# EVOLUTION OF PROTEIN FUNCTION BY DOMAIN SWAPPING

By MARC OSTERMEIER* and STEPHEN J. BENKOVIC†

*Department of Chemical Engineering, Johns Hopkins University, Baltimore, Maryland 21218, and †The Pennsylvania State University, Department of Chemistry, University Park, Pennsylvania 16802

I. Introduction ..... 29
II. Terminology ..... 30
III. Evolution of Proteins in Nature by Domain Swapping ..... 31
A. Multidomain Proteins ..... 33
B. Tandem Duplication ..... 36
C. Domain Recruitment ..... 39
D. Circular Permutations ..... 42
IV. Domain Swapping for Protein Engineering ..... 43
A. Engineering of Allosteric Regulation ..... 44
B. Creation of Activators and Inhibitors ..... 45
C. Improvement in Stability or Expression ..... 45
D. Modification of Substrate Specificity ..... 47
E. Improvement of Catalytic Efficiency ..... 51
F. Alteration of Multimodular Synthetases ..... 51
G. Improvement of Therapeutic Properties ..... 55
H. Creation of Molecular Biosensors ..... 56
I. Creation of Novel Enzymes ..... 58
V. Methodologies ..... 59
A. Non-Combinatorial ..... 59
B. Combinatorial ..... 60
VI. Perspective ..... 67
References ..... 68

## I. Introduction

Obtaining proteins and enzymes with desired properties and activities is an important goal of biotechnology. However, even though recombinant DNA technology and molecular biology techniques allow the creation of a protein with any desired amino acid sequence, our current understanding of proteins falls short of that necessary for *de novo* design of protein function. Thus, two alternatives remain: (1) discovery of novel enzymes through scrutiny of natural resources or through recombinant genetic libraries (Short, 1997) and (2) remodeling of proteins already existing in nature. A balanced approach utilizing both strategies is likely to be the most successful, partly because new proteins discovered using (1) will create a larger database for performing (2).



ADVANCES IN
PROTEIN CHEMISTRY, Vol. 55


0065-3233/00 $35.00

Protein remodeling by rational design utilizes our current understanding of how proteins function and, guided by computer modeling, identifies specific changes designed to impart the desired function onto the existing protein. Since our current understanding of enzymes is less than complete, combinatorial methods for protein remodeling that mimic evolutionary processes have become increasingly popular. Such remodeling of existing proteins can be performed by creating random libraries by methods such as error-prone PCR, cassette mutagenesis, DNA shuffling (Stemmer, 1994), molecular breeding (Crameri *et al.,* 1998), or by swapping elements of two or more existing enzymes to create hybrid enzymes. For the latter, the elements being exchanged may be individual residues, secondary structural elements, subdomains, domains, or whole proteins (i.e., fusion proteins) (Nixon *et al.,* 1998).

This chapter explores the use of large exchanges of structure (subdomains, domains, and whole proteins) to evolve new function in proteins. Both rationally designed and combinatorial exchanges are explored, as both mirror methods nature employs to evolve new function. Judicious application of both strategies will likely prove to be important for advances in the field.

## II. Terminology

The terms ''domain'' and ''subdomain'' are used to describe proteins and fragments thereof. The term domain has been defined as ''a subregion of the polypeptide chain that is autonomous in the sense that it possesses all the characteristics of a complete globular protein'' (Schultz and Schirmer, 1979). The term subdomain usually refers to units smaller than a domain that can be described as having a particular function or structure (i.e., some rationale for being grouped together). What distinguishes the two from each other can often be confusing and the terms are frequently used interchangeably.

The problems in the use of the terms domain and subdomain result from their meaning being context dependent. What constitutes a subdomain for folding may not constitute a subdomain for a catalytic unit. For instance, the portion of a transcription factor responsible for binding DNA is most commonly called a ''DNA binding domain'' as it can bind DNA on it own, even though it lacks the complete function of a transcription factor because it is missing the activator domain. However, for an enzyme that can be divided into separate segments that are responsible for binding substrate and catalysis, these segments are commonly referred to as subdomains, even though the substrate binding subdomain may bind substrate on its own. The difference seems to

come down to more intimate contact and functional coupling between subdomains than between domains. To avoid confusion, this chapter will use the term domain very loosely to refer to both domains and subdomains.

The term ''domain swapping'' refers to genetic rearrangement of proteins by combining pieces of genes that code for domains or subdomains. This is distinguished from the term ''3D domain swapping'' that describes the process in which one domain of a multidomain protein breaks its noncovalent bonds with the other domains and is replaced by the same domain of an identical protein chain (Schlunegger *et al.,* 1997).

