Computational Challenges in Combinatorial Library Design for Protein Engineering

Gregory L. Moore and Costas D. Maranas
Dept. of Chemical Engineering, The Pennsylvania State University, University Park, PA 16802

Keywords: enzyme engineering, directed evolution, high-throughout screening, protein design, DNA recombination, combinatorial design, error-prone PCR

Introduction

Through the processes of natural selection and co-option, nature has crafted an astounding array of proteins with a remarkable repertoire ranging from catalysis, signaling, recognition and regulation to compartmentalization and repair. Despite this plethora of functionalities and exquisite specialization, many biotechnological tasks require proteins to operate under conditions that were not selected for in nature, such as enhanced thermostability, altered substrate specificity, different cofactor (i.e., NADH, ATP, etc.) dependence, nonaqueous environments and, often, combinations of the above. Unlike many of the systems engineered by people, proteins through evolution had to acquire the inherent ability to change and assume over time subtly, or even dramatically, different roles in living organisms. This amazing plasticity has enabled bioengineers to design or more often redesign proteins more attuned to specific tasks. Protein engineering, however, remains a formidable challenge. Proteins are much larger (i.e., over 50 residues) than nonbiological catalysts, and exhibit complex networks of dynamic interaction necessary for function. Given the residue composition of a protein, the task of *de novo* identifying its three-dimensional (3-D) structure is nontrivial and only limited successes (Bradley et al., 2003) are currently available. On top of this, even complete structure resolution does not mean that function is always truly elucidated. In many cases, functionality and nonfunctionality are separated by differences of only fractions of Angstroms in the position of certain key atoms, an accuracy threshold well beyond the current modeling state-of-the-art. These daunting challenges have led to protein engineering paradigms that involve the synthesis and subsequent screening of multiple protein candidates (from tens to billions) as a way of hedging against the imprecise knowledge of sequence-structure-function relations.

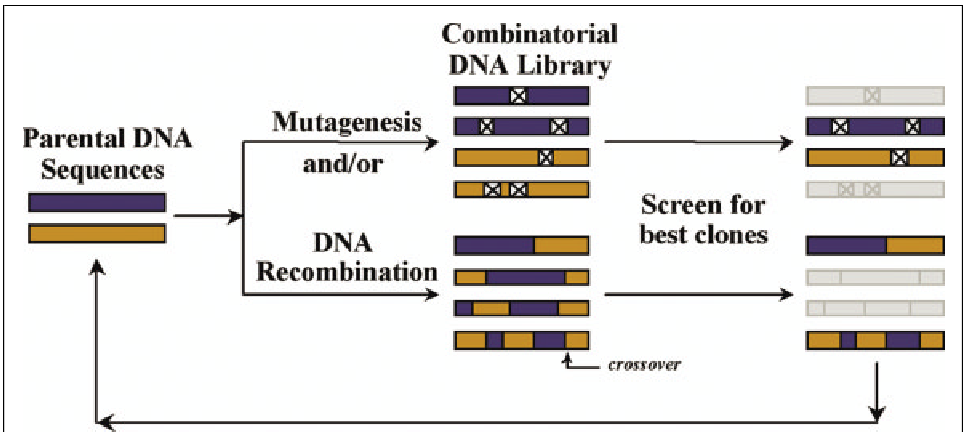

Figure 1. Representation of the key steps of directed evolution experiments.
Crossovers are defined as the junction points between segments from different parental sequences.

This juxtaposition of repeated library generation and screening has emerged as the *directed evolution* design paradigm. Directed evolution methods mimic the process of Darwinian evolution and selection to produce proteins or even entire metabolic pathways with improved properties. These methods (see Figure 1) typically begin with the infusion of diversity into a small set of parental nucleotide sequences through mutagenesis and/or DNA recombination.

Correspondence concerning this article should be addressed to C. D. Maranas at costas@psu.edu. G. L. Moore's e-mail address is glm113@psu.edu.

The resulting combinatorial DNA library is transformed into an appropriate host (e.g., *E. coli*) and then is subjected to a high-throughput screening or selection procedure. The best variants are isolated for another round of mutagenesis or recombination. The cycles of mutagenesis/recombination, screening and isolation continue until a protein with the desired level of improvement is found.

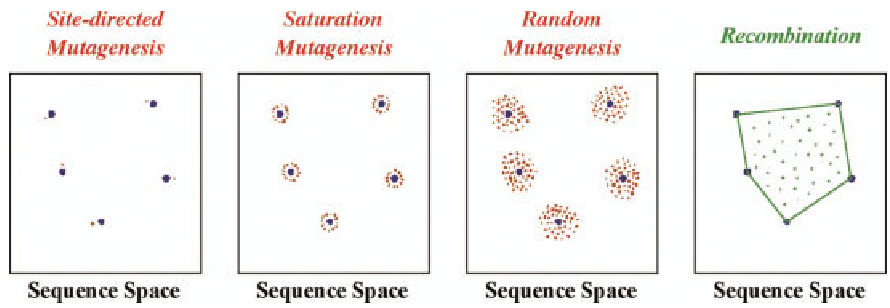

Figure 2. Depiction of the sequence space explored by mutagenesis (site-directed, saturation, and random) and recombination.
Large blue dots represent parental sequences, while smaller red (mutagenesis) and green (recombination) dots represent combinatorial DNA library members.

In the past few years, a wide range of success stories of directed evolution for many different applications has been reported (Petrounia and Arnold, 2000; Brakmann, 2001; Schmidt-Dannert, 2001; Bacher et al., 2002; Dalby, 2003). For example, Schneider et al. (2003) reengineered retroviruses used in gene therapy to greatly enhance their spreading efficiency through human fibrosarcoma cells. Schmidt-Dannert et al. (2000) used directed evolution to engineer a novel biosynthetic pathway in *E. coli* for the production of carotenoids, a diverse class of natural pigments that are of interest for pharmaceuticals and food colorants, while also playing a role in the prevention of cancer and chronic disease. Boder et al. (2000) generated single-chain antibodies that bind essentially irreversibly (femtomolar binding constant) with potential future implications for improved cancer and viral therapeutics. Bessler et al. (2003) enhanced the alkaline pH activity of an α -amylase that can be used to improve the starch removal capability of household detergents. Improved xylanases for wood pulp treatment (Burk, 2003) have led to substantial reduction in the use of bleaching agents, reducing their overall environmental impact. Briefly, other successes include many-fold improvements in enzyme activity and thermostability (Miyazaki et al., 2000; Baik et al., 2003), improved enantioselectivity (Reetz et al., 2001; Carr et al., 2003; Horsman et al., 2003), enhanced bioremediation (Wackett, 1998; Bruhlmann and Chen, 1999; Furukawa, 2000), and even the design of genetic circuits (Yokobayashi et al., 2002) and vaccines (Patten et al., 1997; Marzio et al., 2001; Whalen et al., 2001). It is increasingly becoming apparent, however, that it is vital to be able to assess and then "steer" diversity toward the most promising regions of sequence space (Moore et al., 1997). This is because only an infinitesimally small fraction of the diversity afforded by DNA and protein sequences can be examined regardless of the efficiency of the screening procedure. For example, a 500-nucleotide gene implies $4^{500} \approx 10^{301}$ alternatives, but even the most efficient screening methods can query only up to 10^{12} sequences (Olsen et al., 2000a; Chen and Georgiou, 2002; Lin and Cornish, 2002). Therefore, it is desirable to know how diversity is generated (see second section) and allocated (see third section) in the combinatorial DNA library and what sequence permutations are the most promising in terms of preserving protein structure and activity (see fourth section).

In the November 2003 issue, Lee and Reardon (2003) highlighted progress in the emerging field of proteomics, the system-wide analysis of protein sets. In this article, the engineering of specific proteins through combinatorial library design is examined. Different ways are described for generating library diversity through DNA manipulation, the advantages and disadvantages of various mutagenesis and recombination methods (including recent developments in nonhomologous and synthetic oligonucleotide recombination) are discussed, the computational challenges and progress at the level of combinatorial library generation are highlighted, and efforts are described to discern sequence composition vs. functionality trends at the protein level.

