**Protein Engineering, Design & Selection** vol. 17 no. 7 pp. 589–594, 2004
Published online August 25, 2004 doi:10.1093/protein/gzh067

# Site-directed protein recombination as a shortest-path problem

**Jeffrey B.Endelman[1,2], Jonathan J.Silberg[3], Zhen-Gang Wang[2,3] and Frances H.Arnold[1,2,3]**

[1]Bioengineering Option and [3]Division of Chemistry and Chemical Engineering, California Institute of Technology, Mail Code 210-41, Pasadena, CA 91125-4100, USA

[2]To whom correspondence should be addressed.
E-mail: endelman@caltech.edu; zgw@cheme.caltech.edu; frances@cheme.caltech.edu

**Protein function can be tuned using laboratory evolution, in which one rapidly searches through a library of proteins for the properties of interest. In site-directed recombination, $n$ crossovers are chosen in an alignment of $p$ parents to define a set of $p(n+1)$ peptide fragments. These fragments are then assembled combinatorially to create a library of $p^{n+1}$ proteins. We have developed a computational algorithm to enrich these libraries in folded proteins while maintaining an appropriate level of diversity for evolution. For a given set of parents, our algorithm selects crossovers that minimize the average energy of the library, subject to constraints on the length of each fragment. This problem is equivalent to finding the shortest path between nodes in a network, for which the global minimum can be found efficiently. Our algorithm has a running time of $O(N^3p^2 + N^2n)$ for a protein of length $N$. Adjusting the constraints on fragment length generates a set of optimized libraries with varying degrees of diversity. By comparing these optima for different sets of parents, we rapidly determine which parents yield the lowest energy libraries.**

*Keywords*: dynamic programming/laboratory evolution/optimization/protein design/recombination

## Introduction

Protein design seeks the amino acid sequence that encodes a protein with a desired set of properties (DeGrado, 2001). One successful strategy, called laboratory evolution, involves searching through a library of proteins for the properties of interest (Arnold, 2000). These libraries are often created by recombining and/or mutating parental proteins with similar structures (Neylon, 2004). *Library* design is complementary to *sequence* design, in which a single, novel protein is created *de novo* or by making specific, model-guided changes to a parental protein (Dahiyat and Mayo, 1997; DeGrado *et al*., 1999; Kuhlman *et al*., 2003; Looger *et al*., 2003).

Sequence design is sometimes called rational design, but many aspects of laboratory evolution are also rationally designed (Kamtekar *et al*., 1993; Kono and Saven, 2001; Voigt *et al*., 2001; Moore and Maranas, 2002), including the library 'diversity.' By diversity we mean how proteins in the library differ from the parents and from each other. The design goals and library creation method should dictate library diversity, but our understanding of this subject is still very limited. In several studies the number of mutations $m$ has been correlated with functional change (Zaccolo and Gherardi, 1999; Daugherty *et al*., 2000; Ostermeier, 2003; Otey *et al*., 2004). We use the corresponding library average $\langle m \rangle$ as an example of an empirically useful diversity measure.

Although diversity is needed to effect changes in protein function, it is at odds with the need for stably folded proteins (a prerequisite for most functions). Since most mutations are neutral or disruptive to protein structure, the fraction of stably folded proteins in a library tends to decrease with diversity (Daugherty *et al*., 2000; Guo *et al*., 2004). Equivalently, for an energy function that scores protein stability (Gordon *et al*., 1999; Saraf *et al*., 2004), the average energy of all proteins in a library $\langle E \rangle$ tends to increase with diversity. This type of tradeoff is common in design problems with conflicting performance objectives. Our design strategy involves finding libraries on the optimal energy–diversity tradeoff surface.

We apply this strategy to site-directed recombination (SDR), in which $n$ crossovers are chosen in an alignment of $p$ structurally-related parents to define a set of $p(n+1)$ peptide fragments. These fragments are then assembled combinatorially to create a library of $p^{n+1}$ chimeric proteins (Figure 1). SDR is a relatively new approach to recombination in laboratory evolution (Richardson *et al*., 2002; Hiraga and Arnold, 2003; Meyer *et al*., 2003). Other strategies do not involve a specific choice of crossovers (Stevenson and Benkovic, 2002). Instead, they attempt to generate all possible crossovers using techniques from molecular biology. In practice, however, these methods are often limited in either the number or locations of crossovers (Joern *et al*., 2002).

