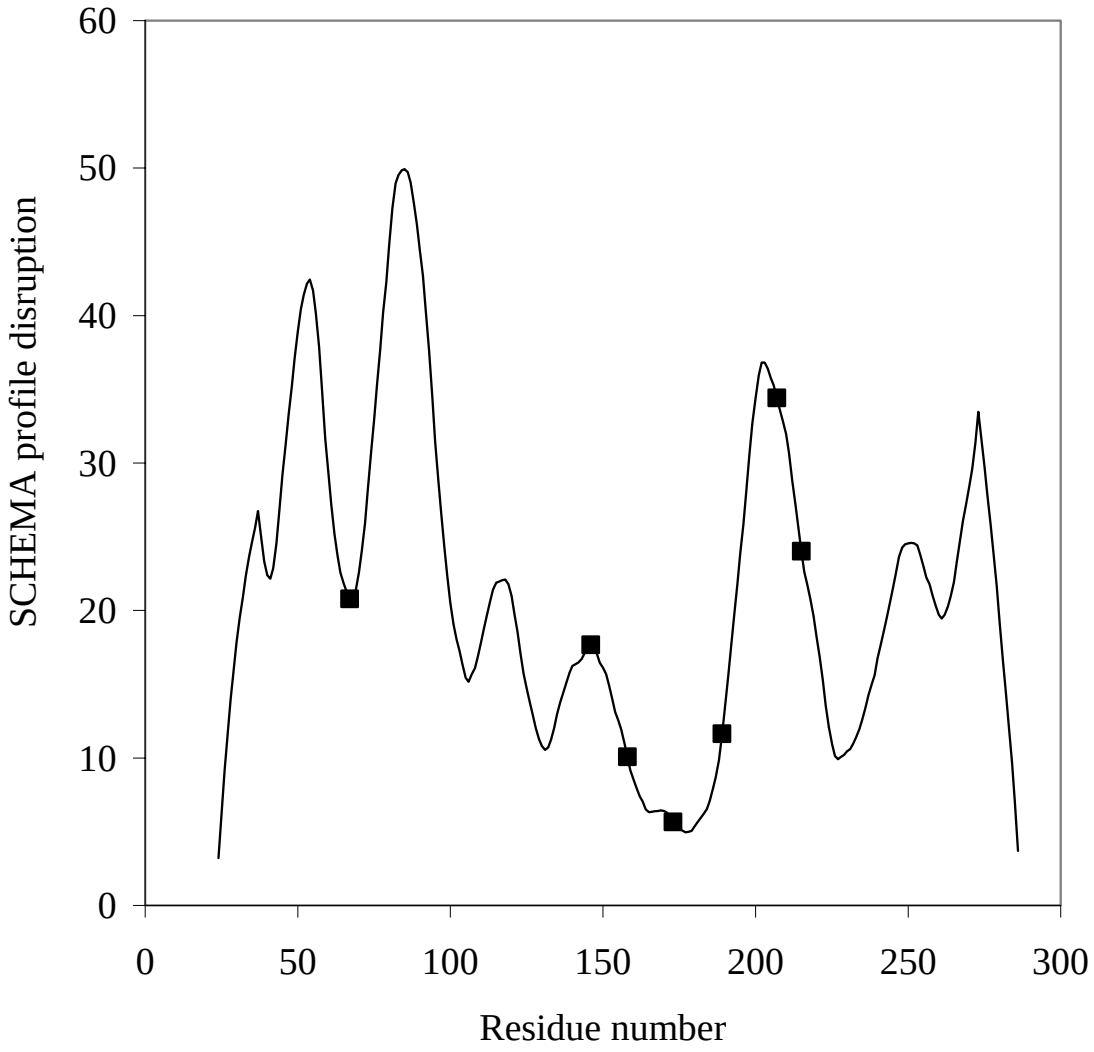

Fraction folded
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
RASPP
▲ SCHEMA
0
10
20
30
40
50
Mutation

Figure II-7. Crossovers chosen by RASPP vs. the SCHEMA profile. The seven minima of the SCHEMA profile for beta-lactamases TEM-1/PSE-4/SED-1 lie directly after residues R41, S68, V106, T131, D177, W227, and T261 in the TEM-1 sequence. A library with crossovers at all seven minima is expected to be 0.1% folded and to contain 46 mutations per folded protein. The RASPP library for these parents with 46 mutations is 3% folded (see Figure II-6), and its crossovers (shown as filled squares) lie after residues M67, L146, T158, N173, R189, D207, and A215.


60
50
40
30
20
10
0
SCHEMA profile disruption
0
50
100
150
200
250
300
Residue number

Chapter III

Comparing the predictive accuracy of pairwise residue potentials

The SCHEMA energy was proposed by Voigt *et al.* to score the stability of chimeric sequences for a target fold (Voigt *et al.*, 2002). SCHEMA uses a crystal structure and parental sequence alignment to count the number of contacts disrupted when portions of the sequence are inherited from different parents. A contact is considered disrupted if the pair of residues found in the chimera is not present in any of the parents (Silberg *et al.*, 2004). One might expect the simplicity of SCHEMA to hinder its effectiveness for protein design. Since RASPP, and even the SCHEMA algorithm, can use any pairwise residue potential, it is important to determine whether SCHEMA is competitive with more detailed alternatives.

A variety of schemes have been used to evaluate energy function accuracy. One involves checking whether the energy of a natural sequence is lower on its native structure than on a set of decoys (Gilis, 2004). It is also possible to compare the change in computational energy upon mutation with experimental measurements of the free energy (Gilis & Rooman, 1997). Others have compared the ability of different potentials to classify as either neutral or deleterious the mutants of several well-studied proteins (Saunders & Baker, 2002). These binary data (neutral = 1, deleterious = 0) resemble those generated by screening site-directed chimeras for their ability to function and/or fold (folded/functional = 1, not = 0).

Binary folding data can be analyzed with information theory to evaluate the predictive accuracy of energy functions. In a sample of folded and unfolded proteins, one cannot predict with certainty whether a randomly chosen sequence is folded. This uncertainty, or entropy (Adami, 2004), can be reduced by knowing the energy of each sequence if proteins with higher energy are less likely to be folded. The decrease in entropy equals the mutual information between folding and energy. An energy function with higher mutual information is better able to predict folding in the sample and presumably in future libraries as well, making it desirable for computational protein design. Within this framework, I compare SCHEMA against other pairwise residue potentials using data from cytochrome P450 and beta-lactamase SDR libraries.

Methods

Cytochrome P450 library

A seven-crossover cytochrome P450 library, designed before the development of RASPP, was selected *in silico* after two rounds of enumeration. Using the heme domains of *Bacillus* homologs CYP102A1, CYP102A2 and CYP102A3 as parents (see Chapter II), 5,000 seven-crossover libraries were evaluated. For each library, crossovers were chosen randomly with a minimum fragment size of 20 residues. Based on folding data from a set of 17 individually constructed two-crossover chimeras (Otey *et al.*, 2004), and by analogy with the beta-lactamase folding model used by Voigt *et al.* (Voigt *et al.*, 2002), the probability of P450 folding was assumed to change abruptly from 1 to 0 at SCHEMA energy $E = 30$. SCHEMA energy calculations, as well as the fraction folded

and expected number of mutations per folded protein, were carried out as described in Chapter II.

From the set of 5,000 randomly generated libraries, only those with a fraction folded greater than 25% were selected for further study. Fourteen crossovers appeared in greater than 40% of these libraries. There are 3,432 possible ways to choose 7 crossovers from this set of 14, all of which were evaluated. The final library was selected for its high fraction folded (40%) and large number of mutations per folded protein (65). The crossovers lie directly after the following residues: Glu64, Ile122, Tyr166, Val216, Thr268, Ala328, and Gln404, numbered from the N-terminus of CYP102A1.

The chimeras were screened for their ability to fold and bind heme using carbon monoxide difference spectroscopy (Otey, 2003). From a random sampling of several thousand colonies, 628 full length P450 sequences were identified, of which 287 bind heme (C. Otey, personal communication). Additional sequencing of folded P450s yielded an expanded data set containing 806 chimeras (including the three parents), of which 465 bind heme. These data are listed in Appendix B.

Beta-lactamase library

A RASPP-curve for beta-lactamase homologs PSE-4, SED-1, and TEM-1 (parents 1, 2, and 3, respectively; see Chapter I) was generated with the SCHEMA energy as described in Chapter II. From this curve a library with high average mutation and low average SCHEMA energy was selected, and the second crossover was moved to be compatible with the experimental protocol used in library construction (Hiraga & Arnold,

2003). The final crossover locations lie directly after the following residues: Arg63, Lys71, Thr147, Arg159, Asp174, Leu188, and Gly216, numbered from the N-terminus of TEM-1. A random sample of 553 sequences contained 111 functional beta-lactamases, as determined by screening for ampicillin resistance on agar plates (M. Meyer, personal communication). Additional sequencing of functional proteins yielded 605 chimeras, of which 163 are functional. These data are listed in Appendix B.

Mutual information between folding and energy

The uncertainty about folding (F) in a set of N chimeras, only N_f of which are folded, was quantified by the Shannon entropy (Adami, 2004). If $p = N_f/N$ denotes the fraction folded, then the entropy in bits per chimera is

$$H(F) = -\left[p \log_2 p + (1-p)\log_2 (1-p)\right] \geq 0. \qquad \text{[III-1]}$$

Given no other information besides the ratio p, the likelihood of correctly predicting the folding status of every chimera is 2^{-NH}. Systems with lower entropy are thus easier to predict, i.e., there is less uncertainty. The conditional entropy $H(F|E)$, which must be less than or equal to $H(F)$, measures uncertainty when the chimeric energies are known, and the difference $H(F) - H(F|E)$ equals the mutual information $I(F:E)$. The conditional entropy $H(F|E)$ is an average over energy values,

$$H(F \mid E) = \sum_{E_k} p(E_k) H(F \mid E_k), \qquad \text{[III-2]}$$

where $p(E_k)$ is the fraction of chimeras with energy E_k. $H(F|E_k)$ is the conditional entropy associated with knowing whether a chimera has energy E_k. It was computed from

Equation III-1 by replacing p with the conditional probability $p(F|E_k)$, which is the probability that a chimera with energy E_k is folded.

In the original work of Voigt *et al.* (Voigt *et al.*, 2002), and for the P450 library described above, the probability of folding was assumed to decrease abruptly with SCHEMA energy at some threshold. This step model was eventually proven inadequate by Meyer *et al.*, who noticed that the probability of folding decreased exponentially with SCHEMA energy (Meyer *et al.*, 2003). To capture the wide range of behaviors expected for energy functions besides SCHEMA, an even more general folding model is needed. With the exponential model it is impossible for the mutual information between folding and energy to approach its maximum value. High mutual information scores require a sharp transition between high and low probability. With a probability model of the form

$$p(F|E) = \frac{1}{c + e^{bE+a}}, \qquad \text{[III-3]}$$

both exponential ($c = 0$, $a = 0$) and sigmoidal ($c = 1$) curves are possible, as well as intermediate behaviors. All three parameters were fit to the binary folding data using maximum likelihood, subject to the constraints $b \geq 0$, $0 \leq c \leq 1$.

CVHclash

Saraf *et al.* have proposed using the biophysical properties of amino acids to restrict which novel residue contacts are counted as disruptive or "clashing" (Saraf & Maranas, 2003). I have created a similar energy function called CVHclash, which counts

the number of contacting residue pairs with charge (C), volume (V), or hydrophobicity (H) outside the range spanned by the parents:

$$E_{CVH} = \sum_i \sum_{j>i} C_{ij} \delta_{ij} \ . \qquad \text{[III-4]}$$

The contact matrix C_{ij} = 1 if the Cβ atoms (Cα for gylcine) of residues *i* and *j* are within 8 Å in the parental structure; otherwise C_{ij} = 0. The delta function δ_{ij} considers whether the residue pair has parent-like CVH properties. The charge (Klein *et al.*, 1984), volume (Krigbaum & Komoriya, 1979), and hydrophobicity (Cid *et al.*, 1992) parameters for each amino acid were taken from the AAindex database (Kawashima *et al.*, 1999). The additive property (C, V, or H) for each contacting pair is the sum of the values for the two amino acids involved (Saraf *et al.*, 2004). If any of the three properties for a chimeric pair lies outside the range spanned by the parents, δ_{ij} = 1 and one clash is counted. Increasing the range of nondisruptive CVH values by 10 or 20% beyond that spanned by the parents did not significantly change the results.

Results

SCHEMA predicts beta-lactamase folding better than P450 folding

The mutual information between folding and energy was used to evaluate the predictive accuracy of SCHEMA. Mutual information ranges from zero up to the Shannon entropy, which quantifies the uncertainty about folding (see Methods). When half the sequences are folded in a sample, the entropy is at its maximum of 1 bit per chimera. As the fraction folded deviates from 0.5, it becomes easier to predict the folding

status of a randomly chosen chimera, and hence the entropy decreases. The beta-lactamase data set contains 605 chimeras, 27% of which are folded and functional, which means the entropy per chimera is 0.84 bits. The P450 entropy is 0.98 bits per chimera because 58% of the 806 sequences are folded.

Although more information is available for the P450s, less is captured by SCHEMA compared to the beta-lactamases. As shown in Figure III-1, the mutual information per chimera is 0.34 bits for the beta-lactamases, compared to only 0.07 bits for the P450s. These results are not sensitive to how structural contacts are defined. The above scores use the original definition of Voigt *et al.*, who considered whether any pair of heavy side-chain atoms or backbone carbon atoms between two residues is within 4.5 Å (Voigt *et al.*, 2002). The structure prediction competition CASP defines two residues as contacting if their Cβ atoms (Cα for glycine) are within 8Å (Aloy *et al.*, 2003). When this standard is used, denoted as $SCHEMA_2$, the mutual information per chimera remains essentially unchanged for both proteins (0.01 bit increase). Increasing or decreasing the contact distance by 1Å also has negligible effect (0.01 bit decrease).

My attempts to improve SCHEMA by weighting each contact with its spatial or sequence separation have been unsuccessful. Inversely weighting each contact by the Cβ distance between atoms decreases the mutual information with beta-lactamase folding by 0.05 bits per chimera. Weighting each contacting residue pair by the number of amino acids separating them along the primary sequence, i.e., their contact order (Plaxco *et al.*, 1998), is one way to emphasize tertiary over secondary contacts. This modification

leaves the mutual information with P450 folding unchanged (+0.01 bit) and slightly decreases the score on beta-lactamase (−0.03 bit).

