

# Protein Design

## Zhilei Chen

*Center for Biophysics and Computational Biology, University of Illinois, Urbana, Illinois, U.S.A.*

## Huimin Zhao

*Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, Illinois, U.S.A.*

## INTRODUCTION

Protein design refers to the ability to alter protein structure to achieve the desired protein function. Of the two classes of proteins (binding proteins and catalytic proteins), catalytic proteins, or so-called enzymes, are of particular importance to chemical processing because they can be used as chemical catalysts. Enzymes are able to catalyze a broad range of chemical reactions with exquisite specificity and selectivity (stereo-, regio-, and chemo-). In addition, most enzymes are quite efficient catalysts and operate at mild conditions, resulting in less energy consumption. In the past several decades, enzymes have been increasingly employed to synthesize chemicals and materials in the pharmaceutical, chemical, food and agriculture industries. However, the number and diversity of the applications are modest compared to the total number of enzymes identified so far (~4000 enzymes). One main reason for this discrepancy is that naturally occurring enzymes are often not functionally optimal under process conditions in terms of their activity, stability, specificity, and selectivity. To overcome this limitation, tailor-made biocatalysts must be developed by protein design. This entry discusses the molecular tools of protein design and their applications in engineering naturally occurring enzymes into commercially viable biocatalysts for chemical processing. However, their applications in the development of therapeutic proteins and monoclonal antibodies are not discussed.

## PROTEIN DESIGNER'S TOOLBOX

Prior to the advent of recombinant DNA technology, the ability to design proteins was limited to chemical modification methods in which specific residues in a protein are modified at the protein level by chemical agents. Different strategies such as atom replacement and segment reassembly have been used to alter enzyme substrate specificity, activity, cofactor requirement, and stability.<sup>[1]</sup> These methods can introduce a diverse

range of functionality that does not occur in natural proteins. However, because only a few protein residues can be selectively modified chemically and the modifying process is rather tedious and time-consuming, these tools have not been widely used for the development of commercial enzymes.

The advent of recombinant DNA technology and polymerase chain reaction (PCR) technology has greatly changed the landscape of protein design. Numerous powerful protein design techniques have been developed in the past two decades, all of which target the modifications at the DNA level. Consequently, these protein design methods have been classified as genetic methods to distinguish them from the above chemical methods. Apart from this classification, all current protein design methods can be categorized into two general strategies: rational design and directed evolution. Rational design involves rational alterations of selected residues in a protein to cause predicted changes in function. It usually requires detailed knowledge of enzyme structure, function, and catalytic mechanism, which represents a bottom-up approach. In comparison, directed evolution, sometimes called irrational design, mimics the natural evolution process in the laboratory and involves repeated cycles of generating a library of different protein variants and selecting the variants with desired functions. Due to its combinatorial nature, it does not require any detailed structural and functional understanding of the target enzymes. Nonetheless, further characterization of the isolated variants may provide insights into protein structure and function. Thus, directed evolution represents a top-down approach. It should be noted that both rational design and directed evolution have been widely used in protein design.

## Rational Design

With the aid of site-directed mutagenesis and the availability of enzyme structures solved by x-ray crystallography, many attempts involving rationally designing proteins were made in the 1980s. Early successful

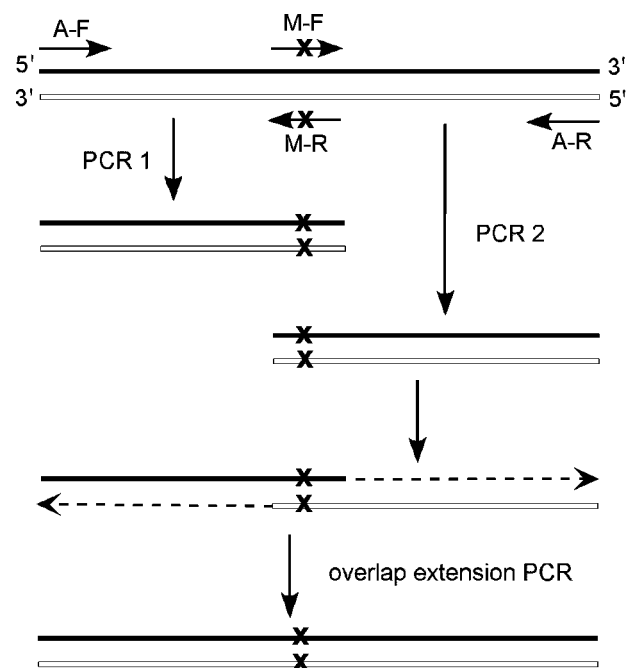
examples such as rational design of tyrosyl-transfer RNA synthetase and  $\beta$ -lactamase demonstrated the power of this approach in understanding enzyme catalysis and engineering enzyme activity and specificity.<sup>[2,3]</sup> The numerous studies that followed these early examples have led to the formation of a new field—protein engineering.<sup>[4]</sup> Notable examples include the design of a T4 lysozyme with increased stability by engineering disulfide bonds into the protein, the redesign of a lactate dehydrogenase into a malate dehydrogenase by introducing a single point mutation, and the design of subtilisin enzymes with increased activity and altered pH profiles by engineering surface charge.<sup>[5–7]</sup> However, owing to the sieving effect of publication (only the successful cases are published), rational design may look more “rational” than it actually is. The truth is that our current ability of rationally designing protein is still rather limited. It became apparent that, owing to the intricate and complex relationship between protein structure and function, rationally introduced mutations often have unexpected disastrous effects on enzyme stability and activity. Thus, to circumvent the unpredictable effects of individual residues, methods that are able to assess mutations at single sites or multiple sites in a large-scale format were developed in the early 1990s. This new trend ultimately led to the establishment of directed evolution as a powerful approach for protein design. The past 10 yr have witnessed many more publications on directed evolution than those on rational design. Nonetheless, with recent advances in genomics, bioinformatics and structural proteomics, our knowledge about proteins is rapidly expanding. Computational techniques are also playing an increasingly important role in protein rational design, especially with the availability of faster and cheaper computers. Thus, it is foreseeable that the ability to rationally design proteins will be significantly improved.