SDR benefits from the unique properties of recombination as an evolutionary search strategy. Recombination of homologs is highly effective at neutral evolution because the mutations it introduces have been selected by nature to be

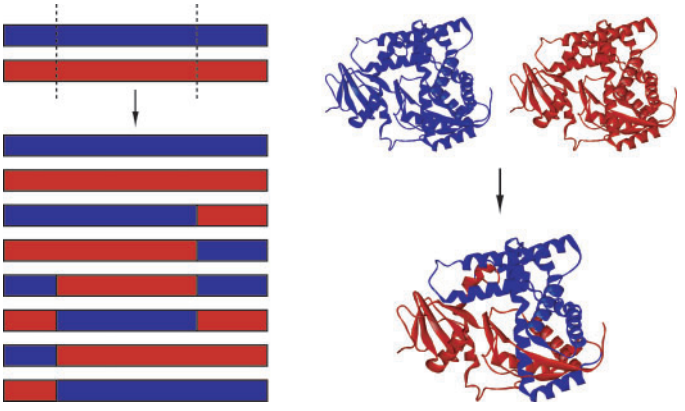

**Fig. 1.** Site-directed recombination of two parental proteins (blue and red) with similar structures. Crossovers (dashed lines) define peptide fragments that are assembled combinatorially to create a library with eight members. One of the chimeric sequences is threaded on to the parental structure.

compatible with the protein structure, albeit in a different genetic background. A recent SDR experiment identified several functional β-lactamases with over 50 mutations (Meyer *et al.*, 2003). More significantly, recombination appears well suited to adaptive evolution in theory (Holland, 1992), computation (Mitchell, 1996; Deem and Bogarad, 1999; Voigt *et al.*, 2001) and laboratory experiments with proteins (Stevenson and Benkovic, 2002; Otey *et al.*, 2004).

To optimize SDR libraries for a given set of parents and fixed number of crossovers $n$, we minimize the average energy of all chimeras $\langle E \rangle$, subject to constraints on the length $L$ of each peptide fragment:

$$\min_{(X_1, X_2, \ldots, X_n)} \langle E \rangle \quad \text{subject to } L_{\min} \leqslant L \leqslant L_{\max}, \tag{1}$$

where $(X_1, X_2, \ldots, X_n)$ are the crossover locations. We prove below that for a protein of length $N$, Equation 1 is equivalent to finding the shortest path in a network with $O(nN)$ nodes. The resulting library design algorithm RASPP (Recombination as a Shortest-Path Problem) has a running time of $O(N^3p^2 + N^2n)$. Adjusting the constraints on fragment length $[L_{\min}, L_{\max}]$ generates a set of optimized libraries with varying degrees of diversity. By comparing these optima for different sets of parents or different numbers of crossovers, we rapidly determine which designs yield the lowest-energy libraries.

## Materials and methods

### *Theoretical energies*

As with other global optimization algorithms for protein design (Dahiyat and Mayo, 1996; Gordon and Mayo, 1999), RASPP can use any energy function with single and pairwise interactions between residues:

$$E = \sum_i e(\sigma_i) + \sum_i \sum_{j>i} e(\sigma_i, \sigma_j) \tag{2}$$

where the amino acid at position $k$, $\sigma_k(S_k)$, is determined by the parent $S_k$ inherited at that position. For a fixed set of crossovers, the average energy of all chimeras in the library can be written as a sum over inheritance patterns:

$$\langle E \rangle_{(X_1, X_2, \ldots, X_n)} = \frac{1}{p^{n+1}} \sum_{S_{0,X_1}} \sum_{S_{X_1,X_2}} \cdots \sum_{S_{X_n,N}} E \tag{3}$$

where $S_{i,j}$ denotes the parent from whom the peptide fragment $[i + 1, j]$ (residues $i + 1$ to $j$, inclusive) is inherited.

Consider two libraries, one with $k-1$ crossovers and the other with $k$ crossovers, whose first $k-1$ crossover locations are identical. The difference in average energy between these libraries is

$$\begin{aligned}
&\langle E \rangle_{(X_1, X_2, \ldots, X_{k-1}, X_k)} - \langle E \rangle_{(X_1, X_2, \ldots, X_{k-1})} \\
&= \frac{1}{p^{k+1}} \sum_{S_{0,X_1}} \cdots \sum_{S_{X_{k-2},X_{k-1}}} \left[ \sum_{S_{X_{k-1},X_k}} \sum_{S_{X_k,N}} - p \sum_{S_{X_{k-1},N}} \right] \\
&\quad \times \sum_{r=1}^{N} \sum_{t=r+1}^{N} e(\sigma_r, \sigma_t)
\end{aligned} \tag{4}$$

$$\begin{aligned}
&= \frac{1}{p^{k+1}} \sum_{S_0, X_1} \cdots \sum_{S_{X_{k-2}, X_{k-1}}} \left[ \sum_{S_{X_{k-1}, X_k}} \sum_{S_{X_k, N}} - p \sum_{S_{X_{k-1}, N}} \right] \\
&\quad \times \sum_{r=X_{k-1}+1}^{X_k} \sum_{t=X_{k+1}}^{N} e(\sigma_r, \sigma_t)
\end{aligned} \tag{5}$$

$$= \frac{1}{p^2} \left[ \sum_{S_{X_{k-1}, X_k}} \sum_{S_{X_k, N}} - p \sum_{S_{X_{k-1}, N}} \right] \sum_{r=X_{k-1}+1}^{X_k} \sum_{t=X_{k+1}}^{N} e(\sigma_r, \sigma_t) \tag{6}$$