Comparing SCHEMA with other energy functions

As a baseline for comparison, I computed the mutual information between folding and mutation to the closest parent. Although mutation is not an energy function in the conventional sense, it can be used to predict folding. Not surprisingly, mutation is less effective than SCHEMA for both the P450s and beta-lactamases, with scores of 0.02 and 0.14 bits per chimera, respectively. Meyer *et al.* also concluded that SCHEMA was superior to mutation by comparing the energies of functional and nonfunctional beta-lactamases at a fixed level of mutation (Meyer *et al.*, 2003). The functional chimeras had statistically significant lower energies.

SCHEMA's use of the parental sequences, however minimal, contributes to its predictive accuracy. Consider an energy function WPS (Without Parental Sequences) that counts as disruptive any contact between residues from different parents, even if one or both positions are completely conserved. SCHEMA only penalizes a subset of these amino acid pairs, specifically those not present in any of the parents. WPS is less predictive than SCHEMA for both proteins. As shown in Figure III-2, it decreases the mutual information per chimera by 0.06 bits for the beta-lactamases and by 0.04 bits for the P450s.

. Whereas SCHEMA considers all sequence changes equally disruptive, Maranas and co-workers have proposed using the biophysical properties of amino acids to restrict

which novel residue pairs are counted (Saraf & Maranas, 2003; Saraf *et al.*, 2004). Inspired by their work, I developed the energy function CVHclash, which considers whether the additive charge, volume, or hydrophobicity (CVH) of a chimeric residue pair lies outside the range of values observed among the parents (see Methods). Unlike SCHEMA, CVHclash is able to recognize conservative substitutions and avoid penalizing novel chimeric pairs with parent-like CVH properties. Nonetheless, CVHclash does not predict folding better than SCHEMA (see Figure III-2).

I also evaluated the accuracy of the Miyazawa-Jernigan (MJ) potential, which assigns a unique energy (derived from the Protein Data Bank) to each residue-residue contact based on the identity of the interacting amino acids (Miyazawa & Jernigan, 1996). Given its easy implementation, the MJ potential has been widely used (over 800 citations) to explore folding and designability, often with lattice proteins. In their 1996 paper, Miyazawa and Jernigan showed that the potential could identify the native structure of 73 of the 88 proteins tested. In my hands the MJ potential predicts P450 and beta-lactamase folding poorly, with mutual information scores even worse than mutation (see Figure III-2).

Discussion

Despite many attempts, I have been unable to find a pairwise residue potential with better performance than SCHEMA. It is surprising that none of the biophysical refinements, particularly CVHclash, lead to improvement. By using binary contacts, SCHEMA seems to robustly account for the short-range, nearest-neighbor interactions

important for protein stability (Chen & Stites, 2001). SCHEMA then considers the structural context of each residue pair by checking whether it has been tested before by nature. This simple insight makes SCHEMA much more effective than a contact potential such as Miyazawa-Jernigan, which considers only the identity of the amino acids (Vendruscolo & Domany, 1998; Khatun *et al.*, 2004).

By weighting amino acid pairs differently depending on their structural context, I believe SCHEMA is able to partially mimic the resolving power of rotamer-based potentials (Gordon *et al.*, 1999; Mendes *et al.*, 2002). After all, a rotamer-rotamer interaction is a context-dependent, residue-residue interaction. While I expect a rotamer potential could be used to predict cytochrome P450 folding better than SCHEMA, it is challenging to evaluate the energy of several hundred, 450-residue proteins bound to a heme cofactor. Preliminary attempts to score P450 chimeras with the rotamer potential used by ORBIT (Mooers *et al.*, 2003) found that a fixed backbone approximation was too severe to compute realistic energies (E. Zollars, personal communication). The sensitivity of rotamer potentials requires a good homology model for each sequence, whereas SCHEMA is coarse enough to score all chimeras using the same parental structure.

Of the several possible reasons why SCHEMA more accurately predicts beta-lactamase folding than cytochrome P450 folding, I believe parental sequence divergence is the most significant. Whereas the beta-lactamase parents have pairwise sequence identities around 40%, the P450 parents have 65% sequence identity. As a result, the beta-lactamase substitutions are less conservative and more deeply buried in the protein,

making the structural disruptions counted by SCHEMA more likely to be deleterious. This is evident in Figure III-3, which plots the probability of folding vs. SCHEMA energy for the two proteins. Fewer novel contacts are needed to reach the $p = ½$ mark for beta-lactamase, and the transition between high and low probability is sharper, leading to higher mutual information. This parental divergence hypothesis could be tested by building another P450 library with more distantly related parents. If the mutual information between SCHEMA and folding is still low, structural differences could be responsible, but I would also suspect factors unrelated to the proteins, e.g., the assays used to score proteins as folded/functional. This possibility will be discussed further in Chapter IV.

Figure III-1. Mutual information between SCHEMA and folding. The maximum information available for the P450 data, represented by the total height of the column, is 0.98 bits per chimera. SCHEMA only captures 0.07 bits of this information, shown in solid. Because the beta-lactamase data is less balanced between folded and unfolded chimeras, less information is available (0.84 bits per chimera). Nonetheless, SCHEMA captures more of it (0.34 bits per chimera).

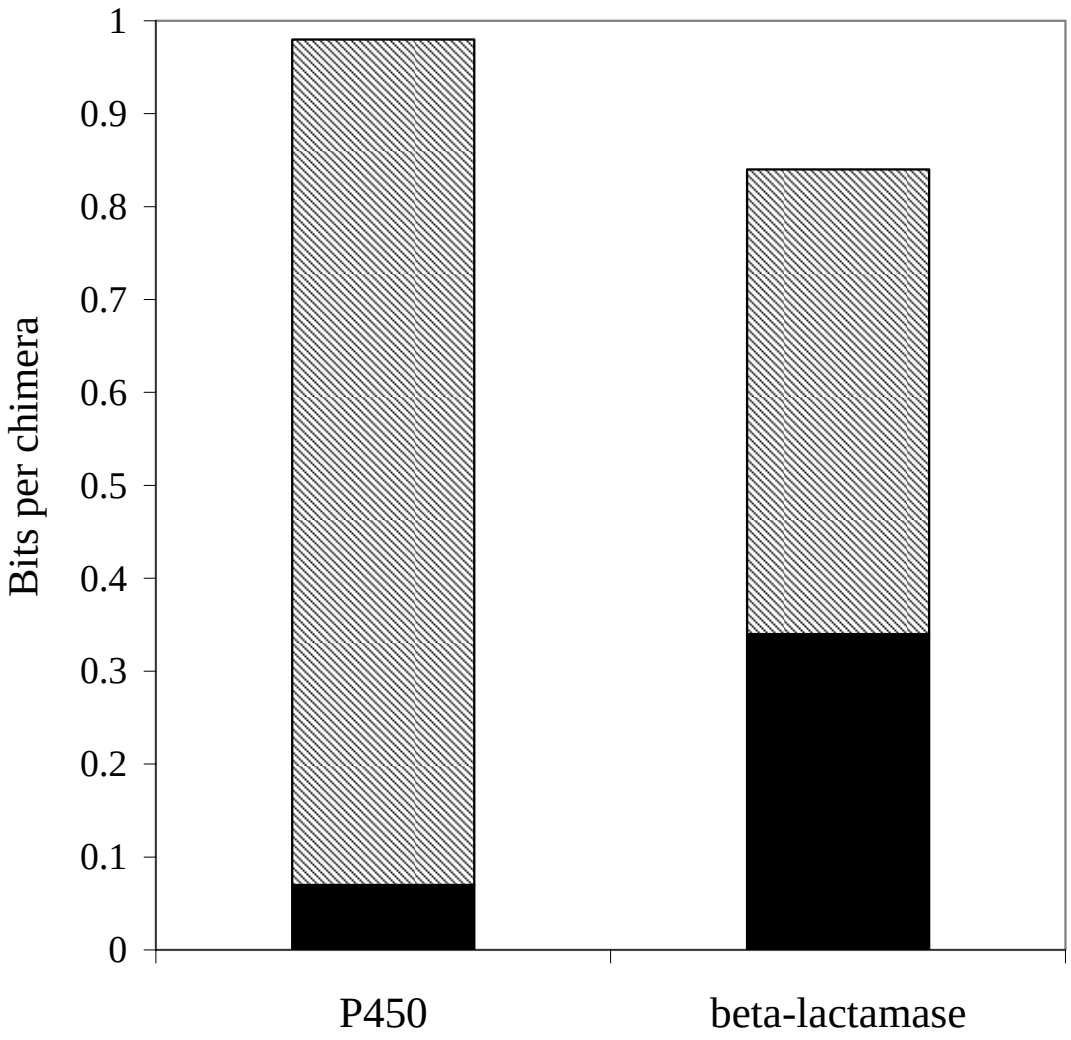

Bits per chimera
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P450
beta-lactamase

Figure III-2. Comparing the mutual information with folding for different energy functions. SCHEMA seems to use just the right amount of sequence information. By counting as disruptive all contacts between fragments from different parents (WPS, third from left), or by counting only those pairs with charge, volume, or hydrophobicity outside the range spanned by the parents (CVHclash, fourth from left), the mutual information decreases. The Miyazawa-Jernigan (MJ) contact potential has zero information on the P450 data.

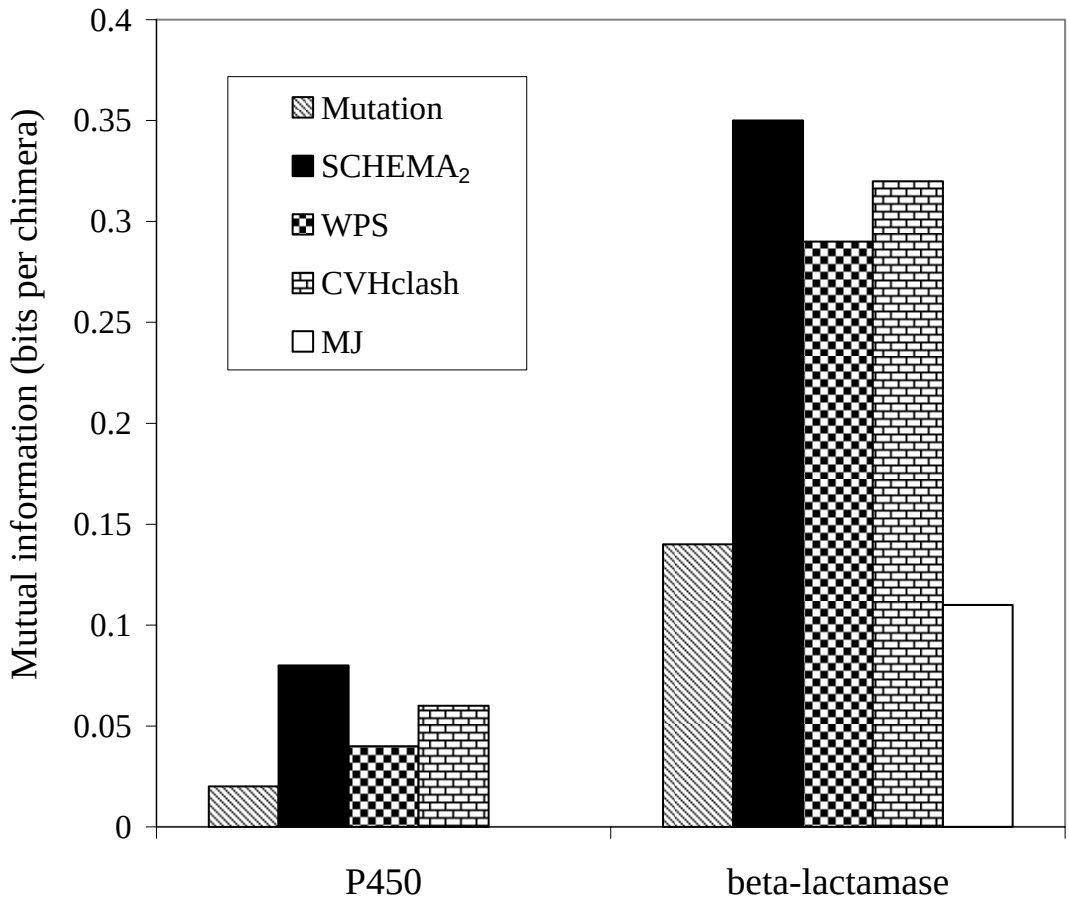
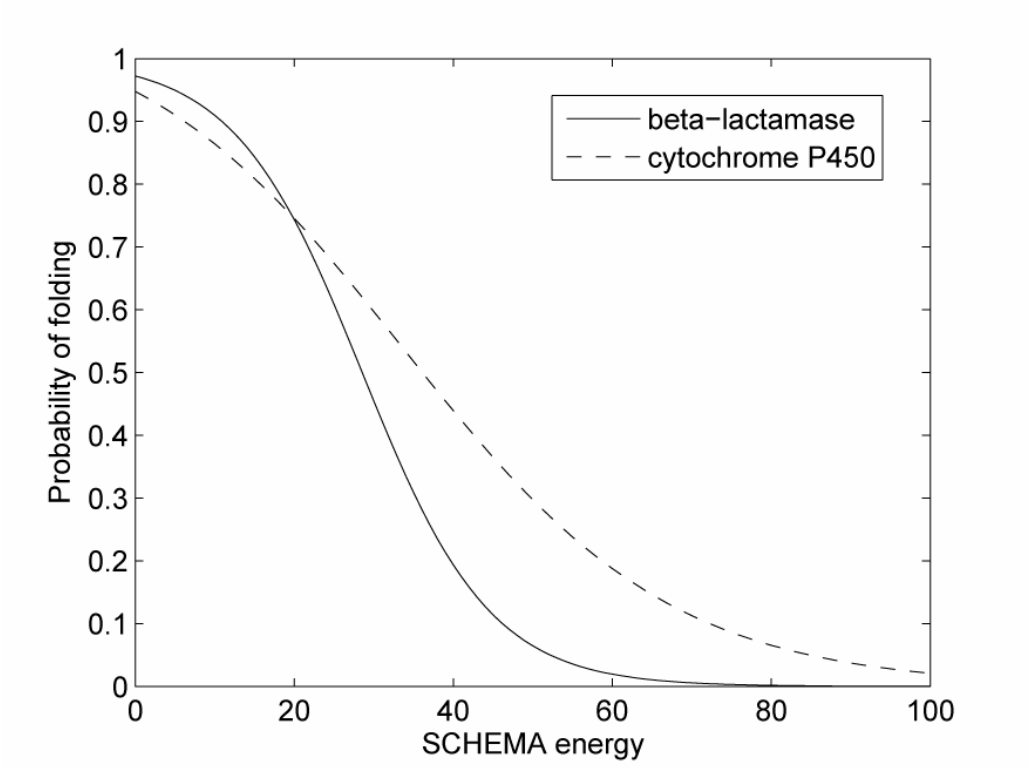

0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Mutual information (bits per chimera)
Mutation
SCHEMA2
WPS
CVHclash
MJ
P450
beta-lactamase

Figure III-3. Probability of folding vs. SCHEMA energy. The curves are maximum likelihood fits to Equation III-3. The best-fit parameters for P450 (dashed line) are $a = -2.1$, $b = 0.059$, $c = 0.93$. For beta-lactamase (solid line), the parameters are $a = -3.6$, $b = 0.12$, and c hits its upper bound at 1. The beta-lactamase fit is more step-like and thus has higher mutual information with folding.


beta−lactamase
cytochrome P450
Probability of folding
SCHEMA energy
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
20
40
60
80
100

Chapter IV

Inferring interactions from an alignment of folded and unfolded protein sequences

Chapter III illustrated the difficulty of predicting which chimeras retain the parental function and/or fold. For example, SCHEMA captures less than 10% of the maximum information available in the cytochrome P450 data. This was the best score of any energy function I tried, but there could be others with better performance. On the other hand, the missing information may simply be impossible to capture with a pairwise potential. To assess the limits of the pairwise approximation, in this chapter I fit an empirical energy function to each data set using logistic regression, an analog of linear regression for binary data (e.g., 1 = folded, 0 = not folded) that is widely used in the medical and social sciences (Hosmer & Lemeshow, 2000; Menard, 2002). The energy function, derived from an alignment of folded and unfolded proteins, is a concise way of representing sequence-function relationships. Such relationships have been generated from residue-residue correlations in alignments of naturally occurring (and hence folded) proteins by a variety of methods (Gobel *et al.,* 1994; Thomas *et al.,* 1996; Larson *et al.,* 2000; Saraf *et al.,* 2003; Suel *et al.,* 2003).