Experimental Techniques for DNA Library Generation

Methods for combinatorial library generation in directed evolution can be broadly classified depending on whether they utilize *mutagenesis* or *recombination* (see Figure 1) as the primary mechanism for generating diversity. Mutagenesis-based methods are deployed to (a) randomly distribute nucleotide mutations throughout the length of the parental DNA sequence(s) (*random mutagenesis*), (b) exhaustively generate all possible mutations at a particular sequence locus (*saturation mutagenesis*), or (c) produce specific nucleotide substitutions at predetermined locations (*site-directed mutagenesis*). Because it is often unclear which residues should be mutated (i.e., counterintuitive mutations distal to the active site frequently enhance activity/stability), the successful use of saturation and site-directed mutagenesis has so far been infrequent. More commonly, random mutagenesis has been used to generate libraries of mutated DNA sequences. It is typically performed by amplifying the initial parental DNA sequence(s) via the error-prone PCR reaction (Leung et al., 1989; Cadwell and Joyce, 1994; Lin-Goerke et al., 1997), which involves spiking the PCR reaction mixture with $MnCl_2$ to increase the mutation rate (other similar methods are described by Matsumura and Ellington (2002)). Another way to generate randomly distributed mutations is by transforming the parental DNA sequence(s) into one of many commercially available bacterial mutator strains (Greener et al., 1996). In all cases, the mutation rate must be carefully tuned to achieve a balance between progressing through sequence space at a "snail's pace" (low mutation rate) and a widespread loss of function in the library through a buildup of deleterious mutations (high mutation rate). Typically, an average rate of one to two amino acid changes per directed evolution cycle has been found to allow steady experimental progress (Arnold and Moore, 1997). Random mutagenesis methods are relatively inexpensive and easy to set up in the laboratory and have produced improved variants with nonobvious mutations absent from any known homologous sequences (Horsman et al., 2003). However, it is important to remember that only sequence diversity adjacent to the parental sequence(s) is probed (see Figure 2). Functioning distant sequence diversity is unlikely to be encountered given that this requires the sampling of an unbroken chain of continually improving point mutations. Moreover, after a few directed evolution cycles, mutational bias could be a factor in the sequence library. Due to redundancies in the codon representation (i.e., 64 codons for only 20 amino acids), a mutated nucleotide may not necessarily code for a different amino acid (silent mutations). Thus, amino acids with larger codon sets tend to mutate less often.

In addition to the use of point mutations for generating library diversity, DNA recombination is used to construct hybrids containing *crossovers*, defined as the junction points at which the sequence switches from one parent to another (see Figure 1). This allows, in principle, the sampling of sequences contained within the convex polytope defined by the vertices representing the parental sequences (see Figure 2). The key idea of recombination is to exchange proven diversity present in existing sequences. The use of DNA recombination for directed evolution was pioneered with the development of DNA shuffling (Stemmer, 1994), which relies on a PCR-like reaction for the reassembly of randomly fragmented parental sequences. Later, family DNA shuffling (Crameri et al., 1998; Ness et al., 1999) was demonstrated by recombining large sets of parental sequences simultaneously. A large number of related protocols such as StEP (Zhao et al., 1998), RACHITT (Coco et al., 2001), and single-stranded shuffling (Kikuchi et al., 2000) have also been developed. In all of these methods, crossover generation relies on the annealing and extension of complementary single-stranded fragments originating from different parental sequences (i.e., heteroduplex formation), which tends to bias crossover positions toward stretches of near perfect sequence identity. This, in turn, tends to give rise to biased combinatorial DNA libraries or, even worse, libraries with no additional diversity over the parental one.

In general, a severe bias toward the reassembly of parental sequences (i.e., no recombination) is observed when sequences with less than 60% sequence identity are recombined with annealing-based protocols (Stemmer, 1994; Moore et al., 2001). Given the fact that protein structure is more frequently conserved than DNA homology, annealing-based methods for recombining genes may potentially exclude solutions to protein engineering problems. The need for a recombination protocol capable of freely exchanging genetic diversity without sequence identity limitations motivated the development of the Incremental Truncation for the Creation of Hybrid EnzYmes (ITCHY) (Ostermeier et al., 1999a) and Sequence Homology-Independent Protein RECombination (SHIPREC) (Sieber et al., 2001) protocols. These protocols are capable of generating libraries from low sequence identity parents with crossovers evenly distributed along the length of the sequence (see analysis in Ostermeier (2003b)). However, ITCHY and SHIPREC are limited to constructing single crossover hybrids between only two parental sequences. Recent protocol design efforts have concentrated on overcoming this limitation by generating multiple crossovers per sequence without homology restrictions. The SCRATCHY protocol (Lutz et al., 2001b) generates multiple crossovers by applying DNA shuffling to ITCHY libraries, redistributing the prepositioned ITCHY crossovers throughout the newly reassembled sequences. The number of crossovers generated by SCRATCHY can be boosted even further by enriching the library via PCR amplification of crossover-containing sequence sections (Kawarasaki et al., 2003). The recently developed Sequence-Independent Site-Directed Chimeragenesis (SISDC) (Hiraga and Arnold, 2003), GeneReassembly (Richardson et al., 2002), and Structure-based COmbinatorial Protein Engineering (SCOPE) (O'Maille et al., 2002) protocols are fundamentally different from ITCHY/SCRATCHY and SHIPREC in that the crossover points must be predetermined prior to the recombination step. For these protocols, fragments have been shown to recombine independently without any sequence bias. A key advantage is the flexibility that they afford to predetermine the number and positions of "smart" crossover sites (Bogarad and Deem, 1999) that hopefully preserve functionality throughout the library.

All DNA recombination methods described so far involve the swapping and concurrent reassembly of parental nucleotide *segments* either obtained through DNA fragmentation or synthesis (GeneReassembly, SCOPE). However, using only nucleotide segments for diversity generation causes blocks of closely spaced polymorphisms to be swapped as a group, limiting library diversity (Ostermeier, 2003a). Synthetic oligonucleotide (nucleotide fragments with lengths of about 20–100 bases) recombination methods overcome this restriction by incorporating degenerate oligonucleotides into the reassembly procedure. The term *degenerate* refers to the synthesis of a mixture of oligonucleotides with different nucleotides (i.e., degeneracies) at certain prespecified positions. The oligonucleotides are designed to include coding information for the polymorphisms present in the parental set, while also including "customized" sequence identity enabling annealing-based recombination between the oligonucleotides. So far, degenerate oligonucleotides have been reassembled by PCR-based reactions (synthetic shuffling (Ness et al., 2002) and Assembly of Designed Oligonucleotides (ADO) (Zha et al., 2003)), as well as a single sequence of annealing, gap-filling, and ligation steps (degenerate homoduplex recombination (DHR) (Coco et al., 2002)). In all of these methods, increasing the corresponding oligonucleotide population in the mixture can boost the occurrence of rare mutations. Furthermore, the oligonucleotides can be designed to be consistent with the codon usage of a specific host organism. Synthetic oligonucleotide recombination can yield a very high crossover density (up to 1 crossover per 12.4 bp (Coco et al., 2002)); however, there is some concern that the high crossover density may disrupt vital interactions throughout the structure. In fact, a lower average library activity has been observed when comparing a synthetic shuffling library with one generated by family DNA shuffling (Ness et al., 2002). In general, the use of synthetic oligonucleotides has been more expensive and time-consuming than the recombination of parental DNA sequences.