Equation 5 follows because the operator in brackets, which is the difference of two inheritance sums, is only non-zero 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 6.

### *Computational energies*

To generate computational results we used SCHEMA disruption, an instance of Equation 2 that counts the number of pairwise interactions broken by recombination (Voigt *et al.*, 2002; Silberg *et al.*, 2004):

$$E = \sum_i \sum_{j>i} C_{ij} \Delta_{ij} \tag{7}$$

The contact matrix $C_{ij}$ depends solely on structural information, while $\Delta_{ij}$ uses only the parental sequence alignment. Specifically, $C_{ij} = 1$ 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 $\sigma_i(S_i)$ and $\sigma_j(S_j)$ that are found in the chimera are also present at homologous positions in any single parent. Otherwise, the $i - j$ interaction is considered broken and $\Delta_{ij} = 1$.

SCHEMA disruption has proven useful in guiding SDR of β-lactamases (Hiraga and Arnold, 2003; Meyer *et al.*, 2003) and cytochromes P450 (Otey *et al.*, 2004), the two systems explored in this study. For SDR of β-lactamases, we used the crystal structure of TEM-1 (1BTL.pdb) (Jelsch *et al.*, 1993) and generated a structural alignment of 263 residues with its homolog PSE-4 (1G68.pdb) (Lim *et al.*, 2001) using Swiss-Pdb Viewer (Guex and Peitsch, 1997). TEM-1 and PSE-4 have 43% sequence identity. For SDR of cytochromes P450, we used the crystal structure of CYP102A1 (1JPZ.pdb) (Haines *et al.*, 2001) from *Bacillus megaterium*, commonly known as P450 BM-3, and generated sequence alignments with *Bacillus subtilis* homologs CYP102A2 (ORF BG12871) and CYP102A3 (ORF BG12299) using CLUSTALW (Thompson *et al.*, 1994). Pairwise sequence identities for these P450 homologs are around 65% based on 456 residues. Cytochrome P450 residues are numbered as in 1JPZ.pdb.

### *Measuring length*

Since library composition is invariant with respect to crossover location within a contiguous region of conserved residues, we do not consider conserved residues as potential crossover locations. This effectively reduces the length of the protein ($N$) by the number of conserved residues. To be consistent we measure fragment length $L$ by the number of residues not conserved across all parents.

### *Measuring diversity*

The mutation level $m$ of each chimera is the minimum number of amino acid changes needed to convert the protein into one of

the parents. The average number of mutations $\langle m \rangle$ in a library with $p^{n+1}$ chimeras is $(\Sigma_i m_i)/p^{n+1}$.

## Results

### *Equation 1 is a shortest-path problem*

Every feasible $n$-crossover library, i.e. one that satisfies the length constraints in Equation 1, can be represented as an $n$-path from node 0 to column $n$ in the directed graph of Figure 2. The node $X_k$ visited in column $k \leq n$ corresponds to the position of the $k$th crossover. To create a one-to-one correspondence between the set of all $n$-paths and the set of all feasible $n$-crossover libraries, nodes in adjacent columns are selectively connected. 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), which has length $X_1$. Thus node 0 is connected to all nodes in the first column that satisfy $L_{\min} \leq X_1 \leq L_{\max}$. 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 $[X_1 + 1, X_2]$, which has length $X_2 - X_1$. Thus node $X_1$ is connected to node $X_2$ if and only if $L_{\min} \leq X_2 - X_1 \leq L_{\max}$. 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 $X_n$ if and only if $L_{\min} \leq X_n - X_{n-1} \leq L_{\max}$ and $L_{\min} \leq N - X_n \leq L_{\max}$.