### Site-directed mutagenesis

Site-directed mutagenesis is the most powerful and widely used rational design approach, which entails the precise modification of specific residue(s) in a given protein at the DNA level. These residue(s) are typically identified from the three-dimensional protein structure obtained by x-ray crystallography or NMR methods. In the absence of structure, a structural model of the target protein can be built from the three-dimensional structure of a related protein sharing high sequence homology with the target protein using homology-modeling programs such as Insight II (Accelrys Inc., San Diego, CA) and SYBYL<sup>®</sup>/Base (Tripos, Inc., St. Louis, MO). If these options are not available, sequence analysis programs such as BLAST and

CLUSTALW (<http://workbench.sdsc.edu/>) can be used to identify the residues that are conserved among a family of homologous proteins, and are therefore presumed to be functionally important.

A large number of experimental methods have been devised for site-directed mutagenesis. The methods developed in the early 1980s all involve the use of single-stranded bacteriophage DNA molecules carrying a target gene and chemically synthesized complementary oligonucleotides containing the desired nucleotide substitutions. After DNA annealing and synthesis, the newly synthesized DNA strand contains the target gene with the desired mutations. Although these methods are conceptually simple, the generation of single-stranded DNA is time-consuming and the mutagenesis efficiency (the frequency of the target genes with mutations) is relatively low (less than 50%). Thus, a few PCR-based mutagenesis methods have been developed to overcome these limitations. One of the most widely used methods is the overlap extension PCR mutagenesis method.<sup>[8]</sup> As illustrated in Fig. 1, four primers, A-F, A-R, M-F and M-R, are used. A-F and A-R correspond to the beginning and the end of the target gene sequence, respectively. M-F and M-R cover the region where a point mutation is desired. First, two independent rounds of PCRs are performed using the target gene as a template.



**Fig. 1** Site-directed mutagenesis by splicing overlap extension (SOEing). Both the target gene and the PCR products are shown in double strands. Primers are shown as arrows and mutations in primers and products as  $\times$ . Dashed arrows indicate the directions of DNA extension by DNA polymerases.

Primers A-F and M-R are used as a pair to amplify the 5' end of the target gene, and this amplified product will contain a point mutation near the 3' end. Primers M-F and A-R are used together to amplify the 3' end of the target gene, leaving a point mutation near the 5' end of the product. Polymerase chain reaction products from both the reactions are purified and combined for another round of PCR without additional primers. This reaction is called overlap extension because the 3' end of the first PCR product is complementary to the 5' end of the second PCR product, resulting in these two DNA ends priming each other. The resulting target gene with a point mutation will be further amplified by primers A-F and A-R and subcloned into a vector for protein expression.

### Domain swapping

Novel protein functions are needed for many applications. Although site-directed mutagenesis is effective in altering protein functions, it often results in incremental improvement of protein functions rather than dramatically improved or novel protein functions because it can only modify protein sequences at single or multiple sites. It is expected that the creation of novel protein functions may require dramatic changes in protein structures. To address this limitation, one simple genetic method to introduce novel protein functions is to combine protein domains from different proteins, or so-called domain swapping. A protein domain is typically defined as a folding and/or functional unit of a protein. Because many large proteins contain several functional domains and domains tend to fold independently, the fusion of domains with different functions may result in a multifunctional protein with unique features.

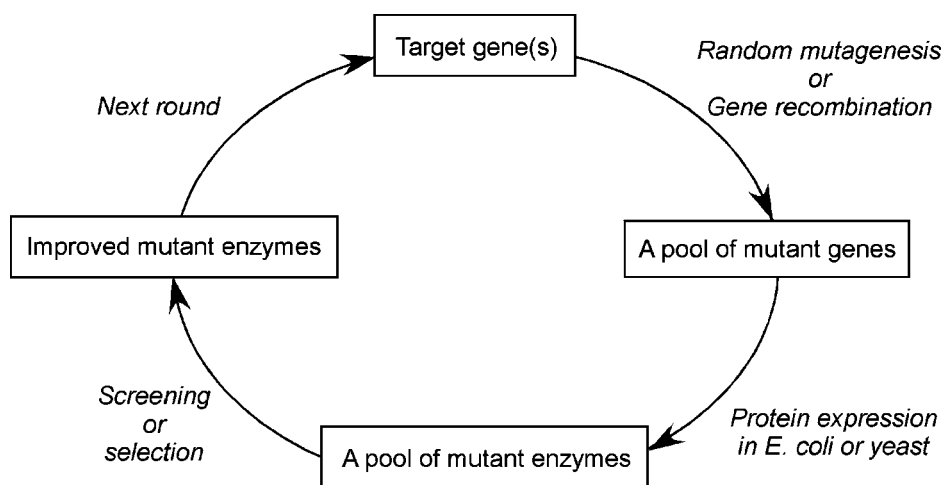
One recent example of domain swapping is the design of a ligand-regulated DNA recombinase in which the ligand binding domain of human estrogen receptor was linked to the Cre recombinase.<sup>[9]</sup> The Cre recombinase can integrate a foreign gene into the *LoxP* site on the host genome while the human estrogen receptor is a nuclear hormone receptor that regulates the action of estrogen in different tissues and organs. The human estrogen receptor contains at least three distinct domains: a DNA-binding domain, a transactivation domain, and a ligand binding domain. The binding of estrogenic compounds to the ligand binding domain of the estrogen receptor will cause conformational changes in the estrogen receptor. The Cre recombinase and the estrogen receptor are two completely different proteins, yet the fusion protein created by domain swapping exhibits a protein function different from either parental protein: the recombinase activity is dependent on the ligand binding of the estrogen receptor. Such a fusion protein was

used to reduce the toxicity to the proliferating *Drosophila* cells caused by the chronic expression of the Cre recombinase. It should be noted that for a domain swapping experiment to be successful, it is very important to define the exact domain boundaries within a protein. Two main methods have been developed: one is through sequence alignment of a pool of homologous proteins, and the other is through deletion experiments. These two methods usually give a similar but not identical definition of domain boundary. In many cases, because of the lack of complete understanding of protein functions, these slightly different definitions may lead to different results.