Table 1 summarizes some of the advantages and disadvantages of each of the protocol types discussed. Recent develop-

Table 1. Summary of Methods for Combinatorial DNA Library Generation

Library Generation Method	Advantages	Disadvantages
Saturation Mutagenesis	√ Complete assessment of all possible mutations at a particular residue position	• Must predetermine residue position • Very limited exploration of sequence diversity
Random Mutagenesis *Error-prone PCR, mutator strains*	√ Easy, inexpensive setup	• Sequence diversity explored only near parental sequences • Biased mutational frequencies
Annealing-based Recombination *DNA shuffling, StEP, RACHITT, single-stranded shuffling*	√ Straightforward PCR-based protocol √ Large sets of parental sequences can be recombined	• Crossover positions biased toward stretches of sequence homology • Severe bias toward parental sequence reassembly when parents have less than 60% sequence identity
Nonhomologous Recombination *ITCHY, SHIPREC, SCRATCHY, SISDC, SCOPE, GeneReassembly*	√ No bias toward regions of sequence identity √ Multiple crossovers possible with SCRATCHY, SISDC, SCOPE, and GeneReassembly √ Can predetermine crossover sites for SISDC, SCOPE, GeneReassembly	• More complicated protocols • Only single-crossover hybrids generated with ITCHY and SHIPREC
Synthetic Oligonucleotide Recombination *Synthetic shuffling, ADO, DHR*	√ Crossovers can occur between closely spaced mutations √ Rare mutations can be boosted with added oligonucleotides √ Codon usage can be modified to comply with a particular host	• Average library activity can be lower due to broken couplings • Generally more expensive, time-consuming to design oligonucleotides

ments in experimental techniques have made it clear that, given sufficient resources, a protocol can be set up to create the desired level of diversity. However, what is less clear is what is the optimal level and type of diversity for a given protein engineering task. Although diversity is required to discover new variants, the average activity of the library tends to drop off as diversity increases (Ness et al., 2002; Ostermeier, 2003a). Ultimately, screening capacity limits and defines the optimal library diversity that needs to be considered. Recently, many exciting advances in high-throughput screening technologies have been made (see excellent reviews by Olsen et al. (2000a), Chen and Georgiou (2002), Lin and Cornish (2002)). For instance, phage display (Fernandez-Gacio et al., 2003) and ribosome display (Dower and Mattheakis, 2002) systems can be used to screen libraries with as many as 10^{12} members. The use of Fluorescence-Activated Cell Sorting (FACS) coupled with the cell-surface display of proteins and customized, Fluorescence Resonance Energy Transfer (FRET)- enabled substrates can be used to sort library members on the basis of k_{cat} or K_m at a rate of 10^9 per hour (Olsen et al., 2000b).

Computational Challenges at the DNA Level

Although the screening step in directed evolution probes for enhanced protein variants, the diversity generation step (i.e., combinatorialization) is performed via DNA manipulation. Without sufficient diversity in the underlying combinatorial DNA library, the encoded diversity within the protein library will be lacking as well, and the often expensive and labor-intensive screening step will underperform. Thus, being able to predict how alternate protocol setups affect the level and type of diversity generated can ultimately determine the success or failure of a directed evolution project. In this section, we describe efforts at developing predictive modeling frameworks for error-prone PCR and DNA shuffling protocols, followed by methods for optimizing combinatorial DNA library generation to target desired regions of sequence space.

Models for error-prone PCR have focused on predicting mutation rate for a given PCR setup (e.g., cycle number, annealing temperature, primer/template concentrations). This requires the consideration of (a) the plateau effect (where replication efficiency diminishes as the cycle number increases), (b) the propagation of mutations over a number of PCR cycles with nucleotide-dependent frequencies, and (c) the ability of nucleotides to back mutate to their original identity given that mutation rates are typically high in error-prone PCR. Some success has been achieved in modeling the plateau effect using kinetic parameters (Weiss and von Haeseler, 1995; Stolovitzky and Cecchi, 1996; Schnell and Mendoza, 1997a,b; Valikanov and Kapral, 1999). Moore and Maranas (2000) tracked mutations from cycle to cycle considering nucleotide-dependent mutation rates while allowing back mutation, but only with constant replication efficiency. Weiss and von Haeseler (1997) tracked mutations in combination with the plateau effect but did not include back mutation. Wang et al. (2000) developed a model that utilizes a branching process to track mutations and incorporates empirical information on the plateau effect. While quite a bit of progress has been achieved towards modeling error-prone PCR, a truly predictive model is still lacking.

Moving next to DNA recombination, Sun first considered models for DNA shuffling of parental sequences with single (Sun, 1998) and multiple (Sun, 1999) point mutations. However, these models did not consider sequence information, and their applicability was limited. Work in our group (Moore et al., 2001) examined for the first time how fragmentation length, annealing temperature, sequence identity, and number of shuffled parental sequences affect the number, type, and distribution of crossovers along the length of full-length reassembled sequences. In the *e*Shuffle framework, annealing events during reassembly were modeled as a network of reactions, and equilibrium thermodynamics along with complete nucleotide sequence information was employed to quantify their conversions and selectivities. Comparisons of *e*Shuffle predictions against experimental data revealed good agreement (Moore et al., 2001), particularly in light of the fact that there were no adjustable parameters. Specifically, we found that reducing fragmentation length boosted crossover numbers and annealing temperature and that crossovers tend to aggregate in regions of near perfect sequence identity. The customization of *e*Shuffle for the SCRATCHY protocol led to the *e*SCRATCHY framework (Lutz et al., 2001b). Using *e*SCRATCHY we found that in SCRATCHY libraries (a) fragmentation length used for reassembly does not influence the number or location of crossovers generated in full-length sequences, (b) the crossover distribution is shaped by the crossover statistics of the ITCHY library, and (c) crossovers are spread evenly throughout the crossover region. The need to safeguard against the formation of reassembled sequences with either truncated or duplicated domains motivated us to further extend the *e*Shuffle framework to consider out-of-sequence annealing events (Moore and Maranas, 2002b). Instead of "locking" fragments into their alignment positions, the annealing free energy change was used to determine the likelihood of duplex formation, allowing the prediction of the relative frequency that fragments from different sequence regions will anneal during reassembly.

Subsequent work by Maheshri and Schaffer (2003) further advanced the level of detail of DNA shuffling computational models with the development of a simulation-based model using nucleotide annealing kinetics and thermodynamics. This simulation approach has the advantage of tracking and recording the sequences of a computational ensemble of fragments through multiple rounds of shuffling, and tracks the fate of all reassembled fragments whether or not they are of parental length. A three-step reassembly process was used: (a) single-stranded fragments randomly collide; (b) on collision, a decision is made whether the molecules will hybridize and, if so, in what arrangement; and (c) duplexes are extended. This process is repeated until the fraction of unhybridized fragments remains unchanged; this constitutes a round of shuffling. Tracking the entire fragment pool allowed for the quantification of the trade-off between reassembly efficiency (i.e., the fraction of fragments that have reached parental length) and crossover frequency while simultaneously following the production of sequences with missing or repetitive regions. This work represented an important step in optimizing the recovery of diverse, full-length reassembled sequences from a DNA shuffling reaction mixture.

In addition to predictive frameworks for quantifying the allocated library diversity for a given protocol setup, a number of approaches have focused on the inverse problem. Specifically, how should we adjust the protocol setup to achieve the desired statistics of parental composition in the combinatorial libraries? In our group, we have explored the possibility of boosting or even specifically redirecting the formation of crossovers in DNA shuffling by exploiting the inherent redundancy in the codon representation (*e.g.*, isoleucine has the following three synonymous codon representations: ATA, ATC and ATT), while complying with host preferences for specific patterns of codon usage (Moore and Maranas, 2002a). The key motivation here is that it is possible to optimize the underlying parental DNA sequence codon representation for increasing and/or shaping diversity while at the same time preserving the parental amino acid encodings in the generated combinatorial protein libraries. To this end, the framework named *e*CodonOpt was developed for exploring the limits of performance that can be achieved through codon optimization.