Once the arc connections are specified, arc lengths are assigned so that the total length of each $n$-path equals the average energy of the corresponding library:

$$\sum_{k=1}^{n} A(X_{k-1}, X_k) = \langle E \rangle_{(X_1, X_2, \ldots, X_n)} \quad (8)$$

where $A(X_{k-1}, X_k)$ is the arc length from node $X_{k-1}$ to node $X_k$ ($X_0 = 0$). Arc lengths from node 0 equal the average energy of a library with one crossover at residue $X_1$:

$$A(0, X_1) = \langle E \rangle_{(X_1)} \quad (9)$$

For arc lengths between columns, we assign the incremental change in energy associated with the next crossover:

$$A(X_{k-1}, X_k) = \langle E \rangle_{(X_1, X_2, \ldots, X_{k-1}, X_k)} - \langle E \rangle_{(X_1, X_2, \ldots X_{k-1})}, \quad k \geq 2 \quad (10)$$

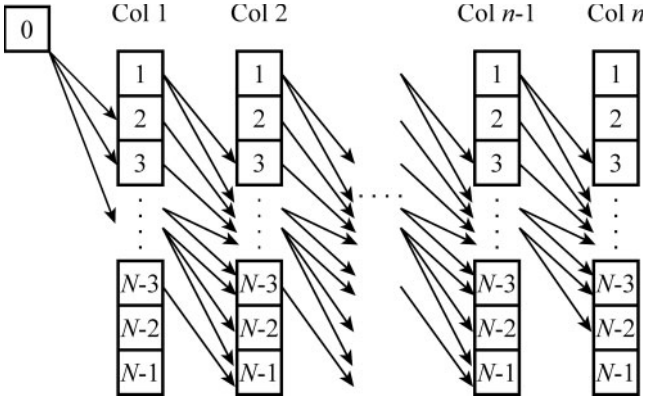


**Fig. 2.** SDR as a shortest-path problem. Every feasible $n$-crossover library can be represented as an $n$-path from node 0 to column $n$. The node visited in column $k$ corresponds to the position of the $k$th crossover, shown here for a protein of length $N$. To constrain the length of each peptide fragment, nodes in adjacent columns are selectively connected. Arc lengths are assigned so that the total path length of each $n$-path equals the average energy of the corresponding library (see Results).

which for pairwise-decomposable energy functions is independent of all crossovers except $X_{k-1}$ and $X_k$ (cf. Equation 6). The first evaluation of Equation 6 requires $O(N^2p^2)$ pairwise energy calculations, but only $O(Np^2)$ are needed to compute each subsequent arc length if the crossover moves by one or two residues. This makes it possible to construct all $O(N^2)$ distinct arc lengths with time complexity $O(N^3p^2)$.

In summary, there is a one-to-one correspondence between feasible libraries and $n$-paths, and the total length of each $n$-path equals the average energy of the corresponding library. Therefore, finding the shortest $n$-path is equivalent to solving Equation 1.

### *Complexity of finding the shortest path*

Shortest-path problems can be solved efficiently because of their recursive structure (Lawler, 1976; Korte and Vygen, 2002). In the case of Figure 2, 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 $k-1$:

$$U_j^k = \min_i \left( U_i^{k-1} + A(i,j) \right) \quad (11)$$

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 11 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 $O(N^2n)$.

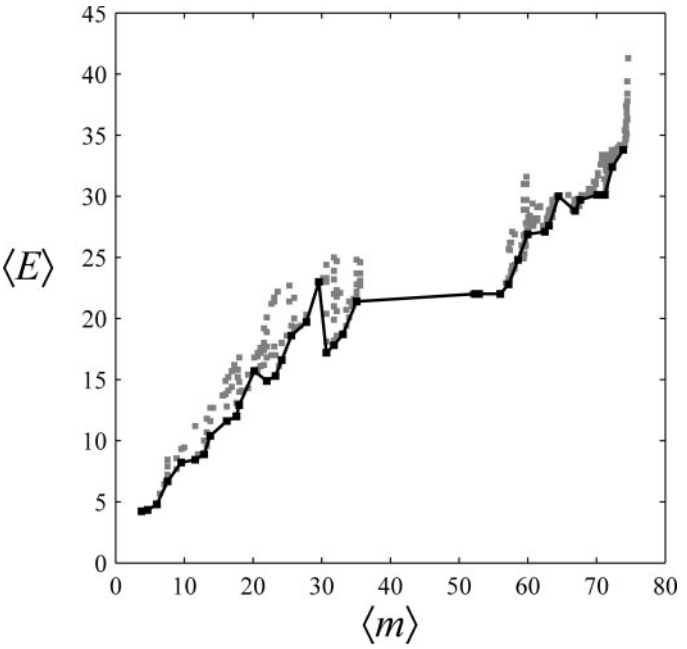


**Fig. 3.** The energy–diversity tradeoff for RASPP libraries. Equation 1 was solved with cytochrome P450 homologs CYP102A1/A2/A3 ($n$ = 7 crossovers, $N$ = 197 non-conserved residues) for length constraints $L_{\min} = 1$ to $N/(n+1)$ and $L_{\max} = N/(n+1)$ to $N - nL_{\min}$. A plot of $\langle E \rangle$ versus $\langle m \rangle$ 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).