By including unfolded proteins in the multiple sequence alignment (MSA), however, it becomes easier to distinguish between two kinds of correlations. It is well known that interactions between residues lead to correlations. It is much less appreciated that residues with strong but independent influences on protein stability can also be

correlated. This correlation arises because, when inherited together, the residues are more likely to lower (or raise) the free energy beyond the threshold needed for thermodynamic stability (Wintrode & Arnold, 2001; Bloom *et al.*, 2005).

To test how well different algorithms handle the confounding effect of the stability threshold, I created a library of fictitious proteins that fold according to a specified energy model. The proteins have one of three possible fragments at each of eight variable positions. While clearly inspired by the SDR libraries from Chapter III, this fictitious library represents other combinatorial strategies equally well (Moore & Maranas, 2004), such as synthetic shuffling of designed oligonucleotides (Hayes *et al.*, 2002). In the hypothetical energy model, each peptide fragment makes an individual, or "one-body," contribution, as well as seven "two-body" interactions with the other positions (Russ & Ranganathan, 2002). The physical interpretation of these terms depends on what the positions represent. For a single residue, one-body terms include interactions with the solvent, with the backbone, and with conserved residues. For a block of residues, one-body terms also include residue-residue interactions within the block. Two-body terms represent interactions between two non-conserved residues or, for the block-level alignment, between all non-conserved residues from two blocks. It is assumed that when the total energy of a fictitious protein is above an arbitrary threshold of zero, it is unfolded; otherwise it is folded.

In addition to logistic regression, three other algorithms are compared for their ability to predict which fragments interact, using only the MSA of fictitious proteins. The first two methods, contingency table (Larson *et al.*, 2000; Kass & Horovitz, 2002;

Fodor & Aldrich, 2004) and statistical coupling (Lockless & Ranganathan, 1999; Suel *et al.*, 2003) analysis, were developed for natural protein families and hence use only the folded subset of the MSA. The third method is excess information analysis, which, like logistic regression, makes use of both the folded and unfolded sequences. Excess information is based on the mutual information between folding and a pair of positions, which is different than (but related to) the mutual information between two positions in an MSA of folded proteins (Atchley *et al.*, 2000; Fodor & Aldrich, 2004; Saraf *et al.*, 2004). Of all four methods, logistic regression is the only one to correctly predict interactions in the hypothetical energy model.

When applied to the real samples of cytochromes P450 and beta-lactamases, logistic regression proposes several sequence-function relationships consistent with the protein structures.

Methods

Construction of the fictitious library

The fictitious library contains 6,561 sequences, representing all combinations of 3 fragments at 8 positions. Fragment *i.x* refers to fragment *x* at position *i*. The total energy of each sequence is the sum of 8 one-body terms (ε_1) and 28 two-body terms (ε_2):

$$E = \sum_{i=1}^{8} \varepsilon_1(i.x) + \sum_{i=1}^{8} \sum_{j=i+1}^{8} \varepsilon_2(i.x, j.y). \qquad \text{[IV-1]}$$

There are 3 one-body parameters for every position (one per fragment) and 3×3 = 9 two-body parameters for every pair of positions (one per fragment-fragment combination).

Of the 8 positions and 28 pairs, 7 positions and 1 pair were arbitrarily selected to make energetic contributions. The parameters for these variables, listed in Tables IV-1 and IV-2, were chosen randomly from the standard normal distribution and constrained to have zero average energy. All other energy parameters were set equal to zero. The larger the differences between parameters for a particular position or pair, the more strongly it affects folding. With a stability threshold of zero energy, roughly half the library is folded (3,272 out of 6,561 sequences).

Contingency table analysis

Several variations on contingency table analysis have been used to detect correlated residues in natural protein families (Larson *et al.*, 2000; Kass & Horovitz, 2002; Fodor & Aldrich, 2004). For each pair *i-j*, first I tallied the number of times each fragment-fragment combination *i.x-j.y* was observed in the folded subset of the hypothetical library. Then I calculated the number expected if the two fragments were inherited independently (Bernstein & Bernstein, 1999). The chi-square statistic quantifies the significance of the differences between the observed and expected values:

$$\chi_{ij}^2 = \sum_x \sum_y \frac{[\text{Observed}(i.x\text{-}j.y) - \text{Expected}(i.x\text{-}j.y)]^2}{\text{Expected}(i.x\text{-}j.y)}. \qquad \text{[IV-2]}$$

The double sum is taken over all nine fragment combinations. Larger χ^2 values indicate greater deviations from the hypothesis that the positions are inherited independently.

Statistical coupling analysis

Statistical coupling analysis, developed by Ranganathan and co-workers to measure energetic coupling in natural protein families (Lockless & Ranganathan, 1999; Suel *et al.,* 2003), was adapted for the folded subset of the combinatorial library. The statistical coupling between positions *i* and *j* measures the response at position *i* when the MSA is perturbed at position *j* ($\Delta\Delta\boldsymbol{G}_{i,j}$) or vice versa ($\Delta\Delta\boldsymbol{G}_{j,i}$). In general these energy vectors will be different. $\Delta\Delta\boldsymbol{G}_{i,j}$ is the difference between the conservation energy for position *i* ($\Delta\boldsymbol{G}_i$) and a perturbed energy vector $\Delta\boldsymbol{G}_{i,\delta j}$. The ΔG_i^x component of $\Delta\boldsymbol{G}_i$ measures the probability of finding fragment *i.x* relative to a reference probability:

$$\frac{\Delta G_i^x}{kT} = \frac{1}{N}\ln\frac{P_i^x}{P^*}. \qquad \text{[IV-3]}$$

P_i^x is the binomial probability of fragment *i.x* appearing N_i^x times in a set of *N* folded proteins. Assuming a reference state where all three fragments are equally likely,

$$P_i^x = \binom{N}{N_i^x}\left(\frac{1}{3}\right)^{N_i^x}\left(\frac{2}{3}\right)^{N-N_i^x}. \qquad \text{[IV-4]}$$

The reference probability *P** obeys Equation IV-4 with the substitution *N*/3 for N_i^x.

The three components of the perturbed energy are defined similarly to Equation IV-3:

$$\frac{\Delta G^x_{i,\delta j}}{kT} = \frac{1}{N}\ln\frac{P^x_{i,\delta j}}{P^*_{\delta j}}, \qquad \text{[IV-5]}$$

except now N is the number of sequences in a subalignment containing only those folded chimeras with fragment j.1. (Equivalently, one could perturb with respect to fragment 2 or fragment 3.) The binomial probabilities $P^{x}_{i,\delta j}$ and $P^{*}_{\delta j}$ follow Equation IV-4 with the appropriate parameters from the subalignment.

Excess information analysis

The Shannon entropy $H(F)$ measures the uncertainty about folding in a set of chimeras:

$$H = -\left[p \log_2 p + (1-p)\log_2 (1-p)\right], \qquad \text{[IV-6]}$$

where p denotes the fraction folded. This uncertainty can be reduced by knowing the chimeric energies, as discussed in Chapter III, or by knowing a specific sequence feature. The conditional entropy $H(F|j.y)$, which must be less than or equal to $H(F)$, measures uncertainty when the presence or absence of fragment $j.y$ is known. It is defined by Equation IV-6 when p is replaced with the conditional probability $p(F|j.y)$, which is the fraction of chimeras with fragment $j.y$ that are also folded. When averaged over all three fragments, the conditional entropy for position j is written as

$$H(F \mid j) = \sum_{y} p(j.y) H(F \mid j.y). \qquad \text{[IV-7]}$$

The change in entropy $H(F) - H(F|j)$, which defines the mutual information $I(F{:}j)$ between folding and position j, represents how much the uncertainty about folding is reduced by knowing the sequence at position j.

The mutual information between folding and pair *i-j* is defined similarly. Given that a sequence contains fragments *i.x* and *j.y*, its probability of folding is $p(F|i.x,j.y)$, which when substituted into Equation IV-6 gives the conditional entropy $H(F|i.x, j.y)$. The conditional entropy for pair *i-j* is the average over all nine fragment combinations:

$$H(F \mid i, j) = \sum_{x} \sum_{y} p(i.x, j.y) H(F \mid i.x, j.y). \qquad \text{[IV-8]}$$

The mutual information between folding and pair *i-j* is $I(F{:}i, j) = H(F) - H(F|i, j)$. The excess information for a pair, defined as the difference between the mutual information for the pair and the mutual information of its constituent positions, $I(F{:}i, j) - I(F{:}i) - I(F{:}j)$, was used to predict interactions.

Logistic regression analysis

Both logistic (Hosmer & Lemeshow, 2000; Menard, 2002) and linear regression are special cases of the statistical methodology known as generalized linear modeling (McCullagh & Nelder, 1989; Agresti, 2002). There are three components to a generalized linear model. The random component specifies a response variable *Y* and its probability distribution. The systematic component specifies a predictor variable

$$\eta = \sum_{i} X_i \beta_i , \qquad \text{[IV-9]}$$

which is a linear combination of explanatory variables (X_i). The use of a linear predictor variable does not preclude modeling interaction effects; each interaction is simply another explanatory variable. The third component of a generalized linear model is the link function $g(\cdot)$, which specifies the relationship between the mean of the response

variable, E[Y] = μ, and the predictor variable via $g(\mu) = \eta$. The choice of link function depends on the probability distribution of the response variable. Linear regression deals with normally distributed variables, for which the link is the identity function. In logistic regression, the response variable is binary and follows the Bernoulli (binomial) distribution, for which the logit function is the appropriate link:

$$\eta = \log\left(\frac{\mu}{1-\mu}\right). \qquad \text{[IV-10]}$$

Inverting Equation IV-10 expresses the mean, which equals the probability of observing $Y = 1$, in terms of the predictor variable:

$$p(Y=1) = \mu = \frac{1}{1+e^{\eta}}. \qquad \text{[IV-11]}$$

This framework was used to model whether a protein is folded (F = 1) or not (F = 0) in the fictitious library. Each of the 3×8 = 24 fragments and 9×28 = 252 fragment-fragment pairs has a corresponding binary explanatory variable to model its presence (= 1) or absence (= 0). If the regression coefficients of these variables (β_i in Equation IV-9) are interpreted as one- and two-body energy terms (cf. Equation IV-1), then Equation IV-11 models the probability of folding $p(F|E)$ as a sigmoidally decreasing function of the energy:

$$p(F|E) = \frac{1}{1+e^{E}}. \qquad \text{[IV-12]}$$

The significance of every position and every position pair was computed relative to a reference model that includes all the one-body terms plus a constant. The energy parameters were fit by maximizing the likelihood function L, which is equivalent to

maximizing the mutual information between folding and energy (defined in Chapter III) as well as minimizing the deviance function $D = -2 \ln L$. Upon removing a position from the reference model, i.e., constraining its three one-body parameters to equal zero, the minimum deviance must increase because there are fewer parameters to fit the data. The magnitude of this increase asymptotically follows the chi-square distribution with two degrees of freedom, which was used to compute a p-value for each position (i.e., the likelihood ratio test was applied). Conversely, adding a position pair to the reference model lowers the minimum deviance, and the significance of this change was computed from the chi-square distribution with four degrees of freedom.

As just mentioned, although each position has three one-body parameters (one for each fragment), there are only two degrees of freedom because the reference energy for each position is arbitrary. To uniquely determine the one-body parameters, the average energy for each position was set equal to zero:

$$\sum_x \varepsilon_1(i.x) = 0 . \qquad \text{[IV-13]}$$

Similarly, there are only four degrees of freedom for each position pair despite the presence of $3 \times 3 = 9$ two-body parameters. The two-body terms were uniquely determined by requiring the average over each fragment index to equal zero:

$$\sum_x \varepsilon_2(i.x, j.y) = \sum_y \varepsilon_2(i.x, j.y) = 0 . \qquad \text{[IV-14]}$$

The five linearly independent constraints in Equation IV-14 were derived by considering the extent to which the two-body terms can be reconstructed with one-body parameters:

$$\varepsilon_2(i.x, j.y) = \omega(i.x) + \omega(j.y). \qquad \text{[IV-15]}$$

The space of matrices completely decomposable according to Equation IV-15 is five-dimensional, and Equation IV-14 makes the matrix of two-body terms perpendicular to this space. Thus I have chosen a two-body representation in which one-body effects are minimized.

Minimizing the deviance subject to the linear constraints in Equations IV-13 and IV-14 is a convex optimization problem in the one- and two-body energy parameters, which means local optimization algorithms converge to the global minimum (Boyd & Vandenberghe, 2004). I used the algorithm MINOS through the NEOS server for optimization (Czyzyk *et al.*, 1998).