While in *e*CodonOpt, the objective was to find a *single*-codon representation for each of the parental protein sequences, Wang and Saven (2002) designed instead an *ensemble* of nucleotide sequences that best "matches" a given set of amino acid probabilities. These probabilities can be derived from a multiple sequence alignment of protein family members (e.g., Pfam database (Bateman et al., 2002)) or statistical mechanics approaches that identify protein sequences likely to fit a given protein backbone (discussed in the next section). A two-term objective function was used to score the degree of correlation between the desired amino acid probability distribution and the distribution expected from the nucleotide ensemble. This objective accounts for (a) the absolute difference between desired and designed probabilities (based on the χ^2 function) and (b) a relative entropy term for quantifying the "distance" between the two distributions (Wang and Saven, 2002). The formulation can also be adapted to generate solutions in accordance with a particular host organism's codon preferences. Significant progress towards predicting and subsequently steering the statistics of unselected combinatorial DNA libraries has been achieved in the last few years. Additional improvements will require a more accurate description of hybridization kinetics and rates of polymerase mediated DNA extensions.

Computational Challenges at the Protein Level

Currently, two different paradigms are being pursued to computationally aid the design and composition of combinatorial protein libraries. The first involves the *a priori* design of a protein or collection of proteins that best fits a given protein fold. In this case, protein(s) are designed "from scratch" with little guidance from protein family sequence data. The second paradigm aims at elucidating what combinations of parental sequence fragments to include or exclude from the recombination mixture to create a combinatorial library that is both diverse and highly active. Proven diversity encoded here in the form of functional parental sequences is used to assess how well hybrid sequences fit the fold of interest.

Ab initio design of a protein or collection of proteins involves finding the amino acid sequence that best fits a given protein fold. The protein fold is represented by the Cartesian coordinates of its backbone atoms, which are usually fixed in space so that the degrees of freedom associated with backbone movement are neglected (some notable exceptions to the "fixed backbone" design paradigm include the work of Harbury et al. (1995), Harbury et al. (1998), Keating et al. (2001), Larson et al. (2002), Klepeis et al. (2003), and Kraemer-Pecore et al. (2003)). Candidate protein designs are generated by selecting amino acid side chains (at atomistic detail) along the backbone design scaffold. For simplicity, side chains are usually only permitted to assume a discrete set of statistically preferred conformations called *rotamers* (see (Dunbrack Jr., 2002) for a review of current rotamer libraries). Thus, a protein design consists of both a residue *and* rotamer assignment. To evaluate how well a possible design fits a given fold, rotamer/backbone and rotamer/rotamer interaction energies for all of the rotamers in the chosen library are tabulated. These potential energies can then be approximated using any of many standard force fields (e.g., CHARMM (MacKerell et al., 1998); DREIDING (Mayo et al., 1990); AMBER (Cornell et al., 1995); GROMOS (Scott et al., 1999)). Alternatively, energy/scoring functions that have been customized for protein design (Chiu and Goldstein, 1998; Kuhlman and Baker, 2000; Looger and Hellinga, 2001) are used. Protein design potentials (see Gordon et al., (1999) for a review) typically include van der Waals interactions, hydrogen bonding, electrostatics, solvation, and even entropy-based penalties for flexible side-chains (e.g., arginine).

Even for a small 50-residue protein, an enormous number (i.e., $153^{50} \approx 10^{109}$ assuming the Lovell et al., (2000) 153-rotamer library) of designs are possible. Both stochastic and deterministic search strategies have been used to tackle the computational challenge of finding the best design within this vast search space. Because activity level is very difficult to assess computationally, an alternative surrogate for hybrid fitness, namely stability, is employed in most studies. The key justification here is that stability is a prerequisite, although not necessarily a monotonic descriptor of functionality. Use of this indirect objective further necessitates the need of designing a combinatorial library, rather than a single design to improve the chances of success. Stochastic strategies search through the space of feasible designs by making a series of random and/or directed moves. Monte Carlo (Kuhlman and Baker, 2000; Kuhlman et al., 2002; Dantas et al., 2003), genetic algorithms (Desjarlais and Handel, 1995; Johnson et al., 1999; Raha et al., 2000), simulated annealing (Jiang et al., 2000; Xu and Farid, 2001), and many other heuristics (Wernisch et al., 2000; Jaramillo et al., 2002; Ogata et al., 2003) have been used in protein design with various levels of success. Although stochastic techniques can be used for problems of very large complexity with relatively small CPU/memory requirements, they are not guaranteed to converge to the optimal solution and require extensive tuning of parameters controlling the convergence rate (Desjarlais and Clarke, 1998; Voigt et al., 2000).

Conversely, deterministic algorithms are guaranteed to converge to the global minimum energy conformation; however, they tend to be long-running and become intractable for large-scale design problems. The most frequently used deterministic technique is dead-end elimination (Desmet et al., 1992), a pruning method in which rotamers and rotamer pairs that cannot be part of the optimal protein design are eliminated over a number of computational cycles. Recent innovations to accelerate rotamer elimination include the use of upper-bounding information (Gordon and Mayo, 1999), conformational splitting (Pierce et al., 2000), the "magic bullet" metric (Gordon and Mayo, 1998), and background optimization (Looger and Hellinga, 2001). Dead-end elimination has been used to design the full sequence of a 28-residue zinc finger (Dahiyat and Mayo, 1997); the cores of T4 lysozyme (26 residues) (Mooers et al., 2003), thioredoxin (32 residues) (Bolon et al., 2003), and the $\alpha M\beta 2$ integrin I domain (45 residues) (Shimaoka et al., 2000); small molecule receptors based on periplasmic binding proteins (Looger et al., 2003); and metal binding proteins (Dwyer et al., 2003).

In practice, more important than finding the mathematical solution to the protein design problem is the ability to generate *in silico* an ensemble of computational designs that subsequently will form the basis for constructing the combinatorial protein library. Furthermore, because the most active proteins are often only marginally stable, examining sub-optimal designs can yield greater insight into a fold's plasticity. Sub-optimal designs may be collected by storing intermediate steps of stochastic searches (e.g., Monte Carlo as in (Hayes et al., 2002)); however, the top 10^5 or even 10^6 designs are not sufficient to completely characterize the vast sequence space associated with large proteins. Alternatively, statistical mechanics based methods can be used to construct, equilibrate, and query ensembles of all possible residue/rotamer states (see Saven (2001) for a review). Mean-field theory allows the extraction of individual rotamer site probabilities (first-order; (Koehl and Delarue, 1994; Lee, 1994; Mendes et al., 1999; Voigt et al., 2001)) or rotamer-rotamer joint probabilities (second-order; (Moore and Maranas, 2003)) after the free energy of the ensemble is minimized. The probabilities represent how well a particular rotamer (or rotamer pair) fits at a particular sequence position (or pair of positions). Equivalently, Saven and co-workers have introduced a method for extracting rotamer site probabilities from a maximal-entropy ensemble (Zou and Saven, 2000; Kono and Saven, 2001).

The methods described so far followed the first paradigm that aims to design proteins and/or libraries "from scratch" that best fit the fold of interest. However, directed evolution experiments have a natural starting point—the original parental sequences. Following the second paradigm, a number of strategies have been developed that utilize the sequence and structure information encoded in the parental sequences to guide the design of combinatorial protein libraries. Typically, this involves the scoring of libraries of hybrid protein sequences against the parental sequences. This idea was first demonstrated with the SCHEMA algorithm (Voigt et al., 2002), which hypothesized structural disruption whenever a contacting residue pair (within 4.5 Å) in a hybrid has differing parental origins. Hybrids are scored for stability by counting the number of disruptions. SCHEMA also uses the information on residue pair disruptions to partition the protein into blocks that should not be interrupted by crossovers (analogous to the schema theory of genetic algorithms (Holland, 1975)). The algorithm was then used to show that crossover distributions in a number of experiments were preferentially allocated to avoid disrupting these blocks (Voigt et al., 2002). Alhough quite successful so far, this approach cannot differentiate between hybrids with different directionality also known as "mirror" chimeras (i.e., A-B vs. B-A arrangement of segments), which have been shown to often have very different functional crossover profiles (Lutz et al., 2001b).