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 exactly correspond to the solution of Equation 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 11 is solved $O(N^2)$ times as before. This can be repeated for all values $k \leqslant n$ with a total running time of $O(N^2n)$, the same as a single iteration of RASPP.

Although RASPP libraries are provably optimal when diversity is measured by fragment length, often we are interested in optimality with respect to other diversity measures, such as the average level of mutation $\langle m \rangle$. To use fragment length as a surrogate for $\langle m \rangle$, we vary the length constraints over the entire range of feasible libraries: $L_{\min} = 1$ to $N/(n + 1)$ and $L_{\max} = N/(n + 1)$ to $N - nL_{\min}$, solving Equation 1 $O(N^2/n)$ times. This runs quickly since the arc lengths are not recalculated for each iteration. The ranges for $L_{\min}$ and $L_{\max}$ can easily be modified to accommodate experimental constraints arising from the library creation method.

### *RASPP libraries are optimized with respect to* $\langle m \rangle$

To illustrate RASPP, consider using three cytochrome P450 homologs (CYP102A1/A2/A3, $N = 197$ non-conserved residues) for the laboratory evolution of novel catalytic properties (Otey *et al.*, 2004). By varying the length constraints for $n = 7$ crossovers, we generated 2052 libraries, of which 391 are distinct (Figure 3). As $L_{\min}$ increases and $L_{\max}$ decreases, the crossovers become more evenly spaced, resulting in libraries with higher $\langle E \rangle$ and higher $\langle m \rangle$. 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 ($3^8 = 6561$ chimeras) and sufficiently high levels of mutation for laboratory evolution (nearly $\langle m \rangle = 75$) based on data from previous experiments (Otey *et al.*, 2004).

The lowest-energy RASPP libraries at increasing values of $\langle m \rangle$ define a 'RASPP curve' (Figure 3). To determine how well RASPP curves approximate the optimal energy–diversity tradeoff surface, we enumerated all four-crossover libraries for cytochromes P450 CYP102A1/A2 and β-lactamases TEM-1/PSE-4 ($25 \times 10^6$ and $20 \times 10^6$ libraries, respectively; Figure 4). 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. around $\langle m \rangle = 30$ for the cytochromes P450. A similar mutation gap can be seen in Figure 3 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 3 and 4 (which differ in both respects).

The pattern of optimal crossovers varies dramatically along a RASPP curve. Figure 5 shows the elements of secondary structure for CYP102A1 (Ravichandran *et al.*, 1993) corresponding to crossovers along the RASPP curve of Figure 3. 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.

Many of the fragments chosen are not intuitive – RASPP frequently cuts through secondary structure motifs. For example, the most commonly chosen crossover region (residues 214–217, which shows up as a long horizontal black line in Figure 5) 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. Previous