Logistic regression analysis of the cytochrome P450 and beta-lactamase data

For both data sets I first calculated chi-square p-values for the 8 blocks and 28 block pairs as described above. In the P450 data set, only eight of the nine possible fragment-fragment combinations are present for block pairs 1-4 and 4-5, so three degrees of freedom were used instead of four. The likelihood ratio test identified as significant six variables for P450 and five variables for beta-lactamase, which were collected into a second-round reference model for further analysis. Upon removing each variable from this model, the chi-square p-value was again calculated. For the P450 data, I also scored the change in deviance using tenfold cross-validation. Initially the data were split into ten random partitions of equal size. For each partition, the energy model was fit to the other 90% of the data and scored by its deviance on the remaining 10%. Unlike the likelihood ratio test, this cross-validation score does not necessarily increase upon

removing each variable. When the score is positive, it indicates the variable is significant. The average increase ± standard deviation for the ten partitions is reported.

Results

Analysis of the fictitious library

The folding status for each of the 3^8 = 6,561 sequences in the fictitious library was determined according to the hypothetical energy model summarized in Figure IV-1A. The diagonal entries of this 8×8 matrix represent the individual, or one-body, contributions of the 8 positions, and the off-diagonal entries represent the interaction, or two-body, strengths of the position pairs. In order of decreasing one-body strength, the positions are 4, 5, 7, 8, 3, 1, 6, and 2. The only nonzero two-body interaction is between positions 2 and 7.

Panels B through E in Figure IV-1 show the predictions of four different algorithms. Except for contingency table analysis (panel B), which does not score the one-body terms, the algorithms make qualitatively correct predictions about the relative importance of the individual contributions made by the positions.

However, there are problems with the predicted two-body interactions for all algorithms except logistic regression (panel E). The contingency table (panel B), statistical coupling (panel C), and excess information (panel D) algorithms all score pair 4-5 as having the strongest interaction. These positions do not interact in the energy model, but their fragment frequencies are strongly correlated because fragments 4.1 and 5.3 are highly stabilizing at their respective positions (see Table IV-1; fragment *i.x* refers

to fragment *i* at position *x*). The stronger the individual contribution of a position, the more frequently it is predicted to make spurious interactions in panels B, C, and D of Figure IV-1. Statistical coupling analysis appears most susceptible to this error, followed by the contingency table and excess information algorithms. The one true interaction in the energy model, between positions 2 and 7, is barely in the top ten picked by statistical coupling analysis but fares better with the other two algorithms. All three methods assign pair 2-7 a lower score than at least one spurious pair, making it is impossible to separate true interactions from false ones.

This problem does not plague logistic regression, as is clear from a comparison of Figure IV-1E with the "answer key" in Figure IV-1A. No other pair even comes close to the score given 2-7. The success of logistic regression stems not only from its use of the unfolded sequences, a characteristic it shares with the excess information analysis depicted in Figure IV-1D. Its distinguishing feature is the ability to directly test for energetic coupling with a sigmoidal folding model (Equation IV-12). The other three algorithms are only able to test for probabilistic coupling, i.e., whether the probability distributions for two positions are independent. This is not a reliable indicator of energetic coupling when folding is a nonlinear function of energy.

For the real protein data sets, only a subset of the library is available for analysis (806 P450s, 605 beta-lactamases; see Appendix B), which makes it more difficult to infer significant interactions. This effect was simulated with the fictitious library by applying logistic regression to the same 806 chimeric patterns, e.g., 11221233, for which P450 data are available. The 8×8 matrix of p-values looks just like a rescaled version of Figure

IV-1E, indicating that all hypothetical interactions are predicted correctly. The least significant true prediction is position 6, with a p-value of 10^{-9}, and the most significant spurious prediction is pair 2-8, with a p-value of 10^{-3}. These p-values represent lower and upper bounds on the significance threshold that should be used with the real data to avoid false positives (much different than $p < 0.05$!).

Logistic regression analysis of the cytochrome P450 and beta-lactamase data

Figure IV-2 shows the hierarchy of p-values from logistic regression analysis of the cytochrome P450 data. Blocks 5, 1, 7, and the pair 1-7 are clearly significant, with p-values less than 10^{-16}. A handful of other variables are marginally significant, with p-values in between the thresholds of 10^{-9} and 10^{-3} established with the fictitious library. The top two from this marginal list, pairs 1-5 and 5-8, were grouped with blocks 1, 5, 7, and pair 1-7 to create an energy model with six variables. The significance of each variable in this model was recalculated using tenfold cross-validation and the likelihood ratio test, which confirmed 1, 5, 7, and 1-7 but ruled out 1-5 and 5-8.

Logistic regression analysis of the beta-lactamase data, shown in Figure IV-3, identified five variables as strongly significant (1, 2, 3, 8, 1-8) and three others as marginally significant (5, 1-7, 2-8). Based on my experience with the P450 data, I only subjected the first five to further testing. When the p-values of blocks 1 and 8 were recalculated relative to a model that includes pair 1-8, their significance diminished considerably ($p = 0.5$ and 4×10^{-3}, respectively). This result stands in contrast to the behavior of pair 1-7 in the P450 protein, whose constituent blocks remained highly

significant even in the presence of the pair. The variables 2, 3, and 1-8 remained significant after the second round of testing.

Discussion

Comparison with studies of natural protein families

My results with the fictitious library may help explain the difficulty of detecting structural contacts from correlations in a sequence alignment of natural proteins. In their pooled analysis of 224 families, Fodor and Aldrich found that for even the best algorithms, at least 75% of the predicted interactions were between non-contacting residues (Fodor & Aldrich, 2004). Larson *et al.* had greater success, with error rates between 30% and 50% for several of the 15 families they studied, although far fewer predictions were made (Larson *et al.*, 2000). Whereas Fodor and Aldrich included sequences with as high as 90% pairwise identity in the alignment, Larson *et al.* used a diversity cap of 40% identity. In hindsight, the large number of spurious predictions in these studies is not surprising. I expect many of the correlations detected were the result of a stability threshold rather than residue-residue interactions.

Reconciling logistic regression energies with protein structure

One would expect the two significant block pairs identified by logistic regression, beta-lactamase pair 1-8 and cytochrome P450 pair 1-7, to be important from an examination of the parental structures. In both proteins the blocks form two halves of a beta-sheet (see Figures IV-4A and IV-5), and the beta-lactamase pair is also coupled

through a pair of adjacent alpha helices (red and green in Figure IV-5). Tables IV-3 and IV-4 show the 3×3 matrix of regression energies for P450 and beta-lactamase, respectively. In both cases the on-diagonal entries are the most stabilizing (negative), which means chimeras that inherit the blocks from the same parent, rather than different parents, are more likely to be folded/functional. This result supports the core hypothesis of SCHEMA, which assigns the best possible score (zero) to these wild-type interactions.

Among the significant blocks, beta-lactamase block 2 is the most readily understandable. It contains eight residues, including a conserved serine critical to the enzyme's catalytic function (Matagne *et al.*, 1998). In total, four positions are conserved across the parents, shown in black in the alignment of Figure IV-6. Of the remaining four positions, site-directed mutagenesis of parent TEM-1 has shown that two are tolerant to the amino acids found in SED-1 and PSE-4 (Huang *et al.*, 1996). These residues are colored green in Figure IV-6. At the positions colored red, TEM-1 is not tolerant to the amino acids found in SED-1 but is identical to PSE-4. These residues may explain why the one-body energy for SED-1 is destabilizing relative to TEM-1 and PSE-4 at block 2 (see Table IV-5).

The importance of cytochrome P450 block 5 may reflect its role in the dynamics of the enzyme, as suggested by a comparison of the crystal structures for the substrate-bound (Li & Poulos, 1997) and unbound (Ravichandran *et al.*, 1993) forms of parent CYP102A1. The F and G helices of CYP102A1 identified in Figure IV-4B close down over the substrate when it binds, and block 5 (colored green) sits at the "hinge" of this flap. On average its residues move 3.6 Å, compared with 1 to 2 Å for most of the other

blocks. The one-body energies in Table IV-6 show that CYP102A1 is destabilizing compared to A2 and A3, but it is difficult to pinpoint which residues are responsible. Unlike beta-lactamase block 2, which contains 4 variable positions, there are 17 in cytochrome P450 block 5, many of which form contacts within the block and with conserved residues outside the block. The instability of A1 may simply be due to fragment length, since the A1 fragment is one amino acid shorter than the other two.

Mutual information between logistic regression energy and folding

In Chapter III the predictive accuracy of the SCHEMA energy was quantified by its mutual information with folding. Mutual information is also appropriate for evaluating the energy model produced by logistic regression. For the sample of 806 fictitious proteins, 43% of which are folded, the maximum mutual information is 0.98 bits per sequence, all of which is captured by logistic regression. Since folding in the fictitious library was determined by a pairwise energy model of the same form fit by logistic regression, it makes sense that perfect information is attained.

Substantially less information is captured by logistic regression of the real proteins, as shown in Figure IV-7. The maximum mutual information scores for the beta-lactamase and P450 samples are 0.84 and 0.98 bits per chimera, respectively. The beta-lactamase sample has lower entropy because only 27% of the chimeras are folded, compared with 58% of the P450s. When limited to statistically significant terms, the dramatic differences between the two proteins observed in Chapter III disappear. Logistic regression captures 0.38 and 0.31 bits per chimera for beta-lactamase and

cytochrome P450, respectively, compared with a mutual information differential of 0.34 (beta-lactamase) to 0.07 (P450) with SCHEMA.

I now return to the questions raised at the end of Chapter III and beginning of Chapter IV. Although the mutual information between beta-lactamase folding and SCHEMA energy is only 40% of the maximum possible, SCHEMA performs as well as could be expected for a pairwise potential. Even by fitting an energy model to the data, logistic regression only captures 45% of the maximum information. Since the pairwise energy approximation is equally limiting for the two proteins, it is tempting to conclude that the poorer performance by SCHEMA on the P450 data is due to biophysical factors rather than artifacts from the folding assay. Two other observations suggest otherwise. For one, it is striking that all the predictors tested in Chapter III, including mutation, scored higher on the beta-lactamase data set. In addition, when the significance constraints in logistic regression are relaxed to include all one- and two-body terms (which leads to overfitting), the beta-lactamase data are 100% predicted by the empirical energy function. In contrast, the mutual information with P450 folding is still only 0.56 bits per chimera, or 57% of maximum.

Future work

One can envision applications of logistic regression analysis other than those presented here. For protein families without structural information, logistic regression could help test putative interactions or otherwise refine a homology model. Because the sigmoid used by logistic regression to link sequence and function (Equation IV-12) is

quite general, other properties besides protein folding could be modeled, including the fitness effects of deleterious mutations (Saunders & Baker, 2002). For the cytochrome P450 and beta-lactamase proteins, the sequence determinants of enzymatic catalysis could be studied by screening the libraries for activity on a particular substrate.

Table IV-1. One-body terms in the hypothetical energy model.

Position	Fragment 1	Fragment 2	Fragment 3
1	0.56	0.21	-0.77
2	0	0	0
3	0.83	-1.00	0.17
4	-2.24	-0.45	2.69
5	1.10	1.40	-2.50
6	-0.00	-0.46	0.46
7	-1.61	0.03	1.58
8	-1.23	0.80	0.43

Table IV-2. Two-body terms for pair 2-7 in the hypothetical energy model. All other two-body terms are zero.

	7.1	**7.2**	**7.3**
2.1	0.46	0.38	-0.84
2.2	0.63	-1.01	0.38
2.3	-1.09	0.63	0.46

Table IV-3. Two-body energies for cytochrome P450 pair 1-7 (parent *x* = CYP102A*x*).

	7.1	**7.2**	**7.3**
1.1	-0.9	1.3	-0.4
1.2	0.1	-1.3	1.2
1.3	0.8	-0.0	-0.8

Table IV-4. Two-body energies for beta-lactamase pair 1-8 (parents 1, 2, 3 = PSE-4, SED-1, TEM-1, respectively).

	8.1	**8.2**	**8.3**
1.1	-1.2	1.7	-0.5
1.2	-0.1	-2.8	2.8
1.3	1.3	1.0	-2.3

Table IV-5. One-body energies for the significant beta-lactamase blocks (ε_0 = 3.045).

Block	Parent 1 (PSE-4)	Parent 2 (SED-1)	Parent 3 (TEM-1)
2	-0.5	1.1	-0.5
3	0.6	1.1	-1.7

Table IV-6. One-body energies for the significant cytochrome P450 blocks (ε_0 = 0.084).

Block	CYP102A1	CYP102A2	CYP102A3
1	0.5	-1.0	0.5
5	1.4	-0.8	-0.6
7	0.3	1.0	-1.3

Figure IV-1. Logistic regression is the only algorithm to correctly predict pairwise interactions in a library of fictitious proteins. **(A) Energy model.** This symmetric 8×8 matrix summarizes the hypothetical energy model used to fold the fictitious proteins (Tables IV-1 and IV-2). The diagonal entries reflect the individual contribution of each position, and the off-diagonal entries represent pairwise interactions. The shade of each square encodes the standard deviation across energy parameters, which is a measure of energetic "strength." Because the average energy for each variable is arbitrarily constrained to zero, the standard deviation is $\sqrt{\frac{1}{3}\sum_{x}\left[\varepsilon_1(i.x)\right]^2}$ for position *i* and $\sqrt{\frac{1}{9}\sum_{x}\sum_{y}\left[\varepsilon_2(i.x, j.y)\right]^2}$ for pair *i*-*j*. **(B) Contingency table analysis.** The shade of each off-diagonal entry encodes the chi-square statistic with four degrees of freedom. Larger values indicate the positions are not inherited independently, but this does not mean they interact energetically. This is clear from a comparison with the energy model in panel A. The diagonal entries are blank because contingency table analysis does not score one-body terms. **(C) Statistical coupling analysis.** Conservation energies are shown on the diagonal and coupling energies are shown off the diagonal. The shade of each square represents the magnitude of the corresponding statistical energy vector, in units where kT = 100. Although the one-body effects are predicted well, the two-body interactions are not. **(D) Excess information analysis**. The shade of each diagonal entry represents the mutual information with folding for that position, in bits per sequence. The off-diagonal entries reflect the excess information for each pair in bits per sequence. Although this algorithm uses both folded and unfolded sequences, there is little improvement over the contingency table analysis in panel B. **(E) Logistic regression analysis.** The shade of each square represents the significance p-value from the likelihood ratio test, reported as $-\log_{10}(p)$ so that higher numbers are more significant. Panels A and E are nearly indistinguishable, which means logistic regression is able to determine all the important energy terms and their relative strengths.