### Directed Evolution

Although rational design has been demonstrated to be an effective protein design strategy, its success rate is not very high because of our limited understanding of protein folding, structure, function, and dynamics. Moreover, it is rather time-consuming owing to the need of a high-quality three-dimensional protein structure for experimental guidance. In contrast, directed evolution does not use any preconceived ideas about what is important and only relies on a very simple algorithm that nature has successfully been using for eons: diversification coupled with selection. In essence, directed evolution mimics natural Darwinian evolution in a laboratory environment. As shown in Fig. 2, genetic diversity is first introduced into a target gene through random mutagenesis and/or recombination. The library of mutant genes is then transformed into host cells in which the mutant genes are converted into their corresponding proteins. Functionally improved mutant proteins are identified through an appropriate selection or screening strategy. The same process will be repeated until the goal is achieved or no further improvement is possible. It should be noted that the host cells used for protein expression in a directed evolution experiment are laboratory microorganisms such as *Escherichia coli* and *Saccharomyces cerevisiae*. Expression of foreign proteins in these host organisms is enabled by recombinant DNA technology. These microorganisms are favored because of their rapid growth rates, the availability of many genetic engineering tools, and their well-known genetics.

It should also be noted that both screening methods and selection methods have their own advantages and disadvantages. Screening involves physically or chemically interrogating every mutant protein in a library individually, and is often implemented in a 96-well plate format using plate readers. As a result, its throughput is relatively low (the size of the library that can be screened is limited to  $\sim 10^4$ ). However, because the screens are done in vitro with whole cells, cell

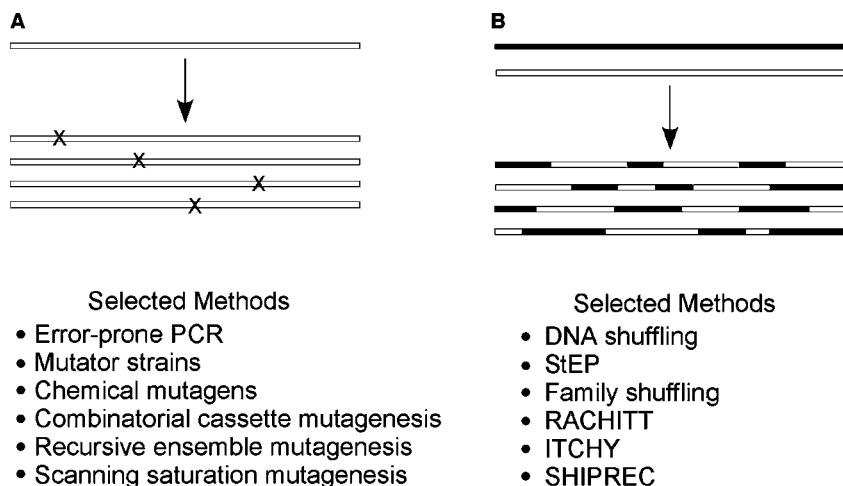


**Fig. 2** The general scheme of directed evolution.

lysates, or partially purified enzymes, and often in much the same way the enzymes are traditionally assayed, the experimental conditions can be easily tailored to meet the reaction constraints such as non-natural environment or substrates, and the screens can be implemented very quickly. Moreover, multiple measurements can be made on each sample to check several key enzyme properties. Thus, screening is the most flexible sorting method in directed evolution. In comparison, selection involves linking the survival or growth of the microorganism to the target protein so that only the microorganisms possessing the desired enzyme function can grow. Consequently, a much larger library of enzyme mutants ( $>10^6$ ) can be assessed by selection, and the size of the library is limited only by the cell transformation efficiency. Unfortunately, it is not an easy task to devise a selection method for a given protein in most cases because the desired enzyme function is often nonnatural and cannot be coupled to the growth and survival of the host organism. Even when a selection is available, because of the redundancy and complexity of genetic regulatory

networks, host organisms can often create solutions that are not related to the targeted enzyme function. Thus, extra care must be taken to ensure that the positives are indeed the result of the mutations in the targeted enzyme. It is often advised to combine selection and screening if both methods are available.

A successful directed evolution experiment involves two key components: creating genetic diversity and developing a screening or selection method. In the past several years, many experimental methods have been developed to introduce genetic diversity into the target gene, all of which can be grouped into two categories: methods of random mutagenesis and methods of gene recombination. As shown in Fig. 3, random mutagenesis starts from a single parent gene and introduces new nucleotide substitutions randomly in the progeny genes, or inserts or deletes one or more nucleotides at random positions in the progeny genes. In contrast, gene recombination usually starts from a pool of mutants from a single gene or a pool of closely related parental genes of different origin and creates blockwise exchange of sequence information among the parent



**Fig. 3** The comparison between random mutagenesis methods and gene recombination methods. Random mutagenesis methods create a library of variants containing point mutations or insertions/deletions (represented by  $\times$ ) from a single parental gene, whereas gene recombination methods create a library of chimeric variants via blockwise exchange of sequence information among the parental genes. A few representative methods that have been developed so far are listed.

genes. Recombination in a genetic sense means the breaking and rejoining of DNA fragments in new combinations. Both random mutagenesis and gene recombination are important natural evolutionary processes. A few of the most commonly used techniques for generation of diversity will be discussed below. For additional and more detailed discussions on various evolutionary methods, interested readers are referred to a recent review contributed by Zhao and Zha.<sup>[10]</sup>

