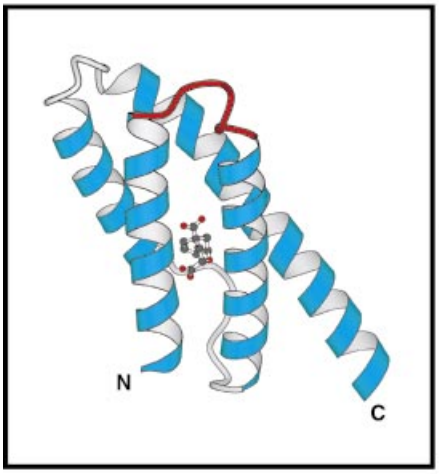
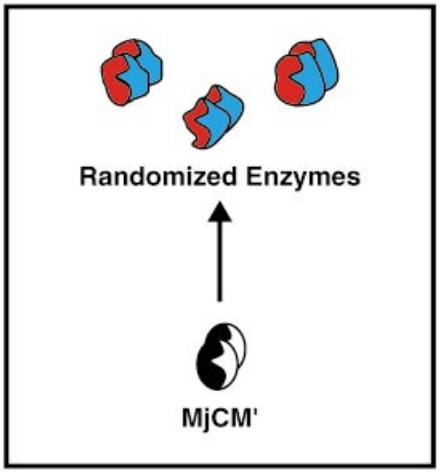
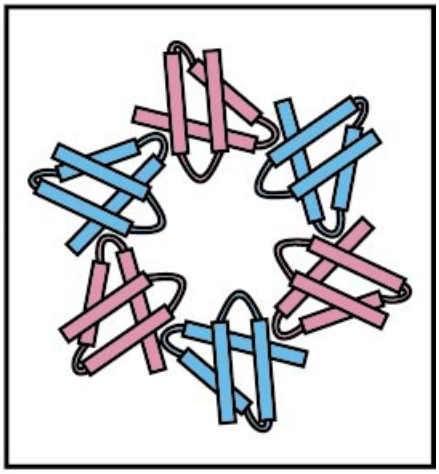
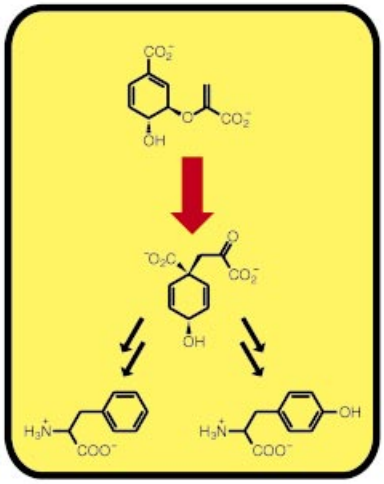
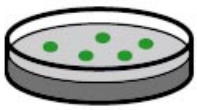
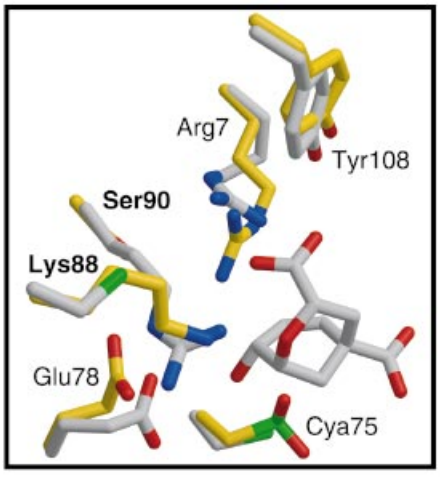


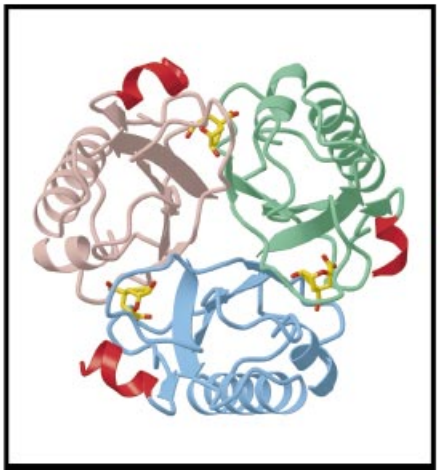


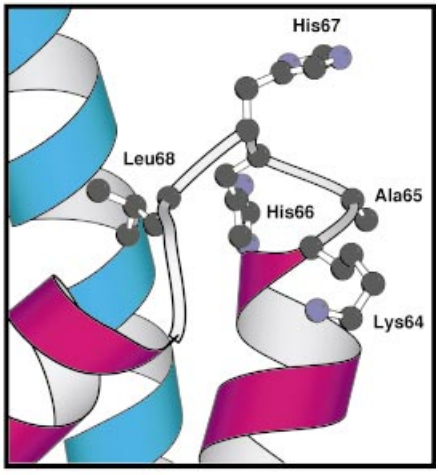












Investigating and Engineering Enzymes by Genetic Selection

Sean V. Taylor, Peter Kast, and Donald Hilvert*

Natural enzymes have arisen over millions of years by the gradual process of Darwinian evolution. The fundamental steps of evolution—mutation, selection, and amplification—can also be exploited in the laboratory to create and characterize protein catalysts on a human timescale. In vivo genetic selection strategies enable the exhaustive analysis of protein libraries with 10^{10} different members, and even larger ensembles can be studied with in vitro methods. Evolutionary approaches can consequently yield statistically meaningful insight into the complex and often subtle interactions that influence protein folding, structure, and catalytic mechanism. Such methods are also being used increasingly as an adjunct to design, thus providing access to novel proteins with tailored catalytic activities and selectivities.

Keywords: directed evolution · enzyme catalysis · enzymes · genetic selection · protein design · protein structures

1. Introduction

Evolution is the slow and continual process by which all living species diversify and become more complex. Through recursive cycles of mutation, selection, and amplification, new traits accumulate in a population of organisms.[1] Those that provide an advantage under prevailing environmental conditions are passed from one generation to the next. Since ancient times, man has exploited evolution in a directed way to produce plants and animals with useful characteristics. The cross-breeding of individuals with favorable traits successfully harnesses sexual recombination, one of the most powerful evolutionary strategies to generate new variants. From these crossings, progeny with improved features are chosen for additional breeding cycles, thus channeling the course of development.

Biologists and chemists have recently begun to use evolutionary strategies to tailor the properties of individual molecules rather than whole organisms. An array of methods has been developed to generate diversity in populations of molecules. Depending on the experiment, mutagenesis might entail degenerate oligonucleotide-directed or error-prone DNA synthesis,[2, 3] shuffling of mutant DNA fragments,[4, 5] or combinatorial syntheses of chemical compound libraries.[6] From the resulting molecular ensembles, desirable members must be identified by selection or screening procedures. In the amplification step, the self-replicating properties of the evolving molecules can be exploited (for example, if these are nucleic acids). Alternatively, more of the desired compound can be prepared by large-scale chemical synthesis.

In practice, the sorting step is the most critical part of any laboratory evolution experiment. How can rare but useful variants be efficiently isolated from complex mixtures of less desirable molecules? A chemist seeking the optimal octameric peptide inhibitor for a particular enzyme might start by making systematic substitutions at a single site in a known peptide. If allowed substitutions are restricted to the 20 standard amino acids, assaying the 20 different octameric peptides individually to find the highest affinity inhibitor would be relatively trivial. Simultaneously randomizing two positions in the peptide would generate 400 (20^2) different variants. Screening and testing such a library would be substantially more involved, but still manageable. However, it would be a daunting task indeed to identify the best inhibitor in a combinatorial library of fully randomized octameric peptides having 20^8 ($=2.6\times10^{10}$) individual members—unless an extremely efficient and sensitive sorting procedure were available.

Genetic selection is perhaps the most powerful technique currently available for analyzing large populations of biomacromolecules. In this review we focus on its use as a tool in mechanistic enzymology and protein design. Because numerous genetic selection systems have been developed, only a small fraction of which can be covered here, the reader is also referred to other recent reviews on this topic.[7–17]

[*] Prof. D. Hilvert, Dr. S. V. Taylor, Dr. P. Kast
Laboratorium für Organische Chemie
ETH Zürich
8093 Zurich (Switzerland)
Fax: (+41) 1-632-1486
E-mail: hilvert@org.chem.ethz.ch

  1433-7851/01/4018-3311 $ 17.50+.50/0

2. Selection versus Screening

Identifying interesting variants in large combinatorial libraries can be accomplished either by assaying all the members individually or by applying conditions that allow only variants of interest to appear. Unfortunately, the terms "screening" for the former strategy and "selection" for the latter are often confounded in the literature. The basic difference between these approaches is illustrated schematically in Figure 1.

In general, screening strategies require an active search of all variants in the library. This approach is depicted in Figure 1 A and B from the point of view of a bumblebee. Desirable nectar-producing flowers must be detected against a large background of flowers unable to produce nectar (undesired library members). If desired and undesired variants are difficult to distinguish, screening must be conducted one at a time in a "random" or "blind" fashion (Figure 1 A). A "facilitated screening" strategy is sometimes possible if the desired variants have a distinctive appearance or "phenotype" (Figure 1 B). Nevertheless, irrelevant library members are always present, which lowers the signal-to-noise ratio and which compete for resources.

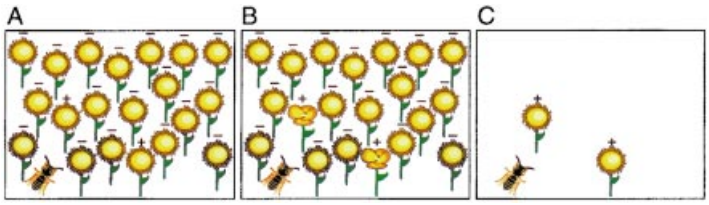


Figure 1. Different search modes for finding nectar-producing flowers (marked with +). A) Random screening: The bumblebee must check each flower individually to differentiate producers from the large number of nonproducers (marked with −). B) Facilitated screening: The bee can distinguish nectar-producing from nonproducing flowers on the basis of their distinct phenotypes. C) Selection: The bee only encounters the desired flowers because nectar production was a prerequisite for plant growth in this case.

In contrast, selection strategies exploit conditions favoring the exclusive survival of desired variants, thereby mimicking the culling process associated with true Darwinian evolution. Figure 1 C shows how the bee easily encounters nectar-producing flowers under selective conditions that eliminate the flowers that do not produce nectar.

2.1. Classical Solutions to the Sorting Problem

Selection and/or screening techniques are widely used in molecular biology, particularly for the cloning of DNA fragments or genes into circular DNA molecules called vectors, which can be replicated inside bacteria by the cellular DNA synthesis machinery.[18, 19] Cloning involves enzymatic

D. Hilvert P. Kast S. V. Taylor

Donald Hilvert received a Ph.D. in chemistry from Columbia University, New York, in 1983. After postdoctoral studies at Rockefeller University, New York, he joined the faculty of the Scripps Research Institute in La Jolla, California, where he pursued his interests in chemical biology. In 1997 he moved to Switzerland and assumed his current position as Professor of Chemistry at the Swiss Federal Institute of Technology (ETH). His research program focuses on understanding how enzymes work and on mimicking the properties of these remarkable catalysts in the laboratory. These efforts have been recognized by a number of awards, including the Pfizer Award in Enzyme Chemistry.

Peter Kast received his Diploma in Natural Sciences (Dipl. Natw. ETH) in 1986 from the Swiss Federal Institute of Technology (ETH) Zürich. He completed his doctoral studies in 1991 at the Institute of Microbiology of the ETH. He worked as a postdoctoral fellow at the ETH from 1991 to 1992 and in the Departments of Chemistry and Molecular Biology at the Scripps Research Institute in La Jolla, California from 1992 to 1995. He was appointed Assistant Professor in the Departments of Molecular Biology and Chemistry at Scripps in 1995. Since December 1997, he has been at the Department of Chemistry (Laboratory of Organic Chemistry) at the ETH Zürich where he continues work on protein engineering through directed molecular evolution using genetically modified bacterial systems.

Sean V. Taylor received his B.S. in chemistry in 1992 from Penn State University. He completed his Ph.D. in 1998, with research on the biosynthesis of thiamin with Prof. Tadhg Begley in the Department of Chemistry at Cornell University. In 1998, he joined Prof. Hilvert's laboratory at the ETH Zürich as a postdoctoral associate, where he focuses on protein design and evolution.

ligation of the ends of the insert fragment with the corresponding complementary ends of an appropriately cut vector.[20] The resulting circular molecule, a plasmid, is then transferred into bacteria (usually the species *Escherichia coli*) by a process called transformation. During transformation, only a small percentage of cells successfully take up DNA. To select for these transformed bacteria, the cell suspension is plated onto agar medium in petri dishes that contains a particular antibiotic. Since the vector also carries a gene that specifies resistance against this antibiotic, only those cells that have received the original vector or the new plasmid survive, multiply, and form a colony on the agar medium. Because every one of the approximately 10^7 cells in the colony contains identical DNA, they all belong to the same clone.

After this very effective initial selection step, the molecular biologist is left with the problem of finding those clones that have the desired plasmid rather than the original vector molecule lacking the inserted gene.[19] Depending on the complexity of the cloning strategy and the efficiency of the ligation step, relatively few of the antibiotic-resistant clones may actually possess the desired DNA fragment. The examples in Figure 2 show how an in vivo selection approach, which employs a positive-selection cloning vector, is superior to either random or facilitated screening in detecting clones with the desired genetic makeup (or "genotype").

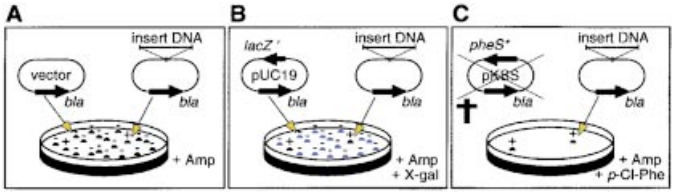


Figure 2. Search strategies for differentiating colonies with plasmids containing an inserted gene (marked with +) from those with only vector DNA (marked with −) after plating on nutrient agar plates following transformation.[20] The *bla* gene encodes β-lactamase, which renders transformants resistant to the antibiotic ampicillin. A) Random screening: The phenotype of all colonies is identical. DNA from many randomly picked clones must be isolated and analyzed to find the desired plasmid. B) Facilitated screening: The cloning vector pUC19[223] carries a fragment (*lacZ'*) of the *lacZ* gene encoding β-galactosidase. The encoded polypeptide (the α-fragment of the enzyme) can associate with the product of a cellular *lacZ* fragment to form functional β-galactosidase,[224] which cleaves 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) resulting in the precipitation of a blue indigo derivative.[225] Vector-containing colonies consequently have a blue color on agar plates containing X-gal. In contrast, the desired plasmid-containing clones form colorless colonies because the cloned gene was inserted into the coding region for the α-fragment, thereby abolishing β-galactosidase activity. The clones with a white colony phenotype (shown here as black spots) must be picked from the background of blue colonies to find the correct genotype. C) Selection: The positive-selection vector pKSS carries a mutant *pheS* gene (*pheS**), which confers relaxed substrate specificity to the encoded phenylalanyl–tRNA synthetase α subunit.[226] In the presence of the phenylalanine analogue *p*-chlorophenylalanine, bacteria expressing the mutant *pheS* gene on pKSS cannot survive because incorporation of the analogue into cellular protein causes general failure of protein function. However, transformed cells can form colonies on nutrient agar plates containing ampicillin and *p*-chlorophenylalanine if the conditionally lethal *pheS** gene on pKSS is destroyed by insertion of a cloned DNA fragment. Thus, all colonies growing after plating the transformation mixture have the correct genotype.

2.2. Advantages and Limitations of Selection

The main advantage of selection over screening is that many more library members can be analyzed simultaneously. This is because uninteresting variants are never seen. As a consequence, surveying libraries is much faster and can be carried out with higher throughput. In the best screening protocols currently available, which take advantage of fluorogenic or chromogenic substrates,[21] the maximum number of library members that can be assayed is about 10^5.[22] In contrast, up to 10^{10} clones can be assessed in a single experiment using genetic selection in *E. coli* cells. This number is an upper limit based on transformation efficiencies obtainable with conventional protocols.[23] The absolute limit, at least in principle, is only dependent on the scale of the transformation.

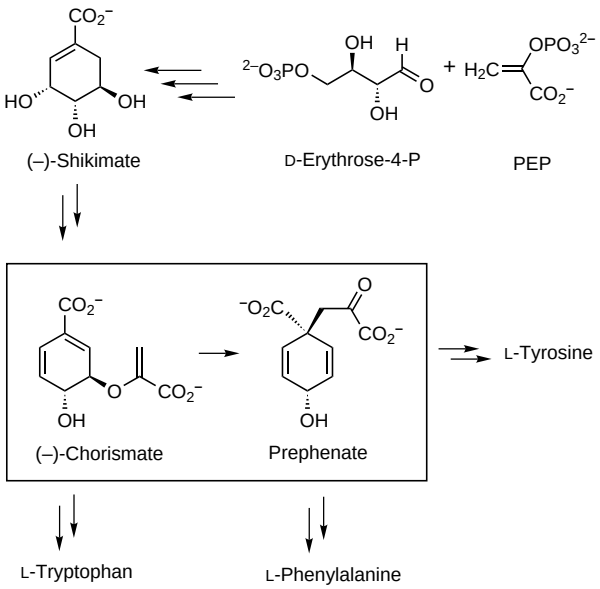
Nevertheless, it is challenging to develop suitable selection strategies for every desired catalytic task. In particular, in vivo selection schemes may be difficult or even impossible to devise if enzymes that work in non-natural environments, such as organic solvents, are desired.[24, 25] Coupling the target reaction to survival in the selection step may require development of complex, nontrivial and intelligent assays. With every selection system, and in particular with very complex ones, there always exists the possibility that viable, but unanticipated or undesired solutions to the posed survival problem will surface.[26, 27] If these false positives become too abundant, an efficient screening step may be necessary, or the system may need to be redesigned to eliminate this unwanted background by introduction of an additional selection step.

It is in the realm of very large combinatorial libraries that selection rather than screening gains crucial importance. As the focus shifts from randomizing an eight-residue peptide to a 100 amino acid protein (the typical size of a small functional domain), the number of sequence permutations rises to an astronomical 20^{100}. The ability to assay even a tiny fraction of this sequence space in directed molecular evolution experiments demands selection, even though initial development of an appropriate system may be considerably more involved than the setup of a screening procedure.