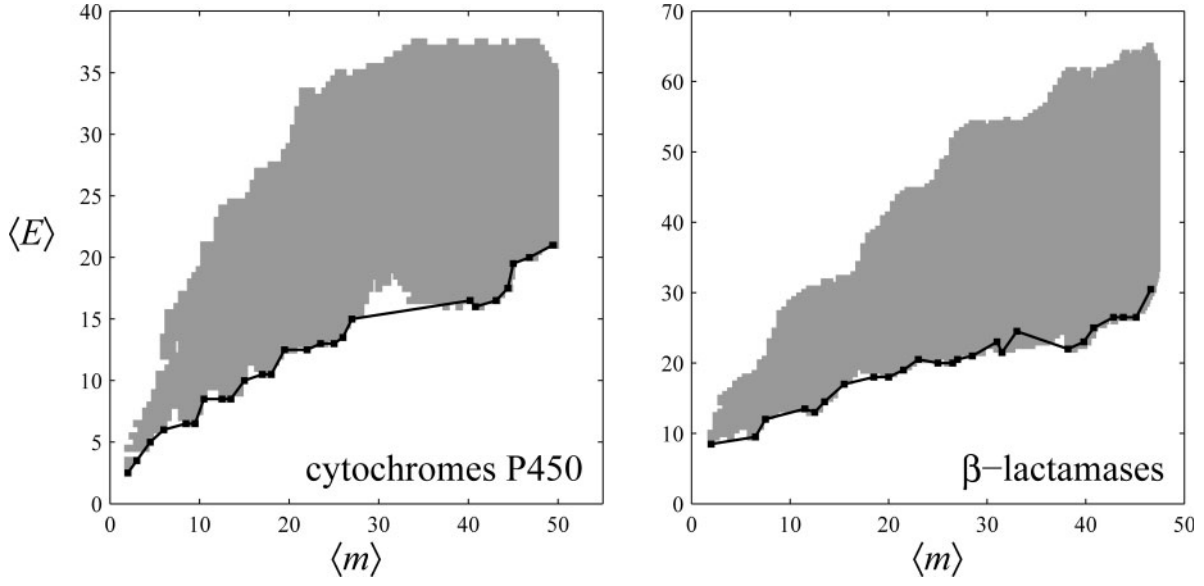


**Fig. 4.** RASPP curves approximate the optimal tradeoff surface. All four-crossover libraries (gray) were enumerated for cytochromes P450 CYP102A1/A2 and β-lactamases TEM-1/PSE-4 ($25 \times 10^6$ and $20 \times 10^6$ libraries, respectively). In both cases, the RASPP curve (black line) closely approximates the optimal energy–diversity tradeoff surface at most values of mutation. One glaring exception is around $\langle m \rangle = 30$ for the cytochromes P450, in which the RASPP curve substantially underestimates the minimum energy. This happens because there are no RASPP libraries nearby (as was true for $40 < \langle m \rangle < 50$ in Figure 3).

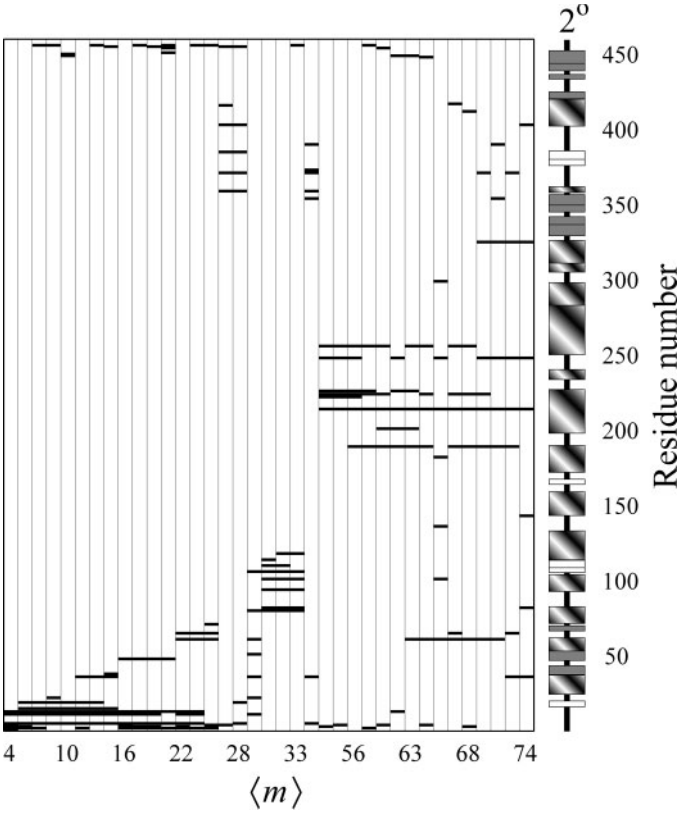


**Fig. 5.** Crossovers along the RASPP curve. Crossover locations (dark horizontal bars) are shown for every library along the RASPP curve of Figure 3 (CYP102A1/A2/A3, $n = 7$ crossovers). When crossovers fall in a contiguous region of conserved residues, they are depicted at the position closest to 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 (numbered as in 1JPZ.pdb). The second axis, labeled as 2°, 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.