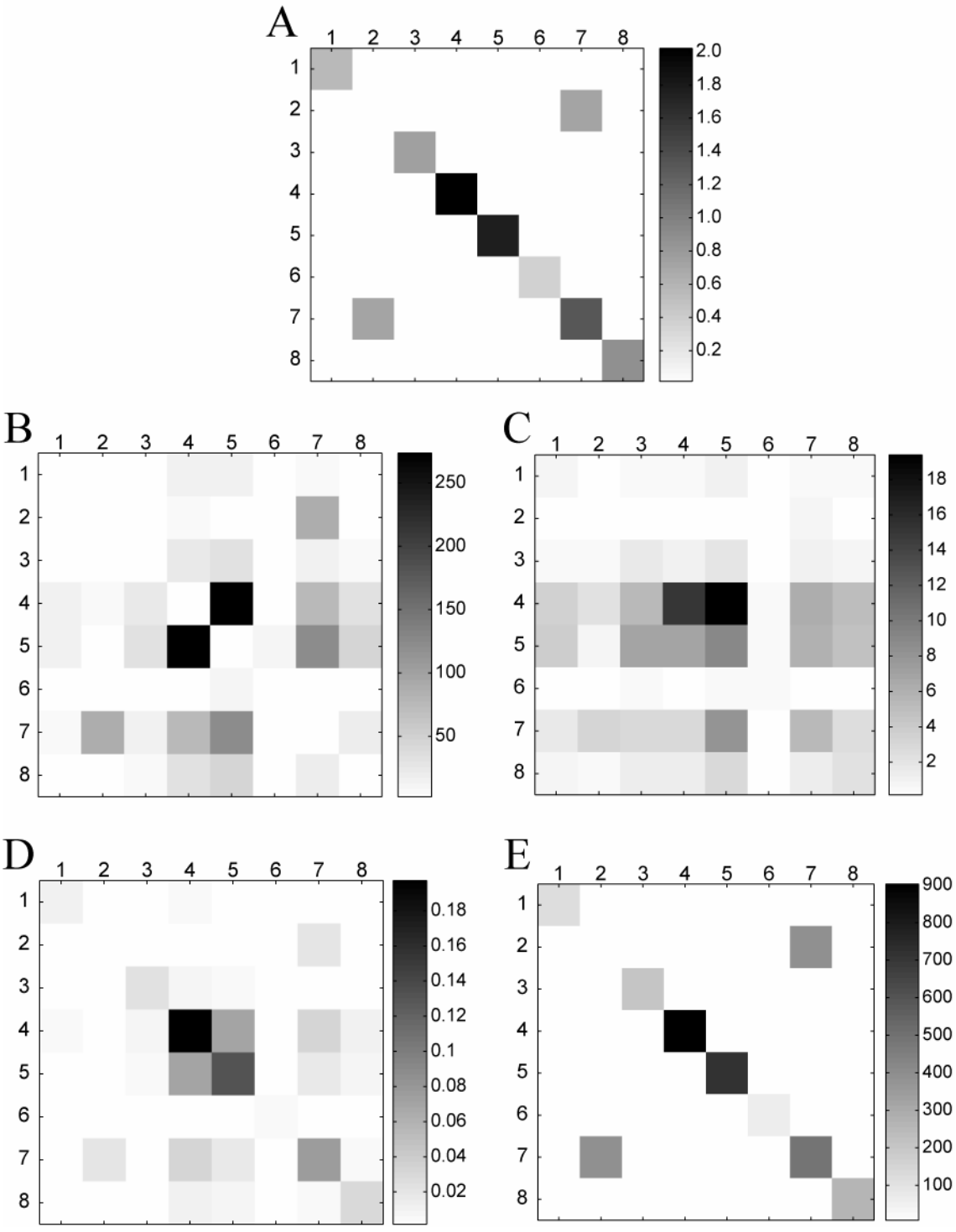

A
1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8
2.0 1.8 1.6 1.4 1.2 1.0 0.8 0.6 0.4 0.2
B
1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8
250 200 150 100 50
C
1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8
18 16 14 12 10 8 6 4 2
D
1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8
0.18 0.16 0.14 0.12 0.1 0.08 0.06 0.04 0.02
E
1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8
900 800 700 600 500 400 300 200 100

Figure IV-2. Logistic regression analysis of 806 cytochrome P450 chimeras. The significance of each block and block pair, reported as $-\log_{10}(p)$, is shown relative to a model with all one-body terms. Blocks 1, 5, 7, and pair 1-7 are the most significant with this test; many other variables appear marginally significant, in particular pairs 1-5 and 5-8. All six were grouped into a second round model for further testing (data not shown), which suggested pairs 1-5 and 5-8 should not be included. Their p-values were 0.01 and 0.02, respectively, with cross-validation scores of 0 ± 3 and 1 ± 2 (positive means significant). By contrast the least significant variable included was block 5, with $p < 10^{-7}$ and a cross-validation score of 3 ± 4.

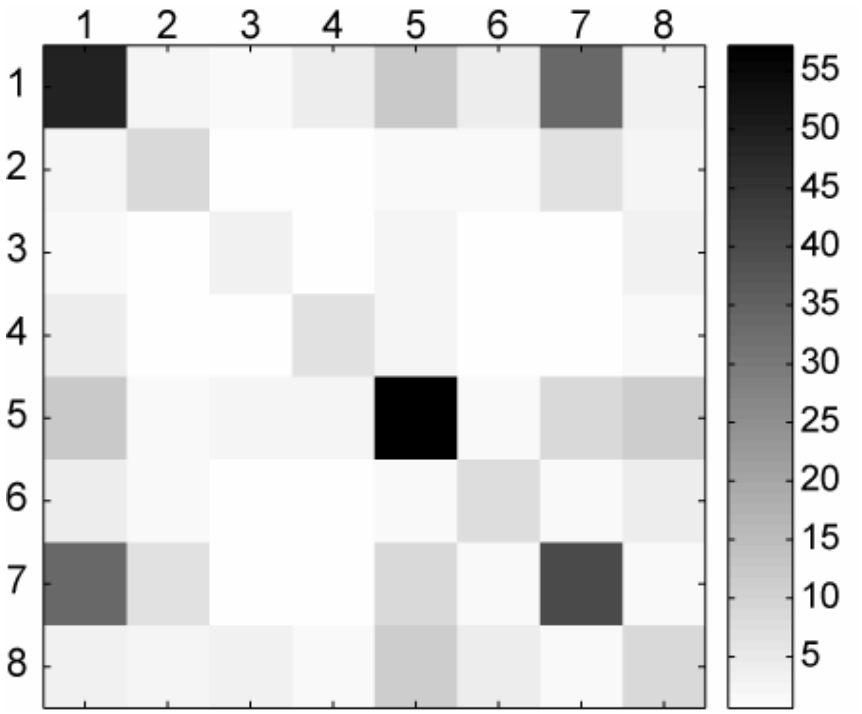

1
2
3
4
5
6
7
8
1
2
3
4
5
6
7
8
55
50
45
40
35
30
25
20
15
10
5

Figure IV-3. Logistic regression analysis of 605 beta-lactamase chimeras. The significance of each block and block pair, reported as $-\log_{10}(p)$, is shown relative to a model with all one-body terms. Blocks 1, 2, 3, 8, and pair 1-8 are the most significant with this test. All five were grouped into a second-round model for further testing (data not shown), which suggested blocks 1 and 8 should not be included ($p = 0.5$ and 4×10^{-3}, respectively).

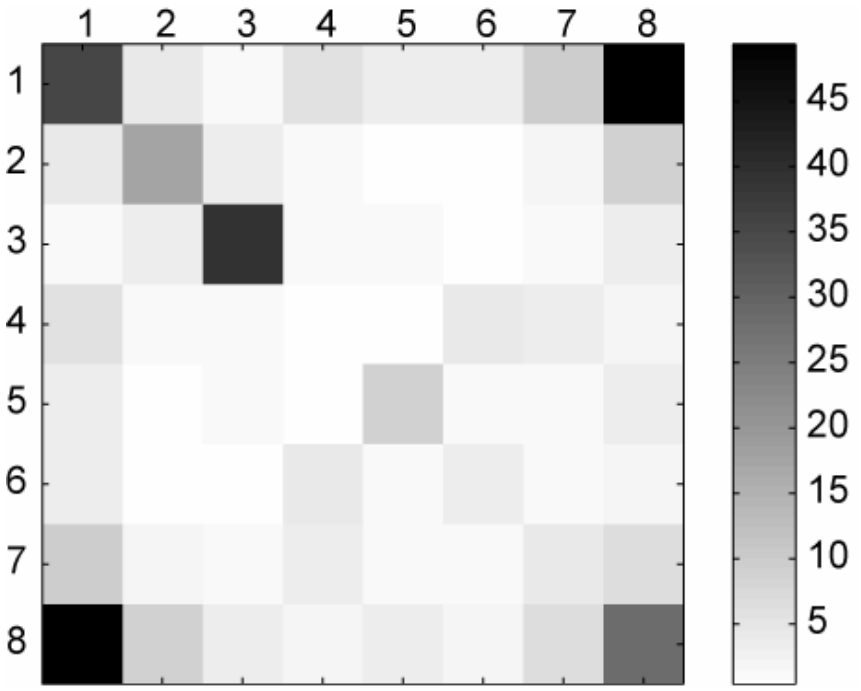

1
2
3
4
5
6
7
8
1
2
3
4
5
6
7
8
45
40
35
30
25
20
15
10
5

Figure IV-4. Structural analysis of pair 1-7 and block 5 for cytochrome P450 parent CYP102A1 (Ravichandran *et al.*, 1993). **(A)** Blocks 1 (pink) and 7 (blue) make extensive contacts as two halves of a beta-sheet. **(B)** The F and G helices form a flap that folds down over the substrate, and block 5 (green) sits at the hinge of this flap.

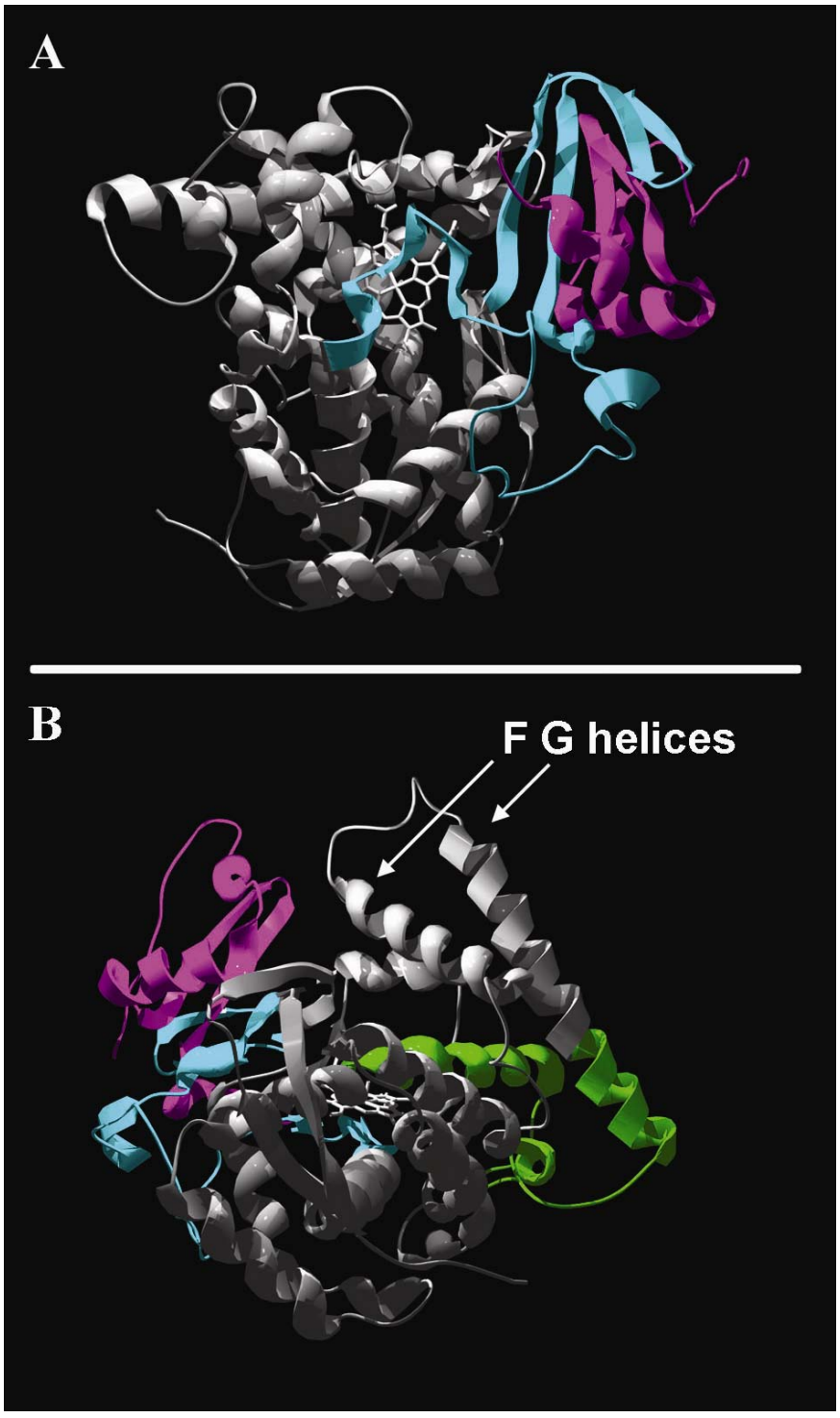

A
B
F G helices

Figure IV-5. Structural context of the eight blocks in the beta-lactamase library. The blocks are color-coded and mapped onto the structure of TEM-1 (Jelsch *et al.*, 1993): 1 = red, 2 = pink, 3 = dark blue, 4 = yellow, 5 = orange, 6 = gray, 7= light blue, 8 = green. Blocks 1 and 8 make extensive contacts through their terminal alpha helices and joint beta-sheet. The small molecule shown in black is an inhibitor bound to the active site.

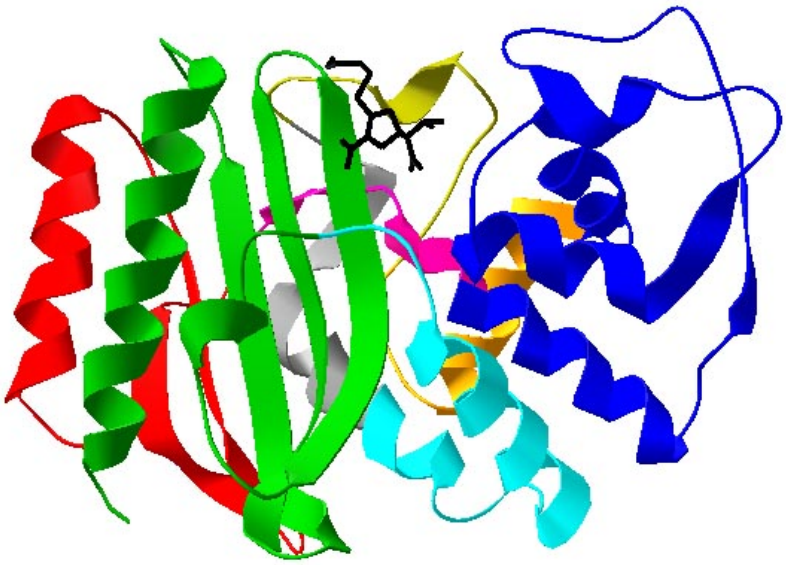


Figure IV-6. Alignment of the beta-lactamase parents for block 2. Four of the eight residues in block 2 are variable (colored) and four are conserved (black). Site-directed mutagenesis (Huang *et al.*, 1996) has shown that TEM-1 can tolerate the amino acids found in SED-1 at green but not red positions. This may explain why SED-1 is destabilizing in the regression energy model of Table IV-5.