3. Genetic Selection of Novel Chorismate Mutases

The utilization of evolutionary strategies in the laboratory can be illustrated with proteins that catalyze simple metabolic reactions. One of the simplest such reactions is the conversion of chorismate to prephenate (Scheme 1), a [3,3] sigmatropic rearrangement. This transformation is a key step in the shikimate pathway leading to aromatic amino acids in plants and lower organisms.[28, 29] It is accelerated more than a million-fold by enzymes called chorismate mutases.[30]

Chorismate mutases have been intensively studied because of the scarcity of other enzyme-catalyzed pericyclic reactions. Mutases from different organisms exhibit similar kinetic properties, though they may share little sequence similarity.[30] This dissimilarity extends to their tertiary and quaternary structures, as shown by comparison of chorismate mutases from *Bacillus subtilis* (BsCM),[31, 32] *E. coli* (EcCM),[33] and the



Scheme 1. Shikimate pathway for the biosynthesis of aromatic amino acids in plants and lower organisms. The [3,3] sigmatropic rearrangement of chorismate into prephenate is shown in the box. PEP = phosphoenolpyruvate.

yeast *Saccharomyces cerevisiae* (ScCM)[34, 35] (Figure 3). BsCM, a member of the AroH class of chorismate mutases, is a symmetric homotrimer, packed as a pseudo-α/β-barrel. It has three identical active sites located at the subunit interfaces (Figure 3 A). In contrast, EcCM and ScCM (Figure 4 B and C, respectively) are distantly related homodimeric members of the *all*-α-helical AroQ class of chorismate mutases.[36, 37] Homology searching in sequence databases has yielded other proteins that belong to each of these structural classes.[36–38] Catalytic antibodies with modest chorismate mutase activity have also been generated.[39–41] They have typical immunoglobulin folds, as shown in Figure 3 D for antibody 1F7.[42]

The architectural diversity that nature exploits to promote this relatively simple chemical transformation is impressive and raises many intriguing questions. How do the individual enzymes work? What similarities and what differences are present? Why are natural chorismate mutases 10^4 times more effective than the catalytic antibody? Can mechanistic insights be used to improve the latter? More generally, what are the underlying chemical determinants of each structure? The individual catalytic domains of the BsCM, EcCM, and antibody proteins are each approximately 100 residues long. Why does the polypeptide adopt a mixed α/β conformation in one case, an *all*-α-helical structure in another, and an *all*-β topology in the third? The relationship between sequence, structure, and function can be rapidly and effectively examined by using random mutagenesis and genetic selection in vivo.

3.1. The Selection System

Over the past decade we have engineered several strains of yeast[43, 44] and *E. coli*[45, 46] that lack chorismate mutase. A typical bacterial selection system is depicted in Scheme 2. It is

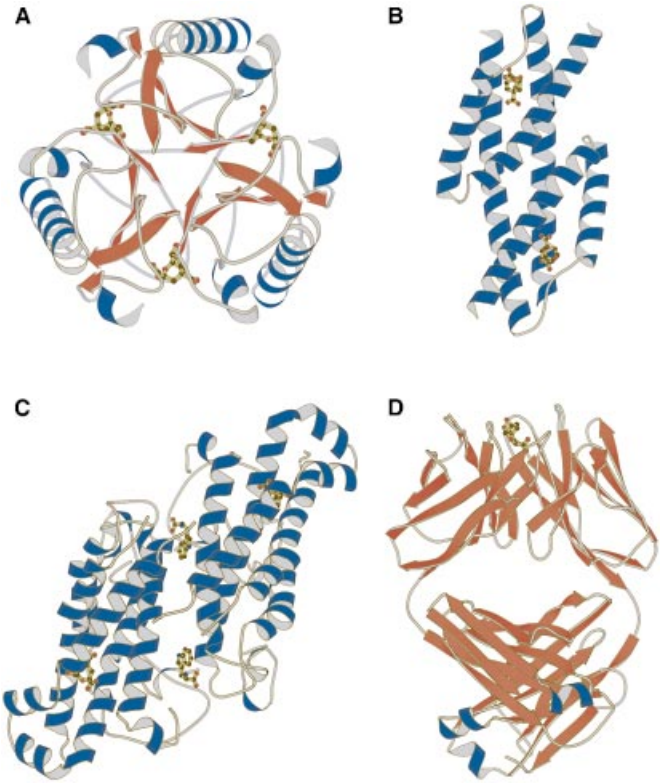
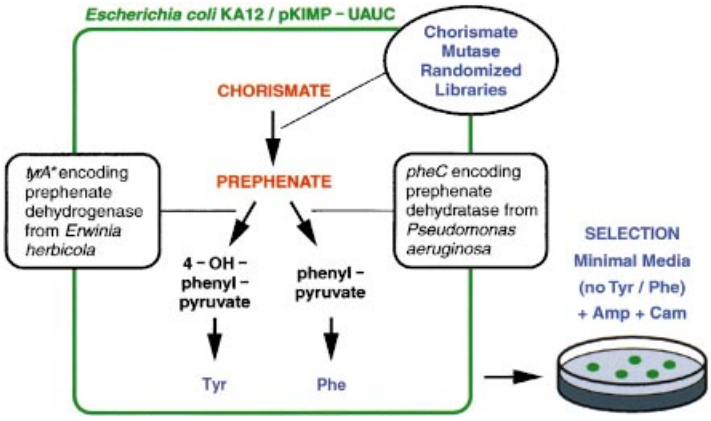


Figure 3. Ribbon diagrams of three naturally occurring chorismate mutases and a catalytic antibody. Helices are shown in blue and β-sheets in red. A transition state analogue inhibitor (**1**, Scheme 3) is shown in a ball-and-stick representation at each active site, with carbon atoms black and oxygen atoms red. A) Monofunctional chorismate mutase from *B. subtilis* (BsCM).[31, 32] B) Chorismate mutase domain of the bifunctional chorismate mutase/prephenate dehydratase from *E. coli* (EcCM).[33] C) Allosterically regulated chorismate mutase from the yeast *S. cerevisiae* (ScCM).[34] Tryptophan, which activates the enzyme, is shown bound at the allosteric sites located at the dimer interface. D) Catalytic antibody 1F7,[42] which was generated by using **1** as a hapten.[39]



Scheme 2. An engineered *E. coli* selection system lacking endogenous chorismate mutase activity.[45] The genes encoding the bifunctional enzymes chorismate mutase/prephenate dehydrogenase and chorismate mutase/prephenate dehydratase were deleted, and monofunctional versions of the dehydrogenase and dehydratase were supplied on plasmid pKIMP-UAUC. Potential chorismate mutases from random gene libraries are evaluated on the basis of their ability to complement the genetic defect and allow the cells to grow in the absence of added tyrosine (Tyr) and phenylalanine (Phe). The library plasmid and pKIMP-UAUC carry genes for ampicillin (Amp) and chloramphenicol (Cam) resistance, respectively.

based on *E. coli* strain KA12,[46] which has deletions of the chromosomal genes for both bifunctional chorismate mutases (chorismate mutase/prephenate dehydrogenase and chorismate mutase/prephenate dehydratase). Monofunctional versions of prephenate dehydratase[47] and prephenate dehydrogenase[48] from other organisms are supplied by the plasmid pKIMP-UAUC, which leaves the cells deficient only in chorismate mutase activity.[45]

As a consequence of the missing enzyme, KA12/pKIMP-UAUC bacteria are unable to grow in the absence of exogenously added tyrosine and phenylalanine. Their metabolic defect can be thought of as an interrupted circuit. Supplying the cells with an additional plasmid encoding a natural chorismate mutase can repair this genetic defect. Alternatively, a mutant of the natural enzyme or even an unrelated polypeptide can be introduced into the cell. If the protein is able to catalyze the conversion of chorismate to prephenate, the cells will be able to produce their own tyrosine and phenylalanine and grow under selective conditions (that is, without tyrosine and phenylalanine supplementation). If the protein does not accelerate the rearrangement, the cells will not grow at all. As a consequence, attention can be focused on functional clones without wasting time and resources on nonproductive members of the library. Additionally, the linkage of the desired phenotype (catalyst-enabled growth under selective conditions) with the responsible genotype (the gene encoding the catalyst) is automatic in this system, and allows easy isolation, diversification, and further evolution of interesting library variants.

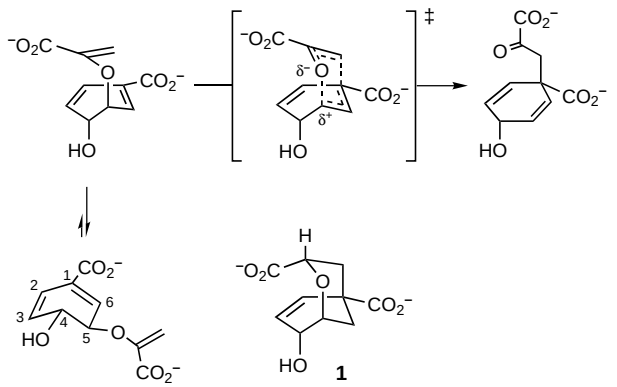
Since misplacement of catalytic residues by even a few tenths of an angstrom can mean the difference between full activity and none at all, enzymatic activity represents an extremely stringent criterion for selection. Insightful answers to many of the questions posed in the previous section can, therefore, be obtained by searching intelligently designed protein libraries for active chorismate mutases. The utility of this strategy in areas ranging from mechanistic enzymology to protein design is best illustrated through a few examples.

3.2. Mechanistic Studies

3.2.1. Active-Site Residues

Our understanding of the chorismate rearrangement derives largely from mechanistic studies of the uncatalyzed reaction[49–52] and from computation studies.[53–59] In order to rearrange, chorismate, which normally has a pseudo-diequatorial conformation in solution,[51] must first adopt a pseudodiaxial conformation (Scheme 3). It then reacts exergonically[60] via a transition state with a chairlike geometry.[50] The process appears to be concerted but asynchronous, with formation of the C–C bond lagging behind cleavage of the C–O bond.[49]

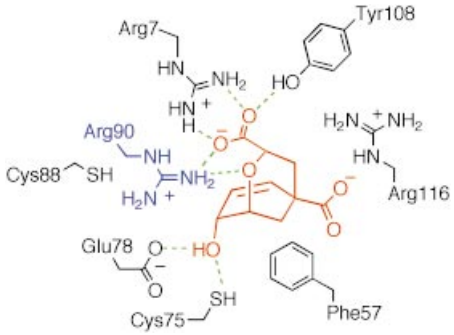
The transition state for the enzymatic reaction has been shown to have a chairlike geometry as well,[61] and conformationally constrained compounds that mimic this structure, such as the oxabicyclic dicarboxylic acid **1** (Scheme 3), are good inhibitors of chorismate mutase enzymes.[62–64] How a



Scheme 3. Chorismate prefers a pseudodiequatorial conformation in solution. It must adopt a disfavored pseudodiaxial conformation to reach the pericyclic transition state. The conformationally constrained oxabicyclic dicarboxylic acid **1**, which mimics the transition state, is a potent inhibitor of natural chorismate mutases.[62] Antibodies raised against this compound also catalyze the reaction, albeit 100 to 10000 times less efficiently than their natural counterparts.[39, 41]

protein might stabilize this high-energy species has been a matter of some debate. Recently, heavy atom isotope effects were used to characterize the structure of the transition state bound to BsCM.[65] A very large ^{18}O isotope effect at O(5) (ca. 5%), the site of bond cleavage, and a small, normal ^{13}C isotope effect at C(1) (ca. 0.6%), the site of bond formation, show that the rearrangement step is significantly rate determining for this enzyme. These findings also confirm that the enzymatic reaction proceeds through a concerted but asymmetric transition state. Comparison with theoretical isotope effects obtained by Becke3LYP/6-31G* calculations[65] suggests, however, that the enzymatic transition state might be more highly polarized than its solution counterpart, with more C–O bond cleavage and less C–C bond formation.

Examination of the residues that line the active site of BsCM (Scheme 4)[31, 32] suggests a plausible explanation for the greater polarization of the bound transition state. The positively charged guanidinium group of Arg90 is located within hydrogen-bonding distance of the ether oxygen atom of the bound substrate, where it can stabilize the partial negative charge that builds up at this site in the transition



Scheme 4. Schematic representation of the BsCM active site. An extensive array of hydrogen bonding and electrostatic interactions is used to bind transition state analogue **1** (red). The side chain of Arg90 (blue) is oriented so that its guanidinium group is placed within hydrogen-bonding distance of the ether oxygen atom of the ligand.

state. The contribution of such a residue to catalysis would normally be investigated by site-directed mutagenesis. For instance, Arg 90 might be replaced with lysine or methionine to examine what happens when the guanidinium is replaced with a different cation or when the positive charge is removed completely. However, since a selection system is available, all 20 natural amino acids can be evaluated at position 90 simultaneously. Those amino acids that yield a functional enzyme can be quickly identified by sequencing the genes able to complement the chorismate mutase deficiency. When this experiment was performed, the results were dramatic.[45] Only arginine at position 90 yielded an active enzyme. Even the conservative replacement of arginine with lysine gave an enzyme that was inactive in vivo. Kinetic characterization of representative mutants showed that removal of the arginine side chain reduces catalytic efficiency by five to six orders of magnitude.[38]

If the role of Arg 90 is electrostatic stabilization of the developing negative charge on the ether oxygen atom of chorismate in the transition state, one might wonder why complementation was not observed when arginine was replaced with another positively charged residue such as lysine. One possibility is that the Arg 90Lys variant is produced poorly in *E. coli*. The growth phenotype depends on the total amount of chorismate mutase activity present in the cell, which is determined by the specific activity of the catalyst and also its concentration. If the catalyst is active but present only at extremely low concentrations, no growth will be observed. While the BsCM mutants appear to be produced at comparable levels within the selection system,[45] exceptions are possible.

A potentially more interesting explanation is that the active site of the enzyme places intrinsic structural constraints on possible substitutions. The lysine side chain is shorter than that of arginine and it is conceivable that it cannot reach far enough into the active site to place its ammonium group within hydrogen-bonding distance of the critical ether oxygen atom. If this is true, additional mutations at other positions within the active site might allow replacement of Arg 90 without loss of function.

To test this idea, we simultaneously mutated two residues—Arg 90 and Cys 88—within the binding pocket of BsCM.[45] The side chain of Cys 88 is about 7-Å distant from the bound ligand, but it lies against the side chain of Arg 90 and might influence the conformation of the latter (Scheme 4). Combinatorial mutagenesis of these two residues gives 400 variants, which can be rapidly evaluated in the chorismate mutase deficient selection strain. The results of this experiment proved quite informative (Figure 4).[45] They showed first that Cys 88 is not essential for catalysis. It can be replaced by large, medium-sized, and small amino acids; as long as an arginine residue is present at position 90, an active enzyme is obtained. Active double mutants with an additional Arg 90Lys mutation were found if a small residue (glycine or alanine, for example) replaces Cys 88, which shows that the guanidinium group is not crucial for catalysis. Perhaps the most interesting result to emerge from this study is that a cation at position 90 can be eliminated altogether, provided that Cys 88 is replaced with a lysine residue.

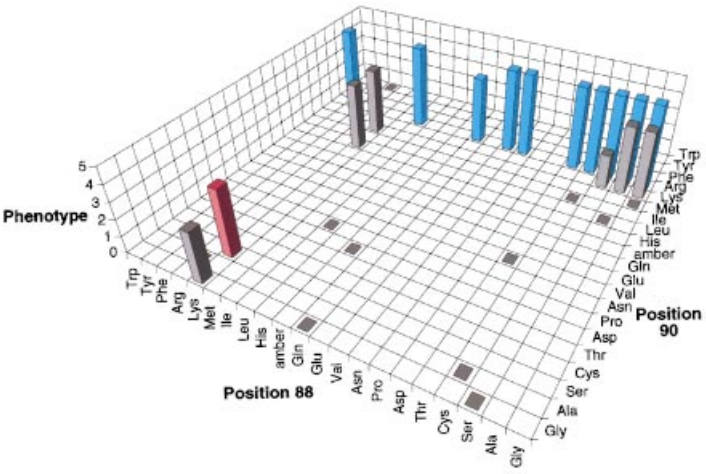


Figure 4. Outcome of a combinatorial mutagenesis and selection experiment involving simultaneous randomization of positions 88 and 90 in BsCM.[45] The amino acids are ordered according to their side-chain volume. Active variants that were sequenced and a few inactive controls are shown with their in vivo activity, or phenotype. The phenotype is reflected by colony size, which is rated from 0 (no growth) to 5 (wild-type growth). Blue columns show variants with wild-type Arg 90. The red column is variant Cys 88 Lys/Arg 90 Ser, which was analyzed in more detail (see Figure 5).

Kinetic studies on the Cys 88Lys/Arg 90Ser double mutant show that a lysine residue at position 88 restores three of the five to six orders of magnitude in catalytic efficiency lost upon removal of the arginine side chain at position 90.[38] Crystallographic studies of this double mutant show that the side chain of Lys 88 extends into the active site and orients its ammonium group, like the guanidinium group of Arg 90 in the wild-type enzyme, within hydrogen-bonding distance of the ether oxygen atom of the bound ligand (Figure 5).[38] Thus, combinatorial mutagenesis and selection resulted in a redesigned active site, which afforded a novel solution to the chemical problem of catalyzing the chorismate mutase reaction.

The results of these selection experiments are mechanistically significant insofar as they support a critical role for a cation in the mechanism of chorismate mutase. No active catalysts were found that lacked a cation in the vicinity of the

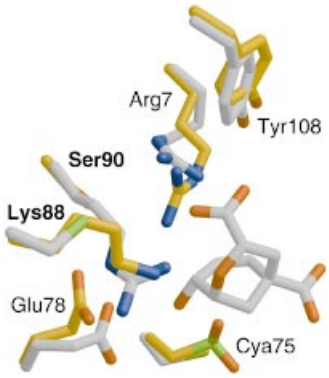
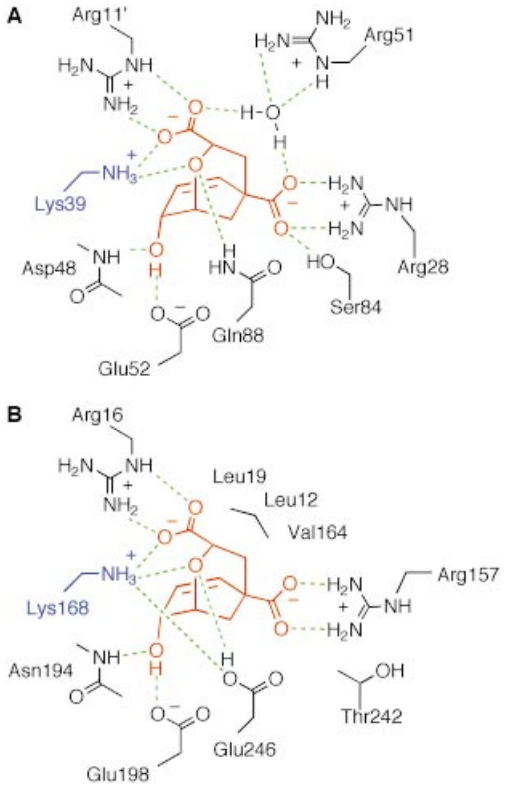


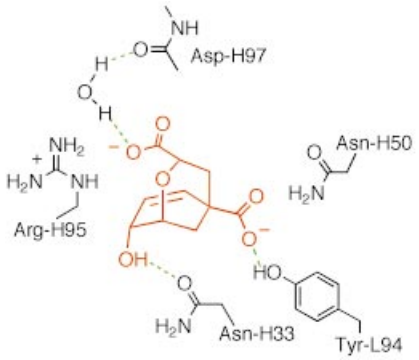
Figure 5. Superposition of residues from unbound Cys 88 Lys/Arg 90 Ser (carbon atoms in yellow, residues labeled) and wild-type BsCM complexed with inhibitor **1** (grey carbon atoms).[38] Cys 75 is partially oxidized to the sulfinate (Cya 75) in the crystal of the mutant.

substrate's ether oxygen atom. The conclusion that a cation is crucial for high chorismate mutase activity can thus be made with much greater confidence than would have been possible following a conventional mutagenesis experiment in which only single substitutions were considered.[66] Of course, we cannot exclude the possibility that other, equally effective arrangements of catalytic groups lacking a cation might be found in the future through more extensive mutagenesis experiments. However, it is worth noting that chorismate mutases from both *E. coli*[33] and yeast[34, 35] have a cationic lysine residue in an equivalent position in their respective active sites (Scheme 5), even though their tertiary structures are otherwise unrelated to BsCM. This lysine residue has been shown by mutagenesis to be catalytically essential in EcCM.[67, 68]



Scheme 5. Schematic representation of the active sites of EcCM (A) and ScCM (B) complexed with transition state analogue **1**. The lysine counterparts of Arg90 in BsCM are shown in blue.