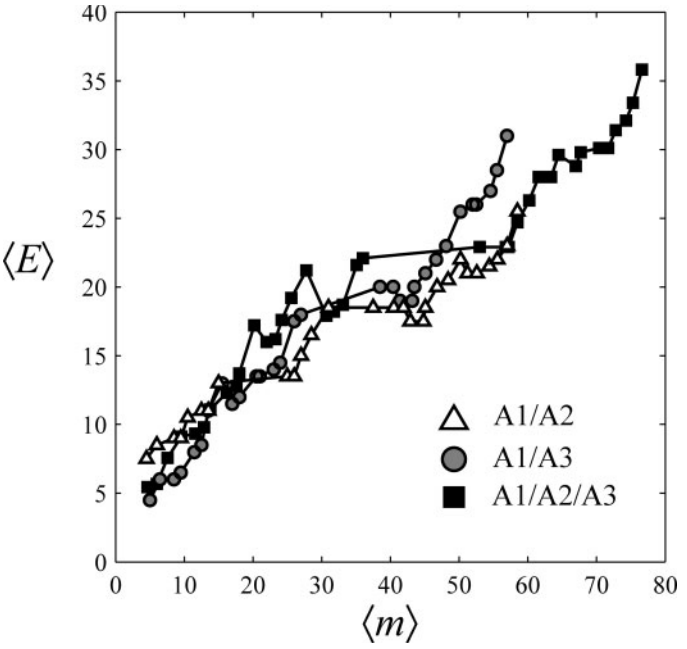


**Fig. 6.** RASPP curves guide the choice of parents. Three alternative sets of cytochromes P450 are compared: CYP102A1/A2, A1/A3 and A1/A2/A3, each with nine crossovers. The RASPP curves, computed as described in Figure 3, represent the lowest energy libraries possible for each set of parents. The optimal 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 $\langle m \rangle > 60$.

computation-guided experiments have verified that site-directed recombination within secondary structure elements often yields folded proteins (Voigt *et al*., 2002; Meyer *et al*., 2003).

### *RASPP curves for parental design*

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 which combination of cytochrome P450 homologs (A1/A2, A1/A3 or A1/A2/A3) is best for laboratory evolution. 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 reveal significant differences at mutation levels $\langle m \rangle > 40$ (Figure 6). 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 their non-conserved residues have the same surface accessibility on average (1.5 contacts per non-conserved residue). 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

We have presented computational results using a residue-based energy function called SCHEMA disruption because it is simple yet effective at identifying folded chimeras (Meyer *et al*., 2003; Otey *et al*., 2004). RASPP is compatible with all pairwise-decomposable, residue-based energy functions, including rotamer-based energies (Gordon *et al*., 1999) that have been averaged to derive residue–residue interactions. Energy functions more sophisticated than SCHEMA may have greater success at predicting the stability of chimeric proteins (Saraf *et al*., 2004). These energy functions will enable RASPP to design libraries closer to the optimal folding–diversity tradeoff surface.

Equation 6 is our key theoretical result which shows that dynamic programming can be used for SDR library design. In this respect, Equation 6 is analogous to dead-end elimination (DEE) theorems (Desmet *et al*., 1992; Goldstein, 1994; Looger and Hellinga, 2001), which have led to many successes in protein sequence design (Dahiyat and Mayo, 1996; Looger *et al*., 2003). However, Equation 6 and the DEE theorem have very different consequences for computational protein design. RASPP finds the global energy minimum (for Equation 1) in $O(N^3p^2 + N^2n)$ operations for a protein of length $N$, making it

efficient in theory and practice (Papadimitriou and 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 and Winfree, 2002). By averaging over the library, we transform protein design from a hard problem to an easy one.

Our tractable formulation of SDR library design also depends on constraining *fragment* diversity to effect changes in *library* diversity. We have described RASPP using fragment length, but RASPP can use any measure of fragment diversity compatible with the arc connections in Figure 2. We have shown that constraints on fragment length are effective when library diversity is measured by the average number of mutations $\langle m \rangle$. As we learn more about the effects of library diversity on protein evolution, other measures of fragment diversity will be needed.

## Acknowledgements

We thank Costas Maranas, Matt DeLisa, Jeff Saven and Niles Pierce for their comments on the manuscript. This work was supported by National Institutes of Health Grant R01 GM068664-01, Army Research Office Contract DAAD19-03-D-0004, the W. M. Keck Foundation, a National Defense Science and Engineering Graduate Fellowship (J.B.E.) and NIH Fellowship F32 GM64949-01 (J.J.S.).