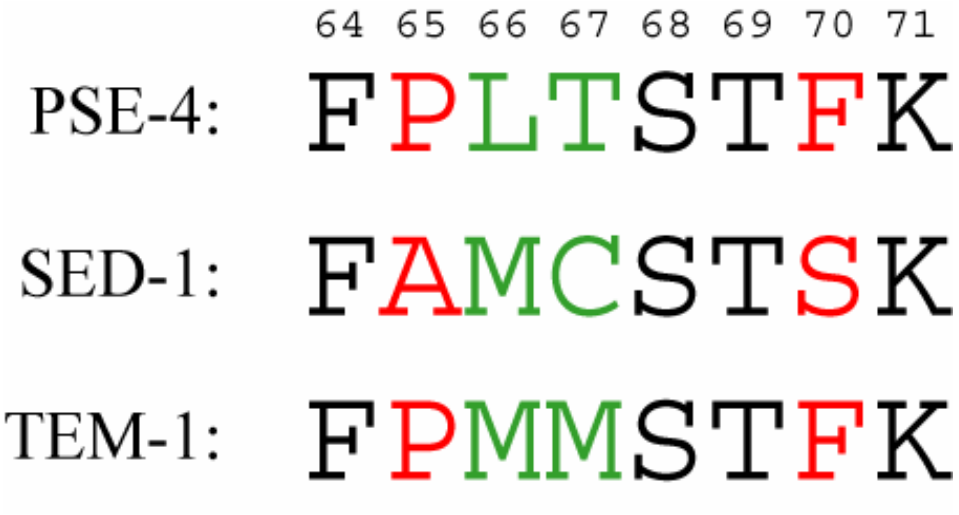
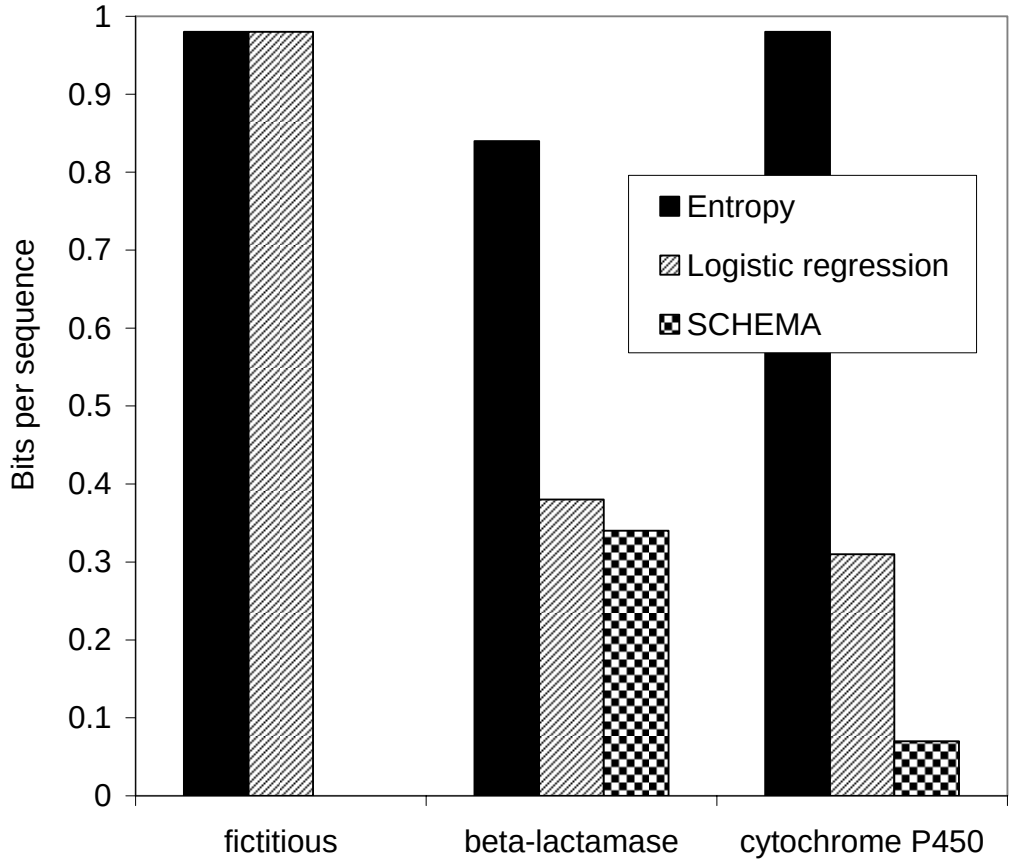

64 65 66 67 68 69 70 71
PSE-4: FPLTSTFK
SED-1: FAMCSTSK
TEM-1: FPMMSTFK

Figure IV-7. Comparing logistic regression with SCHEMA. The maximum mutual information, or Shannon entropy, is shown with the solid bars. Logistic regression (striped bars) captures 100% of the information for the fictitious library compared to 44% for the beta-lactamases and 32% for the P450s. The SCHEMA energy, shown with checkered bars for the real proteins, captures nearly as much mutual information with beta-lactamase folding/function as logistic regression.


Bits per sequence
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Entropy
Logistic regression
SCHEMA
fictitious
beta-lactamase
cytochrome P450

Appendix A

Chimeric sequence design is NP-hard

In Chapter II a contrast was drawn between the difficulty of finding the site-directed recombination (SDR) library with minimum energy (Equation II-1) and that of finding the amino acid sequence with minimum energy. The former can be solved by dynamic programming in $O(N^3)$ operations, where N is the length of the protein, while the latter is NP-hard (Pierce & Winfree, 2002). By NP-hard I mean the corresponding decision problem (Does there exist an amino acid sequence with energy less than some constant?) is NP-complete. NP-complete problems have been proven to be as hard as any other in the complexity class NP, which includes most problems of practical interest (Papadimitriou, 1994). For this reason it is unlikely (but not yet disproven) that polynomial-time algorithms exist. Exact solutions are best found with algorithms that intelligently traverse a search tree, e.g., branch-and-bound (Papadimitriou & Steiglitz, 1998), or one can look for approximate solutions (Hochbaum, 1997). RASPP can be viewed as an approximation algorithm for finding which SDR library has the highest fraction folded in a given range of library diversity. The nonlinear relationship between folding and SCHEMA energy, e.g., a single exponential, makes this problem hard, but I have been unable to prove anything about it.

I have proven, however, that within the context of sequence design, choosing crossovers between two parents can be as hard as choosing amino acids from the full

alphabet, i.e., both are NP-hard. Consider finding the chimeric sequence with minimum energy, subject to constraints on its level of mutation:

$$\min_{(x_1,x_2,\ldots,x_n)} \quad E \qquad \text{[A-1]}$$

$$\text{subject to} \quad m_{min} \leqslant m \leqslant m_{max}.$$

The decision problem for Equation A-1, called CSD for Chimeric Sequence Design, is whether there exists a chimera with energy $E \leqslant K$ and $m_{min} \leqslant m \leqslant m_{max}$. To prove CSD is NP-complete, I show that any algorithm which solves CSD in polynomial time could be used to solve another NP-complete problem in polynomial time (Sipser, 1997).

The problem I reduce to CSD is that of finding a balanced cut in an undirected graph (Bui & Jones, 1992). For an undirected graph G with N vertices, an α-balanced cut is a partition of the vertices into two disjoint sets V_1 and V_2 such that neither contains less than αN vertices, where α is some specified fraction. The cost of this cut is the number of edges with one endpoint in V_1 and the other endpoint in V_2. It is NP-complete to decide if there exists an α-balanced cut with cost $C \leqslant K$.

To construct a polynomial-time reduction to CSD, assign each vertex in G to a residue position in an alignment of two parents with zero sequence identity. When G is partitioned, every residue whose vertex is assigned to V_1 is inherited from parent 1, and every residue whose vertex is assigned to V_2 is inherited from parent 2. Under this mapping, the number of vertices in the smaller partition equals the mutation level m of the corresponding chimera. Choose an energy function such that if two vertices are connected in G, the interaction energy between the corresponding residues is 1 when they

are inherited from different parents and 0 when inherited from the same parent. All other pairwise energies are zero, and there are no one-body energies. With this prescription, the cost C of a cut equals the energy E of the chimera, and the α-balanced cut problem reduces to CSD with $m_{min} = \alpha N$ and $m_{max} = N/2$.

Appendix B

Folding data for cytochromes P450 and beta-lactamases

Table B-1. Folding status of 806 cytochrome P450 chimeras (C. Otey, personal communication). Chimeras were scored as folded (= 1) if and only if they bind a heme cofactor. Homologs CYP102A1, CYP102A2, and CYP102A3 are parents 1, 2, and 3, respectively (Nelson, 2005). The eight blocks are defined by crossovers after residues Glu64, Ile122, Tyr166, Val216, Thr268, Ala328, and Gln404, numbered from the N-terminus of CYP102A1.

11112123	0
11112212	1
11113223	0
11113233	1
11131313	1
11132223	0
11132232	0
11132323	0
11133231	1
11212333	1
11213133	1
11213231	1
11231232	0
11232111	0
11232232	1
11232323	0
11232333	1
11311233	1
11312233	1
11313223	0
11313233	1
11313333	1

11331123	0
11331312	1
11331333	1
11332221	0
11332233	1
11332333	1
11333122	0
11333212	1
11333323	0
12112333	1
12133223	0
12211222	0
12211232	1
12211333	1
12212112	1
12212211	0
12212223	0
12212332	1
12231231	0
12232111	0
12232232	1
12232233	1

12232332	1
12233112	0
12233323	0
12311333	1
12313331	1
12322333	1
12331123	1
12331211	0
12331221	0
12331333	1
12332123	0
12332223	1
12332233	1
12332333	1
12333331	1
12333333	1
13132223	0
13132322	0
13132333	0
13133323	0
13212122	0
13212321	0

13213131	1
13231332	0
13232123	0
13232311	0
13232323	0
13233133	0
13233212	1
13233233	0
13233322	0
13311311	0
13331123	0
13331333	0
13332223	0
13332332	0
13332333	1
13333122	1
13333123	0
13333131	1
13333222	0
13333223	0
13333233	0
13333323	0

21111112	0
21111212	1
21111312	0
21111321	1
21111322	0
21111323	1
21111333	1
21112123	1
21112212	1
21112222	1
21112232	1
21112312	1
21112322	1
21113111	1
21113112	1
21113212	1
21113221	1
21113222	0
21113223	1
21113322	1
21131111	0
21131121	1
21131212	0
21131321	0
21132112	1
21132113	1
21132121	0
21132222	1
21132311	1
21132313	1
21132323	1
21133123	0
21133131	1
21133212	1
21133222	1
21133223	1
21133232	1
21133233	1
21133313	1
21133321	1
21133322	1
21133331	1

21133332	1
21211113	0
21211122	0
21211211	0
21211222	0
21211223	1
21211321	1
21212112	0
21212122	1
21212123	1
21212212	0
21212231	1
21212333	1
21213121	1
21213212	1
21213231	1
21222112	1
21231233	1
21232112	1
21232122	1
21232132	1
21232212	0
21232222	1
21232231	1
21232233	1
21232321	1
21232332	1
21233112	0
21233132	1
21233212	1
21233221	1
21233233	1
21233312	1
21233322	0
21311111	0
21311122	1
21311223	1
21311311	0
21311331	0
21311333	0
21312111	1
21312112	1

21312121	0
21312123	1
21312212	0
21312222	1
21312223	1
21312321	0
21312322	1
21312323	1
21313112	1
21313122	1
21313231	1
21313312	1
21313313	1
21313322	1
21331111	0
21331112	0
21331131	0
21331223	1
21331312	0
21331313	0
21331332	1
21331333	1
21332113	1
21332131	1
21332212	1
21332223	1
21332231	1
21332233	1
21332323	1
21332331	1
21332332	1
21332333	1
21333121	0
21333132	1
21333133	0
21333212	1
21333221	1
21333223	1
21333233	1
21333312	1
21333331	0
21333332	0

22111223	1
22111231	0
22111332	1
22112111	1
22112223	1
22112321	1
22112323	1
22113132	0
22113223	1
22113233	1
22113313	1
22113323	1
22121331	0
22131111	0
22131112	0
22131132	0
22131133	0
22131221	1
22131222	0
22131321	0
22132233	1
22132312	1
22132323	1
22132331	1
22133112	1
22133211	1
22133212	1
22133232	1
22133312	1
22133323	1
22211132	0
22211222	0
22211331	0
22211332	0
22212232	1
22212312	1
22212322	1
22213111	1
22213112	1
22213222	1
22213223	1
22213312	1

22213321	1
22231113	0
22231122	0
22231211	0
22231212	0
22231221	1
22231223	1
22231232	0
22231311	0
22231312	0
22231333	0
22232112	1
22232122	1
22232212	1
22232213	0
22232222	1
22232223	1
22232233	1
22232312	1
22232322	1
22232323	1
22232333	1
22233112	1
22233221	1
22233222	1
22233223	1
22233323	1
22233331	0
22233332	1
22233333	0
22311121	0
22311123	1
22311212	0
22311231	0
22311332	0
22312123	1
22312132	0
22312211	1
22312221	1
22312222	1
22312223	1
22312231	1

22312232	1
22312312	1
22312322	1
22312333	1
22313221	1
22313222	1
22313232	1
22313233	1
22313323	1
22313331	1
22313333	0
22331121	0
22331123	1
22331133	0
22331211	0
22331212	0
22331221	1
22331222	0
22331223	1
22331321	0
22331323	1
22331332	0
22332112	1
22332113	1
22332123	1
22332132	1
22332211	1
22332222	1
22332223	1
22332232	1
22332233	1
22332312	1
22332321	1
22332322	1
22332332	1
22333111	0
22333122	1
22333132	1
22333133	1
22333212	1
22333221	1
22333222	1