In this context, it is also interesting that the 10^4-fold less-efficient catalytic antibody 1F7[39, 40] lacks a comparable strategically placed cation (Scheme 6).[42] This fact is not terribly surprising given the structure of the transition state analogue used to generate this catalyst (see Scheme 3).[39] Nevertheless, it does suggest possible strategies for producing antibodies with much higher activity. For example, additional negative charges could be designed into a hapten to elicit new antibodies that contain the catalytically essential cation. Alternatively, a cation might be directly engineered into the 1F7 active site by using site-directed mutagenesis. A third approach that utilizes selection may ultimately be the most effective strategy for optimizing antibody activity, however.



Scheme 6. Schematic representation of the active site of catalytic antibody 1F7 with bound **1**. Arg-H95 is not within hydrogen-bonding distance of the ligand's ether oxygen atom.[42]

The Fab fragment of 1F7 has already been shown to function in the cytoplasm of a chorismate mutase deficient yeast strain.[43, 44] The catalytic antibody, when produced at a sufficiently high level, is able to replace the missing enzyme and weakly complement the metabolic defect. Conceivably, therefore, it can be placed under selection pressure to identify variants that have higher catalytic efficiency. Preliminary results from such experiments appear quite promising.[69]

3.2.2. Random Protein Truncation

Identifying appropriate residues for mutational analysis is difficult when structural information for a protein is lacking. In such cases, random mutagenesis coupled with selection can provide a powerful means of analysis. We have used such an approach to examine how the 17 C-terminal residues of BsCM contribute to enzyme efficiency.[70]

BsCM is 127 amino acids long. The first 115 residues adopt a well-defined structure, which ends in a 3_{10} helix near the active site (Figure 3A).[31, 32] Residues beyond this point are not seen in X-ray structures of the free enzyme,[31] which implies that they are highly disordered or have multiple accessible conformations. Some ordering of the C-terminus is observed crystallographically upon binding of the ligand to BsCM.[31] In addition, ligand-induced conformation changes have been inferred from Fourier-transform infrared (FT-IR) spectroscopic studies.[71] These observations raise the possibility that the flexible C-terminal tail might play some role in enzyme function. For example, it might serve as a lid to the active site that sequesters the substrate from bulk solution and/or contributes stabilizing interactions to the transition state.

To address these questions we developed a strategy involving random C-terminal truncation of the enzyme followed by selection of functional clones in the KA12/pKIMP-UAUC system.[70] By using a polymerase chain reaction (PCR) procedure with a partially randomized oligonucleotide, we constructed two gene libraries in which the BsCM codons corresponding to residues 116–127 or residues 111–127 (including the crystallographically observed 3_{10} helix) were randomized. The libraries were designed to optimize the frequency of stops at each mutagenized codon. In this way, a nested set of randomly truncated proteins was

created. Since only functional genes survive selection, dispensable portions of the encoded proteins can be rapidly and directly identified. Non-native sequences are also explored by this method, so residues that might fulfill a crucial structural or mechanistic role can also be discovered.

After construction and transformation into the bacterial selection strain, both libraries contained 10^4-10^5 members. Essentially all members of the library in which the 12 C-terminal residues were mutagenized were viable on selection medium. These results demonstrate that none of the 12 C-terminal residues are absolutely required for enzymatic activity. They can be mutated and the last 11 residues can even be removed without significantly impairing activity in vivo. In contrast, when the last 17 residues were mutagenized, only about 25% of the library was able to grow on selection plates. The additional five amino acids form a well-defined 3_{10} helix and provide contacts with the rest of the protein and its ligands. Sequencing data show that mutations are tolerated at these positions, although a clear preference for wild-type Lys111, Ala112, Leu115, and Arg116 is observed. Moreover, active enzymes shorter than 116 amino acids were not found. Apparently, residues in the 3_{10} helix cannot be removed without killing the enzyme.

Kinetic characterization of several selected BsCM variants shows that truncation or mutation of the C-terminal tail has little effect on the turnover number (k_{cat}) of the enzyme (Table 1). When chorismate is bound to the active site of the variants, it is converted into prephenate nearly as efficiently as with wild-type BsCM. However, a substantial reduction in the k_{cat}/K_m value is evident (Table 1). This finding indicates that the C-terminus, while not directly involved in the chemical transformation of bound ligand, does contribute to enzymatic efficiency by uniformly binding the substrate and transition state.

Table 1. Catalytic parameters of truncated BsCM variants.[70]

BsCM variant	C-terminal amino acid sequence	k_{cat} [s^{-1}]	k_{cat}/K_m [$M^{-1}s^{-1}$]
5–8	KAVVLR	26	2.8×10^3
1–3	NSNVLRP	30	2.0×10^3
V-7	KAVVLLT	26	1.6×10^3
5–11	KAVVLRPN	23	5.6×10^3
wild-type	KAVVLRPDLSLTKNTEL	46	6.9×10^5

3.3. Structural Studies

Selection methods are ideally suited for examining factors that influence protein structure and stability. The informational content of protein sequences is notoriously non-uniform, with many residues being highly tolerant to a wide range of substitutions, whereas others cannot be altered without dramatic consequences for folding or function.[72, 73] Analysis of protein structure is thus a combinatorial problem requiring efficient methods for the simultaneous evaluation of many different alternative sequences. An experimentally accessible population of 10^{10} molecules is miniscule compared with the 20^{100} ($=10^{130}$) possibilities for a polypeptide with a length of 100 amino acids. Nevertheless, libraries of this size can provide statistically meaningful insights into intrinsic secondary structural preferences in proteins,[74, 75] constraints on segments that link secondary structural elements,[76–78] and optimal packing arrangements of residues in the interior of a protein.[79–82]

3.3.1. Constraints on Interhelical Loops

The role of interhelical turns in determining protein structure has been explored in relatively simple four-helix-bundle proteins by generating combinatorial libraries and screening them for functional variants. For example, Hecht and co-workers mutagenized a three-residue interhelical turn in cytochrome *b*-562.[76] All 31 variants isolated from the random library adopted stable nativelike structures as judged by the characteristic red color associated with heme complexes of cytochrome *b*-562. Similar results have been obtained with mutagenized turns in Rop,[77, 83] a dimeric RNA-binding four-helix-bundle protein, although detailed analysis of individual mutants showed some variation in thermostability.[84] Collectively, these studies suggest that there are relatively few constraints on sequence or length in interhelical segments.

How general are such conclusions? Do they apply to more complex proteins? Are there more elusive features that influence turn preferences? Selection for catalytic activity, an extremely sensitive probe of structural integrity, has the potential to reveal more subtle sequence preferences than simple screening protocols. We tested this premise using the EcCM enzyme (Figure 3B) as a template.[85] The two identical polypeptides that make up this helical bundle protein consist of three helices joined by two loop segments (Figure 6A).[33] The L2 loop connecting the H2 and H3 helices lies farthest from the active site and was targeted for analysis.

Initially, three solvent-exposed residues in the turn (Ala65–His66–His67, Figure 6B) were randomized and catalytically active variants identified by selection in vivo. The starting ensemble contained 8000 (20^3) distinct members, of which more than 63% were found to complement the chorismate mutase deficiency of the bacterial host strain, albeit with widely varying growth rates. The high percentage of complementing clones clearly confirms that there are few restrictions on this loop segment. Nevertheless, the range of growth rates observed for complementing clones suggests that substitutions in the turn can affect protein production, stability, or catalytic activity. In fact, careful sequence comparison uncovered a statistically meaningful preference for hydrophilic residues at these solvent-exposed positions in the variants affording the fastest growth. Although proline-containing turns with alternative backbone conformations are also tolerated, such sequences were only found in partially active clones.

Randomization of the same Ala65–His66–His67 tripeptide in combination with Lys64, which is part of the adjacent H2 helix (Figure 6), enabled investigation of the extent to which proximal secondary structure influences the allowable loop substitution patterns. A library containing all possible 160000 (20^4) members was constructed and subjected to selection. Again, a large fraction of these sequences (>50%) could functionally replace the native sequence, which shows

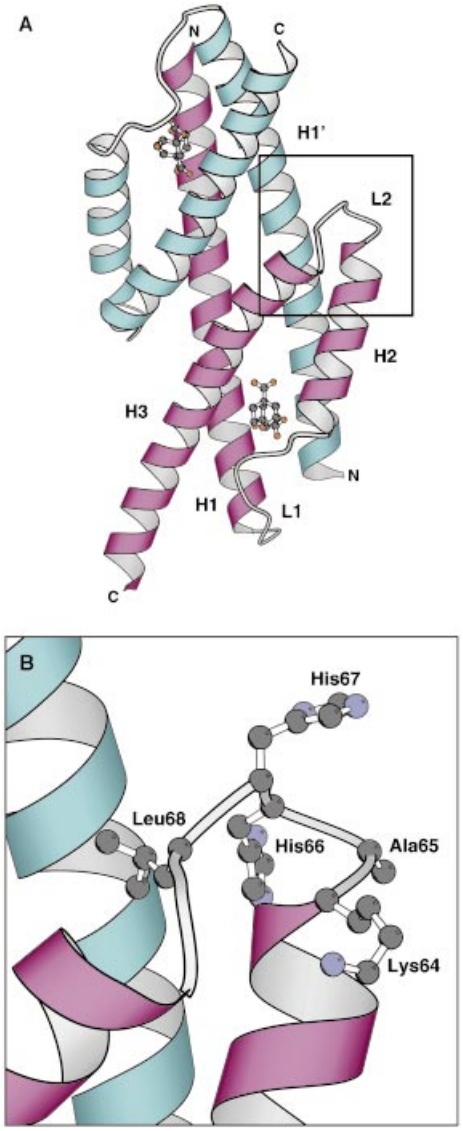


Figure 6. A) Ribbon structure and B) L2 loop of EcCM. Random mutagenesis of the L2 loop followed by selection for chorismate mutase activity in vivo showed little sequence constraint on solvent-exposed turn residues, aside from a modest bias in favor of hydrophilic amino acids.[85] In contrast, long-range tertiary contacts impose a strict requirement for hydrophobic aliphatic amino acids at position 68.

that the secondary structure at position 64 does not limit the range of side chains allowed at this site, and that its randomization does not affect residue preferences in the region of the neighboring turn.

Quite different results were obtained when the buried turn residue Leu 68 was randomized together with Ala 65 – His 66 – His 67. Unlike the other turn residues, the side chain of Leu 68 is buried at the interface between L2 and H1′ (helix H1 of the other subunit) and thus sequestered from solvent (Figure 6). In this case, only 6 – 7 % of the 160 000 possible sequences complemented the chorismate mutase deficiency, and sequence analysis indicated an extremely strict requirement for hydrophobic aliphatic amino acids at position 68. Thus, in contrast to the weak influence of a residue involved in a secondary structure element, long-range tertiary interactions between residues in an interhelical turn and residues elsewhere in the protein can impose strong constraints on allowable sequence substitutions.

The EcCM study indicates the key advantages of genetic selection over screening methods. First, much larger libraries can be evaluated with little extra effort. Screening 10^5 individual clones in the absence of a convenient spectroscopic assay would be an arduous undertaking, but selection readily sorts the entire mixture into active, partially active, and inactive enzymes. Second, the greater stringency imposed by the requirement that selected variants be catalytically active allows nonobvious features to be discerned, such as the statistically meaningful preference for hydrophilic residues at solvent-exposed positions, or the identification of turns with altered backbone conformations. These finely tuned details of protein structure might be missed with less stringent assay criteria.

3.4. Altering Protein Topology

All proteins are composed of a limited set of secondary structure elements. Yet, we are just beginning to learn how such building blocks can be combined to yield well-defined tertiary and quaternary structures. Evolutionary strategies, which allow the simultaneous evaluation of huge numbers of alternatives on the basis of a selectable phenotype, such as catalytic activity, can greatly aid such efforts.

3.4.1. New Quaternary Structures

The re-engineering of existing protein scaffolds represents a manageable first step toward the larger goal of designing functional enzymes de novo. Alteration of the oligomerization state of various proteins, for example by converting monomers into multimers[86] or multimers into monomers,[87–92] has received particular attention. Such systems may shed light on the evolutionary origins of multimeric proteins.[93, 94]

Converting an oligomeric protein into a catalytically functional monomer is particularly difficult when the individual polypeptides must be untangled, as in the intricately entwined homodimeric EcCM.[33] Nevertheless, it seemed possible that the insertion of a flexible "hinge – loop" into the long H1 helix that spans the EcCM dimer would alter the enzyme's quaternary structure. The important role of hinge – loops in determining the oligomerization state of a variety of proteins has been noted previously.[93, 94] Although the active sites of EcCM are constructed from residues contributed by both polypeptide chains, such a segment might allow the N-terminal half of H1 to bend back on itself, thus displacing the second polypeptide, to yield a monomeric and catalytically active four-helix-bundle (Figure 7). Depending on the loop length and composition, other oligomeric states could also conceivably be populated.

To test these possibilities, random segments of four to seven residues were inserted into the middle of the EcCM H1 helix.[95] The individual libraries (designated L4, L5, L6, and L7) have a maximum theoretical diversity of 160 000 (20^4), ca. 3.2×10^6 (20^5), ca. 6.4×10^7 (20^6), and ca. 1.28×10^9 (20^7) distinct members, respectively. In each case, transformation of

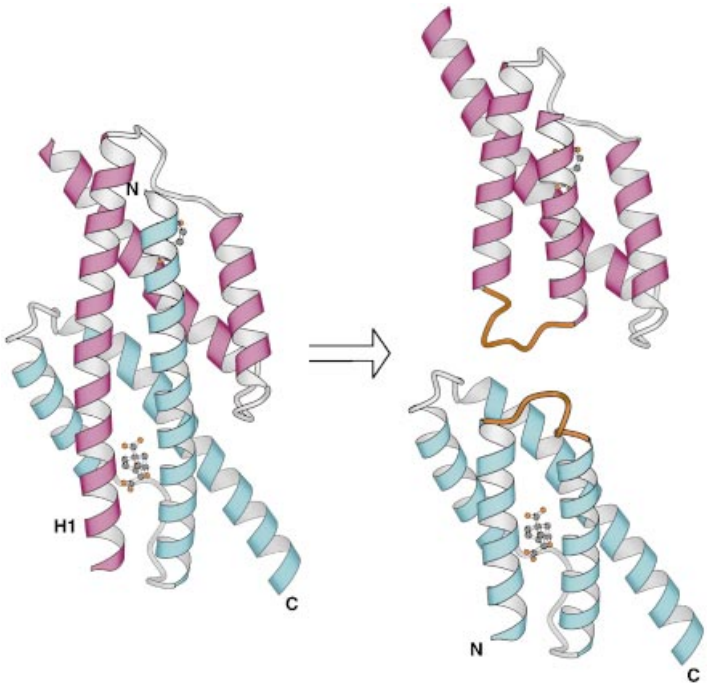
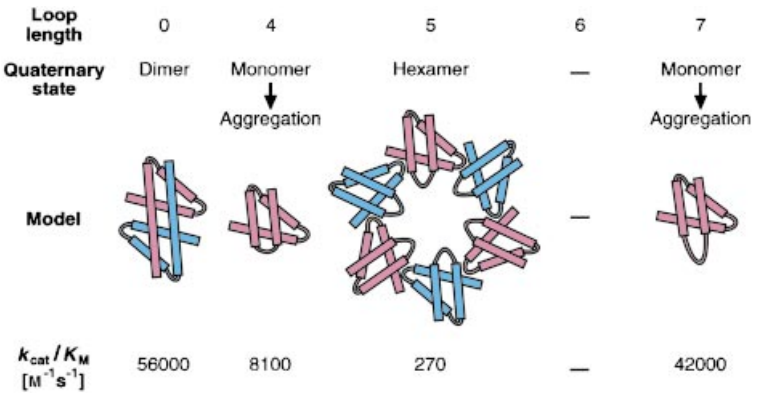


Figure 7. Topological conversion of dimeric EcCM into a monomer. Insertion of a flexible loop into the dimer-spanning H1 helix would allow the N-terminal portion of the helix to bend back on itself, thus displacing the other polypeptide to form a functional four-helix-bundle catalyst.

chorismate mutase deficient bacteria yielded roughly 10^7 clones, to give fully diverse and redundant coverage of the L4 and L5 libraries, about 10 % sequence coverage of the L6 library, and about 1 % coverage of the L7 library.

No functional clones were found in the L6 library under selection conditions, but 0.05 %, 0.002 %, and 0.5 % of the L4, L5, and L7 libraries, respectively, complemented the genetic defect. Several active enzymes from the libraries were characterized biochemically (Scheme 7). Insertion of 4 or 7 amino acid segments into the middle of the H1 helix yielded monomeric enzymes with near wild-type activity, but these proteins were unstable and tended to aggregate. In contrast, insertion of a five amino acid segment (Cys-Phe-Pro-Trp-Asp) yielded a well-behaved hexameric chorismate mutase (as judged by analytic ultracentrifugation) that is about 200-fold less active than the wild-type protein.



Scheme 7. Models for the structural organization of hinge–loop variants of EcCM showing the number of residues in the inserted hinge loop, the oligomeric state of active variants, and experimentally determined k_{cat}/K_m values.[95]

The properties of the hexameric species, when compared with those of the monomers, suggest that protein stability is an important driving force in the evolution of oligomeric proteins.[95] However, additional work is needed to elucidate how the subunits are organized and how quaternary structure influences function.

On a more general note, loop length appears to play a more important structural role than might have been anticipated from other experiments.[83] The low percentage of active proteins recovered in these selection experiments indicates that relatively few loop sequences permit a change in the quaternary structure without affecting the structure of the active site. Moreover, the scarcity of functional clones (1 out of every 50 000 L5 library members, for example) also means that it would have been extremely difficult to find active enzymes by screening with an in vitro chorismate mutase assay alone.