## III. Evolution of Proteins in Nature by Domain Swapping

The study and understanding of the natural evolution of protein function clearly has important implications for the design of *in vitro* evolution strategies. Because nature is blind to the rules and relationships of sequence, structure, and function, strategies that seem most applicable are combinatorial methods. Sequence and structural space are enormously large yet likely to be sparse in function. Nature must explore these spaces in a manner that preferentially examines areas relatively high in function. Thus, an understanding of how proteins evolve is an important step in rationally designing combinatorial protein engineering strategies.

It is generally accepted that nature evolves proteins for novel function by redesigning existing protein frameworks. This is supported by a number of lines of evidence: (1) the occurrence of gene duplications and recombination of genes and gene fragments is well established, (2) the relatedness of interspecies homologs observed by sequence and structural comparisons, (3) the distribution of motifs related by sequence and structure among disparate proteins, (4) the limited number of protein folds, with similar folds found in functionally dissimilar proteins, and (5) the deduction that sequence space is largely empty of function; hence, the *de novo* design of new proteins is too restrictive and inefficient a strategy to have evolved the current diversity and complexities of proteins, given the age of life on earth.

The creation of new function depends on minor and major changes in the sequence of existing genes. Minor changes include point mutations, deletions, and insertions of a few amino acids. Such minor changes can produce no effect, totally inactivate a protein, or lead to small or large changes in functionality. However, major changes in the sequences of genes through insertions of long sequences, tandem duplication, and circular permutations provide opportunities for major structural re-

arrangements and the evolution of entirely novel function. Major changes allow a protein to escape a local energy minimum and activate, in combination with other sequences, latent functionality. The importance of domain swapping for evolution is clearly illustrated by the evolution of *Escherichia coli* from *Salmonella* (Lawrence, 1997). None of the phenotypic differences between the two species can be attributed to point mutations. All the phenotypic changes have arisen by domain swapping.

The following sections explore nature's use of domain swapping to evolve new function. These include the formation of multifunctional proteins, tandem duplication, domain recruitment, and cicular permutation (Fig. 1). The evolution of several enzymes in the purine (Fig. 2) and pyrimidine (Fig. 3) *de novo* biosynthetic pathways, as well as other enzymes, are discussed as illustrative examples.

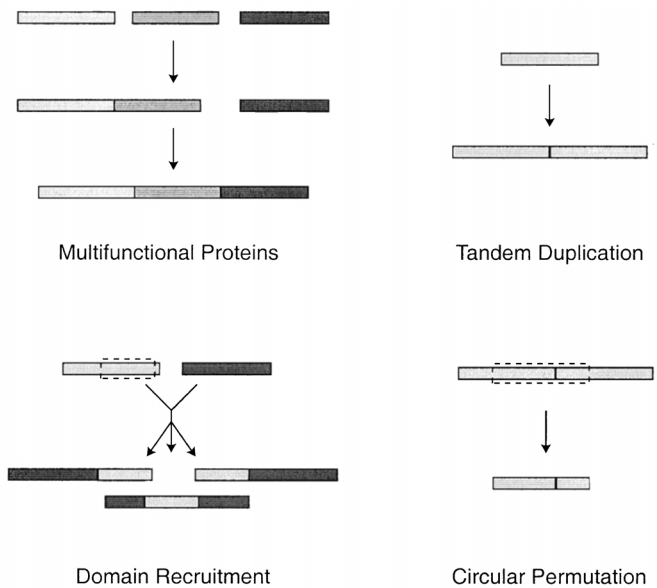



FIG. 1. Schematics of evolutionary mechanisms of domain swapping in nature. Multifunctional proteins arise from the fusion of the genes coding for individual enzymes. Often the individual domains of multifunctional proteins catalyze successive steps in metabolic pathways. In tandem duplication, a gene is duplicated and the 3′ end of one copy is fused in-frame to the 5′ end of the second copy. In domain recruitment, a functional unit (whole gene or gene fragment) from one gene is either inserted within or fused to an end of a second gene. Circular permuted genes are believed to arise via tandem duplication followed by introduction of new start and stop codons (Ponting *et al.,* 1995).