22333223	1
22333231	1
22333313	1
22333323	1
22333332	1
23111112	0
23111212	0
23112123	0
23112213	0
23112221	1
23112222	0
23112233	1
23112323	1
23112333	1
23113111	1
23113121	1
23113212	1
23113311	1
23113312	1
23113323	1
23122212	1
23131323	1
23131332	0
23132111	0
23132121	1
23132212	1
23132221	1
23132231	1
23132232	1
23132233	1
23132322	0
23132323	1
23133112	1
23133121	1
23133233	1
23133311	1
23133312	0
23133321	1
23133333	1
23211121	0
23211131	0
23211132	1

23211222	0
23211311	0
23211332	0
23212112	1
23212212	1
23212231	1
23212312	0
23212332	1
23212333	1
23213123	1
23213223	1
23213231	0
23213232	1
23213311	1
23213322	1
23213333	1
23231121	0
23231212	0
23231233	1
23231323	0
23232211	1
23232212	1
23232221	0
23232223	0
23232233	1
23232323	1
23233221	1
23233231	1
23233232	1
23233322	0
23233333	1
23311112	0
23311221	0
23311222	0
23311233	1
23311313	0
23311323	1
23312122	1
23312131	1
23312223	1
23312311	1
23312312	1

23312323	1
23313121	0
23313133	1
23313212	1
23313222	1
23313231	0
23313232	1
23313233	1
23313322	0
23313323	1
23313333	1
23331112	0
23331212	0
23331232	0
23331233	1
23331323	1
23332221	1
23332222	1
23332223	1
23332231	1
23332311	1
23332322	0
23332323	1
23332331	1
23333111	1
23333122	0
23333123	1
23333131	1
23333211	1
23333213	1
23333222	1
23333223	1
23333232	1
23333233	1
23333323	1
31111233	1
31112121	0
31112333	1
31113132	1
31113222	1
31113321	0
31113323	1

31113331	1
31113332	1
31131233	1
31131312	0
31131323	0
31132221	0
31132223	0
31132232	1
31132311	0
31132312	0
31132333	1
31133112	0
31133123	0
31133212	0
31133233	1
31133331	1
31211122	0
31211132	0
31211211	0
31211232	1
31211312	0
31212112	1
31212113	0
31212132	0
31212211	0
31212232	1
31212321	1
31212333	1
31213122	0
31213223	0
31213232	1
31213233	1
31213323	1
31213332	1
31231211	0
31231311	0
31231323	0
31232231	1
31232312	1
31232322	0
31232332	1
31232333	1

31233111	0
31233122	0
31233133	0
31233221	1
31233222	0
31233233	1
31233333	1
31311112	0
31311122	0
31311212	0
31311231	1
31311233	1
31311312	0
31311332	1
31312212	1
31312221	1
31312222	1
31312231	1
31312233	1
31312323	1
31312332	1
31312333	1
31313111	1
31313123	0
31313131	1
31313132	1
31313133	1
31313222	0
31313223	1
31313232	1
31313233	1
31313321	0
31313333	1
31331221	0
31331222	0
31331223	0
31331331	1
31331332	0
31331333	1
31332112	0
31332131	1
31332132	0

31332133	1
31332221	0
31332232	1
31332233	1
31332312	1
31332322	1
31332323	1
31332333	1
31333112	0
31333222	0
31333223	0
31333232	0
31333233	1
31333311	0
31333322	1
31333332	1
31333333	1
32111112	0
32111121	0
32111123	0
32111211	0
32111311	0
32111333	1
32112212	1
32112232	0
32112311	0
32112321	1
32113112	0
32113233	1
32131133	1
32131212	0
32131311	0
32132211	0
32132212	0
32132221	0
32132232	1
32132233	1
32132331	1
32133111	1
32133113	0
32133122	0
32133212	0

32133223	0
32133232	1
32133233	1
32133311	0
32133312	0
32133321	0
32133323	0
32133331	1
32211111	0
32211112	1
32211133	0
32211212	0
32211323	1
32212111	0
32212122	0
32212133	1
32212231	1
32212232	1
32212233	1
32212311	0
32212323	1
32212333	1
32213123	1
32213132	1
32213311	0
32213312	0
32213333	1
32231122	0
32231222	0
32231332	0
32232111	0
32232133	0
32232212	0
32232213	0
32232221	0
32232222	0
32232322	1
32232331	1
32232333	1
32233112	0
32233122	0
32233123	0

32233222	1
32233232	0
32233233	0
32233332	1
32311131	1
32311132	0
32311212	0
32311221	0
32311322	1
32311323	1
32312212	1
32312231	1
32312233	1
32312311	1
32312323	1
32312331	1
32312332	1
32312333	1
32313122	0
32313133	1
32313212	0
32313222	0
32313231	1
32313232	1
32313233	1
32313313	1
32313332	1
32313333	1
32331111	0
32331112	0
32331122	0
32331132	0
32331212	0
32331221	0
32331222	0
32331311	0
32331313	0
32332111	0
32332112	0
32332123	0
32332133	1
32332211	0

32332212	0
32332222	0
32332223	1
32332232	1
32332323	1
32332331	1
32332333	1
32333122	0
32333212	0
32333223	1
32333232	1
32333233	1
32333311	0
32333312	1
32333322	0
32333323	1
32333333	1
33111122	0
33111212	0
33111222	0
33111312	0
33112112	0
33113111	1
33113121	0
33113212	1
33113223	0
33113233	1
33131122	0
33131123	0
33131332	0
33131333	1
33132222	0
33132223	0
33132322	0
33133121	0
33133131	1
33133223	0
33133233	0
33133321	0
33133323	0
33133332	0
33133333	1

33211112	0
33211211	0
33211312	0
33211321	0
33212213	1
33212222	0
33212311	1
33212312	0
33212313	0
33212333	1
33213112	0
33213211	1
33213232	1
33213333	1
33231212	0
33231221	0
33231312	0
33231333	0
33232112	0
33232122	0
33232123	0
33232222	0
33232223	0
33232233	1
33232312	1
33232322	0
33232323	0
33232333	1
33233112	0
33233131	1
33233221	0
33233222	0
33233223	0
33233233	1
33233323	0
33233333	1
33311122	0
33311223	0
33311231	1
33311311	0
33311312	0
33311322	0

33311332	0
33312233	0
33312312	0
33312322	1
33312323	0
33312333	1
33313122	0
33313223	0
33313233	1
33313311	0
33313323	1
33313333	1

33331122	0
33331133	0
33331223	0
33331232	1
33331233	1
33331311	0
33331321	0
33331331	0
33331333	1
33332112	0
33332121	0
33332123	0

33332131	1
33332133	1
33332211	0
33332221	0
33332223	0
33332232	1
33332233	1
33332323	1
33332333	1
33333133	0
33333222	0
33333223	0

33333231	1
33333232	1
33333233	1
33333311	0
33333321	0
33333323	1
33333332	0
11111111	1
22222222	1
33333333	1

Table B-2. Folding status of 605 beta-lactamase chimeras (M. Meyer, personal communication). Chimeras were scored as folded (= 1) if and only if they conferred resistance to ampicillin. Homologs PSE-4 (Lim *et al.*, 2001), SED-1 (Petrella *et al.*, 2001), and TEM-1 (Jelsch *et al.*, 1993) are parents 1, 2, and 3, respectively. The eight blocks are defined by crossovers after residues Arg63, Lys71, Thr147, Arg159, Asp174, Leu188, and Gly216, numbered from the N-terminus of TEM-1.

11111222	0	11332133	1	13131333	0	13333332	0
11113312	0	11332322	0	13133132	0	21133332	0
11121312	0	11333112	0	13213312	0	21211332	0
11121322	0	11333212	0	13223132	0	21212233	0
11123212	0	11333222	0	13223232	1	21213332	0
11123312	0	11333312	0	13223332	0	21233112	0
11131133	0	11333322	0	13233233	0	21312213	0
11131333	0	12123132	0	13233332	0	21313132	1
11132133	0	12123333	0	13311112	0	21313222	1
11133222	0	12131323	0	13311332	0	21313232	1
11133322	0	12133332	0	13312223	0	21321132	1
11211333	0	12211322	0	13313212	0	21322322	1
11223222	0	12211332	0	13313323	0	21323222	1
11223333	0	12211333	0	13313332	0	21323232	1
11233332	0	12212133	0	13322333	0	21332132	1
11311332	0	12222322	0	13323312	0	21333232	1
11312312	0	12233322	0	13331232	0	22223322	1
11312332	0	12233332	0	13331333	0	22233112	0
11313112	0	12313233	0	13332122	0	22311322	0
11313133	1	12313332	0	13332212	0	22313222	1
11313232	0	12321331	0	13332222	0	22313322	1
11313322	0	12323132	0	13332312	0	22313332	0
11321132	0	12323212	0	13332323	0	22323122	0
11323132	0	12323322	1	13333112	0	22332323	0
11323222	0	12332222	0	13333212	0	22333213	0
11331212	0	12332232	0	13333232	0	22333232	0
11331232	0	13111233	0	13333233	1	22333332	0
11331332	0	13123332	0	13333322	0	22333333	0

23132332	0
23133232	1
23133323	0
23212313	0
23213222	0
23221332	0
23222332	0
23311332	0
23323212	1
23331232	1
23333132	1
23333222	1
23333312	0
31111212	0
31111333	0
31112333	1
31113212	0
31113312	0
31113322	0
31122113	0
31123133	0
31133333	1
31213322	0
31222322	0
31233333	0
31311313	1
31312233	1
31313232	0
31322312	0
31322333	1
31323112	0
31323223	1
31323332	0
31331122	0
31331332	0
31332112	0
31332333	1
31333233	1
32122333	0
32123232	0
32123333	0
32133233	0

32133332	0
32222312	0
32223333	0
32311233	0
32313122	0
32313222	0
32321132	0
32321222	0
32321333	0
32323212	0
32323322	0
32333322	0
33113322	0
33122312	0
33123332	0
33133333	0
33213332	0
33222333	0
33223332	0
33311132	0
33311233	0
33312133	1
33313232	0
33313312	0
33313332	0
33313333	1
33321113	1
33321123	1
33321132	0
33321312	0
33321333	1
33323212	0
33323232	0
33333122	0
32313322	1
31213332	0
11112212	0
11112332	0
11113232	0
11121132	0
11121223	0
11122212	0

11131332	0
11133332	0
11211132	0
11212312	0
11213332	0
11213333	1
11221313	0
11222332	0
11223312	0
11223332	0
11233132	0
11233312	0
11311133	1
11311232	0
11311322	0
11313222	0
11313312	0
11313332	0
11313333	1
11321332	0
11322312	0
11323133	0
11323213	0
11323232	0
11323312	0
11323322	0
11323332	0
11331112	0
11331113	0
11331312	0
11332332	0
11332333	1
11333132	0
11333133	1
11333232	0
11333332	0
12113332	0
12122332	0
12123232	0
12123322	0
12123332	0
12132322	0

12133132	0
12133233	0
12212233	0
12212333	0
12213132	0
12221232	0
12223312	0
12223332	0
12231212	0
12233112	0
12233312	0
12311122	0
12311332	0
12312233	0
12313132	0
12313212	0
12313222	0
12313232	0
12313312	0
12313322	0
12313323	0
12321333	0
12331123	0
12331313	0
12332332	0
12333132	0
12333312	0
13122313	0
13123232	0
13131212	0
13132212	0
13212233	0
13213332	0
13221113	0
13222213	0
13222233	0
13231233	0
13231333	0
13232332	0
13232333	0
13311123	0
13311132	0

13311232	0
13312132	0
13313112	0
13313122	0
13313312	0
13313322	0
13313333	1
13322233	1
13331312	0
13331332	0
13332232	0
13332333	1
13333132	0
13333323	0
21133232	0
21233312	0
21233332	0
21311333	0
21313312	0
21313322	1
21313333	0
21322132	1
21323322	1
21323332	1
21331132	0
21331232	0
21331312	0
21333132	1
21333222	1
21333322	1
21333332	1
22133132	0
22311332	0
22331333	0
22332312	0
22333132	0
23123332	1
23132322	1
23133321	0
23133322	0
23213112	0
23223212	0

23223333	0
23313332	1
23322332	1
23323322	1
23332222	1
23332332	1
23333112	1
23333212	1
23333322	1
23333332	1
31113232	0
31123332	0
31133132	0
31133332	0
31211213	0
31212232	0
31213333	0
31222132	1
31223132	0
31223133	0
31223322	0
31223332	0
31233122	0
31233132	0
31233312	0
31233322	0
31311122	0
31311133	1
31311222	0
31311323	0
31311332	0
31311333	1
31312132	0
31312133	1
31312333	1
31313113	1
31313122	0
31313132	0
31313233	1
31313312	0
31313333	1
31322212	0

31323132	0
31323232	0
31323333	1
31332313	1
31332322	0
31333322	0
31333332	0
31333333	1
32111113	0
32123212	0
32133232	0
32213332	0
32221232	0
32222233	0
32222332	0
32222333	0
32223322	0
32223332	0
32231332	0
32311122	0
32311123	0
32311232	0
32312312	0
32312322	0
32312333	1
32313132	0
32313133	1
32313312	0
32313332	0
32313333	1
32321223	0
32322132	0
32322322	0
32322332	0
32322333	0
32323122	0
32323222	0
32323332	0
32331312	0
32331333	0
32333332	0
33123322	0

33222232	0
33223312	0
33223323	0
33311133	1
33311312	0
33311313	1
33311332	0
33313122	0
33313322	0
33322313	0
33322333	1
33323112	0
33323312	0
33323322	0
33332312	0
33332333	1
33333212	0
33333232	0
33333332	0
12333332	0
31313332	0
11121333	0
11122332	0
11132333	0
11221322	0
11221333	0
11222233	0
11231132	0
11232332	0
11233333	0
11311333	0
11312331	1
11312333	0
11313331	1
11321333	0
11322332	0
11323333	0
11331133	0
11331331	1
11331333	0
11332131	0
11332132	0

11332331	1
11333231	1
11333333	1
12121333	0
12122333	0
12131333	0
12212331	0
12231333	0
12232333	0
12311131	0
12311133	0
12311233	0
12311333	0
12312131	1
12312132	0
12312331	0
12322333	0
12331333	0
12332131	1
12332331	0
12332333	0
12333121	1
13132333	0
13133332	0
13133333	0
13221132	0
13231223	0
13231332	0
13233333	1
13311323	0
13311331	0
13311333	1
13312131	0
13312312	0
13312333	1
13321331	0
13331231	0
13331331	0
13332133	0
13332233	1

13333333	1
21131331	0
21231331	0
21231332	0
21232333	0
21311311	0
21312331	0
21331321	0
21331331	0
21332133	0
21332232	1
21332313	0
21332332	1
21332333	0
22113332	0
22122332	0
22131333	0
22222333	0
22232333	0
22311333	0
22312333	0
22332321	1
22332331	0
22332332	0
22333323	0
23221333	0
23222333	0
23312332	1
23321122	1
23332131	0
31121213	0
31121332	0
31123333	0
31212332	0
31231312	0
31232333	0
31311312	0
31312332	0
31321332	0
31323233	1

31331132	0
31331313	1
31331333	1
31332233	1
32122332	0
32221332	0
32312332	0
32313232	0
32322331	0
32332233	1
33121323	0
33131331	0
33211213	0
33222133	0
33222332	0
33231133	0
33311333	1
33312333	1
33313133	1
33322332	0
33331333	1
33332233	1
33332331	0
33333333	1
11322331	0
11312133	1
11331231	1
11332231	1
11332233	1
11333233	1
11333331	1
13311133	1
13312133	1
13313133	1
13321233	1
13322332	1
21312132	1
21312332	1
21313332	1
21331333	1

21332222	1
21333122	1
22222222	1
22313312	1
23132333	1
23313112	1
23313132	1
23313232	1
23313322	1
23321232	1
23322222	1
23323222	1
23323232	1
23323332	1
23331222	1
23332212	1
23332232	1
23333122	1
23333232	1
33323333	1
31113333	1
31311233	1
31311331	1
31313112	1
31313133	1
32213312	1
32311133	1
32321233	1
32322233	1
32333333	1
33312213	1
33312233	1
33312312	1
33313233	1
33322233	1
33331233	1
11111111	1

References

Adami, C. (2004) Information theory in molecular biology. *Phys. Life Rev.* **1**, 3−22.