3.4.2. *Stable Monomeric Mutases*

Although the topological conversion of the EcCM homodimer to a monomer preserves the enzyme active site, it yields a much less stable protein. A cluster of hydrophobic residues between the two active sites appears to stabilize the dimer,[37] but these interactions would be lost in the monomer. Exposure of these buried apolar groups to solvent presumably favors aggregation as well. Since the highly charged binding pocket is located in the interior of the helical bundle,[33] the monomer itself lacks a conventional hydrophobic core and, hence, the usual driving force for protein folding.[96]

A thermostable dimer was therefore considered as an alternative starting point for the design of a stable monomeric mutase. A large number of EcCM sequence homologues, some from thermophilic organisms, are known. For example, the hyperthermophilic archaeon *Methanococcus jannaschii* produces a chorismate mutase (MjCM) that is 25 K more stable than EcCM.[37] Despite a sequence identity of only 21 %, six prominent residues that line the active site are strictly conserved and the two enzymes have comparable activities. Since the hydrophobic core of MjCM is very similar to that of EcCM, interactions distant from the dimer interface must be responsible for its additional stability. These same interactions were expected to stabilize the desired monomer.

Topological variants of MjCM were prepared by inserting a flexible loop into the middle of the H1 helix,[97] as described in the previous section for EcCM (see Figure 7). In this case, though, the turn segment was modeled on a known helix–turn–helix motif in *E. coli* seryl-tRNA synthetase.[98] Two residues in helix H1 (Leu 22 and Lys 23, numbered according to the homologous EcCM sequence) were duplicated and six residues of random sequence were inserted between the repeated amino acids. Two point mutations (Leu 22a to Glu and Ile 79 to Arg) were also included as an element of negative design to disfavor dimer formation and to minimize aggre-

gation of the monomers by reducing the exposed hydrophobic surface area. This design yields a library with approximately 6.4×10^7 (20^6) individual members. The library was constructed at the genetic level, used to transform the chorismate mutase deficient *E. coli* strain ($>10^8$ transformants), and evaluated by selection.

In contrast to the selection experiments described above which were carried out on solid media, the library was grown in liquid culture lacking tyrosine and phenylalanine. This process allows a direct competition for resources and amplification of the small fraction of the original library containing a functional chorismate mutase (0.7%). After three days of selection, more than 80% of the clones in the library complemented the genetic defect, which corresponds to over a 100-fold enrichment.

Selection greatly facilitates the search for a tractable monomer by eliminating inactive or weakly active clones. However, function, not topology, is the basis for selection. For this reason, representative clones were subsequently screened by size-exclusion chromatography. Only one of the 26 clones analyzed in this way had a retention time expected for a monomer (Figure 8). This protein had the six-residue insert

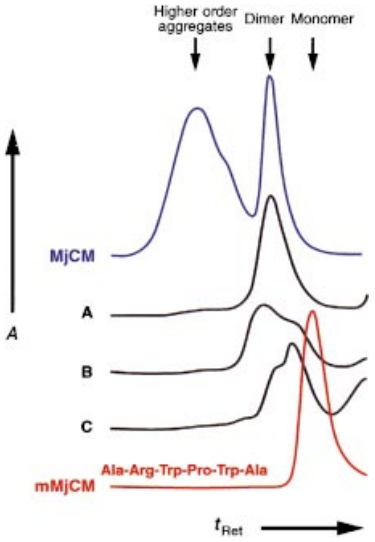


Figure 8. Analytical size-exclusion chromatography distinguishes between dimeric and monomeric chorismate mutases.[97] Wild-type MjCM (top trace) was found to be a mixture of dimer and higher order aggregates. Most selected variants (for example, traces A – C) were dimers or mixtures of dimers and monomers. Only one of 26 variants tested (mMjCM, bottom trace) eluted as a monomer. It had the six amino acid insert shown. Analytical ultracentrifugation confirmed that this protein is monomeric in solution.

Ala-Arg-Trp-Pro-Trp-Ala. Its monomeric nature was confirmed by analytical ultracentrifugation. It is also highly helical, undergoes cooperative thermal and chemical denaturation ($\Delta G_U(H_2O) = 2.7$ kcal mol^{-1}), and has significant catalytic activity. Its k_{cat} value for the rearrangement of chorismate is identical to that of MjCM, while its K_m value is elevated only threefold.[97]

The combination of selection (>100-fold enrichment) and screening (1 in 26) shows that fewer than 0.05% of the possible turn sequences are capable of yielding well-behaved, monomeric proteins. This result again contradicts the simple expectation that most interhelical turn sequences are functionally equivalent.[76] It also underscores in dramatic fashion the advantage of the selection approach. Individual screening of 1000 – 10000 proteins to find the one with the desired activity and topology would represent a major experimental undertaking.

Another important advantage of selection is exemplified by the observation that cells harboring the monomer grow much more effectively in the absence of tyrosine and phenylalanine than cells containing the wild-type MjCM dimer.[97] Although the precise reasons for this difference are unknown, it may reflect the fact that the MjCM dimer forms catalytically inactive aggregates when overproduced in *E. coli*, whereas the monomer shows no tendency to do so (Figure 8). Not only does the process of genetic selection maximize catalytic activity, but problems associated with protein misfolding, aggregation, or toxicity are simultaneously minimized. This feature may be of value in the production of biocatalysts for commercially practical applications, where stable, easily produced enzymes are essential.

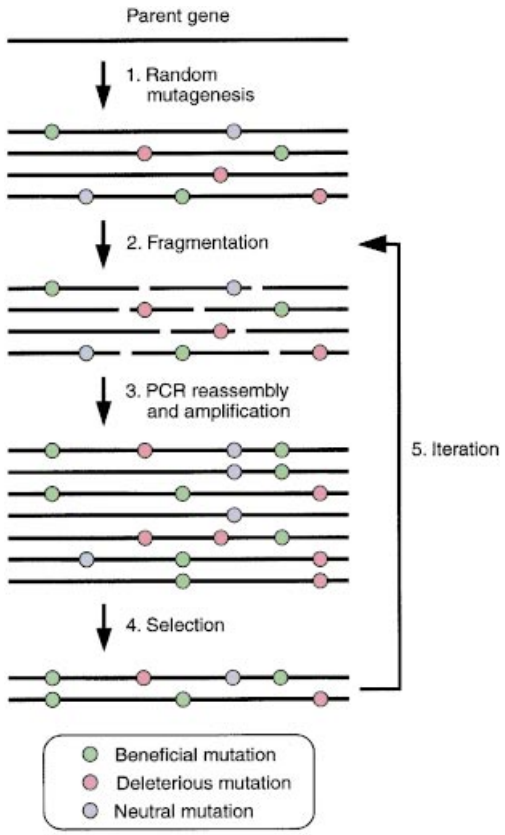
3.5. Augmenting Weak Enzyme Activity

True Darwinian evolution involves multiple cycles of mutation and selection. This process can be mimicked in a laboratory setting to optimize the properties of an inefficient enzyme. Such an endeavor is often termed directed evolution.[10]

The hexameric but weakly active chorismate mutase[95] described in Section 3.4.1 has been improved in this way.[99] Mutations were introduced into the gene encoding the hexamer subunit by DNA shuffling (Scheme 8),[5, 100] which mimics sexual recombination in vitro. Improved variants were selected, as before, by their ability to complement the chorismate mutase deficiency in bacteria. Plasmid DNA was isolated from the fastest growing cells and the entire procedure was repeated.

After two rounds of mutation and selection, several clones were found that grew at wild-type levels. The corresponding enzymes were isolated and characterized biochemically.[99] One variant, which contained three mutations (Ser15Asp/Leu79Phe/Thr87Ile), had a 9-fold improvement in the k_{cat} value and a 35-fold improvement in the k_{cat}/K_m value. Size-exclusion chromatography also showed that it is still considerably larger than the wild-type dimer, but more detailed biophysical studies will be needed to determine its precise oligomerization state and subunit organization. This catalyst was subjected to a further round of mutagenesis and selection, and additional clones with somewhat improved k_{cat}/K_m values have been identified. Interestingly, many of these are missing the C-terminal histidine tag, which had been introduced to facilitate purification, plus an additional five to eight amino acids. Eliminating the C-terminal extension of the H3 helix could conceivably improve packing interactions between the subunits in the higher order oligomer.

In principle, directed evolution procedures could be repeated indefinitely until any desired activity has been attained. At some point, however, the catalyst will be



Scheme 8. DNA shuffling.[5, 100] Techniques such as error-prone PCR can be used to mutagenize a parent gene. The resulting pool of homologous genes contains beneficial, deleterious, or neutral point mutations. Treatment of this pool with DNaseI gives rise to a set of random fragments. Novel recombinations accompany the reassembly of the random fragments into full-length genes. Selection for variants with improved function provides the starting point for another round of shuffling. Point mutations arise during the recombination process itself, which makes every round a genuine Darwinian evolution cycle.

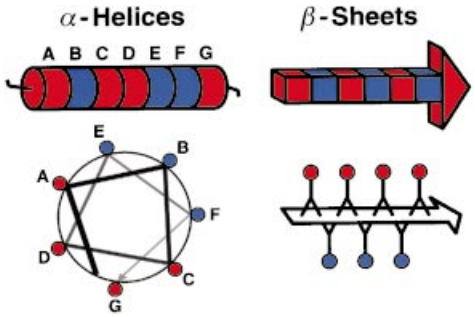
sufficiently active that the host cell grows like the wild-type strain, thus making selection for further improvement difficult. This is true for the modified hexamer even though it is still an order of magnitude less efficient than the homodimeric MjCM.[37] Since total activity depends on the catalyst concentration as well as specific activity, reducing the available catalyst concentration can further increase selection pressure. In practice, intracellular protein concentrations can be lowered in a variety of ways, for example, by the use of low copy plasmids,[101] weak promoters[102], and inefficient ribosome binding sites.[103]

3.6. Protein Design

The de novo design of *functional* proteins represents a far more ambitious goal than topological redesign of an existing enzyme. A number of small protein scaffolds have been successfully designed from first principles,[104–110] often with the aid of computer algorithms.[111, 112] Nevertheless, conferring function to these molecules—the ability to recognize another molecule with high selectivity or to catalyze a chemical reaction—remains an enormous challenge.

In principle, the same selection methods that have been successfully used to characterize the mechanistic and structural determinants of enzymes and to alter their topology should aid the design of new protein catalysts. For example, one could imagine directly selecting active catalysts from large random protein libraries.[113] However, sequence space is infinitely vast. A library with the mass of the earth (5.98×10^{27} g) would contain at most 3.3×10^{47} different sequences. This number represents but a miniscule fraction of the total diversity available to even a small protein, yet far exceeds what is accessible to experimental investigation. Unless catalysts are present in ensembles of random variants at an unexpectedly high frequency, even the most sensitive and efficient selection procedure will be fruitless.

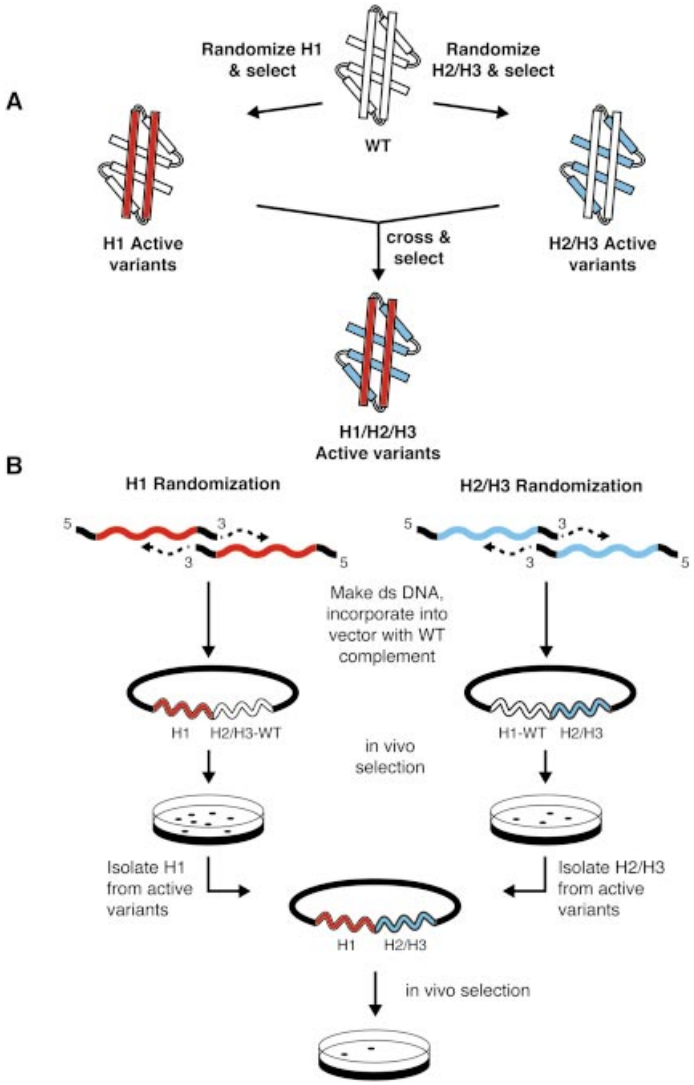
An attainable compromise between these two approaches might utilize basic structural information, such as the sequence preferences of helices and sheets or the tendency of hydrophobic residues to be buried in the protein interior, to design focused random libraries from which functional catalysts can be selected. Binary patterning of polar and nonpolar amino acids,[75, 114, 115] for example, is a potentially general strategy for protein design that exploits the nonrandom distribution of amino acids in folded structures (Scheme 9). It has been successfully exploited to create compact helical bundles that share many of the properties of native proteins, such as protease resistance and cooperative folding.[116–118]



Scheme 9. Amphiphilic secondary structures. Amphiphilic helical or β-sheet structures can be designed by specifying the locations of polar (red) and nonpolar (blue) residues through simple binary patterns.[227]

We have systematically replaced the secondary structural elements in a homodimeric chorismate mutase (Scheme 10) with binary-patterned units of random sequence to explore the feasibility of such an approach for the design of active catalysts. Genetic selection was then used to assess the catalytic capabilities of the proteins in the resulting libraries, to provide quantitative information about the robustness of this particular protein scaffold and insight into the subtle interactions needed to form a functional active site.[119]

In the first stage of our experiment modules corresponding to the H1 and H2/H3 helices were combined individually with the appropriate complementary wild-type helical segments of the thermostable MjCM′, a protease-resistant version of the dimeric *M. jannaschii* chorismate mutase.[37] This amounted to randomizing 37 % and 42 % of the entire protein for the H1 and H2/H3 libraries, respectively. At each randomized



Scheme 10. Generation of large binary-patterned libraries of AroQ chorismate mutase genes. A) A two-stage strategy was adopted involving separate randomization and selection of functional variants of the H1 (red) and H2/H3 (blue) helices, followed by combination of functional binary-patterned segments from the initial libraries and reselection. In the final constructs, approximately 80% of the protein was randomized. B) The binary-patterned genes were constructed from synthetic random oligonucleotides corresponding to the segments encoding the H1 or H2/H3 helices (red or blue, respectively) and assembled in vector fragments (black) as shown. In the initial step, the randomized segments were combined with their wild-type counterparts H2/H3-WT and H1-WT (white). Selection was carried out using the system shown in Scheme 2. WT, wild-type MjCM′; ds = double stranded.

position, only four polar (asparagine, aspartate, glutamate, and lysine) or four nonpolar (isoleucine, leucine, methionine, and phenylalanine) residues were permitted. The loops and conserved catalytic residues were kept constant in the initial design.

The chorismate mutase deficient bacterial strain was transformed with the binary-patterned libraries and the ability of the partially randomized proteins to fold into ordered, catalytically active structures was assessed by complementation of the chorismate mutase deficiency. Roughly 0.02% of the H1 library clones were active. In other words, about 1 out of 4500 binary-patterned enzymes is a viable catalyst. A somewhat smaller fraction of the H2/H3 library, about 0.006% (1 out of 17500 variants), was found to complement the genetic deficiency. Given the apparent abundance of catalysts, it would have been possible to find active enzymes even if the three active site residues in each of the two libraries (Arg11, Arg28, and Lys39 in H1, and Arg51, Glu52, and Gln88 in H2/H3) had been additionally mutagenized. Assuming that only a single amino acid is tolerated at each active site position, the probability of finding a catalyst would only have decreased by a factor of 20^3. Thus, active chorismate mutases would still have been isolable from experimentally accessible libraries containing as few as 10^8 members ($20^3 \times 10^4$). Binary patterning is clearly useful for catalyst design.

Catalysts from both the H1 and H2/H3 libraries were isolated and characterized.[119] Despite the restricted set of building blocks used in the randomized segments, the enzymes are remarkably active, with modestly elevated K_m values and k_{cat} values in the range 0.2 to 2.3 s^{-1}, which compare favorably with the value of 5.7 s^{-1} for wild-type MjCM.[37] While most positions in the protein are relatively tolerant to substitution, sequence analysis has identified a few sites as quite restrictive. For instance, all active clones have a strictly conserved Asn–Lys dyad at positions 84 and 85; such a preference is not evident in unselected clones. In retrospect, this finding can be rationalized by examination of the EcCM structure.[33] The homologue of Asn84 contributes a hydrogen bond to one of the carboxylate groups of the bound ligand, and the methylene groups of the adjacent lysine residue probably make van der Waals interactions with the ligand. Apparently, none of the other amino acids allowed at these positions would have been able to make such interactions.

The preselected H1 and H2/H3 modules were combinatorially crossed in much the same way as immunoglobulin heavy and light chains are combined in the immune system to create diverse antibody structures.[120] In the resulting H1/H2/H3 binary-patterned libraries, about 80% of the MjCM′ residues were randomized and >90% of the protein was constructed from only eight amino acids (Scheme 10A). Interestingly, only about 1 out of every 10000 possible combinations yields a functional catalyst, even though each of the preselected H1 and H2/H3 components is individually functional in a wild-type context. The native MjCM′ H1 and H2/H3 helices are clearly much more effective than the selected segments in templating correctly folded, active enzymes, presumably because they are better able to tolerate destabilizing changes in the varied portions of the protein.

As observed for the individual H1 and H2/H3 libraries, certain positions are highly conserved in active H1/H2/H3 variants. In addition to the Asn84/Lys85 dyad seen previously, the majority of complementing clones contain the residues Ile14, Asp15, and Asp18. These amino acids, which are not as highly conserved in the H1 library, probably help stabilize the active site. The residue at position 14 in EcCM interacts directly with a bound ligand through van der Waals contacts, whereas residues 15 and 18 are involved in second-sphere interactions with the catalytically important residues Arg51 and Arg28.

Biochemical characterization of one of the selected H1/H2/H3 variants confirms that it is a helical homodimer that undergoes cooperative thermal denaturation ($T_m \approx 44\,^{\circ}C$). It

catalyzes the chorismate to prephenate rearrangement with a k_{cat} value only 15-fold lower than that of the *M. jannaschii* enzyme and a K_m value that is 40-fold higher.[119] By combining binary patterning with two evolutionary cycles, we have thus created a novel helix-bundle protein from a simplified set of building blocks that has biophysical and functional properties similar to those of natural chorismate mutases. Optimization of this and related catalysts from the library through additional rounds of mutagenesis and selection will be useful for identifying key interactions that influence folding, function, and stability.