### Error-prone PCR

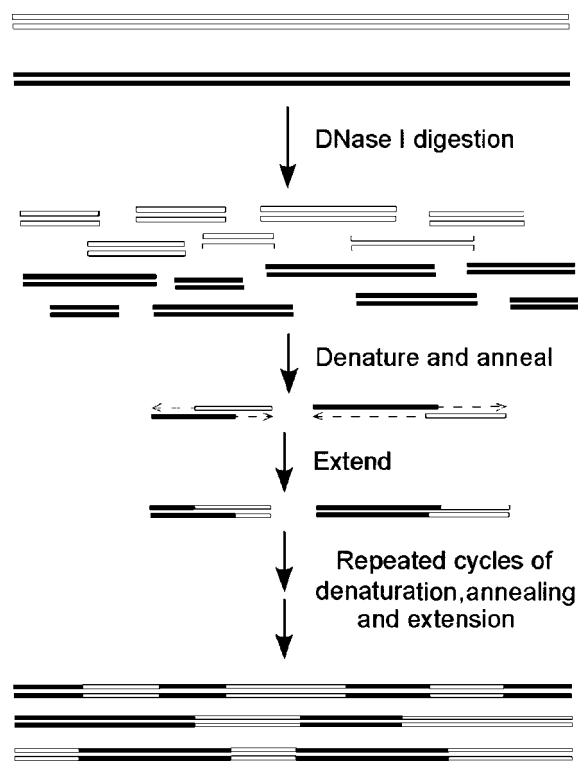
Because of its simplicity and efficiency, error-prone PCR is the most widely used random mutagenesis method. It is essentially a variation of the standard PCR with slightly modified reaction conditions. The reaction buffer used for standard PCR contains an equal molar amount of each dNTP and a certain concentration of  $Mg^{2+}$  (typically 1.5 mM).  $Mg^{2+}$  is very important for the activity of *Taq* DNA polymerase because it is directly involved in catalysis. During the DNA amplification process in a standard PCR, the chance of incorporating wrong nucleotides into the progeny genes by the *Taq* DNA polymerase is very low. However, this very small error rate can be dramatically increased by modifying the reaction buffer so as to force the *Taq* DNA polymerase to incorporate more incorrect nucleotides during amplification. These changes include: 1) use of unbalanced concentrations of dNTPs; 2) addition of  $MnCl_2$  in addition to  $MgCl_2$ ; and 3) use of a high concentration of  $Mg^{2+}$  (typically 7 mM). Owing to the limitations on library sorting imposed by screening or selection and the observation that most of the mutations are deleterious and beneficial mutations are rare, the mutation rate must be tuned to the power of the sorting method. Thus, only one or two amino acid substitutions are usually introduced to the target protein to increase the possibility of finding mutant proteins with the desired function. Fortunately, it was found that the error rate of error-prone PCR can be easily and precisely controlled by the  $Mn^{2+}$  concentration, which makes this method even more appealing.<sup>[11]</sup> The main disadvantage of error-prone PCR is that it can only access about six amino acid substitutions at a given residue position because of the degeneracy of the genetic code, thus reducing the potential diversity significantly. Moreover, the mutations are not truly random. For example, a common bias of error-prone PCR is the high occurrence of AG substitutions.

### DNA shuffling and family shuffling

The method of DNA shuffling, also known as “sexual PCR,” is the first and most widely used

gene recombination method, developed in 1994 by Stemmer.<sup>[12]</sup> As shown in Fig. 4, a pool of selected closely related genes containing point mutations are randomly digested with enzyme DNase I to obtain small double-stranded DNA fragments (20–50 bp). These DNA fragments are purified and reassembled into a full-length gene in a PCR-like reaction (without primers). Recombinogenic events occur when fragments derived from different parental genes prime one another. This reassembly mixture is then used as a template for a standard PCR reaction with primers flanking the gene of interest. The final amplified product will consist of a library of full-length genes containing recombined mutations from different parental genes.

It is noteworthy that these closely related genes are often mutants derived from a single parental gene. However, naturally occurring homologous genes sharing relatively high sequence identity (>70%) can also be recombined using DNA shuffling under modified reaction conditions, which is called “family shuffling.”<sup>[13]</sup> It has been demonstrated that family shuffling can significantly accelerate the rate of functional enzyme improvement in comparison with



**Fig. 4** DNA shuffling. For simplicity, only two homologous genes are shown. These two double-stranded genes are mixed in equal molar ratio followed by random fragmentation with DNase I. These short fragments are reassembled into full-length chimerical genes in a PCR-like process. The full-length genes may be amplified by a standard PCR and subcloned to an appropriate vector.

DNA shuffling or error-prone PCR. The power of family shuffling may arise from its ability to sparsely sample a larger portion of sequence space that is functionally rich because the parental genes have been preselected in nature to be functional and useful. Generally speaking, compared to error-prone PCR, the main advantages of DNA shuffling or family shuffling include its ability to rapidly accumulate beneficial mutations and remove deleterious mutations, and its ability to explore a larger portion of protein sequence space.

### Nonhomologous DNA recombination

DNA shuffling or family shuffling relies on relatively high levels of sequence identity (more than 70%) to recombine genes in vitro. DNA shuffling also tends to generate crossovers only in regions of the highest sequence identity. In general, if the sequence identity is less than 70%, most of the shuffled progeny genes will be the same as parental genes. Given that many proteins having similar three-dimensional structures share low or no discernible sequence homology, homology-dependent methods for recombining genes may potentially limit the solutions to protein design problems. In fact, many lines of evidence from rational design and computational studies have indicated that functional proteins can be obtained by recombining genes with low sequence homology.<sup>[12]</sup>

Recently, several methods have been developed to recombine nonhomologous genes. One of them is the so-called sequence homology-independent protein recombination (SHIPREC) method.<sup>[14]</sup> As illustrated in Fig. 5, two parental nonhomologous genes of similar size are fused together through a DNA linker containing multiple restriction sites. This dimer gene is digested with DNase I to produce a pool of random-length fragments. Fragments of a length corresponding to either of the parental genes are isolated and treated with nuclease S1 to form blunt ends. These blunt-ended fragments will then be circularized by blunt-end ligation, followed by linearization with restriction digestion in the linker region to create a library of chimerical genes. The chimerical genes are subsequently cloned into an appropriate expression vector and transformed into a suitable host for further screening or selection. This method was used to recombine a membrane-associated human cytochrome P450 (1A2) and the heme domain of a soluble bacterial P450 (BM3), which resulted in variants of 1A2 enzymes that are properly folded and more soluble in the bacterial cytoplasm than the wild-type 1A2 enzymes.<sup>[14]</sup> One limitation of this method is that only two parental genes are shuffled and there is only one crossover in every chimerical gene.