In our group, we have reevaluated the effect of having contacting residue pairs with different parental origins. Instead of always counting them as unfavorable, we view such pairs as places where potential clashes may occur between contacting residues. In the Second-order mean-field Identification of Residue-residue Clashes in protein Hybrids (SIRCH) (Moore and Maranas, 2003) procedure for evaluating protein hybrids, an extended, *second-order* mean-field description is used to elucidate the probabilities of all possible residue-residue combinations in a minimum Helmholtz free energy ensemble. The pairwise substitution patterns uncovered by the second-order mean-field description are then used to detect clashes in potential hybrids. SIRCH has been used to analyze pairwise substitution patterns in the dihydrofolate reductase (DHFR) enzyme and to assess the result of the recombination of *E. coli* and human glycinamide ribonucleotide (GAR) transformylases (Ostermeier et al., 1999b; Lutz et al., 2001ab). Results demonstrate that experimentally determined functional crossover positions for the GAR transformylases are consistent with the predicted residue-residue clashes. Analysis of these predicted clashes revealed that they primarily arise due to (a) the introduction of repulsive residue pairs such as +/+ or -/-, (b) the disruption of hydrogen bonds due to the formation of donor/donor or acceptor/acceptor pairs, and (c) the generation of steric clashes or cavities (Saraf and Maranas, 2003).

SCHEMA, SIRCH, and residue clash maps are increasingly being used to predict "smart" crossover sites (Meyer et al., 2003) for experimental protocols that require preset crossover positions, such as SISDC, Gene Reassembly, and synthetic oligonucleotide recombination methods. In addition, clash map information can be used in conjunction with protein design algorithms to suggest site-directed mutagenesis strategies for alleviating clashes in either parental sequences (upstream) or promising hybrids (downstream).

Future Perspectives

As we enter the post-genomic era, we have in our hands an abundance of protein designs, experimental techniques, and computational approaches. By creatively applying the ever-growing palette of molecular biology techniques, a variety of protocols are currently available for constructing combinatorial libraries with customized statistics of mutations and/or parental fragments. Future protocol developments are likely to be driven by the need to navigate around the increasingly complicated intellectual property landscape. To this end, the use of synthetic oligomers, taking advantage of substantial reductions in price, is likely to dominate, thus providing the means for exquisite control of combinatorial library diversity.

These enabling technology developments, along with the emerging trend of recombining more distant homologues, will further stress the need to computationally assess protein hybrids for stability and even functionality. The key dilemma of computational developments lies at establishing the proper trade-off between modeling accuracy and evaluation speed. Force fields are increasingly becoming more elaborate and customized to the task of protein engineering. However, there is almost unanimous agreement that their accuracy is still limited. For instance, an adequate and computationally tractable description of electrostatics remains elusive. Notable contributions in this direction include the recent work of Hellinga's group (Wisz and Hellinga, 2003). In response to the inherent difficulty of designing potentials with a firm grounding on biophysics fundamentals, a number of researchers are increasingly developing and successfully making use of scoring functions heavily parameterized to predict existing folds (Kuhlman and Baker, 2000). A recent impressive contribution along these lines is the *in silico* design and verification of a novel fold by Baker's group (Kuhlman et al., 2003).

Even though ample experimental evidence shows that proteins have not evolved to maximize their stability, most computational approaches have aimed to design proteins with this as an objective. This is primarily a manifestation of our inability to *a priori* predict functionality rather than an affirmation that stability and functionality are always correlated. Clearly, there is a need to move beyond stability as a monolithic surrogate for functionality. To this end, sequence information gleaned from protein family databases (e.g., Pfam (Bateman et al., 2002)) can indirectly provide some answers. In the same way that protein structures in the Protein Data Bank (Berman et al., 2000) have been used to design potential energy functions for protein design, protein family sequence data, spanning all of nature's known solutions, can be used to constrain the solutions for various protein engineering problems. In fact, Lockless and Ranganathan (1999) have found that statistical sequence database-derived coupling energies correlate with thermodynamic coupling free energies (i.e., $\delta\delta G$ from double mutant cycle analysis) in a small protein domain.

Furthermore, it is important to stress that current protein design methods rely on a static picture for proteins. However, it is increasingly being accepted that proteins require the coordinated motion of an extensive network of interacting residues for correct catalytic function (see Benkovic and Hammes-Schiffer (2003) for review). Hybrid quantum-classical molecular dynamics (MD) simulations of wild-type and mutant dihydrofolate reductases uncovered a network of coupled promoting motions that occur as the wild-type hydride transfer reaction progresses (Agarwal et al., 2002). The network was found to be disrupted in the mutant, reflecting its reduced reaction rate. In addition, recent MD simulations have revealed a link between thermostability and the fluctuations of surface loops away from the native state (Wintrode et al., 2003). Incorporating dynamic information into protein design frameworks is likely to be challenging but may prove necessary to design proteins with novel functions.

The ever-accelerating rate of searching sequence space, driven by increased computational speed and clever algorithm design, is likely to continue. Particularly promising will be methods that can effectively combine the ability of stochastic methods (e.g., genetic algorithms and simulated annealing) to scan vast amounts of sequence space with deterministic algorithms (e.g., dead-end elimination) that can produce provably optimal solutions. Motivated by the need to design protein-based therapeutics and proteins with novel functionalities, exciting developments are likely to be forthcoming fueled by the inventiveness and constrained only by the imagination of experimentalists and theoreticians.

Acknowledgments

The authors would like to thank Professor Stefan Lutz and Dr. Alexander Horswill for useful suggestions. Funding by the National Science Foundation grant BES0331047 is gratefully acknowledged.

Literature Cited

Agarwal, P. K., S. R. Billeter, P. T. Rajagopalan, S. J. Benkovic and S. Hammes-Schiffer, "Network of Coupled Promoting Motions in Enzyme Catalysis," *Proc. Natl. Acad. Sci. USA*, **99**, 2794 (2002).

Arnold, F. H., and J. C. Moore, "Optimizing Industrial Enzymes by Directed Evolution," *Adv. Biochem. Eng. Biotechnol*,. **58**, 1 (1997).

Bacher, J. M., B. D. Reiss, and A. D. Ellington, "Anticipatory Evolution and DNA Shuffling," *Genome Biol.*, **3**, REVIEWS1021 (2002).

Baik, S. H., T. Ide, H. Yoshida, O. Kagami, and S. Harayama, "Significantly Enhanced Stability of Glucose Dehydrogenase by Directed Evolution," *Appl. Microbiol. Biotechnol.*, **61**, 329 (2003).

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer, "The Pfam Protein Families Database," *Nucleic Acids Res.*, **30**, 276 (2002).

Benkovic, S. J., and S. Hammes-Schiffer, "A Perspective on Enzyme Catalysis," *Science*, **301**, 1196 (2003).

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, **28**, 235 (2000).

Bessler, C., J. Schmitt, K. H. Maurer, and R. D. Schmid, "Directed Evolution of a Bacterial Alpha-amylase: Toward Enhanced pH-Performance and Higher Specific Activity," *Protein Sci.*, **12**, 2141 (2003).

Boder, E. T., K. S. Midelfort, and K. D. Wittrup, "Directed Evolution of Antibody Fragments with Monovalent Femtomolar Antigen-Binding Affinity," *Proc. Natl. Acad. Sci. USA*, **97**, 10701 (2000).

Bogarad, L. D., and M. W. Deem, "A Hierarchical Approach to Protein Molecular Evolution," *Proc. Natl. Acad. Sci. USA*, **96**, 2591 (1999).

Bolon, D. N., J. S. Marcus, S. A. Ross, and S. L. Mayo, "Prudent Modeling of Core Polar Residues in Computational Protein Design," *J. Mol. Biol.*, **329**, 611 (2003).

Bradley, P., et al., "Rosetta Predictions in CASP5: Successes, Failures, and Prospects for Complete Automation," *Proteins* **53 Suppl 6**, 457 (2003).

Brakmann, S., "Discovery of Superior Enzymes by Directed Molecular Evolution," *ChemBioChem*, **2**, 865 (2001).

Bruhlmann, F., and W. Chen, "Tuning Biphenyl Dioxygenase for Extended Substrate Specificity," *Biotechnol. Bioeng.*, **63**, 544 (1999).