Experiments of this kind can provide valuable insights into the chemical determinants of protein structure. The results support speculations that ancient proteins might have been constructed from small numbers of amino acids.[121–123] Nevertheless, the relatively low abundance of active enzymes in the combinatorial libraries underscores the difficulty of generating helical scaffolds capable of catalysis—even when both the position and identity of all critical active site residues are specified in advance.

We estimate[119] that finding a catalyst among binary-patterned but fully randomized AroQ templates would have required a library containing at least 10^{24} members. Experience suggests that other folds will be substantially more difficult to engineer.[124] Since libraries of this size are experimentally inaccessible, incremental approaches will undoubtedly be needed for the de novo design of tailored enzymes. Initially, a stable scaffold might be developed,[228] binding and catalytic groups subsequently introduced, and the entire ensemble finally optimized for the desired activity. It is clear that having an efficient selection system will be invaluable, if not absolutely essential, for the success of such endeavors.

4. Other Applications of in vivo Selection

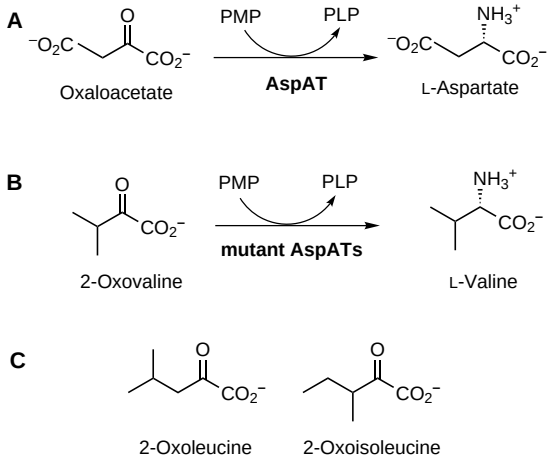
The principles of genetic selection, illustrated above with the chorismate mutase system, are broadly applicable to the study of protein structure and function. A few additional examples are presented below to highlight applications not considered in the previous sections.

4.1. Altering Selectivity

As a consequence of their exceptional efficiency, high selectivity, and mild reaction conditions, enzymes are valuable tools in organic synthesis, particularly for transformations that cannot be carried out with more conventional methods or those involving structurally complex and biologically important substances such as carbohydrates, nucleic acids, and proteins.[125] For chemists seeking general synthetic methods, however, the narrow substrate specificity of many natural enzymes may represent a significant limitation. For this reason, considerable effort has been devoted to the rational redesign of enzyme binding sites, but, with a few notable exceptions,[126–128] success has been meager.

Randomization plus screening or selection offers an attractive alternative to rational redesign as a means of altering substrate specificity. Conversion of a homodimeric aspartate aminotransferase (AspAT) into an efficient valine aminotransferase furnishes a striking example of what can be achieved by in vivo selection.[129]

Wild-type AspAT catalyzes the interconversion of oxaloacetate and aspartate (Scheme 11), but has barely detectable activity for hydrophobic β-branched amino acids. The parent



Scheme 11. A) Aspartate aminotransferase (AspAT) catalyzes the transamination of oxaloacetate to L-aspartate, concomitant with conversion of the cofactor pyridoxamine phosphate (PMP) into pyridoxal phosphate (PLP). The reaction is readily reversible. B) Mutagenesis and selection in an auxotrophic *E. coli* strain yielded AspAT variants that also promote the interconversion of β-branched 2-oxovaline and L-valine.[129] C) 2-Oxoleucine and 2-oxoisoleucine are also substrates for the mutant AspATs.

gene was mutagenized by DNA shuffling (see Scheme 8) to alter its substrate preference, and variants were selected for their ability to complement an auxotrophic *E. coli* strain lacking the gene for the branched-chain amino acid aminotransferase. This enzyme is required for the synthesis of valine, leucine, and isoleucine. Consequently, the mutant strain can only grow in the absence of valine if the activity of AspAT for the interconversion of 2-oxovaline to valine increases.

Five rounds of mutagenesis and selection were performed. Each round yielded 10^6 to 10^7 transformants, and the plasmid DNA of 90 – 100 of the fastest growing clones was used for the next round of shuffling. The stringency of selection for mutant AspATs able to utilize 2-oxovaline as a substrate was increased at each round: first by shortening the incubation time on selective medium, then by omitting 2-oxovaline from the growth media (thus lowering its intracellular concentration), and finally by lowering the expression level of the AspAT mutant genes.

One of the mutants that emerged from this process showed a 10^5-fold increase in the k_{cat}/K_m value for 2-oxovaline and a 30-fold decrease in the k_{cat}/K_m value for acidic 2-oxo acids relative to the wild-type enzyme (Scheme 11 B). It contained 13 amino acid substitutions.[129] The use of an AspAT triple mutant as the starting point for another evolutionary experiment yielded another, closely related variant with 17 amino

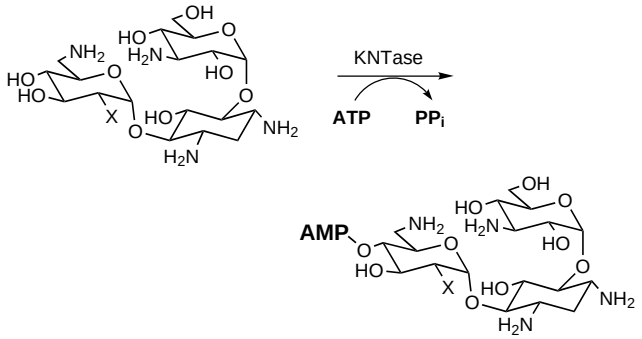
acid substitutions.[130] Its efficiency with β-branched 2-oxo acids was increased 2×10^6-fold. Crystallographic analysis of the latter protein complexed with isovalerate showed that only one of the 17 mutated residues contacts the substrate, and several are more than 10 Å from the active site. The resulting remodeling of the active site and subunit interface, which is ultimately responsible for the change in specificity, is thus a consequence of the subtle and cumulative effects of numerous substitutions distant from the binding pocket. Similar observations have been made in natural systems, for example, during maturation of the immune response to small haptens.[131–133] The difficulty of predicting such second-sphere modifications a priori makes the utility of directed evolution clear.

Successful alteration of the substrate specificities of β-lactamase,[100, 134–139] thymidine kinase,[140] prenyltransferases,[141] oxidosqualene cyclizing enzymes,[142] *N*′-[(5′-phosphoribosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide isomerase (ProFARI),[143] and aminoacyl-tRNA synthetases,[144, 145] to cite only a few recent examples, speaks to the versatility of the selection approach.

4.2. Increasing Thermostability

The moderate stability of enzymes isolated from mesophilic organisms may represent another limitation for many practical applications. For this reason, there has been considerable interest in isolating catalysts from organisms that grow under extreme conditions.[146] Evolutionary approaches also lend themselves to the creation of proteins that tolerate a wider range of experimental operating conditions.

One strategy for creating thermostable enzymes entails introducing the gene for an enzyme from a mesophilic organism into a thermophile and selecting for variants that retain activity at higher growth temperatures.[147] This approach has been used to identify mutant kanamycin nucleotidyltransferases (Scheme 12).[148, 149]



Scheme 12. Kanamycin nucleotidyltransferase (KNTase) transfers a nucleoside monophosphate group from ATP to the 4′-hydroxyl group of the antibiotics kanamycin A (X = OH) and B (X = NH_2). This reaction provides bacterial resistance to aminoglycoside drugs.

Cells of the moderate thermophile *Bacillus stearothermophilus* containing the mesophilic wild-type enzyme are resistant to the antibiotic kanamycin at 47 °C, but not at temperatures above 55 °C. Variants that are kanamycin resistant above 60 °C were obtained following mutagenesis and selection in the thermophile. A single mutation, Asp 80 Tyr, was found to be responsible for the increased resistance of the enzyme to irreversible thermal inactivation.[148, 149] Additional selection experiments at 70 °C identified Thr 130 Lys as a second stabilizing mutation.[148] The Asp 80 Tyr/Thr 130 Lys double mutant has a half-life of more than one hour at 60 °C, whereas the wild-type enzyme is denatured almost immediately under these conditions. The double mutant is also substantially more resistant to proteolytic degradation. An even more stable enzyme ($T_m = 84$ °C), which contains 19 amino acid substitutions, has been obtained by multiple rounds of DNA shuffling and selection in the extreme thermophile *Thermus thermophilus*.[150] Interestingly, the beneficial mutations did not include the amino acids identified in the earlier studies. Moreover, most of the affected residues are located on the surface of the protein, and 5 of the 19 changes involved incorporation of proline.

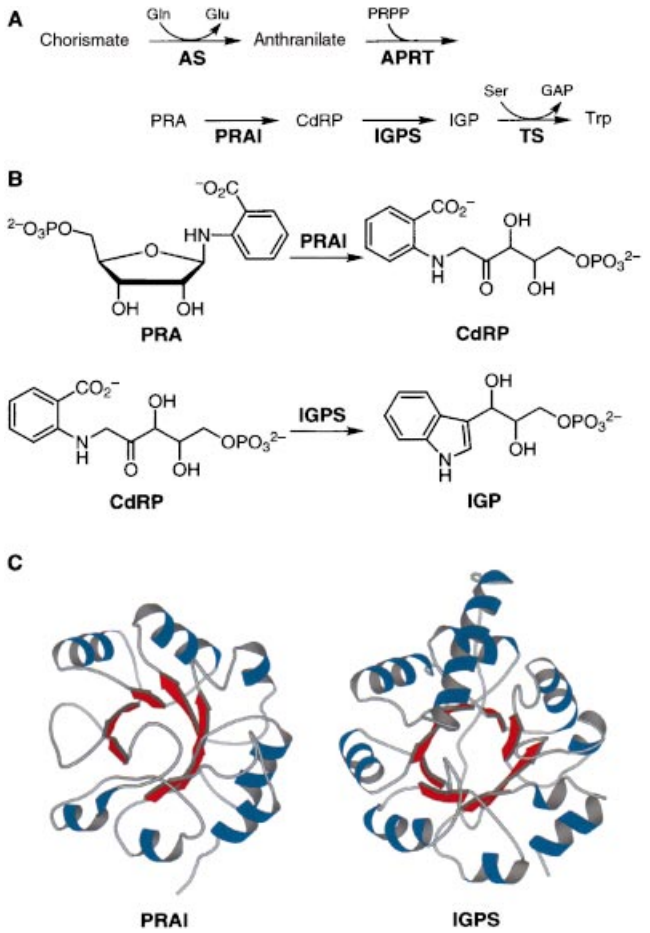
Any enzymatic activity that is selectable in thermophiles should be amenable to this approach, although in vitro strategies using various display formats (see Section 5) may ultimately prove more flexible with respect to stabilization against a wider variety of harsh conditions. In addition to the practical benefits that accrue, such experiments can greatly enhance our understanding of the interactions that contribute to protein stability.[151]

4.3. Altering Function

A powerful strategy for conferring new functions to existing proteins is to mimic the process of divergent evolution. Nature exploits a limited number of protein folds[152] and puts them to many and diverse uses. The generic immunoglobulin structure[153] is a good example since it is exploited, with minor variation, to recognize the universe of foreign antigens. In the same way that chemists have exploited antibody molecules (and the microevolutionary processes inherent to the mammalian immune system) to create a wide range of tailored catalysts,[154–156] many existing protein scaffolds might serve as starting points for the development of new active sites capable of diverse functions. The α/β-barrel fold[157, 158] that constitutes roughly 10 % of all structurally characterized proteins may be particularly versatile in this respect.[161]

The conversion of indole-3-glycerol-phosphate synthase (IGPS) into a highly active phosphoribosylanthranilate isomerase (PRAI) demonstrates the feasibility of this strategy.[159] PRAI and IGPS catalyze sequential steps in the biosynthesis of tryptophan (Scheme 13 A). The first promotes the conversion of *N*-5′-phosphoribosylanthranilate (PRA) to 1′-(2′-carboxyphenylamino)-1′-deoxyribulose 5′-phosphate (CdRP), and the second accelerates the rearrangement of CdRP to indole-3-glycerol phosphate (IGP) with concomitant release of CO_2 and H_2O (Scheme 13 B). Both enzymes are α/β-barrels (Scheme 13 C) that are covalently linked in *E. coli*. Both also bind CdRP.

The strategy to convert IGPS into PRAI involved rational redesign coupled with random mutagenesis and genetic selection. Guided by the crystal structures of the two enzymes,



Scheme 13. A) Biosynthesis of tryptophan from chorismate. AS = Anthranilate synthase; PRPP = phosphoribosylpyrophosphate; APRT = anthranilate phosphoribosyl transferase; PRA = *N*-5′-phosphoribosylanthranilate; PRAI = phosphoribosylanthranilate isomerase; CdRP = 1′-(2′-carboxyphenylamino)-1′-deoxyribulose 5′-phosphate; IGPS = indole-3-glycerol phosphate synthase; IGP = indole-3-glycerol phosphate; TS = tryptophan synthase; GAP = glyceraldehyde 3-phosphate. B) Directed evolution was used to create a novel PRAI, which catalyzes the conversion of PRA into CdRP, from IGPS, which catalyzes the cyclization of CdRP to IGP.[159] C) Both PRAI and IGPS are α/β barrel proteins but have somewhat different loop structures.

the IGPS scaffold was initially redesigned to more closely resemble PRAI. First, a long N-terminal segment, absent in PRAI, was removed. In a second step, a 15 residue loop was replaced with a 4–7 residue loop (PRAI has a 4 residue loop). The variation in loop length and sequence provided a diverse set of variants from which active catalysts could be chosen in later selection steps. Finally, a second loop region was replaced with the PRAI consensus sequence GXGGXGQ, and an aspartate residue was introduced at position 184 to serve as a general base in the remodeled active site.

Active variants were isolated from the resulting library of IGPS mutants by genetic selection in a PRAI-deficient *E. coli* strain that cannot produce its own tryptophan. Approximately 500 out of 30 000 library members were found to grow at low tryptophan concentrations. The genes from the selected clones were then subjected to additional rounds of mutagenesis and selection until several variants were found that could grow in the absence of tryptophan. One of these (ivePRAI) was characterized in detail. The encoded protein was soluble and properly folded. Moreover, it exhibited significant PRAI activity in vitro but no IGPS activity, despite having a sequence identity of only 28% to PRAI and a sequence identity of 90% to IGPS. In fact, the specificity constant (k_{cat}/K_m) for ivePRAI was sixfold higher than for wild-type PRAI, because of an enhanced affinity for the PRA substrate.[159]

The successful interconversion of IGPS and PRAI in the laboratory supports the idea that tryptophan biosynthetic enzymes arose naturally by a process of divergent evolution.[160] In particular, this experiment shows how new catalytic activities can evolve quite rapidly from an existing binding site. Although extension of this approach to reactions of substrates that have no affinity for the template protein may prove more difficult, the promising results with IGPS suggest that the commonly occurring α/β-barrel motif may indeed be a generally useful scaffold for designing new enzymes.[161]

4.4. Optimizing Multistep Processes

As in natural Darwinian evolution, multiple components of a biosynthetic pathway—even whole organisms—can be optimized simultaneously in evolution experiments in the laboratory. A usual prerequisite for a genetic selection system is the construction of an appropriate selection strain through the mutation or elimination of chromosomal genes. However, another approach is to use an existing organism and apply selective conditions that require foreign genes. Optimization of the efficiency of an arsenate detoxification pathway,[162] which has potential value as a tool for bioremediation, is an interesting example of this strategy.

The arsenate resistance operon from the bacterium *Staphylococcus aureus* consists of three genes, *arsR*, *arsB*, and *arsC*, which afford resistance to arsenate and arsenite ions. The entire operon was mutagenized by DNA shuffling (see Scheme 8) and introduced into a standard *E. coli* strain. Arsenate-resistant bacteria were then selected from the population by plating on media containing increasing concentrations of arsenate (16–400 mM). After three rounds of mutagenesis and selection, fast growing clones were isolated. One of these was even able to grow in liquid media containing 500 mM arsenate.

Analysis of the evolved operon region revealed a total of 13 mutations. Ten (seven of which were silent) occurred in *arsB*, which encodes a membrane-associated pump that actively exports arsenite from the cell; these mutations collectively yielded a four- to sixfold increase in arsenite resistance, possibly by increasing the production levels while simultaneously reducing the toxicity of the ArsB protein. Two silent mutations occurred in *arsR*, which encodes the operon repressor protein, and one mutation occurred immediately downstream of the operon. Interestingly, the arsenate reductase gene *arsC* contained no mutations at all. Nevertheless, the overall rate of arsenate reduction was improved 12-fold,

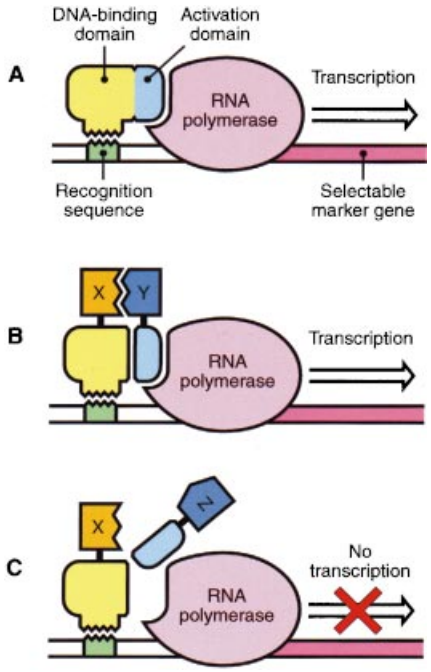
which was explained by an increased level of gene expression as a result of the mutations in the upstream gene *arsB*.[162]

These experiments underscore again the fact that many changes other than those leading to higher kinetic efficiency of a particular enzyme can help an organism survive under selection conditions. It is irrelevant to the selection process whether enhanced survival results from greater catalytic ability, more efficient gene expression, or increased protein stability. In fact, fine-tuning noncatalytic properties is likely to be far easier, in general, than optimizing catalytic activity. The corresponding mutations are presumably more frequent and more easily tolerated in almost all systems.

4.5. Finding Ligands and Inhibitors

Although enzyme structure and function is the primary focus of this review, it is worth mentioning that genetic selection has also emerged as a powerful tool for identifying novel peptide ligands for biological receptors and for characterizing ligand–receptor interactions in vivo. Yeast two-hybrid systems[163–167] are particularly useful in this regard.

The yeast two-hybrid system is depicted schematically in Scheme 14. Gene transcription in yeast relies on the interaction of RNA polymerase and an activator domain of a transcriptional activator. The activator domain is positioned in the DNA promotor region through the activator's DNA binding domain. In the yeast two-hybrid system, the covalent interaction between the two domains of the transcriptional activator is disrupted. Individual libraries of hybrids are formed by linking one or both of the separated domains to ensembles of other molecules (DNA, RNA, proteins, and small compounds are all conceivable candidates). If any two of the linked molecules interact, then the activator and DNA-binding domains will come together and transcription can occur from the promotor proximal to the DNA binding site. This transcription leads to the expression of a reporter gene and, hence, to a selectable or screenable phenotype.