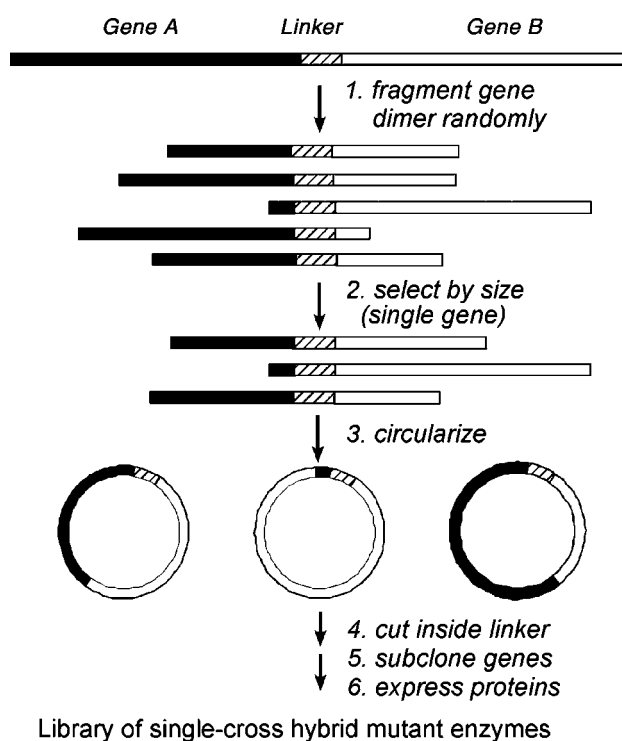


Fig. 5 Schematic representation of the SHIPREC method.

### Hybrid Approaches Combining Rational Design and Directed Evolution

Although directed evolution is a powerful tool for protein design, it is limited by the number of protein variants that can be screened experimentally. As shown in Table 1, for a protein of typical size (300 amino acids), the library size of protein variants with each variant containing three mutations is over  $3 \times 10^{10}$ , which is too large to be examined experimentally. However, it is possible that even more simultaneous mutations are needed in a protein to achieve novel protein activity or drastic improvement of protein functions. On the other hand, although rational design has been used for more than two decades, our

**Table 1** The potential mutant library size of a protein with 300 amino acids

Mutations	Potential library size <sup>a</sup>
1	5700
2	16, 190, 850
3	30, 557, 530, 900
4	43, 109, 036, 717, 175
5	48, 489, 044, 499, 478, 440

<sup>a</sup>The number of enzyme mutants can be calculated by the simple algorithm  $N = 19^M \times 300! / [(300 - M)!M!]$  where  $M$  is the number of simultaneous mutations in a protein.

current understanding of protein structure and function still cannot guarantee the success of rational design approaches. Nonetheless, rational design may enable large jumps in the fitness landscape and create protein variants with novel but poor functions, which can then be optimized by directed evolution (Fig. 6). In particular, owing to their high speed, computational techniques may be used to perform *in silico* screening of protein variants with a library size of  $\sim 10^{80}$  or to search vast regions of sequence space to identify the protein sites for more efficient directed evolution.<sup>[15,16]</sup> Thus, a hybrid approach combining the best of rational design and directed evolution may represent the most powerful approach for protein design.

One such hybrid approach has been recently used to create TEM-1  $\beta$ -lactamase mutants with increased resistance to the antibiotic cefotaxime.<sup>[15]</sup> This approach involves a computational protein optimization algorithm called protein design automation (PDA) to reduce the sequence space by many orders of magnitude followed by experimental screening of these selected sequences. The algorithm starts with the structure of a target protein, selects all or a specified set of residues responsible for a specific protein function and predicts the optimal sequences consistent with the proper fold. In this study, 19 residues near the active site of TEM-1  $\beta$ -lactamase were selected for optimization, which corresponds to a theoretical library of  $20^{19}$  ( $\sim 7 \times 10^{23}$ ) sequences. After computational prescreen by PDA, this library was reduced to 200,000 sequences, which were then constructed and experimentally screened. In a single round, several variants showing a 1280-fold increase in cefotaxime resistance were obtained. These variants contained multiple mutations that have not been discovered before.

## EXAMPLES

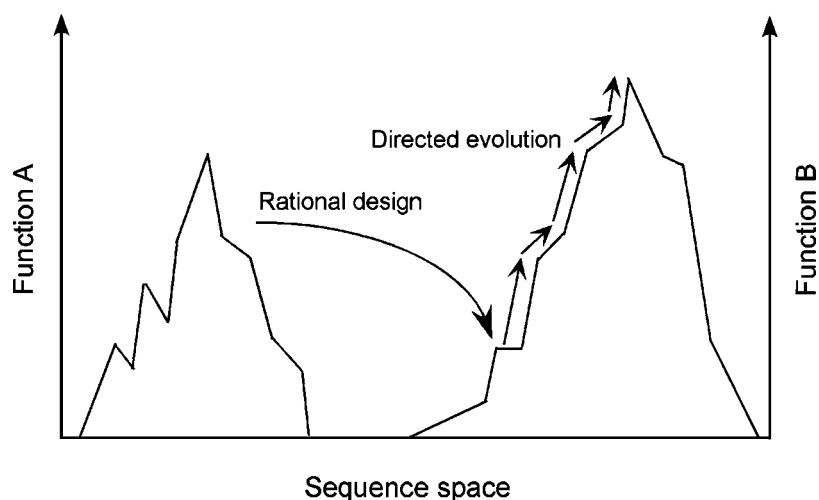
The applications of enzymes as biocatalysts for industrial chemical processes are rapidly expanding. Like chemical catalysts, the ideal biocatalysts are those enzymes with high stability, activity, specificity, and/or enantioselectivity under process conditions. Unfortunately, naturally occurring enzymes are often deficient in one or more of these important aspects, and therefore cannot meet the stringent requirements of process conditions. In addition, because not all the chemical reactions can be catalyzed by enzymes, owing to either the use of unnatural substrates or the limitations in the chemistry of catalysis, the creation of novel enzyme functions is highly desired. Protein design has been used to address these limitations with great success. Some of these examples will be highlighted below.