Burk, M., "Discovery and Optimization of Enzymes for Chemical Transformations through Biodiversity Access and Directed Evolution Technologies," presented at *Enzyme Engineering XVII*, Santa Fe, NM (2003).

Cadwell, R. C., and G. F. Joyce, "Mutagenic PCR," *PCR Methods Appl.*, **3**, S136 (1994).

Carr, R., M. Alexeeva, A. Enright, T. S. Eve, M. J. Dawson, and N. J. Turner, "Directed Evolution of an Amine Oxidase Possessing both Broad Substrate Specificity and High Enantioselectivity," *Angew. Chem. Int. Ed. Engl* **42**, 4807 (2003).

Chen, W., and G. Georgiou, "Cell-Surface Display of Heterologous Proteins: From High-throughput Screening to Environmental Applications," *Biotechnol. Bioeng.*, **79**, 496 (2002).

Chiu, T. L., and R. A. Goldstein, "Optimizing Potentials for the Inverse Protein Folding Problem," *Protein Eng.*, **11**, 749 (1998).

Coco, W. M., et al., "Growth Factor Engineering by Degenerate Homoduplex Gene Family Recombination," *Nat. Biotechnol.*, **20**, 1246 (2002).

Coco, W. M., et al., "DNA Shuffling Method for Generating Highly Recombined Genes and Evolved Enzymes," *Nat. Biotechnol.*, **19**, 354 (2001).

Cornell, W. D., et al., "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids and Organic Molecules," *J. Am. Chem. Soc.*, **117**, 5179 (1995).

Crameri, A., S. A. Raillard, E. Bermudez, and W. P. Stemmer, "DNA Shuffling of a Family of Genes from Diverse Species Accelerates Directed Evolution," *Nature*, **391**, 288 (1998).

Dahiyat, B. I., and S. L. Mayo, "De novo Protein Design: Fully Automated Sequence Selection," *Science*, **278**, 82 (1997).

Dalby, P. A., "Optimising Enzyme Function by Directed Evolution," *Curr. Opin. Struct. Biol.* **13**, 500 (2003).

Dantas, G., B. Kuhlman, D. Callender, M. Wong, and D. Baker, "A Large Scale Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins," *J. Mol. Bio..*, **332**, 449 (2003).

Desjarlais, J. R., and N. D. Clarke, "Computer Search Algorithms in Protein Modification and Design," *Curr. Opin. Struct. Biol.*, **8**, 471 (1998).

Desjarlais, J. R., and T. M. Handel, "De Novo Design of the Hydrophobic Cores of Proteins," *Protein Sci.*, **4**, 2006 (1995).

Desmet, J., M. Demaeyer, B. Hazes, and I. Lasters, "The Dead-End Elimination Theorem and its use in protein side-chain positioning," *Nature*, **356**, 539 (1992).

Dower, W. J., and L. C. Mattheakis, "In vitro Selection as a Powerful Tool for the Applied Evolution of Proteins and Peptides," *Curr. Opin. Chem. Biol.*, **6**, 390 (2002).

Dunbrack Jr., R. L., "Rotamer Libraries in the 21st Century," *Curr. Opin. Struct. Biol.*, **12**, 431 (2002).

Dwyer, M. A., L. L. Looger, and H. W. Hellinga, "Computational Design of a Zn2+ Receptor that Controls Bacterial Gene Expression," *Proc. Natl. Acad. Sci. USA*, **100**, 11255 (2003).

Fernandez-Gacio, A., M. Uguen, and J. Fastrez, "Phage Display as a Tool for the Directed Evolution of Enzymes," *Trends Biotechnol.* , **21**, 408 (2003).

Furukawa, K., "Engineering Dioxygenases for Efficient Degradation of Environmental Pollutants," *Curr. Opin. Biotechnol.*, **11**, 244 (2000).

Gordon, D. B., S. A. Marshall and S. L. Mayo, "Energy Functions for Protein Design," *Curr. Opin. Struct. Biol.*, **9**, 509 (1999).

Gordon, D. B., and S. L. Mayo, "Radical Performance Enhancements for Combinatorial Optimization Algorithms Based on the Dead-End Elimination Theorem," *J. Comp. Chem.*, **19**, 1505 (1998).

Gordon, D. B., and S. L. Mayo, "Branch-and-Terminate: a Combinatorial Optimization Algorithm for Protein Design," *Structure*, **7**, 1089 (1999).

Greener, A., M. Callahan, and B. Jerpseth, "An Efficient Random Mutagenesis Technique Using an E. coli Mutator Strain," *Methods Mol. Biol.*, **57**, 375 (1996).

Harbury, P. B., J. J. Plecs, B. Tidor, T. Alber and P. S. Kim, "High-Resolution Protein Design with Backbone Freedom," *Science*, **282**, 1462 (1998).

Harbury, P. B., B. Tidor, and P. S. Kim, "Repacking Protein

Cores with Backbone Freedom: Structure Prediction for Coiled Coils," *Proc. Natl. Acad. Sci. USA*, **92**, 8408 (1995).

Hayes, R. J., et al., "Combining Computational and Experimental Screening for Rapid Optimization of Protein Properties," *Proc. Natl. Acad. Sci. USA*, **99**, 15926 (2002).

Hiraga, K., and F. H. Arnold, "General Method for Sequence-Independent Site-Directed Chimeragenesis," *J. Mol. Biol.*, **330**, 287 (2003).

Holland, J., *Adaptation in Natural and Artificial Systems*, The University of Michigan Press, Ann Arbor, MI (1975).

Horsman, G. P., A. M. Liu, E. Henke, U. T. Bornscheuer, and R. J. Kazlauskas, "Mutations in Distant Residues Moderately Increase the Enantioselectivity of Pseudomonas fluorescens esterase towards Methyl 3bromo-2-methylpropanoate and Ethyl 3phenylbutyrate," *Chemistry*, **9**, 1933 (2003).

Jaramillo, A., L. Wernisch, S. Hery, and S. J. Wodak, "Folding Free Energy Function Selects Native-like Protein Sequences in the Core but Not on the Surface," *Proc. Natl. Acad. Sci. USA*, **99**, 13554 (2002).

Jiang, X., H. Farid, E. Pistor, and R. S. Farid, "A New Approach to the Design of Uniquely Folded Thermally Stable Proteins," *Protein Sci.*, **9**, 403 (2000).

Johnson, E. C., G. A. Lazar, J. R. Desjarlais and T. M. Handel, "Solution Structure and Dynamics of a Designed Hydrophobic Core Variant of Ubiquitin," *Structure Fold Des.*, **7**, 967 (1999).

Kawarasaki, Y., et al., "Enhanced Crossover SCRATCHY: Construction and High-throughput Screening of a Combinatorial Library Containing Multiple Non-Homologous Crossovers," *Nucleic Acids Res.*, **31**, e126 (2003).

Keating, A. E., V. N. Malashkevich, B. Tidor, and P. S. Kim, "Side-Chain Repacking Calculations for Predicting Structures and Stabilities of Heterodimeric Coiled Coils," *Proc. Natl. Acad. Sci. USA*, **98**, 14825 (2001).

Kikuchi, M., K. Ohnishi, and S. Harayama, "An Effective Family Shuffling Method Using Single-Stranded DNA," *Gene*, **243**, 133 (2000).

Klepeis, J. L., et al., "Integrated Computational and Experimental Approach for Lead Optimization and Design of Compstatin Variants with Improved Activity," *J Am Chem Soc.*, **125**, 8422 (2003).

Koehl, P., and M. Delarue, "Application of a Self-Consistent Mean Field Theory to Predict Protein Side-Chains Conformation and Estimate their Conformational Entropy," *J. Mol. Bio.l*, **239**, 249 (1994).

Kono, H., and J. G. Saven, "Statistical Theory for Protein Combinatorial Libraries. Packing Interactions, Backbone Flexibility, and the Sequence Variability of a Main-Chain Structure," *J. Mol. Biol.*, **306**, 607 (2001).

Kraemer-Pecore, C. M., J. T. Lecomte and J. R. Desjarlais, "A de novo Redesign of the WW Domain," *Protein Sci.*, **12**, 2194 (2003).