Scheme 14. Yeast two-hybrid systems.[163–167] A) Gene expression in yeast involves interaction of RNA polymerase with a DNA-binding transcriptional activator. B) Two different proteins or protein libraries are fused to the DNA-binding and activator domains of a yeast transcriptional activator, respectively. An interaction between two cognate proteins (X and Y) reconstitutes the activator, which leads to transcription of the reporter gene. C) Noncognate library members cannot assemble a functional activator.

Two-hybrid screening is most common,[164–166] where the reporter gene product provides a nonessential function to the cell that is easy to detect. However, two-hybrid genetic selection systems that utilize essential reporter genes are also used. For example, selection based on transcriptional activation has been used to identify peptides that bind to cyclin-dependent kinase 2 (Cdk2) in vivo.[168] The peptides were displayed as a 20-residue insert in the active-site loop of *E. coli* thioredoxin; the best variants had nanomolar affinity for their target. A yeast three-hybrid system for detecting ligand–receptor interactions has also been described.[169] The relatively low transformation efficiencies of yeast limit the size of the libraries that can be exhaustively studied, but *E. coli*-based two-hybrid systems have been developed[170, 171] that could extend the capabilities of such systems.

Genetic selection in yeast has also been used directly to identify peptides that inhibit biological pathways, without recourse to two-hybrid formats.[172] For example, cleverly engineered yeast strains have been used to select inhibitors of the spindle checkpoint and mating pheromone signaling pathways from libraries of peptides displayed on a surface loop of an inactive variant of staphylococcal nuclease. The putative intracellular targets were subsequently identified by a combination of two-hybrid analysis and genetic dissection of the target pathways.[172] A similar genetic selection approach, albeit in *E. coli*, was recently used to identify nine-residue peptides that inhibit intracellular dimerization of HIV-1 protease.[173]

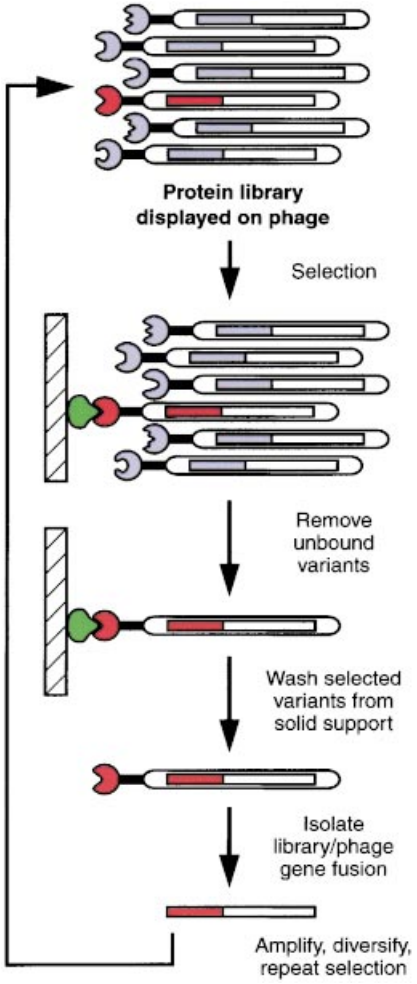
Such studies are advantageous to drug discovery since they provide novel leads and a wealth of information about structure–function relationships in the biological receptor. This information can help guide the design of potent non-peptidic inhibitors. Characterization of the site of action of the selected peptides can also reveal new components of the investigated pathway which may, in turn, represent formidable targets for inhibition by small organic molecules.

5. In vitro Selection Systems

The examples cited in this review show that a wide range of different protein functions can be linked to a genetically selectable phenotype. Nonetheless, establishing such a connection can be nontrivial for many interesting applications and may be impossible in some cases (for example, if nonphysiological conditions are required). In such instances, in vitro selection can often be a practical alternative. Selection in vitro, as in vivo, necessitates linking genetic information with protein function in some fashion. Some of the ways this task can be realized are outlined below.

5.1. Phage and Cell-Surface Display

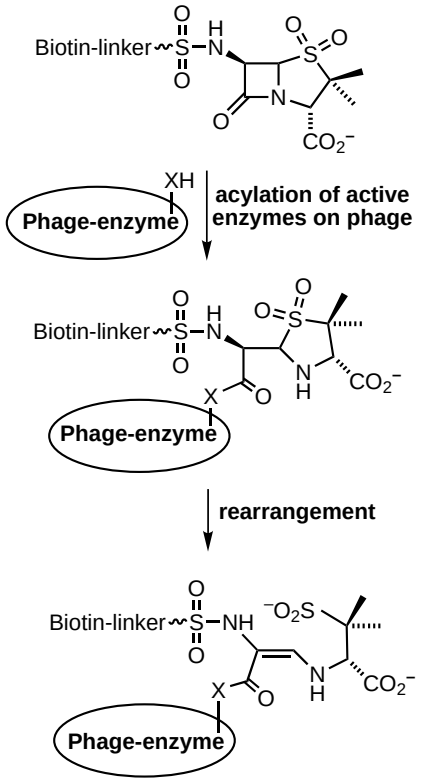
Display of proteins (and peptides) on the surface of cells[174, 175] or bacteriophage[176, 177] is a particularly effective method for coupling phenotype and genotype. The repertoire of proteins to be sampled is fused to a membrane protein or to one of the outer capsid proteins of filamentous phage (Scheme 15). By exposing the resulting library to a ligand that is immobilized (or immobilizable) on a solid support, specific receptors can be isolated by affinity selection in vitro. The gene encoding the functional polypeptide is selected simultaneously since it is housed within the cell or bacteriophage bearing the receptor. It can be recovered from the solid support and amplified in vivo. It can also be diversified and subjected to additional rounds of selection to fine-tune the properties of the encoded protein.



Scheme 15. Selection by phage display.[176, 177] Fusion of the gene encoding an antibody or another protein to the gene for one of the coat proteins of the bacteriophage allows the display of the protein and its variants on the surface of the virus particle. Affinity selection with an immobilized ligand (green) can be used to isolate proteins with desirable binding properties from large populations and, concomitantly, their topologically linked encoding genes (red).

Display technologies have been broadly used to identify peptide ligands for diverse receptors and to select proteins with novel binding activities. Optimization of DNA-binding proteins for the specific recognition of any desired DNA sequence,[178–185] minimization of the immunoglobulin Z-domain of protein A,[186] and evolution of antibody fragments with femtomolar antigen-binding affinity[187] are notable in this regard. The selection of proteins directly on the basis of their catalytic activity has proved more difficult to achieve, but several promising strategies have been reported.

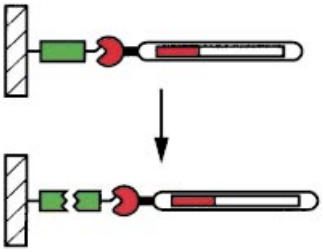
In one approach, reactive ligands[188] and mechanism-based inhibitors[189, 190] have been used to enrich enzymes and catalytic antibodies from phage display libraries. Phage particles with appropriate chemical functionality react covalently with the agent, either directly or after activating it chemically. As a consequence of this chemical event, active phage are captured and immobilized on a solid support (Scheme 16). The same protein group(s) responsible for activating the inhibitor can be subsequently exploited for the catalytic conversion of a substrate. This approach can be useful for the discovery of new catalysts,[189, 190] but it is necessarily indirect. Selection requires irreversible modification of the receptor—a single step—rather than efficient substrate turnover, which may require stabilization of multiple transition states and avoidance of product inhibition. The need to design and synthesize appropriately reactive derivatives and inefficient trapping of the reactive intermediate



Scheme 16. Trapping phage enzymes with mechanism-based inhibitors. Penicillanic acid sulfones connected through a spacer to biotin can be used to trap enzymes displayed on filamentous phage that hydrolyze β-lactams through a nucleophilic mechanism.[189] Enzymes lacking the active site nucleophile (X) are not covalently modified. Biotin binding to immobilized streptavidin allows isolation of the trapped phage.

generated by the catalyst may represent additional practical limitations to this approach.

Another viable strategy for selecting catalysts takes advantage of co-localization of the catalyst and substrate or product. For example, immobilized substrates can be cross-linked to phage in the vicinity of a displayed enzyme (Scheme 17).[191] If the enzyme is active, intramolecular



Scheme 17. Intramolecular conversion of substrate to product provides a means of selecting phage-bearing active enzymes if the reaction results in cleavage of the phage from a solid support.[191]

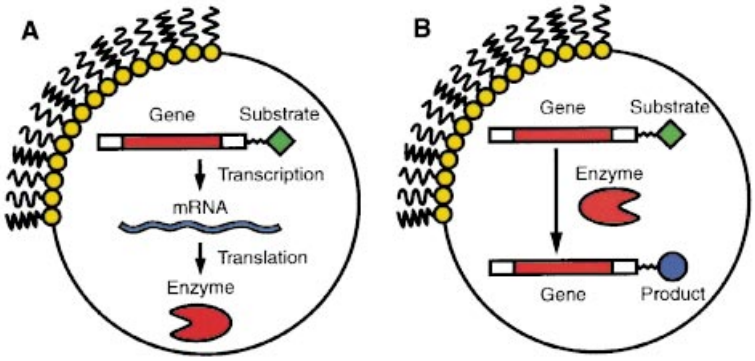
cleavage of the substrate selectively releases the corresponding phage particle from the solid support, which allows isolation and amplification of the encoding gene. Analogous schemes have been devised for synthetic reactions that result in the attachment of a tag to the phage or that produce a tagged product that can subsequently react with the phage particle before diffusing away.[192] The peptide-ligating activity of a modified protease has been optimized by such a scheme: enzyme variants that successfully attached a biotinylated peptide to their own extended N-termini were selectively captured by streptavidin.[193] In another variation of this general approach, antibodies against the product have been used to isolate catalytically active enzymes displayed on phage particles to which the reaction products were anchored.[194]

A very different scheme was recently reported for identifying functional proteases in protein libraries displayed on the surface of a bacterial cell.[195] It relies on electrostatic interactions to retain positively charged peptide substrates and products on the negatively charged cell surface. Cleavage of the substrate, which is covalently linked to fluorescent FRET (fluorescence resonance energy transfer) groups, releases the quenching component, and gives rise to a fluorescent signal. Active catalyst variants can then be isolated by fluorescence-activated cell sorting (FACS). Mutants of the serine protease OmpT with 60-fold improved efficiencies for cleavage of Arg–Val bonds were isolated in this way from libraries of 6×10^5 random variants. This method, which is technically a high-throughput screening procedure rather than genuine Darwinian selection, should be applicable to a variety of proteins displayed on the surface of bacteria and yeast.

5.2. Compartmentalization

Cell-free transcription/translation systems can constitute an entirely in vitro alternative to phage display. Such systems have all the RNA molecules, proteins, and small molecules needed to transcribe a gene and translate it into protein.[196, 197] Genotypes and phenotypes can be topologically linked by encapsulating these components in an artificial cell-like compartment. For example, the ability of a translated enzyme to modify its co-localized encoding gene can be exploited as a means of selection (Scheme 18). In compartments in which an inactive protein is produced, the corresponding genes remain unmodified and thus not amplifiable.

The feasibility of this approach has been demonstrated by using mixtures of genes encoding *Hae*III methyltransferase and another, unrelated enzyme. The methyltransferase methylates DNA and renders it resistant to cleavage by *Hae*III endonuclease. The mixture was encapsulated in water/oil emulsions, which have aqueous compartments with volumes close to those of bacteria.[198] After transcription/translation in the aqueous compartments, genes encoding the methyltransferase were protected by methylation and amplifiable by PCR, whereas those encoding the other protein were



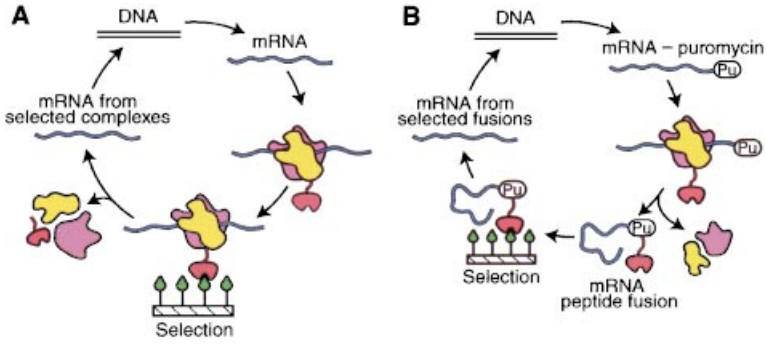
Scheme 18. Gene selection by compartmentalization is based on transcription/translation of single genes in aqueous compartments of a water/oil emulsion (A).[198] Subsequent modification of the gene by the encoded and co-localized enzyme (B) affords a means of selection. Genes encoding an inactive protein will not become modified and can be eliminated in the selection process.

degraded upon treatment with the endonuclease *Hae*III. In this way methyltransferase genes were isolated from a 10^7-fold excess of genes encoding the other enzyme.[198]

If a general strategy for selection can be developed, compartmentalization could prove broadly applicable for molecular evolution experiments since passage through cells is unnecessary. As a consequence, library sizes are potentially much larger than for in vivo experiments and difficulties associated with inefficient microbial production of some proteins can be circumvented.

5.3. Linking mRNA and Protein during Synthesis

Ribosome (polysome) display[199–203] and generation of mRNA–protein fusions[204–206] are two additional methods that can be exploited for the evolution of proteins in vitro. Both rely on cell-free translation of mRNA templates to produce proteins on ribosomes (Scheme 19). To forestall the uncoupling of message (protein) and messenger (mRNA) that would normally occur during protein synthesis in



Scheme 19. Selection by ribosome display or mRNA – peptide/protein fusions.[199–203] A) In vitro translation of mRNA on ribosomes can be halted by the addition of Mg^{2+} ions and incubation at low temperature. Ribosome/mRNA/nascent peptide complexes containing a desired receptor (red) can be selected by binding to an immobilized ligand (green). Recovery and reverse transcription of the retained mRNA, followed by PCR amplification, yields DNA for characterization or the next round of selection. B) An mRNA – protein fusion is formed upon ribosome-catalyzed translation of an mRNA – puromycin conjugate. Upon completion of the protein synthesis, the protein (red) is transferred from tRNA to puromycin (Pu) at the peptidyl transferase center of the ribosome. The folded protein and its encoding mRNA are subsequently amenable to in vitro selection.

the absence of compartmentalization, the system can be rigged so as to prevent dissociation of the translated protein from the ribosome.[199–203] Alternatively, the protein can be joined to its own mRNA through a stable covalent linkage.[204–206]

Complexes of ribosomes, mRNA, and protein can be stabilized against dissociation by low temperatures and addition of magnesium acetate to the buffer.[201] More importantly, eliminating the stop codon that normally signals binding of polypeptide release factors and termination of protein synthesis greatly slows the hydrolysis of the ribosome-bound peptidyl-tRNA.[201] If the displayed proteins fold properly, affinity selection protocols can be used to isolate variants that recognize small molecules, oligonucleotides, or other proteins (Scheme 19A). For example, this technology has been used successfully to increase the stability and ligand affinity of antibody molecules.[201, 207] Addition of ribonuclease inhibitors to the reaction mixture and incorporation of stem – loop structures into the mRNAs to prevent degradation of the messenger have helped to optimize the efficiency of such experiments.[201] Providing molecular chaperones to promote proper folding of the displayed proteins in vitro can also be beneficial.[201]

The generation of covalent fusions between an mRNA and the protein it encodes may ultimately prove to be a more robust method of linking genotype and phenotype.[204–206] Such constructs are formed by in vitro translation of mRNA templates containing the antibiotic puromycin at their 3′-end (Scheme 19B). After the RNA is translated, puromycin can enter the peptidyl transferase site of the ribosome and attach itself to the C-terminus of the nascent protein to form a covalent mRNA – protein conjugate. The translated protein can then be selected as described for the other formats. In contrast to the strategy of Scheme 19A, selection can be performed after removal of the ribosome and other components of the in vitro translation system because the protein is covalently linked to its own mRNA.

In vitro transcription/translation systems are attractive for directed evolution for a number of reasons. First, an expanded range of conditions can be exploited for the selection step itself. Second, unnatural amino acids can be incorporated directly into the proteins through suppressor codons.[208] Finally, libraries of much greater complexity and size can be generated in vitro than in vivo because these systems are not limited by the transformation efficiency of *E. coli*. Libraries containing 10^{11} to 10^{13} different members can be constructed with current protocols, and modest scale-up and improved construction protocols may make ensembles of 10^{15} members accessible.[205, 209]

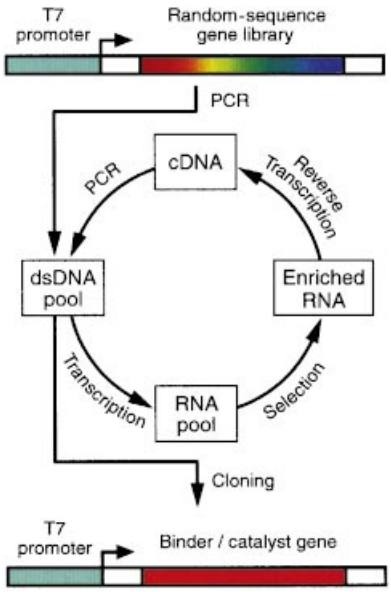
As with phage and cell surface display technologies, these systems are ideally suited for selection schemes based on receptor – ligand interactions. For example, four ATP binders were recently selected using the mRNA – peptide fusion approach starting from a library of 6×10^{12} proteins in which 80 contiguous amino acids were randomized.[210] It has been suggested that libraries constructed from natural sources of genomic or cDNA will also be useful for rapidly finding interacting proteins in organisms where genetic analysis is difficult.[205] The combination of some of the in vitro transcription/translation strategies discussed herein with methods for selecting proteins on the basis of their catalytic activity (see Section 5.1) may also allow the directed evolution of novel enzymatic function.

5.4. RNA Evolution Experiments

The value of having large libraries has been demonstrated by the direct selection of RNA and DNA molecules that function as receptors and catalysts.[210] Such experiments routinely involve generating and screening molecular repertoires with as many as 10^{15} independent members. Since nucleic acids combine structure, function, and genetic information within a single molecule, in vitro recovery and amplification of rare genes for a desired phenotype is straightforward and very efficient (Scheme 20). Consequently, populations of nucleic acids can be easily subjected to multiple cycles of selection and enrichment.

RNA and DNA catalysts have been prepared for a wide variety of reactions, such as the cleavage, ligation, and modification of oligonucleotides, alkylations, acyl transfer reactions, and even abiological Diels – Alder cycloadditions.[210] Although selection of functional nucleotides appears to be easier than functional proteins (see, for example, Section 3.6), the resulting catalysts often have relatively modest efficiency relative to highly evolved enzymes, perhaps because they are constructed from a limited palette of four nucleotides. In some cases, however, extraordinary activities have been achieved.