## References

Arnold,F.H. (ed.) (2000) *Evolutionary Protein Design*. Academic Press, San Diego.
Dahiyat,B.I. and Mayo,S.L. (1996) *Protein Sci.*, **5**, 895–903.
Dahiyat,B.I. and Mayo,S.L. (1997) *Science*, **278**, 82–87.
Daugherty,P.S., Chen,G., Iverson,B.L. and Georgiou,G. (2000) *Proc. Natl Acad. Sci. USA*, **97**, 2029–2034.
Deem,M.W. and Bogarad,L.D. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 2591–2595.
DeGrado,W.F. (2001) *Chem. Rev.*, **101**, 3025–3026.
DeGrado,W.F., Summa,C.M., Pavone,V., Nastri,F. and Lombardi,A. (1999) *Annu. Rev. Biochem.*, **68**, 779–819.
Desmet,J., De Maeyer,M., Hazes,B. and Lasters,I. (1992) *Nature*, **356**, 539–542.
Goldstein,R.F. (1994) *Biophys. J.*, **66**, 1335–1340.
Gordon,D.B. and Mayo,S.L. (1999) *Structure*, **7**, 1089–1098.
Gordon,D.B., Marshall,S.A. and Mayo,S.L. (1999) *Curr. Opin. Struct. Biol.*, **9**, 509–513.
Guex,N. and Peitsch,M.C. (1997) *Electrophoresis*, **18**, 2714–2723.
Guo,H.H., Choe,J. and Loeb,L.A. (2004) *Proc. Natl Acad. Sci. USA*, **101**, 9205–9210.
Haines,D.C., Tomchick,D.R., Machius,M. and Peterson,J.A. (2001) *Biochemistry*, **40**, 13456–13465.
Hiraga,K. and Arnold,F.H. (2003) *J. Mol. Biol.*, **330**, 287–296.
Holland,J.H. (1992) *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, MA.
Jelsch,C., Mourey,L., Masson,J.M. and Samama,J.P. (1993) *Proteins*, **16**, 364–383.
Joern,J.M., Meinhold,P. and Arnold,F.H. (2002) *J. Mol. Biol.*, **316**, 643–656.
Kamtekar,S., Schiffer,J.M., Xiong,H., Babik,J.M. and Hecht,M.H. (1993) *Science*, **262**, 1680–1685.
Kono,H. and Saven,J.G. (2001) *J. Mol. Biol.*, **306**, 607–628.
Korte,B. and Vygen,J. (2002) *Combinatorial Optimization: Theory and Algorithms*. Springer, Berlin.
Kuhlman,B., Dantas,G., Ireton,G.C., Varani,G., Stoddard,B.L. and Baker,D. (2003) *Science*, **302**, 1364–1368.
Lawler,E. (1976) *Combinatorial Optimization: Networks and Matroids*. Holt, Rinehart and Winston, New York.
Lim,D., Sanschagrin,F., Passmore,L., De Castro,L., Levesque,R.C. and Strynadka,N.C. (2001) *Biochemistry*, **40**, 395–402.
Looger,L.L. and Hellinga,H.W. (2001) *J. Mol. Biol.*, **307**, 429–445.
Looger,L.L., Dwyer,M.A., Smith,J.J. and Hellinga,H.W. (2003) *Nature*, **423**, 185–190.
Meyer,M.M., Silberg,J.J., Voigt,C.A., Endelman,J.B., Mayo,S.L., Wang,Z.G. and Arnold,F.H. (2003) *Protein Sci.*, **12**, 1686–1693.
Mitchell,M. (1996) *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA.
Moore,G.L. and Maranas,C.D. (2002) *Nucleic Acids Res.*, **30**, 2407–2416.
Neylon,C. (2004) *Nucleic Acids Res.*, **32**, 1448–1459.
Ostermeier,M. (2003) *Trends Biotechnol.*, **21**, 244–247.
Otey,C.R., Silberg,J.J., Voigt,C.A., Endelman,J.B., Bandara,G. and Arnold,F.H. (2004) *Chem. Biol.*, **11**, 309–318.
Papadimitriou,C.H. and Steiglitz,K. (1998) *Combinatorial Optimization: Algorithms and Complexity*. Dover, Mineola.
Pierce,N.A. and Winfree,E. (2002) *Protein Eng.*, **15**, 779–782.
Ravichandran,K.G., Boddupalli,S.S., Hasemann,C.A., Peterson,J.A. and Deisenhofer,J. (1993) *Science*, **261**, 731–736.
Richardson,T.H., Tan,X., Frey,G., Callen,W., Cabell,M., Lam,D., Macomber,J., Short,J.M., Robertson,D.E. and Miller,C. (2002) *J. Biol. Chem.*, **277**, 26501–26507.
Saraf,M.C., Horswill,A.R., Benkovic,S.J. and Maranas,C.D. (2004) *Proc. Natl Acad. Sci. USA*, **101**, 4142–4147.
Silberg,J.J., Endelman,J.B. and Arnold,F.H. (2004) *Methods Enzymol.*, **388**, 35–42.
Stevenson,J.D. and Benkovic,S.J. (2002) *J. Chem. Soc., Perkin Trans. 2*, 1483–1493.
Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
Voigt,C.A., Kauffman,S. and Wang,Z.G. (2001) *Adv. Protein Chem.*, **55**, 79–160.
Voigt,C.A., Martinez,C., Wang,Z.G., Mayo,S.L. and Arnold,F.H. (2002) *Nat. Struct. Biol.*, **9**, 553–558.
Zaccolo,M. and Gherardi,E. (1999) *J. Mol. Biol.*, **285**, 775–783.

**Received August 12, 2004; accepted August 16, 2004**

**Edited by Stephen Mayo**