## Increasing Enzyme Activity and Stability

Both enzyme stability and activity are major concerns in all biocatalytic processes. Poor enzyme stability requires frequent replacement of catalysts while poor enzyme activity requires the use of large amounts of catalysts, both of them resulting in low productivity and unfavorable process economics. Moreover, a high reaction temperature is usually desired in a chemical process because it can result in high reaction rates, favorable equilibrium, and increased substrate solubility, but it also requires high enzyme stability. As naturally occurring enzymes have been evolved for specific biological functions under biological conditions (i.e., aqueous solutions, neutral pH and mild temperatures for most of the organisms), their intrinsic activity and stability are often not high enough for practical applications that may require enzymes to function at extreme pHs, high temperatures, or in organic cosolvents. Thus, further improvement of enzyme activity and stability may be required.

Many successful examples using rational design to engineer enzyme stability or activity have been reported. One of them is the above-mentioned redesign of a T4 lysozyme with increased stability by engineering disulfide bonds into the protein.<sup>[5]</sup> Another notable example is the design of a subtilisin BPN' mutant containing six site-directed mutations, which was 100-fold more stable than the wild-type enzyme in an aqueous environment and 50-fold more stable than the wild-type in anhydrous dimethylformamide.<sup>[17]</sup> Subtilisin enzymes have been used as additives in laundry detergents for stain hydrolysis and solubilization, and catalysts for the synthesis of peptides and enantioselective transformation of chiral alcohols, acids, and amines.

Directed evolution is also very effective in engineering enzyme stability and activity. Unlike rational design, which tends to improve one enzyme property at a time (in fact, attempts to rationally alter one enzyme property often disrupt other existing important characteristics), directed evolution may improve multiple enzyme properties simultaneously. For example, five rounds of directed evolution consisting of alternate cycles of error-prone PCR and *in vitro* gene recombination coupled with screening led to the isolation of a highly stable and active subtilisin E mutant.<sup>[11]</sup> This mutant contained eight thermo-stabilizing mutations, which were located all over the protein structure. It showed a >200-fold longer thermal inactivation half-life at 65°C, an 18°C higher temperature optima, and a >5-fold higher activity than the wild-type enzyme. Another impressive example is the simultaneous improvement of four distinct enzyme properties of subtilisin, including thermostability, activity in organic solvents, activity at pH 10, and activity at pH 5.5 by



**Fig. 6** Schematic representation of hybrid approaches combining rational design and directed evolution. Rational design may allow a large jump in sequence space from high fitness for function A to a low fitness for novel function B. The low fitness for function B might be readily optimized by directed evolution.

directed evolution.<sup>[18]</sup> Family shuffling was used to recombine 26 homologous subtilisin genes to create a library of chimerical subtilisin genes. Out of 654 active subtilisins, a few mutants showed significant improvement over any of the parental enzymes for each individual enzyme property.

### Changing Substrate Specificity

Natural enzymes are optimized for their intended substrates, and often cannot accept nonnatural substrates that are required for a desired chemical process. In the past two decades, both rational design and directed evolution have been successfully used to alter substrate specificity of a large number of common classes of enzymes such as oxidoreductases, hydrolases, and transferases. One exemplary enzyme of particular interest is aspartate aminotransferase which catalyzes the interconversion of aspartate with its corresponding  $\alpha$ -keto acid using pyridoxal phosphate as a cofactor. *E. coli* aspartate aminotransferase shares 43% sequence identity with *E. coli* tyrosine aminotransferase. Onuffer and Kirsch<sup>[19]</sup> used homology modeling to build a structural model of *E. coli* tyrosine aminotransferase based on the solved crystal structure of *E. coli* aspartate aminotransferase.<sup>[19]</sup> Structural comparison of the active sites of these two enzymes revealed six different positions between them. Mutagenesis of all six residues in aspartate aminotransferase by site-directed mutagenesis of those found in tyrosine aminotransferase successfully altered the substrate specificity of aspartate aminotransferase. The reactivity of the aspartate aminotransferase with phenylalanine was increased by over three orders of magnitude without sacrificing the high transamination activity with aspartate observed for both enzymes. In another case, directed evolution was used to convert *E. coli* aspartate

aminotransferase into a valine aminotransferase.<sup>[20]</sup> A mutant enzyme with 17 amino acid substitutions was generated that shows a  $2.1 \times 10^6$ -fold increase in the catalytic efficiency for a nonnative substrate, valine. The crystal structure of the mutant enzyme indicated a remodeled active site and altered subunit interface caused by the cumulative effects of mutations. Most amazingly, only one of the mutations directly contacts the substrate, which underscores our limited understanding of enzyme substrate specificity. These mutations would be difficult, if not impossible, to be identified and introduced to the mutant enzyme by a rational design approach.

### Improving Enantioselectivity

The production of enantiomerically pure compounds is of great importance to the chemical and pharmaceutical industries. Enzymes are chiral catalysts by nature and they have incredible potential for creating enantiomerically pure products. However, many existing natural enzymes show low degrees of enantioselectivity, which requires further improvement by protein design.

Because the molecular basis of enantioselectivity is poorly understood, directed evolution seems to be an excellent choice for engineering enantioselective biocatalysts. Several impressive examples have been documented. In a classical study, Reetz and coworkers<sup>[21]</sup> used error-prone PCR coupled with a 96-well plate based colorimetric screening method to increase the enantioselectivity of a *Pseudomonas aeruginosa* lipase toward 2-methyldecanoate.<sup>[21]</sup> After several rounds of directed evolution, the enantioselectivity of the lipase increased from  $E = 1.04$  (2% enantiomeric excess) to  $E = 25$  (90–93% enantiomeric excess, ee) ( $E$  is the enantioselectivity factor). Using a similar approach,

Arnold and coworkers even inverted the enantioselectivity of hydantoinase from D-selectivity (40% ee) to moderate L-preference (20% ee at 30% conversion).<sup>[22]</sup> This evolved mutant is now being evaluated in an industrial chemical process at Degussa.