The transfer of phosphoryl groups is particularly amenable to ribozyme catalysis. Direct selection of highly efficient self-ligating ribozymes from large pools of random sequence is one of the most impressive accomplishments in this regard.[211]



Scheme 20. In vitro evolution of RNA catalysts. A combinatorial library of RNAs generated by in vitro transcription is subjected to selection for ligand binding or catalysis. Molecules possessing the desired trait are separated from inactive variants to yield a new RNA pool enriched in functional sequences. The enriched pool is reverse transcribed to cDNA and amplified by the polymerase chain reaction (PCR) to obtain material for another round of selection. The cycle is repeated until the pool is sufficiently enriched in active sequences, which can be cloned for detailed characterization. If an error-prone PCR method is used,[3] each round represents a true Darwinian-type cycle of evolution.

An initial library was designed as a pool of approximately 10^{15} different RNAs with 220 random nucleotides. About 1 in 10^{13} random sequences proved to have the desired ligase activity. Clearly, if the starting library had contained 10^{10} sequences, the chances of finding these molecules would have been vanishingly small. Further diversification and selection was used to evolve the population of ligase ribozymes, one of which was subsequently re-engineered to function as a true catalyst that promotes an intermolecular ligation with multiple turnovers.[212] It exhibits a rate constant of $>1\ s^{-1}$, which corresponds to a rate acceleration approaching 10^9 over the uncatalyzed reaction.

Nucleic acid ligands, receptors, and catalysts are novel analogues of their peptide/protein counterparts with many conceivable applications as drugs or diagnostic agents. The strategies used to generate them, an understanding of how they evolve from random sequence pools, and insight into their relative advantages and disadvantages compared to polypeptides are all likely to improve our ability to create and optimize novel protein catalysts.

6. Summary and Perspectives

Nature has solved the problem of protein design through the mechanism of Darwinian evolution. Every one of the proteins in our cells, from enzymes and receptors to regulatory and structural proteins, has arisen by this process. As the examples presented in this review amply demonstrate, evolutionary strategies can also be successfully exploited in the laboratory to study the structure and function of existing proteins and to engineer new ones.

Site-directed mutagenesis, by which one amino acid in a protein is substituted for another, is an established and invaluable tool in protein chemistry. However, the size and complexity of proteins often precludes unambiguous interpretation of the results of such experiments. This arises from the non-uniform distribution of structural information within a protein sequence—the fact that some residues are tolerant to substitution, whereas others cannot be replaced without deleterious consequences—and to the many energetically similar conformational states available to any polypeptide. These factors, and the extreme sensitivity of catalytic activity to seemingly modest structural perturbation, make characterization and (re)design of enzymes in the laboratory so difficult.

In contrast to methods involving single substitutions, combinatorial methods are ideally suited for evaluating the complex network of interactions in proteins that ultimately determine form and function.[73] As shown in the investigations of chorismate mutase enzymes (Sections 3.2 and 3.3), degenerate protein ensembles can illuminate subtle trends that might be missed in analyses of individual amino acid replacements. This information can be used to verify and refine our understanding of protein structure and function. Evolutionary strategies can also be used in conjunction with data on structure and mechanism to support design efforts,[8] as shown by the creation of novel monomeric and binary patterned mutases (Sections 3.4 – 3.6) and by the alteration of aspartate aminotransferase specificity (Section 4.1) and IGPS function (Section 4.3). The iterative nature of evolutionary approaches is particularly advantageous in this regard, since it allows the rapid optimization of the population of molecules under study toward some desirable goal. Even in the absence of detailed structural information or a clear understanding of how a system works, directed evolution can be used to tailor the properties of proteins for specific practical applications (Section 4.4).

Of course, successful application of evolutionary methods requires an efficient method for sifting through the many library members to identify those that have desirable properties, be they recognition of a particular ligand or catalysis of an interesting chemical transformation. For small libraries, finding the useful variants is readily accomplished by screening. Even with large libraries, depending on the difficulty of the task at hand—and the distance the starting structure is from the desired endpoint in sequence space[213]—screening can often be exploited with great effectiveness. For example, enantioselective esterases and lipases,[214, 215] thermostable hydrolases,[216, 217] and metal-dependent enzymes with novel specificities[218, 219] have been identified through random mutagenesis and screening. Moreover, advances in automation and miniaturization can be expected to extend the utility of screening methods for high-throughput analysis of reasonably large molecular ensembles.[10, 11, 195, 198]

Genetic selection is likely to be the method of choice for problems where statistically meaningful correlations are sought or solutions are extraordinarily rare. In vivo selection schemes allow exhaustive searches of libraries of 10^{10} protein

variants, and significantly larger libraries can be handled with in vitro methods (Section 5.2 – 5.4). The principal challenge in developing an effective selection protocol is the linkage of the macromolecular property of interest to a biological readout, thereby allowing facile recovery of the responsible encoding gene. In principle, any metabolic process can be exploited for this purpose, as can transformations involving the creation or destruction of vital nutrients or toxins (Sections 3 and 4). Clever in vitro schemes further expand the range of possibilities for selecting molecules with interesting binding and catalytic properties (Section 5).

Although laboratory evolution experiments are now well established, improvements in several areas could significantly enhance the utility of such approaches for diverse practical applications. Improved protocols for mutagenesis[134, 220, 221] and for transforming microorganisms,[23, 222] for example, might allow the creation of better-designed and still-larger protein libraries. Selection schemes that are more versatile and general than those currently available, and which would permit selection for any desired function, would also clearly be of great value. Furthermore, we can dream of extending evolutionary principles to a wider range of molecules than nucleic acids and proteins. Realization of such an ambitious goal would require developing novel encoded combinatorial libraries as well as strategies for extracting desirable molecules and their blueprints from such ensembles. To that end, new methods for "autocatalytic" detection and selective amplification of the "best" molecules after each evolutionary cycle will be needed.

Evolutionary approaches will never supplant "rational" approaches for the study of molecules. Rather, they powerfully augment the classical strategies of dissection and design. In fact, directed evolution is likely to be most effective in combination with detailed structural and mechanistic information for exploring and engineering the properties of proteins, nucleic acids, and other encoded molecules. In the post-genomic era, these methods are likely to become increasingly valuable in analyzing and integrating the flood of data that is emerging for so many biological systems. In the future we can expect that intelligent application of evolutionary methods will help to clarify the links between sequence and structure, structure and function, and between interacting macromolecules in the cell. Novel applications in medicine and industry will not be far behind.

The authors wish to thank Kinya Hotta and Gavin MacBeath for assistance in the preparation of graphics. The financial support of the ETH Zurich, the Schweizerischer Nationalfonds, and Novartis Pharma are gratefully acknowledged.

Received: January 25, 2001 [A 446]

[1] R. Dawkins, *The Blind Watchmaker*, Longham, London **1986**.
[2] *In vitro Mutagenesis Protocols* (Ed.: M. K. Trower), Humana Press, Totowa, NJ, **1996**.
[3] R. C. Cadwell, G. F. Joyce, *PCR Methods Appl.* **1992**, *2*, 28 – 33.
[4] W. P. C. Stemmer, *Bio/Technology* **1995**, *13*, 549 – 553.
[5] W. P. C. Stemmer, *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 10747 – 10751.
[6] *Methods Enzymol.* **1996**, *267* (Ed.: J. N. Abelson).
[7] P. Kast, D. Hilvert, *Pure Appl. Chem.* **1996**, *68*, 2017 – 2024.
[8] P. Kast, D. Hilvert, *Curr. Opin. Struct. Biol.* **1997**, *7*, 470 – 479.
[9] R. T. Sauer, *Fold. Des.* **1996**, *1*, R27 – R30.
[10] F. H. Arnold, *Acc. Chem. Res.* **1998**, *31*, 125 – 131.
[11] F. H. Arnold, A. A. Volkov, *Curr. Opin. Chem. Biol.* **1999**, *3*, 54 – 59.
[12] D. C. Demirjian, P. C. Shah, F. Moris-Varas, *Top. Curr. Chem.* **1999**, *200*, 1 – 29.
[13] U. Kettling, A. Koltermann, M. Eigen, *Top. Curr. Microbiol. Immunol.* **1999**, *243*, 173 – 186.
[14] A. Koltermann, U. Kettling, *Biophys. Chem.* **1997**, *66*, 159 – 177.
[15] L. P. Encell, D. M. Landis, L. A. Loeb, *Nat. Biotechnol.* **1999**, *17*, 143 – 147.
[16] A. Skandalis, L. P. Encell, L. A. Loeb, *Chem. Biol.* **1997**, *4*, 889 – 898.
[17] J. Fastrez, *Mol. Biotechnol.* **1997**, *7*, 37 – 55.
[18] B. Lewin, *Genes VII*, Oxford University Press, London, **2000**.
[19] J. D. Watson, M. Gilman, J. Witkowski, M. Zoller, *Recombinant DNA: A Short Course*, W. H. Freeman, San Francisco, **1992**.
[20] J. Sambrook, D. W. Russell, *Molecular Cloning. A Laboratory Manual*, *3rd ed.*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, **2001**.
[21] P. B. Fernandes, *Curr. Opin. Chem. Biol.* **1998**, *2*, 597 – 603.
[22] A. Qureshi-Emili, G. Cagney, *Nat. Biotechnol.* **2000**, *18*, 393 – 397.
[23] D. Hanahan, J. Jessee, F. R. Bloom, *Methods Enzymol.* **1991**, *204*, 63 – 113.
[24] L. You, F. H. Arnold, *Protein Eng.* **1996**, *9*, 77 – 83.
[25] F. H. Arnold, J. C. Moore, *Adv. Biochem. Eng. Biotechnol.* **1997**, *58*, 1 – 14.
[26] C. M. Gates, W. P. C. Stemmer, R. Kaptein, P. J. Schatz, *J. Mol. Biol.* **1996**, *255*, 373 – 386.
[27] R. R. Breaker, A. Banerji, G. F. Joyce, *Biochemistry* **1994**, *33*, 11980 – 11986.
[28] U. Weiss, J. M. Edwards, *The Biosynthesis of Aromatic Amino Compounds*, Wiley, New York, **1980**.
[29] E. Haslam, *Shikimic Acid: Metabolism and Metabolites*, Wiley, New York, **1993**.
[30] B. Ganem in *Comprehensive Natural Products Chemistry*, *Vol. 5* (Ed.: C. D. Poulter), Elsevier, Oxford, **1999**, pp. 343 – 369.
[31] Y. M. Chook, J. V. Gray, H. Ke, W. N. Lipscomb, *J. Mol. Biol.* **1994**, *240*, 476 – 500.
[32] Y. M. Chook, H. Ke, W. N. Lipscomb, *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 8600 – 8603.
[33] A. Y. Lee, P. A. Karplus, B. Ganem, J. Clardy, *J. Am. Chem. Soc.* **1995**, *117*, 3627 – 3628.
[34] N. Sträter, G. Schnappauf, G. Braus, W. N. Lipscomb, *Structure* **1997**, *5*, 1437 – 1452.
[35] Y. Xue, W. N. Lipscomb, R. Graf, G. Schnappauf, G. Braus, *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 10814 – 10818.
[36] W. Gu, D. S. Williams, H. C. Aldrich, G. Xie, D. W. Gabriel, R. A. Jensen, *Microb. Comp. Genomics* **1997**, *2*, 141 – 158.
[37] G. MacBeath, P. Kast, D. Hilvert, *Biochemistry* **1998**, *37*, 10062 – 10073.
[38] P. Kast, C. Grisostomi, I. A. Chen, S. Li, U. Krengel, Y. Xue, D. Hilvert, *J. Biol. Chem.* **2000**, *275*, 36832 – 36838.
[39] D. Hilvert, S. H. Carpenter, K. D. Nared, M.-T. M. Auditor, *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 4953 – 4955.
[40] D. Hilvert, K. D. Nared, *J. Am. Chem. Soc.* **1988**, *110*, 5593 – 5594.
[41] D. Y. Jackson, J. W. Jacobs, R. Sugasawara, S. H. Reich, P. A. Bartlett, P. G. Schultz, *J. Am. Chem. Soc.* **1988**, *110*, 4841 – 4842.
[42] M. R. Haynes, E. A. Stura, D. Hilvert, I. A. Wilson, *Science* **1994**, *263*, 646 – 652.
[43] K. Bowdish, Y. Tang, J. B. Hicks, D. Hilvert, *J. Biol. Chem.* **1991**, *266*, 11901 – 11908.
[44] Y. Tang, J. B. Hicks, D. Hilvert, *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 8784 – 8786.
[45] P. Kast, M. Asif-Ullah, N. Jiang, D. Hilvert, *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 5043 – 5048.
[46] P. Kast, M. Asif-Ullah, D. Hilvert, *Tetrahedron Lett.* **1996**, *37*, 2691 – 2694.
[47] G. Zhao, T. Xia, R. S. Fischer, R. A. Jensen, *J. Biol. Chem.* **1992**, *267*, 2487 – 2493.

[48] T. Xia, G. Zhao, R. S. Fischer, R. A. Jensen, *J. Gen. Microbiol.* **1992**, *138*, 1309 – 1316.
[49] L. Addadi, E. K. Jaffe, J. R. Knowles, *Biochemistry* **1983**, *22*, 4494 – 4501.
[50] S. D. Copley, J. R. Knowles, *J. Am. Chem. Soc.* **1985**, *107*, 5306 – 5308.
[51] S. D. Copley, J. R. Knowles, *J. Am. Chem. Soc.* **1987**, *109*, 5008 – 5013.
[52] J. J. Gajewski, J. Jurayj, D. R. Kimbrough, M. E. Gande, B. Ganem, B. K. Carpenter, *J. Am. Chem. Soc.* **1987**, *109*, 1170 – 1186.
[53] O. Wiest, K. N. Houk, *J. Org. Chem.* **1994**, *59*, 7582 – 7584.
[54] O. Wiest, K. N. Houk, *J. Am. Chem. Soc.* **1995**, *117*, 11628 – 11639.
[55] P. D. Lyne, A. J. Mulholland, W. G. Richards, *J. Am. Chem. Soc.* **1995**, *117*, 11345 – 11350.
[56] M. M. Davidson, I. R. Gould, I. H. Hillier, *J. Chem. Soc. Perkin Trans.* **1996**, *2*, 525 – 532.
[57] M. M. Davidson, I. H. Hillier, *J. Chem. Soc. Perkin Trans.* **1994**, *2*, 1415 – 1417.
[58] D. L. Severence, W. L. Jorgensen, *J. Am. Chem. Soc.* **1992**, *114*, 10966 – 10968.
[59] H. A. Carlson, W. L. Jorgensen, *J. Am. Chem. Soc.* **1996**, *118*, 8475 – 8484.
[60] P. Kast, Y. B. Tewari, O. Wiest, D. Hilvert, K. N. Houk, R. N. Goldberg, *J. Phys. Chem. B* **1997**, *101*, 10976 – 10982.
[61] S. G. Sogo, T. S. Widlanski, J. H. Hoare, C. E. Grimshaw, G. A. Berchtold, J. R. Knowles, *J. Am. Chem. Soc.* **1984**, *106*, 2701 – 2703.
[62] P. A. Bartlett, C. R. Johnson, *J. Am. Chem. Soc.* **1985**, *107*, 7792 – 7793.
[63] P. A. Bartlett, Y. Nakagawa, C. R. Johnson, S. H. Reich, A. Luis, *J. Org. Chem.* **1988**, *53*, 3195 – 3210.
[64] J. V. Gray, D. Eren, J. R. Knowles, *Biochemistry* **1990**, *29*, 8872 – 8878.
[65] D. J. Gustin, P. Mattei, P. Kast, O. Wiest, L. Lee, W. W. Cleland, D. Hilvert, *J. Am. Chem. Soc.* **1999**, *121*, 1756 – 1757.
[66] S. T. Cload, D. R. Liu, R. M. Pastor, P. G. Schultz, *J. Am. Chem. Soc.* **1996**, *118*, 1787 – 1788.
[67] D. R. Liu, S. T. Cload, R. M. Pastor, P. G. Schultz, *J. Am. Chem. Soc.* **1996**, *118*, 1789 – 1790.
[68] S. Zhang, P. Kongsaeree, J. Clardy, D. B. Wilson, B. Ganem, *Bioorg. Med. Chem.* **1996**, *4*, 1015 – 1020.
[69] "Evolutionary Studies with a Catalytic Antibody": Y. Tang, Ph.D. Thesis, The Scripps Research Institute, **1996**.
[70] M. Gamper, D. Hilvert, P. Kast, *Biochemistry* **2000**, *39*, 14087 – 14094.
[71] J. V. Gray, J. R. Knowles, *Biochemistry* **1994**, *33*, 9953 – 9959.
[72] M. H. J. Cordes, A. R. Davidson, R. T. Sauer, *Curr. Opin. Struct. Biol.* **1996**, *6*, 3 – 10.
[73] J. F. Reidhaar-Olson, R. T. Sauer, *Science* **1988**, *241*, 53 – 57.
[74] M. Blaber, X. Zhang, B. W. Matthews, *Science* **1993**, *260*, 1637 – 1640.
[75] H. Xiong, B. L. Buckwalter, H. M. Shieh, M. H. Hecht, *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 6349 – 6353.
[76] A. P. Brunet, E. S. Huang, M. E. Huffine, J. E. Loeb, R. J. Weltman, M. H. Hecht, *Nature* **1993**, *364*, 355 – 358.
[77] L. Castagnoli, C. Vetriani, G. Cesareni, *J. Mol. Biol.* **1994**, *237*, 378 – 387.
[78] J. A. Ybe, M. H. Hecht, *Protein Sci.* **1996**, *5*, 814 – 824.
[79] W. A. Lim, R. T. Sauer, *Nature* **1989**, *339*, 31 – 36.
[80] M. E. Milla, R. T. Sauer, *Biochemistry* **1995**, *34*, 3344 – 3351.
[81] E. P. Baldwin, O. Hajiseyedjavadi, W. A. Baase, B. W. Matthews, *Science* **1993**, *262*, 1715 – 1718.
[82] Z. L. Fredericks, G. J. Pielak, *Biochemistry* **1993**, *32*, 929 – 936.
[83] P. F. Predki, L. Regan, *Biochemistry* **1995**, *34*, 9834 – 9839.
[84] P. F. Predki, V. Agrawal, A. T. Brünger, L. Regan, *Nat. Struct. Biol.* **1996**, *3*, 54 – 58.
[85] G. MacBeath, P. Kast, D. Hilvert, *Protein Sci.* **1998**, *7*, 325 – 335.
[86] S. M. Green, A. G. Gittis, A. K. Meeker, E. E. Lattman, *Nat. Struct. Biol.* **1995**, *2*, 746 – 751.
[87] R. R. Dickason, D. P. Huston, *Nature* **1996**, *379*, 652 – 655.
[88] R. A. Albright, M. C. Mossing, B. W. Matthews, *Biochemistry* **1996**, *35*, 735 – 742.
[89] T. V. Borchert, R. Abagyan, R. Jaenicke, R. K. Wierenga, *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 1515 – 1518.
[90] M. C. Mossing, R. T. Sauer, *Science* **1990**, *250*, 1712 – 1715.
[91] W. Schliebs, N. Thanki, R. Jaenicke, R. K. Wierenga, *Biochemistry* **1997**, *36*, 9655 – 9662.
[92] X. Shao, P. Hensley, C. R. Matthews, *Biochemistry* **1997**, *36*, 9941 – 9949.
[93] M. J. Bennett, S. Choe, D. Eisenberg, *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 3127 – 3131.
[94] M. J. Bennett, M. P. Schlunegger, D. Eisenberg, *Protein Sci.* **1995**, *4*, 2455 – 2468.
[95] G. MacBeath, P. Kast, D. Hilvert, *Protein Sci.* **1998**, *7*, 1757 – 1767.
[96] L. Lins, R. Brasseur, *FASEB J.* **1995**, *9*, 535 – 540.
[97] G. MacBeath, P. Kast, D. Hilvert, *Science* **1998**, *279*, 1958 – 1961.
[98] M. G. Oakley, P. S. Kim, *Biochemistry* **1997**, *36*, 2544 – 2549.
[99] K. U. Walter, S. V. Taylor, G. Mäder, P. Kast, D. Hilvert, unpublished results.
[100] W. P. C. Stemmer, *Nature* **1994**, *370*, 389 – 391.
[101] M. Couturier, F. Bex, W. K. Maas, *Microbiol. Rev.* **1988**, *52*, 375 – 395.
[102] T. A. Y. Ayoubi, W. J. M. VanDeVen, *FASEB J.* **1996**, *10*, 453 – 460.
[103] H. A. DeBoer, A. S. Hui, *Methods Enzymol.* **1990**, *185*, 103 – 114.
[104] G. R. Dieckmann, D. K. McRorie, D. L. Tierney, L. M. Utschig, C. P. Singer, T. V. O'Halloran, J. E. Penner-Hahn, W. F. DeGrado, V. L. Pecoraro, *J. Am. Chem. Soc.* **1997**, *119*, 6195 – 6196.
[105] S. F. Betz, W. F. DeGrado, *Biochemistry* **1996**, *35*, 6955 – 6962.
[106] S. Dalal, S. Balasubramanian, L. Regan, *Nat. Struct. Biol.* **1997**, *4*, 548 – 552.
[107] J. E. Rozzelle, A. Tropsha, B. W. Erickson, *Protein Sci.* **1994**, *3*, 345 – 355.
[108] K. M. Gernert, M. C. Surles, T. H. Labean, J. S. Richardson, D. C. Richardson, *Protein Sci.* **1995**, *4*, 2252 – 2260.
[109] P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, P. S. Kim, *Science* **1998**, *282*, 1462 – 1467.
[110] X. Jiang, E. J. Bishop, R. S. Farid, *J. Am. Chem. Soc.* **1997**, *119*, 838 – 839.
[111] B. I. Dahiyat, S. L. Mayo, *Science* **1997**, *278*, 82 – 87.
[112] S. M. Malakauskas, S. L. Mayo, *Nat. Struct. Biol.* **1998**, *5*, 470 – 475.
[113] A. R. Davidson, R. T. Sauer, *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 2146 – 2150.
[114] M. W. West, M. H. Hecht, *Protein Sci.* **1995**, *4*, 2032 – 2039.
[115] J. U. Bowie, R. T. Sauer, *Proc. Natl. Acad. Sci. USA* **1989**, *86*, 2152 – 2156.
[116] S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, M. H. Hecht, *Science* **1993**, *262*, 1680 – 1685.
[117] S. Roy, G. Ratnaswamy, J. A. Boice, R. Fairman, G. McLendon, M. H. Hecht, *J. Am. Chem. Soc.* **1997**, *119*, 5302 – 5306.
[118] S. Roy, M. H. Hecht, *Biochemistry* **2000**, *39*, 4603 – 4607.
[119] S. V. Taylor, K. U. Walter, P. Kast, D. Hilvert, *Proc. Natl. Acad. Sci.USA*, in press.
[120] E. A. Kabat, *Structural Concepts in Immunology and Immunochemistry*, Holt, Rinehart, and Winston, New York, **1976**.
[121] J. T.-F. Wong, *Proc. Natl. Acad. Sci. USA* **1975**, *72*, 1909 – 1912.
[122] C. Woese, *The Genetic Code*, Harper & Row, New York, **1967**.
[123] F. H. C. Crick, *J. Mol. Biol.* **1968**, *38*, 367 – 379.
[124] W. F. DeGrado, C. M. Summa, V. Pavone, F. Nastri, A. Lombardi, *Annu. Rev. Biochem.* **1999**, *68*, 779 – 819.
[125] C.-H. Wong, G. M. Whitesides, *Enzymes in Synthetic Organic Chemistry*, Elsevier Science, Tarrytown, **1994**.
[126] H. M. Wilks, K. W. Hart, R. Feeney, C. R. Dunn, H. Muirhead, W. N. Chia, D. A. Barstow, T. Atkinson, A. R. Clarke, J. J. Holbrook, *Science* **1988**, *242*, 1541 – 1544.
[127] L. Hedstrom, L. Szilagyi, W. J. Rutter, *Science* **1992**, *255*, 1249 – 1253.
[128] J. J. Onuffer, J. F. Kirsch, *Protein Sci.* **1995**, *4*, 1750 – 1757.
[129] T. Yano, S. Oue, H. Kagamiyama, *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 5511 – 5515.
[130] S. Oue, A. Okamoto, T. Yano, H. Kagamiyama, *J. Biol. Chem.* **1999**, *274*, 2344 – 2349.
[131] P. A. Patten, N. S. Gray, P. L. Yang, C. B. Marks, G. J. Wedemayer, J. J. Boniface, R. C. Stevens, P. G. Schultz, *Science* **1996**, *271*, 1086 – 1091.
[132] G. J. Wedemayer, P. A. Patten, L. H. Wang, P. G. Schultz, R. C. Stevens, *Science* **1997**, *276*, 1665 – 1669.
[133] I. M. Tomlinson, G. Walter, P. T. Jones, P. H. Dear, E. L. L. Sonnhammer, G. Winter, *J. Mol. Biol.* **1996**, *256*, 813 – 817.
[134] A. Crameri, S. A. Raillard, E. Bermudez, W. P. C. Stemmer, *Nature* **1998**, *391*, 288 – 291.
[135] H. Viadiu, J. Osuna, A. L. Fink, X. Soberon, *J. Biol. Chem.* **1995**, *270*, 781 – 787.