Kuhlman, B., and D. Baker, "Native Protein Sequences are Close to Optimal for their Structures," *Proc. Natl. Acad. Sci. USA*, **97**, 10383 (2000).

Kuhlman, B., G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, "Design of a Novel Globular Protein Fold with Atomic-Level Accuracy," *Science*, **302**, 1364 (2003).

Kuhlman, B., J. W. O'Neill, D. E. Kim, K. Y. Zhang, and D. Baker, "Accurate Computer-Based Design of a New Backbone Conformation in the Second Turn of Protein L," *J. Mol. Biol.*, **315**, 471 (2002).

Larson, S. M., J. L. England, J. R. Desjarlais, and V. S. Pande, "Thoroughly Sampling Sequence Space: Large-Scale Protein Design of Structural Ensembles," *Protein Sci.*, **11**, 2804 (2002).

Lee, C., "Predicting Protein Mutant Energetics by Self-Consistent Ensemble Optimization," *J. Mol. Biol.*, **236**, 918 (1994).

Lee, K. H., and K. F. Reardon, "Proteomics: An Exciting New Science, but Where are the Chemical Engineers?," *AIChE J.*, **49**, 2682 (2003).

Leung, D. W., E. Chen, and D. V. Goeddel, "A Method for Random Mutagenesis of a Defined DNA Segment Using a Modified Polymerase Chain Reaction," *Technique*, **1**, 11 (1989).

Lin, H., and V. W. Cornish, "Screening and Selection Methods for Large-Scale Analysis of Protein Function," *Angew. Chem. Int. Ed. Engl.*, **41**, 4402 (2002).

Lin-Goerke, J. L., D. J. Robbins, and J. D. Burczak, "PCR-Based Random Mutagenesis Using Manganese and Reduced dNTP Concentration," *Biotechniques*, **23**, 409 (1997).

Lockless, S. W., and R. Ranganathan, "Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families," *Science*, **286**, 295 (1999).

Looger, L. L., M. A. Dwyer, J. J. Smith, and H. W. Hellinga, "Computational Design of Receptor and Sensor Proteins with Novel Functions," *Nature*, **423**, 185 (2003).

Looger, L. L., and H. W. Hellinga, "Generalized Dead-End Elimination Algorithms Make Large-Scale Protein Side-Chain Structure Prediction Tractable: Implications for Protein Design and Structural Genomics," *J. Mol. Biol.*, **307**, 429 (2001).

Lovell, S. C., J. M. Word, J. S. Richardson, and D. C. Richardson, "The Penultimate Rotamer Library," *Proteins*, **40**, 389 (2000).

Lutz, S., M. Ostermeier, and S. J. Benkovic, "Rapid Generation of Incremental Truncation Libraries for Protein Engineering Using Alpha-Phosphothioate Nucleotides," *Nucleic Acids Res.*, **29**, E16 (2001a).

Lutz, S., M. Ostermeier, G. L. Moore, C. D. Maranas, and S. J. Benkovic, "Creating Multiple Crossover Libraries Independent of Sequence Identity," *Proc. Natl. Acad. Sci. USA*, **98**, 11248 (2001b).

MacKerell, A. D., et al., "CHARMM: The Energy Function and its Parameterization with an Overview of the Program, *The Encyclopedia of Computational Chemistry*, **1**, R. Schleyer, ed., Wiley, Chichester, U.K., p. 271 (1998).

Maheshri, N., and D. V. Schaffer, "Computational and Experimental Analysis of DNA Shuffling," *Proc Natl Acad Sci U S A*, **100**, 3071 (2003).

Marzio, G., K. Verhoef, M. Vink, and B. Berkhout, "In vitro Evolution of a Highly Replicating, Doxycycline-Dependent HIV for Applications in Vaccine Studies," *Proc. Natl. Acad. Sci. USA*, **98**, 6342 (2001).

Matsumura, I., and A. D. Ellington, "Mutagenic Polymerase Chain Reaction of Protein-Coding Genes for in vitro Evolution," *Meth. Mol. Biol.*, **182**, 259 (2002).

Mayo, S. L., B. D. Olafson and W. A. Goddard, "DREIDING-A Generic Force-Field for Molecular Simulations," *J. Phys. Chem.*, **94**, 8897 (1990).

Mendes, J., C. M. Soares, and M. A. Carrondo, "Improvement of Side-Chain Modeling in Proteins with the Self-Consistent Mean Field Theory Method Based on an Analysis of the Factors Influencing Prediction," *Biopolymers*, **50**, 111 (1999).

Meyer, M. M., et al., "Library Analysis of SCHEMA-Guided Protein Recombination," *Protein Sci.*, **12**, 1686 (2003).

Miyazaki, K., P. L. Wintrode, R. A. Grayling, D. N. Rubingh, and F. H. Arnold, "Directed Evolution Study of Temperature Adaptation in a Psychrophilic Enzyme," *J. Mol. Biol.*, **297**, 1015 (2000).

Mooers, B. H., D. Datta, W. A. Baase, E. S. Zollars, S. L. Mayo, and B. W. Matthews, "Repacking the Core of T4 lysozyme by Automated Design," *J. Mol. Biol.*, **332**, 741 (2003).

Moore, G. L., and C. D. Maranas, "Modeling DNA Mutation and Recombination for Directed Evolution Experiments," *J. Theor. Biol.*, **205**, 483 (2000).

Moore, G. L., and C. D. Maranas, "eCodonOpt: a Systematic Computational Framework for Optimizing Codon Usage in Directed Evolution Experiments," *Nucleic Acids Res.*, **30**, 2407 (2002a).

Moore, G. L., and C. D. Maranas, "Predicting Out-of-Sequence Reassembly in DNA Shuffling," *J. Theor. Biol.*, **219**, 9 (2002b).

Moore, G. L., and C. D. Maranas, "Identifying Residue-Residue Clashes in Protein Hybrids by Using a Second-Order Mean-Field Approach," *Proc. Natl. Acad. Sci. USA*, **100**, 5091 (2003).

Moore, G. L., C. D. Maranas, S. Lutz, and S. J. Benkovic, "Predicting Crossover Generation in DNA Shuffling," *Proc. Natl. Acad. Sci. USA*, **98**, 3226 (2001).

Moore, J. C., H. Jin, O. Kuchner, and F. H. Arnold, "Strategies for the in vitro Evolution of Protein Function: Enzyme Evolution by Random Recombination of Improved Sequences," *J. Mol. Biol.*, **272**, 336 (1997).

Ness, J. E., et al., "Synthetic Shuffling Expands Functional Protein Diversity by Alowing Amino Acids to Recombine Independently," *Nat. Biotechnol.*, **20**, 1251 (2002).

Ness, J. E., et al., "DNA Shuffling of Subgenomic Sequences of Subtilisin," *Nat. Biotechnol.*, **17**, 893 (1999).

Ogata, K., et al., "Automatic Sequence Design of Major Histocompatibility Complex Class I Binding Peptides Impairing CD8+ T Cell Recognition," *J. Biol. Chem.*, **278**, 1281 (2003).

Olsen, M., B. Iverson, and G. Georgiou, "High-Throughput Screening of Enzyme Libraries," *Curr. Opin. Biotechnol.*, **11**, 331 (2000a).

Olsen, M. J., D. Stephens, D. Griffiths, P. Daugherty, G. Georgiou, and B. L. Iverson, "Function-Based Isolation of Novel Enzymes from a Large Library," *Nat. Biotechnol.*, **18**, 1071 (2000b).

O'Maille, P. E., M. Bakhtina, and M. D. Tsai, "Structure-Based Combinatorial Protein engineering (SCOPE)," *J. Mol. Biol.* **321**, 677 (2002).

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

Ostermeier, M., "Theoretical Distribution of Truncation Lengths in Incremental Truncation Libraries," *Biotechnol. Bioeng.*, **82**, 564 (2003b).