With the increasing knowledge of the molecular basis of enzyme enantioselectivity, rational design has also achieved some success. In one case, Rotticci et al. used molecular modeling to study the different binding modes for alcohol enantiomers in the active site of *Candida antarctica* lipase B and proposed a model for its enantioselectivity.<sup>[23]</sup> Site-directed mutagenesis was used to alter the active site residues causing unfavorable interactions between the substrate and the enzyme. A single mutation, Ser47Ala, resulted in improvement of the lipase-catalyzed resolution of 1-chloro-2-octanol from  $E = 14$  to  $E = 28$ . In another case, van Den Heuvel, Fraaije, and Van Berkel<sup>[24]</sup> studied the crystal structure of vanillyl-alcohol oxidase and identified a few important residues within the active site that might contribute to the (*R*)-selective formation of (*R*)-1-(4'-hydroxyphenyl) ethanol from 4-ethylphenol. A double mutant was constructed by site-directed mutagenesis, which shows inverted enantioselectivity [(*S*)-selective with 80% ee].

### Creating de novo Catalytic Activity

Although natural enzymes can catalyze numerous chemical transformations such as oxidation, reduction, hydrolytic reactions, and carbon-carbon bond formation reactions, enzymes with novel catalytic activities are still needed for the application of enzymes in many industrial biocatalytic processes. The ultimate goal of protein design is to design de novo catalytic activity such that a biocatalyst can be readily obtained for any given chemical transformation. While most protein design so far has really been protein redesign in which the existing protein functions have been adapted under different regimes, a few successful attempts have been made toward this ultimate goal.

An impressive example is the creation of novel enzyme substrate specificity and activity by the DNA shuffling of two highly homologous triazine hydrolases, AtzA and TriA.<sup>[25]</sup> These two enzymes catalyze the dechlorination and deamination reaction of atrazine and aminoatrazine, respectively. Although they share limited overlap in substrate preference, they only differ by 9 out of 475 amino acids. After one round of DNA shuffling, several variants were found to hydrolyze substrates that were not substrates for either of the parental enzymes.

Another impressive example is the directed evolution of novel ampicillin-resistant activity from a functionally unrelated DNA fragment.<sup>[26]</sup> A DNA

fragment from the genomic DNA library of *Pyrococcus furiosus* was shown to confer very low ampicillin resistance activity ( $\beta$ -lactamase activity) on *E. coli*, while *P. furiosus* itself does not have any  $\beta$ -lactamase activity. This ampicillin resistance activity was significantly enhanced after 50 rounds of DNA shuffling and screening at increasing ampicillin concentrations. The evolved DNA fragments also confer resistance to other drugs that inhibit bacterial cell-wall synthesis.

By taking advantage of the ever-increasing computing power, various computational techniques have been attempted to create de novo protein activity. A particularly impressive example is the creation of catalytic activity in a binding (catalytically inert) protein.<sup>[27]</sup> Mainly owing to its favorable protein expression properties and thermodynamic stability, rather than any structural similarity to a natural enzyme, *E. coli* thioredoxin was used as a protein scaffold to create enzyme activity concerning histidine-mediated nucleophilic hydrolysis of *p*-nitrophenyl acetate. The design strategy is based on the physical and chemical principles governing protein stability and catalytic mechanism. A protein design software ORBIT was used to perform an active site scan and identified two promising catalytic positions and the surrounding active-site mutations required for substrate binding. Two candidate mutants were constructed by site-directed mutagenesis and both of them showed catalytic activity significantly above the background. Although they are not particularly impressive catalysts, such mutants should be adequate starting points for directed evolution.

### CONCLUSIONS

Protein design or engineering is a rapidly growing field of academic research and industrial practice. Its goals include not only addressing fundamental relationships among protein folding, structure, function, and dynamics, but also designing proteins with desired features for applications in pharmaceutical, chemical, agricultural, and food industries. Of particular interest to chemical processing is the application of protein engineering tools to the development of enzyme biocatalysts. Two distinct and yet complementary protein design approaches, rational design and directed evolution, have been successfully developed to engineer biocatalysts with altered or novel stability, activity, substrate specificity, selectivity, cofactor specificity, reaction chemistry, and pH optima. Recently, a third approach combining the best of rational design and directed evolution has also shown great promise in protein engineering, which will attract increasing attention in the near future. With recent advances in

structural genomics and proteomics, and the development of miniaturized and automated high throughput screening technologies, protein engineers will be equipped with more powerful tools to tackle the ever-challenging protein design problems, which will certainly accelerate the widespread adoption of biocatalysts in chemical processing.

## ACKNOWLEDGMENT

We thank the National Science Foundation (grant BES-0348107 to HZ) for supporting our work on protein design of human estrogen receptors.

## ARTICLE OF FURTHER INTEREST

*Biocatalysis*, p. 101.