Agresti, A. (2002) *Categorical Data Analysis*. 2nd ed, John Wiley & Sons, Hoboken.

Aloy, P., Stark, A., Hadley, S. & Russell, R. B. (2003) Predictions without templates: New folds, secondary structure, and contacts in CASP5. *Proteins* **53**, 436−456.

Arnold, F. H., Ed. (2000) Evolutionary Protein Design. Vol. 55. *Adv. Protein Chem.* Academic Press, San Diego.

Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W. & Dress, A. W. (2000) Correlations among amino acid sites in bHLH protein domains: An information theoretic analysis. *Mol. Biol. Evol.* **17**, 164−178.

Bernstein, S. & Bernstein, R. (1999) *Elements of Statistics II: Inferential Statistics*. Schaum's Outline Series, McGraw-Hill, New York.

Bloom, J. D., Silberg, J. J., Wilke, C. O., Drummond, D. A., Adami, C. & Arnold, F. H. (2005) Thermodynamic prediction of protein neutrality. *Proc. Natl. Acad. Sci. USA* **102**, 606−611.

Boyd, S. & Vandenberghe, L. (2004) *Convex Optimization*, Cambridge University Press, Cambridge.

Bui, T. N. & Jones, C. (1992) Finding good approximate vertex and edge partitions is NP-hard. *Inf. Process. Lett.* **42**, 153−159.

Chen, J. M. & Stites, W. E. (2001) Higher-order packing interactions in triple and quadruple mutants of staphylococcal nuclease. *Biochemistry* **40**, 14012−14019.

Cid, H., Bunster, M., Canales, M. & Gazitua, F. (1992) Hydrophobicity and structural classes in proteins. *Protein Eng.* **5**, 373−375.

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

Czyzyk, J., Mesnier, M. & More, J. (1998) The NEOS Server. *IEEE J. Comp. Sci. Eng.* **5**, 68−75.

Dahiyat, B. I. & Mayo, S. L. (1996) Protein design automation. *Protein Sci.* **5**, 895−903.

Daugherty, P. S., Chen, G., Iverson, B. L. & Georgiou, G. (2000) Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. *Proc. Natl. Acad. Sci. USA* **97**, 2029−2034.

DeGrado, W. F. (2001) Introduction: Protein Design. *Chem. Rev.* **101**, 3025−3026.

Drummond, D. A., Silberg, J. J., Wilke, C. O., Meyer, M. M. & Arnold, F. H. (2005) On the conservative nature of intragenic recombination. *Proc. Natl. Acad. Sci. USA* **102**, 5380−5385.

Fodor, A. A. & Aldrich, R. W. (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* **56**, 211−221.

Gilis, D. & Rooman, M. (1997) Predicting protein stability changes upon mutation using database-derived potentials: Solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J. Mol. Biol.* **272**, 276−290.

Gilis, D. (2004) Protein decoy sets for evaluating energy functions. *J. Biomol. Struct. Dyn.* **21**, 725−735.

Gobel, U., Sander, C., Schneider, R. & Valencia, A. (1994) Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309−317.

Goldstein, R. F. (1994) Efficient rotamer elimination applied to protein side-chains and related spin-glasses. *Biophys. J.* **66**, 1335−1340.

Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999) Energy functions for protein design. *Curr. Opin. Struct. Biol.* **9**, 509−513.

Guex, N. & Peitsch, M. C. (1997) SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling. *Electrophoresis* **18**, 2714−2723.

Guo, H. H., Choe, J. & Loeb, L. A. (2004) Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci. USA* **101**, 9205−9210.

Haines, D. C., Tomchick, D. R., Machius, M. & Peterson, J. A. (2001) Pivotal role of water in the mechanism of P450BM-3. *Biochemistry* **40**, 13456−13465.

Hayes, R. J., Bentzien, J., Ary, M. L., Hwang, M. Y., Jacinto, J. M., Vielmetter, J., Kundu, A. & Dahiyat, B. I. (2002) Combining computational and experimental screening for rapid optimization of protein properties. *Proc. Natl. Acad. Sci. USA* **99**, 15926−15931.

Hellinga, H. W. & Richards, F. M. (1991) Construction of new ligand binding sites in proteins of known structure. *J. Mol. Biol.* **222**, 763−785.

Hiraga, K. & Arnold, F. H. (2003) General method for sequence-independent site-directed chimeragenesis. *J. Mol. Biol.* **330**, 287−296.

Hochbaum, D. S., Ed. (1997) Approximation Algorithms for NP-hard Problems. PWS, Boston.

Hosmer, D. W. & Lemeshow, S. (2000) *Applied Logistic Regression*, John Wiley and Sons, New York.

Huang, W. Z., Petrosino, J., Hirsch, M., Shenkin, P. S. & Palzkill, T. (1996) Amino acid sequence determinants of beta-lactamase structure and activity. *J. Mol. Biol.* **258**, 688−703.

Jelsch, C., Mourey, L., Masson, J. M. & Samama, J. P. (1993) Crystal structure of *Escherichia coli* TEM1 beta-lactamase at 1.8 Å resolution. *Proteins* **16**, 364−383.

Kass, I. & Horovitz, A. (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins* **48**, 611−617.

Kawashima, S., Ogata, H. & Kanehisa, M. (1999) AAindex: Amino Acid Index Database. *Nucleic Acids Res.* **27**, 368−369.

Khatun, J., Khare, S. D. & Dokholyan, N. V. (2004) Can contact potentials reliably predict stability of proteins? *J. Mol. Biol.* **336**, 1223−1238.

Klein, P., Kanehisa, M. & Delisi, C. (1984) Prediction of protein function from sequence properties: Discriminant analysis of a database. *Biochim. Biophys. Acta* **787**, 221−226.

Korte, B. & Vygen, J. (2002) *Combinatorial Optimization: Theory and Algorithms*, Springer, Berlin.

Krigbaum, W. R. & Komoriya, A. (1979) Local interactions as a structure determinant for protein molecules .2. *Biochim. Biophys. Acta* **576**, 204−228.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364−1368.

Larson, S. M., Di Nardo, A. A. & Davidson, A. R. (2000) Analysis of covariation in an SH3 domain sequence alignment. *J. Mol. Biol.* **303**, 433−446.

Lawler, E. (1976) *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart & Winston, New York.

Lazaridis, T. & Karplus, M. (2000) Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **10**, 139−145.

Li, H. Y. & Poulos, T. L. (1997) The structure of the cytochrome p450BM-3 haem domain complexed with the fatty acid substrate, palmitoleic acid. *Nat. Struct. Biol.* **4**, 140−146.

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

Lockless, S. W. & Ranganathan, R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295−299.

Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003) Computational design of receptor and sensor proteins with novel functions. *Nature* **423**, 185−190.

Matagne, A., Lamotte-Brasseur, J. & Frere, J. M. (1998) Catalytic properties of class A beta-lactamases: efficiency and diversity. *Biochem. J.* **330**, 581−598.

McCullagh, P. & Nelder, J. A. (1989) *Generalized Linear Models*. 2nd ed, Chapman & Hall, New York.

Menard, S. (2002) *Applied Logistic Regression Analysis*. Quantitative Applications in the Social Sciences (Lewis-Beck, M. S., Ed.), Sage, Thousand Oaks.

Mendes, J., Guerois, R. & Serrano, L. (2002) Energy estimation in protein design. *Curr. Opin. Struct. Biol.* **12**, 441−446.

Meyer, M. M., Silberg, J. J., Voigt, C. A., Endelman, J. B., Mayo, S. L., Wang, Z.-G. & Arnold, F. H. (2003) Library analysis of SCHEMA-guided protein recombination. *Protein Sci.* **12**, 1686−1693.

Miyazawa, S. & Jernigan, R. L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623−644.

Mooers, B. H. M., Datta, D., Baase, W. A., Zollars, E. S., Mayo, S. L. & Matthews, B. W. (2003) Repacking the core of T4 lysozyme by automated design. *J. Mol. Biol.* **332**, 741−756.

Moore, G. L. & Maranas, C. D. (2004) Computational challenges in combinatorial library design for protein engineering. *AIChE J.* **50**, 262−272.

Nelson, D. (2005) http://drnelson.utmem.edu/CytochromeP450.html.

Ostermeier, M., Shim, J. & Benkovic, S. J. (1999) A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat. Biotechnol.* **17**, 1205−1209.

Ostermeier, M. (2003) Synthetic gene libraries: in search of the optimal diversity. *Trends Biotechnol.* **21**, 244−247.

Otey, C. R. (2003) High-throughput carbon monoxide binding assay for cytochromres P450. In *Directed Enzyme Evolution: Screening and Selection Methods* (Arnold, F. H. & Georgiou, G., eds.), Vol. 230, pp. 137−139. Humana Press, Totowa.

Otey, C. R., Silberg, J. J., Voigt, C. A., Endelman, J. B., Bandara, G. & Arnold, F. H. (2004) Functional evolution and structural conservation in chimeric cytochromes P450: Calibrating a structure-guided approach. *Chem. Biol.* **11**, 309−318.

Papadimitriou, C. H. (1994) *Computational Complexity*, Addison-Wesley, Reading.

Papadimitriou, C. H. & Steiglitz, K. (1998) *Combinatorial Optimization: Algorithms and Complexity*, Dover, Mineola.

Petrella, S., Clermont, D., Casin, I., Jarlier, V. & Sougakoff, W. (2001) Novel class A beta-lactamase SED-1 from *Citrobacter sedlakii*: Genetic diversity of beta-lactamases within the *Citrobacter* genus. *Antimicrob. Agents Chemother.* **45**, 2287−2298.

Pierce, N. A., Spriet, J. A., Desmet, J. & Mayo, S. L. (2000) Conformational splitting: A more powerful criterion for dead-end elimination. *J. Comput. Chem.* **21**, 999−1009.

Pierce, N. A. & Winfree, E. (2002) Protein design is NP-hard. *Protein Eng.* **15**, 779−782.

Plaxco, K. W., Simons, K. T. & Baker, D. (1998) Contact order, transition state placement, and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985−994.

Ravichandran, K. G., Boddupalli, S. S., Hasemann, C. A., Peterson, J. A. & Deisenhofer, J. (1993) Crystal structure of hemoprotein domain of P450BM-3, a prototype for microsomal P450's. *Science* **261**, 731−736.

Russ, W. P. & Ranganathan, R. (2002) Knowledge-based potential functions in protein design. *Curr. Opin. Struct. Biol.* **12**, 447−452.

Saraf, M. C. & Maranas, C. D. (2003) Using a residue clash map to functionally characterize protein recombination hybrids. *Protein Eng.* **16**, 1025−1034.

Saraf, M. C., Moore, G. L. & Maranas, C. D. (2003) Using multiple sequence correlation analysis to characterize functionally important protein regions. *Protein Eng.* **16**, 397−406.

Saraf, M. C., Horswill, A. R., Benkovic, S. J. & Maranas, C. D. (2004) FamClash: A method for ranking the activity of engineered enzymes. *Proc. Natl. Acad. Sci. USA* **101**, 4142−4147.

Saraf, M. C., Gupta, A. & Maranas, C. D. (2005) Design of combinatorial protein libraries of optimal size. *Proteins* (in press).

Saunders, C. T. & Baker, D. (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.* **322**, 891−901.

Silberg, J. J., Endelman, J. B. & Arnold, F. H. (2004) SCHEMA-guided protein recombination. *Methods Enzymol.* **388**, 35−42.

Sipser, M. (1997) *Introduction to the Theory of Computation*, PWS, Boston.

Stemmer, W. P. (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**, 389−391.

Stevenson, J. D. & Benkovic, S. J. (2002) Combinatorial approaches to engineering hybrid enzymes. *J. Chem. Soc., Perkin Trans. 2* **9**, 1483−1493.

Suel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* **10**, 59−69.

Thomas, D. J., Casari, G. & Sander, C. (1996) The prediction of protein contacts from multiple sequence alignments. *Protein Eng.* **9**, 941−948.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673−4680.

Vendruscolo, M. & Domany, E. (1998) Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys.* **109**, 11101−11108.

Voigt, C. A., Kauffman, S. & Wang, Z.-G. (2001a) Rational evolutionary design: The theory of *in vitro* protein evolution. *Adv. Protein Chem.* **55**, 79−160.

Voigt, C. A., Mayo, S. L., Arnold, F. H. & Wang, Z.-G. (2001b) Computational method to reduce the search space for directed protein evolution. *Proc. Natl. Acad. Sci. USA* **98**, 3778−3783.

Voigt, C. A. (2002) Computationally optimizing the directed evolution of proteins. Ph.D. thesis, California Institute of Technology.

Voigt, C. A., Martinez, C., Wang, Z.-G., Mayo, S. L. & Arnold, F. H. (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553−558.

Wintrode, P. L. & Arnold, P. H. (2001) Temperature adaptation of enzymes: Lessons from laboratory evolution. *Adv. Protein Chem.* **55**, 161−225.

Zaccolo, M. & Gherardi, E. (1999) The effect of high-frequency random mutagenesis on *in vitro* protein evolution: A study on TEM-1 β-lactamase. *J. Mol. Biol.* **285**, 775−783.