Ostermeier, M., A. E. Nixon, J. H. Shim, and S. J. Benkovic, "Combinatorial Protein Engineering by Incremental Truncation," *Proc. Natl. Acad. Sci. USA*, **96**, 3562 (1999a).

Ostermeier, M., J. H. Shim, and S. J. Benkovic, "A Combinatorial Approach to Hybrid Enzymes Independent of DNA Homology," *Nat. Biotechnol.*, **17**, 1205 (1999b).

Patten, P. A., R. J. Howard, and W. P. Stemmer, "Applications of DNA Shuffling to Pharmaceuticals and Vaccines," *Curr. Opin. Biotechnol.*, **8**, 724 (1997).

Petrounia, I. P., and F. H. Arnold, "Designed Evolution of Enzymatic Properties," *Curr. Opin. Biotechnol.*, **11**, 325 (2000).

Pierce, N. A., J. A. Spriet, J. Desmet, and S. L. Mayo, "Conformational Splitting: A More Powerful Criterion for Dead-End Elimination," *J. Comp. Chem.*, **21**, 999 (2000).

Raha, K., A. M. Wollacott, M. J. Italia, and J. R. Desjarlais, "Prediction of Amino Acid Sequence from Structure," *Protein Sci.*, **9**, 1106 (2000).

Reetz, M. T., S. Wilensek, D. Zha, and K. E. Jaeger, "Directed Evolution of an Enantioselective Enzyme through Combinatorial Multiple-Cassette Mutagenesis," *Angew. Chem. Int. Ed. Engl.*, **40**, 3589 (2001).

Richardson, T. H., et al., "A Novel, High Performance Enzyme for Starch Liquefaction. Discovery and Optimization of a Low pH, Thermostable Alpha-amylase," *J. Biol. Chem.*, **277**, 26501 (2002).

Saraf, M. C., and C. D. Maranas, "Using a Residue Clash Map to Functionally Characterize Protein Recombination Hybrids," *Protein Eng.*, in press (2003).

Saven, J. G., "Designing Protein Energy Landscapes," *Chem. Rev.*, **101**, 3113 (2001).

Schmidt-Dannert, C., "Directed Evolution of Single Proteins, Metabolic Pathways, and Viruses," *Biochemistry*, **40**, 13125 (2001).

Schmidt-Dannert, C., D. Umeno, and F. H. Arnold, "Molecular Breeding of Carotenoid Biosynthetic Pathways," *Nat. Biotechnol.*, **18**, 750 (2000).

Schneider, R. M., et al., "Directed Evolution of Retroviruses Activatable by Tumour-Associated Matrix Metalloproteases," *Gene Ther.*, **10**, 1370 (2003).

Schnell, S. and C. Mendoza, "Enzymological Considerations for a Theoretical Description of the Quantitative Competitive Polymerase Chain Reaction (QC-PCR)," *J. Theor. Biol.*, **184**, 433 (1997a).

Schnell, S., and C. Mendoza, "Theoretical Description of the Polymerase Chain Reaction," *J. Theor. Biol.*, **188**, 313 (1997b).

Scott, W. R., et al., "The GROMOS Biomolecular Simulation Program Package," *J. Phys. Chem. A*, **103**, 3596 (1999).

Shimaoka, M., J. M. Shifman, H. Jing, J. Takagi, S. L. Mayo. and T. A. Springer, "Computational Design of an Integrin I Domain Stabilized in the Open High Affinity Conformation," *Nat. Struct. Biol.*, **7**, 674 (2000).

Sieber, V., C. A. Martinez, and F. H. Arnold, "Libraries of Hybrid Proteins from Distantly Related Sequences," *Nat. Biotechnol.*, **19**, 456 (2001).

Stemmer, W. P., "DNA Shuffling by Random Fragmentation and Reassembly: in vitro Recombination for Molecular Evolution," *Proc. Natl. Acad. Sci. USA*, **91**, 10747 (1994).

Stolovitzky, G., and G. Cecchi, "Efficiency of DNA Replication in the Polymerase Chain Reaction," *Proc. Natl. Acad. Sci. USA*, **93**, 12947 (1996).

Sun, F., “Modeling DNA Shuffling,” *RECOMB ’98, Proc. of the Second Annual International Conference on Computational Molecular Biology*, 251 (1998).
Sun, F., “Modeling DNA Shuffling,” *J. Comput. Biol.*, **6**, 77 (1999).
Valikanov, M. V., and R. Kapral, “Polymerase Chain Reaction: a Markov Process Approach,” *J. Theor. Biol.*, **201**, 239 (1999).
Voigt, C. A., D. B. Gordon, and S. L. Mayo, “Trading accuracy for Speed: A Quantitative Comparison of Search Algorithms in Protein Sequence Design,” *J. Mol. Biol.*, **299**, 789 (2000).
Voigt, C. A., C. Martinez, Z.-G. Wang, S. L. Mayo, and F. H. Arnold, “Protein Building Blocks Preserved by Recombination,” *Nat. Struct. Biol.*, **9**, 553 (2002).
Voigt, C. A., S. L. Mayo, F. H. Arnold, and Z. G. Wang, “Computational Method to Reduce the Search Space for Directed Protein Evolution,” *Proc. Natl. Acad. Sci. USA*, **98**, 3778 (2001).
Wackett, L. P., “Directed Evolution of New Enzymes and Pathways for Environmental Catalysis,” *Ann. NY Acac. Sci.*, **864**, 142 (1998).
Wang, D., C. Zhao, R. Cheng, and F. Sun, “Estimation of the Mutation Rate during Error-Prone Polymerase Chain Reaction,” *J. Comput. Biol.*, **7**, 143 (2000).
Wang, W., and J. G. Saven, “Designing Gene Libraries from Protein Profiles for Combinatorial Protein Experiments,” *Nucleic Acids Res.*, **30**, e120 (2002).
Weiss, G., and A. von Haeseler, “Modeling the Polymerase Chain Reaction,” *J. Comput. Biol.*, **2**, 49 (1995).
Weiss, G., and A. von Haeseler, “A Coalescent Approach to the Polymerase Chain Reaction,” *Nucleic Acids Res.*, **25**, 3082 (1997).
Wernisch, L., S. Hery, and S. J. Wodak, “Automatic Protein Design with all Atom Force-Fields by Exact and Heuristic Optimization,” *J. Mol. Biol.*, **301**, 713 (2000).
Whalen, R. G., R. Kaiwar, N. W. Soong, and J. Punnonen, “DNA Shuffling and Vaccines,” *Curr. Opin. Mol. Ther.*, **3**, 31 (2001).
Wintrode, P. L., D. Zhang, N. Vaidehi, F. H. Arnold, and W. A. Goddard III, “Protein Dynamics in a Family of Laboratory Evolved Thermophilic Enzymes,” *J. Mol. Biol.*, **327**, 745 (2003).
Wisz, M. S., and H. W. Hellinga, “An Empirical Model for Electrostatic Interactions in Proteins Incorporating Multiple Geometry-Dependent Dielectric constants,” *Proteins*, **51**, 360 (2003).
Xu, Z., and R. S. Farid, “Design, Synthesis, and Characterization of a Novel Hemoprotein,” *Protein Sci.*, **10**, 236 (2001).
Yokobayashi, Y., R. Weiss, and F. H. Arnold, “Directed Evolution of a Genetic Circuit,” *Proc. Natl. Acad. Sci. USA*, **99**, 16587 (2002).
Zha, D., A. Eipper, and M. T. Reetz, “Assembly of Designed Oligonucleotides as an Efficient Method for Gene Recombination: a New Tool in Directed Evolution,” *Chembiochem*, **4**, 34 (2003).
Zhao, H., L. Giver, Z. Shao, J. A. Affholter, and F. H. Arnold, “Molecular Evolution by Staggered Extension Process (StEP) in vitro Recombination,” *Nat. Biotechnol.*, **16**, 258 (1998).
Zou, J., and J. G. Saven, “Statistical Theory of Combinatorial Libraries of Folding Proteins: Energetic Discrimination of a Target Structure,” *J. Mol. Biol.*, **296**, 281 (2000).