[136] M. Zaccolo, E. Gherardi, *J. Mol. Biol.* **1999**, *285*, 775 – 783.
[137] J. F. Petrosino, T. Palzkill, *J. Bacteriol.* **1996**, *178*, 1821 – 1828.
[138] W. Huang, J. Petrosino, M. Hirsch, P. S. Shenkin, T. Palzkill, *J. Mol. Biol.* **1996**, *258*, 688 – 703.
[139] C. Cantu, W. Huang, T. Palzkill, *J. Biol. Chem.* **1996**, *271*, 22538 – 22545.
[140] K. M. Munir, D. C. French, D. K. Dube, L. A. Loeb, *Protein Eng.* **1994**, *7*, 83 – 89.
[141] S.-I. Ohnuma, K. Hirooka, H. Hemmi, C. Ishida, C. Ohto, T. Nishino, *J. Biol. Chem.* **1996**, *271*, 18831 – 18837.
[142] E. A. Hart, L. Hua, L. B. Darr, W. K. Wilson, J. H. Pang, S. P. T. Matsuda, *J. Am. Chem. Soc.* **1999**, *121*, 9887 – 9888.
[143] C. Jürgens, A. Strom, D. Wegener, S. Hettwer, M. Wilmanns, R. Sterner, *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 9925 – 9930.
[144] D. R. Liu, P. G. Schultz, *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 4780 – 4785.
[145] D. R. Liu, T. J. Magliery, M. Pastrnak, P. G. Schultz, *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 10092 – 10097.
[146] M. W. W. Adams, R. M. Kelly, *Trends Biotechnol.* **1998**, *16*, 329 – 332.
[147] T. Oshima, *Curr. Opin. Struct. Biol.* **1994**, *4*, 623 – 628.
[148] H. Liao, T. McKenzie, R. Hageman, *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 576 – 580.
[149] M. Matsumura, S. Aiba, *J. Biol. Chem.* **1985**, *260*, 15298 – 15303.
[150] J. Hoseki, T. Yano, Y. Koyama, S. Kuramitsu, H. Kagamiyama, *J. Biochem.* **1999**, *126*, 951 – 956.
[151] M. Matsumura, S. Yasumura, S. Aiba, *Nature* **1986**, *323*, 356 – 358.
[152] Y. I. Wolf, N. V. Grishin, E. V. Koonin, *J. Mol. Biol.* **2000**, *299*, 897 – 905.
[153] D. R. Davies, E. A. Padlan, S. Sheriff, *Annu. Rev. Biochem.* **1990**, *59*, 439 – 473.
[154] P. G. Schultz, R. A. Lerner, *Science* **1995**, *269*, 1835 – 1842.
[155] D. Hilvert, *Annu. Rev. Biochem.* **2000**, *69*, 751 – 793.
[156] R. A. Lerner, S. J. Benkovic, P. G. Schultz, *Science* **1991**, *252*, 659 – 667.
[157] C. Bränden, J. Tooze, *Introduction to Protein Structure*, Garland, New York, **1999**.
[158] C. I. Bränden, *Curr. Opin. Struct. Biol.* **1991**, *1*, 978 – 983.
[159] M. M. Altamirano, J. M. Blackburn, C. Aguayo, A. R. Fersht, *Nature* **2000**, *403*, 617 – 622.
[160] R. Sterner, G. R. Kleemann, H. Szadkowski, A. Lustig, M. Hennig, K. Kirschner, *Protein Sci.* **1996**, *5*, 2000 – 2008.
[161] G. Winter, C. Milstein, *Nature* **1991**, *349*, 293 – 299.
[162] A. Crameri, G. Dawes, E. Rodriguez, Jr., S. Silver, W. P. C. Stemmer, *Nat. Biotechnol.* **1997**, *15*, 436 – 438.
[163] S. Fields, O.-K. Song, *Nature* **1989**, *340*, 245 – 246.
[164] P. Colas, R. Brent, *Trends Biotechnol.* **1998**, *16*, 355 – 363.
[165] R. M. Frederickson, *Curr. Opin. Biotechnol.* **1998**, *9*, 90 – 96.
[166] R. K. Brachmann, J. D. Boeke, *Curr. Opin. Biotechnol.* **1997**, *8*, 561 – 568.
[167] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, J. M. Rothberg, *Nature* **2000**, *403*, 623 – 627.
[168] P. Colas, B. Cohen, T. Jessen, I. Grishina, J. McCoy, R. Brent, *Nature* **1996**, *380*, 548 – 550.
[169] E. J. Licitra, J. O. Liu, *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 12817 – 12821.
[170] J. C. Hu, M. G. Kornacker, A. Hochschild, *Methods* **2000**, *20*, 80 – 94.
[171] G. Karimova, J. Pidoux, A. Ullmann, D. Ladant, *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 5752 – 5756.
[172] T. C. Norman, D. L. Smith, P. K. Sorger, B. L. Drees, S. M. O'Rourke, T. R. Hughes, C. J. Roberts, S. H. Friend, S. Fields, A. W. Murray, *Science* **1999**, *285*, 591 – 595.
[173] S.-H. Park, R. T. Raines, *Nat. Biotechnol.* **2000**, *18*, 847 – 851.
[174] J. A. Francisco, R. Campbell, B. L. Iverson, G. Georgiou, *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 10444 – 10448.
[175] E. T. Boder, K. D. Wittrup, *Nat. Biotechnol.* **1997**, *15*, 553 – 557.
[176] *Phage Display of Peptides and Proteins* (Eds.: B. K. Kay, J. Winter, J. McCafferty), Academic Press, New York, **1996**.
[177] G. P. Smith, V. A. Petrenko, *Chem. Rev.* **1997**, *97*, 391 – 410.
[178] S. Chusacultanachai, K. A. Glenn, A. O. Rodriguez, E. K. Read, J. F. Gardner, B. S. Katzenellenbogen, D. J. Shapiro, *J. Biol. Chem.* **1999**, *274*, 23591 – 23598.
[179] H. A. Greisman, C. O. Pabo, *Science* **1997**, *275*, 657 – 661.
[180] S. A. Wolfe, E. I. Ramm, C. O. Pabo, *Struct. Fold. Des.* **2000**, *8*, 739 – 750.
[181] T. E. Wilson, T. J. Fahrner, M. Johnston, J. Milbrandt, *Science* **1991**, *252*, 1296 – 1300.
[182] R. R. Beerli, D. J. Segal, B. Dreier, C. F. Barbas, *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 14628 – 14633.
[183] A. C. Jamieson, S.-H. Kim, J. A. Wells, *Biochemistry* **1994**, *33*, 5689 – 5695.
[184] E. J. Rebar, C. O. Pabo, *Science* **1994**, *263*, 671 – 673.
[185] Q. Liu, D. J. Segal, J. B. Ghiara, C. F. Barbas, *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 5525 – 5530.
[186] A. C. Braisted, J. A. Wells, *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 5688 – 5692.
[187] E. T. Boder, K. S. Midelfort, K. D. Wittrup, *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 10701 – 10705.
[188] C. Gao, C.-H. Lin, C.-H. L. Lo, S. Mao, P. Wirsching, R. A. Lerner, K. D. Janda, *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 11777 – 11782.
[189] P. Soumillion, L. Jespers, M. Bouchet, J. Marchand-Brynaert, G. Winter, J. Fastrez, *J. Mol. Biol.* **1994**, *237*, 415 – 422.
[190] K. D. Janda, L.-C. Lo, C.-H. L. Lo, M.-M. Sim, R. Wang, C.-H. Wong, R. A. Lerner, *Science* **1997**, *275*, 945 – 948.
[191] H. Pedersen, S. Hölder, D. P. Sutherlin, U. Schwitter, D. S. King, P. G. Schultz, *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 10523 – 10528.
[192] J.-L. Jestin, P. Kristensen, G. Winter, *Angew. Chem.* **1999**, *111*, 1196 – 1200; *Angew. Chem. Int. Ed.* **1999**, *38*, 1124 – 1127.
[193] S. Atwell, J. A. Wells, *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 9497 – 9502.
[194] S. Demartis, A. Huber, F. Viti, L. Lozzi, L. Giovannoni, P. Neri, G. Winter, D. Neri, *J. Mol. Biol.* **1999**, *286*, 617 – 633.
[195] M. J. Olsen, D. Stephens, D. Griffiths, P. Daugherty, G. Georgiou, B. L. Iverson, *Nat. Biotechnol.* **2000**, *18*, 1071 – 1074.
[196] S. A. Lesley, *Methods Mol. Biol.* **1995**, *37*, 265 – 278.
[197] R. J. Jackson, T. Hunt, *Methods Enzymol.* **1983**, *96*, 50 – 74.
[198] D. S. Tawfik, A. D. Griffiths, *Nat. Biotechnol.* **1998**, *16*, 652 – 656.
[199] L. C. Mattheakis, R. R. Bhatt, W. J. Dower, *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 9022 – 9026.
[200] L. C. Mattheakis, J. M. Dias, W. J. Dower, *Methods Enzymol.* **1996**, *267*, 195 – 207.
[201] J. Hanes, A. Plückthun, *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 4937 – 4942.
[202] C. Schaffitzel, J. Hanes, L. Jermutus, A. Plückthun, *J. Immunol. Methods* **1999**, *231*, 119 – 135.
[203] M. He, M. J. Taussig, *Nucleic Acids Res.* **1997**, *25*, 5132 – 5134.
[204] R. W. Roberts, J. W. Szostak, *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 12297 – 12302.
[205] R. Liu, J. E. Barrick, J. W. Szostak, R. W. Roberts, *Methods Enzymol.* **2000**, *318*, 268 – 293.
[206] N. Nemoto, E. Miyamoto-Sato, Y. Husimi, H. Yanagawa, *FEBS Lett.* **1997**, *414*, 405 – 408.
[207] J. Hanes, L. Jermutus, S. Weber-Bornhauser, H. R. Bosshard, A. Plückthun, *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 14130 – 14135.
[208] V. W. Cornish, D. Mendel, P. G. Schultz, *Angew. Chem.* **1995**, *107*, 677 – 689; *Angew. Chem. Int. Ed. Engl.* **1995**, *34*, 621 – 633.
[209] G. Cho, A. D. Keefe, R. Liu, D. S. Wilson, J. W. Szostak, *J. Mol. Biol.* **2000**, *297*, 309 – 319.
[210] D. S. Wilson, J. W. Szostak, *Annu. Rev. Biochem.* **1999**, *68*, 611 – 647.
[211] D. P. Bartel, J. W. Szostak, *Science* **1993**, *261*, 1411 – 1418.
[212] E. H. Ekland, J. W. Szostak, D. P. Bartel, *Science* **1995**, *269*, 364 – 370.
[213] J. C. Moore, H.-M. Jin, O. Kuchner, F. H. Arnold, *J. Mol. Biol.* **1997**, *272*, 336 – 347.
[214] U. T. Bornscheuer, J. Altenbuchner, H. H. Meyer, *Biotechnol. Bioeng.* **1998**, *58*, 554 – 559.
[215] M. T. Reetz, A. Zonta, K. Schimossek, K. Liebeton, K.-E. Jaeger, *Angew. Chem.* **1997**, *109*, 2961 – 2963; *Angew. Chem. Int. Ed. Engl.* **1997**, *36*, 2830 – 2832.
[216] L. Giver, A. Gershenson, P.-O. Freskgard, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 12809 – 12813.

[217] H. Zhao, F. H. Arnold, *Protein Eng.* **1999**, *12*, 47 – 53.
[218] H. Joo, Z. Lin, F. H. Arnold, *Nature* **1999**, *399*, 670 – 673.
[219] T. Kumamaru, H. Suenaga, M. Mitsuoka, T. Watanabe, K. Furukawa, *Nat. Biotechnol.* **1998**, *16*, 663 – 666.
[220] Z. Shao, H. Zhao, L. Giver, F. H. Arnold, *Nucleic Acids Res.* **1998**, *26*, 681 – 683.
[221] H. Zhao, L. Giver, Z. Shao, J. A. Affholter, F. H. Arnold, *Nat. Biotechnol.* **1998**, *16*, 258 – 261.
[222] J. R. Thompson, E. Register, J. Curotto, M. Kurtz, R. Kelly, *Yeast* **1998**, *14*, 565 – 571.
[223] C. Yanisch-Perron, J. Vieira, J. Messing, *Gene* **1985**, *33*, 103 – 119.
[224] A. Ullmann, F. Jacob, J. Monod, *J. Mol. Biol.* **1967**, *24*, 339 – 343.
[225] J. P. Horwitz, J. Chua, R. J. Curby, A. J. Tomson, M. A. Da Rooge, B. E. Fisher, J. Mauricio, I. Klundt, *J. Med. Chem.* **1964**, *7*, 574 – 575.
[226] P. Kast, *Gene* **1994**, *138*, 109 – 114.
[227] E. T. Kaiser, F. J. Kézdy, *Science* **1984**, *223*, 249 – 255.
[228] A. D. Keefe, J. W. Szostak, *Nature* **2001**, *410*, 715 – 718.