## REFERENCES

1. Qi, D.; Tann, C.M.; Haring, D.; Distefano, M.D. Generation of new enzymes via covalent modification of existing proteins. *Chem. Rev.* **2001**, *101* (10), 3081–3111.
2. Winter, G.; Fersht, A.R.; Wilkinson, A.J.; Zoller, M.; Smith, M. Redesigning enzyme structure by site-directed mutagenesis—tyrosyl transfer—RHA synthetase and ATP binding. *Nature* **1982**, *299* (5885), 756–758.
3. Sigal, I.S.; Harwood, B.G.; Arentzen, R. Thiol-beta-lactamase—replacement of the active-site serine of RTEM beta-lactamase by a cysteine residue. *Proc. Natl. Acad. Sci. USA.* **1982**, *79* (23), 7157–7160.
4. Brannigan, J.A.; Wilkinson, A.J. Protein engineering 20 years on. *Nat. Rev. Mol. Cell. Biol.* **2002**, *3* (12), 964–970.
5. Perry, L.J.; Wetzel, R. Disulfide bond engineered into T4 lysozyme: stabilization of the protein toward thermal inactivation. *Science* **1984**, *226* (4674), 555–557.
6. Wilks, H.M.; Hart, K.W.; Feeney, R.; Dunn, C.R.; Muirhead, H.; Chia, W.N.; Barstow, D.A.; Atkinson, T.; Clarke, A.R.; Holbrook, J.J. A specific, highly active malate dehydrogenase by redesign of a lactate dehydrogenase framework. *Science* **1988**, *242* (4885), 1541–1544.
7. Russell, A.J.; Fersht, A.R. Rational modification of enzyme catalysis by engineering surface charge. *Nature* **1987**, *328* (6130), 496–500.
8. Horton, R.M.; Cai, Z.L.; Ho, S.N.; Pease, L.R. Gene splicing by overlap extension: tailor-made genes using the polymerase chain reaction. *Biotechniques* **1990**, *8* (5), 528–535.
9. Heidmann, D.; Lehner, C.F. Reduction of Cre recombinase toxicity in proliferating Drosophila cells by estrogen-dependent activity regulation. *Dev. Genes. Evol.* **2001**, *211* (8–9), 458–465.
10. Zhao, H.; Zha, W. *Enzyme Functionality: Design, Engineering and Screening*; Svendsen, A., Ed.; Marcel Dekker, Inc.: New York, 2003; 353–373.
11. Zhao, H.; Moore, J.C.; Volkov, A.A.; Arnold, F.H. In *Manual of Industrial Microbiology and Biotechnology*, 2nd Ed.; Demain, A.L., Davies, J.E., Eds.; ASM Press: Washington, DC, 1999; 597–604.
12. Stemmer, W.P. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **1994**, *370* (6488), 389–391.
13. Cramer, A.; Raillard, S.A.; Bermudez, E.; Stemmer, W.P. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **1998**, *391* (6664), 288–291.
14. Sieber, V.; Martinez, C.A.; Arnold, F.H. Libraries of hybrid proteins from distantly related sequences. *Nat. Biotechnol.* **2001**, *19* (5), 456–460.
15. Hayes, R.J.; Bentzien, J.; Ary, M.L.; Hwang, M.Y.; Jacinto, J.M.; Vielmetter, J.; Kundu, A.; Dahiyat, B.I. Combining computational and experimental screening for rapid optimization of protein properties. *Proc. Natl. Acad. Sci. USA.* **2002**, *99* (25), 15926–15931.
16. Voigt, C.A.; Mayo, S.L.; Arnold, F.H.; Wang, Z.G. Computational method to reduce the search space for directed protein evolution. *Proc. Natl. Acad. Sci. USA.* **2001**, *98* (7), 3778–3783.
17. Wong, C.-H.; Chen, S.-T.; Hennen, W.J.; Bibbs, J.A.; Wang, Y.-F.; Liu, J.L.C.; Pantoliano, N.W.; Whitlow, M.; Bryan, P.N. Enzymes in organic synthesis—use of subtilisin and a highly stable mutant derived from multiple site-specific mutations. *J. Am. Chem. Soc.* **1990**, *112* (3), 945–953.
18. Ness, J.E.; Welch, M.; Giver, L.; Bueno, M.; Cherry, J.R.; Borchert, T.V.; Stemmer, W.P.; Minshull, J. DNA shuffling of subgenomic sequences of subtilisin. *Nat. Biotechnol.* **1999**, *17* (9), 893–896.
19. Onuffer, J.J.; Kirsch, J.F. Redesign of the substrate specificity of Escherichia coli aspartate aminotransferase to that of Escherichia coli tyrosine aminotransferase by homology modeling and site-directed mutagenesis. *Protein Sci.* **1995**, *4* (9), 1750–1757.
20. Oue, S.; Okamoto, A.; Yano, T.; Kagamiyama, H. Redesigning the substrate specificity of an enzyme by cumulative effects of the mutations

- of non-active site residues. *J. Biol. Chem.* **1999**, *274* (4), 2344–2349.
21. Liebeton, K.; Zonta, A.; Schimossek, K.; Nardini, M.; Lang, D.; Dijkstra, B.W.; Reetz, M.T.; Jaeger, K.E. Directed evolution of an enantioselective lipase. *Chem. Biol.* **2000**, *7* (9), 709–718.
  22. May, O.; Nguyen, P.T.; Arnold, F.H. Inverting enantioselectivity by directed evolution of hydantoinase for improved production of L-methionine. *Nat. Biotechnol.* **2000**, *18* (3), 317–320.
  23. Rotticci, D.; Rotticci-Mulder, J.C.; Denman, S.; Norin, T.; Hult, K. Improved enantioselectivity of a lipase by rational protein engineering. *Chem-biochem.* **2001**, *2* (10), 766–770.
  24. Van Den Heuvel, R.H.; Fraaije, M.W.; Van Berkel, W.J. Direction of the reactivity of Vanillyl-alcohol oxidase with 4-alkylphenols. *FEBS Lett* **2000**, *481* (2), 109–112.
  25. Raillard, S.; Krebber, A.; Chen, Y.; Ness, J.E.; Bermudez, E.; Trinidad, R.; Fullem, R.; Davis, C.; Welch, M.; Seffernick, J.; Wackett, L.P.; Stemmer, W.P.; Minshull, J. Novel enzyme activities and functional plasticity revealed by recombining highly homologous enzymes. *Chem. Biol.* **2001**, *8* (9), 891–898.
  26. Yano, T.; Kagamiyama, H. Directed evolution of ampicillin-resistant activity from a functionally unrelated DNA fragment: a laboratory model of molecular evolution. *Proc. Natl. Acad. Sci. USA.* **2001**, *98* (3), 903–907.
  27. Bolon, D.N.; Mayo, S.L. Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci. USA.* **2001**, *98* (25), 14274–14279.