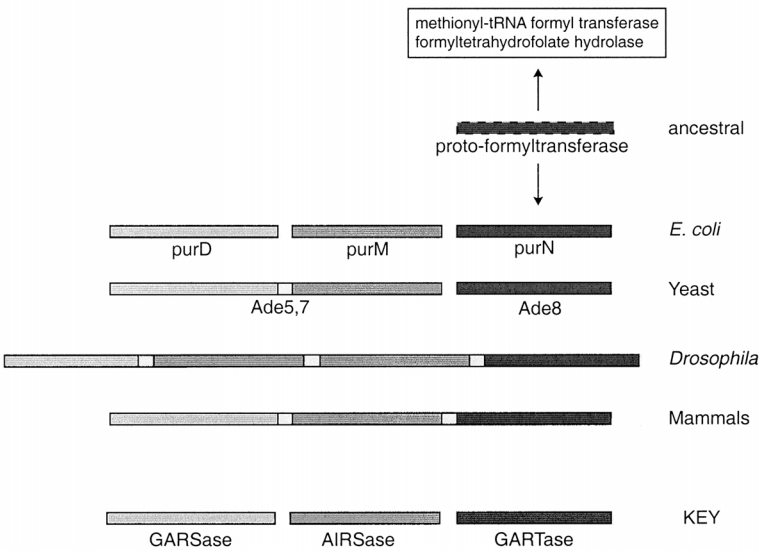



FIG. 2. Gene and domain organization of the glycinamide ribonucleotide synthetase (GARSase), aminoimidazole ribonucleotide synthetase (AIRSase), glycinamide ribonucleotide formyltransferase (GARTase), and related genes from different species. GARSase, AIRSase, and GARTase catalyze the second, fifth, and third steps in the *de novo* purine biosynthetic pathway. In mammals, chicken, and *Drosophila,* all enzymes are encoded by a single gene. The *Drosophila* gene differs by having an internal duplication of the AIRSase domain. In yeast, the GARSase and AIRSase are a single gene (Ade5,7) and GARTase is separate (Ade8). All three activities are encoded by separate genes in *E. coli.* Sequential, structural and functional similarities (see also Fig 4) suggest that fragments of the *E. coli* genes for GARTase, methionyl-tRNA formyltransferase, and formyltetrahydrofolate hydrolase have evolved from an ancestral formyltransferase and were incorporated into these genes by domain recruitment.

### A. *Multidomain Proteins*

Multidomain proteins tend to occur more frequently in eukaryotes than in prokaryotes. Often the eukaryotic counterpart to a set of individual prokaryotic enzymes that catalyze successive reactions is a single, multidomain protein. The theoretical advantages proposed for such an arrangement include (1) a geometry for the direct transfer of substrates from one active site to another, in a process known as substrate channeling, in order to increase the overall flux of the pathway, (2) the protection of intermediates that may be unstable in aqueous environments or may be acted on inappropriately by other enzymes, (3) the facilitation of interactions between domains for purposes of allosteric regulatory functions, and (4) the establishment of a fixed stoichiometric ratio of the

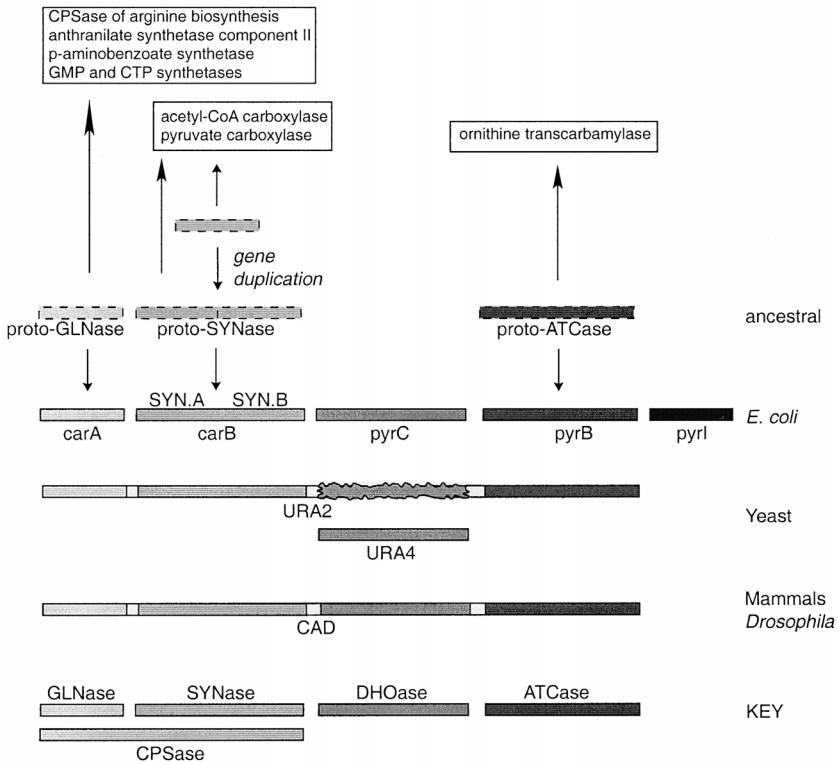



FIG. 3. Gene and domain organization of the carbamoyl-phosphate synthetase (CPSase), dihydroorotate dehydrogenase (DHOase), aspartate carbamoyltransferase (ATCase), and related genes from different species. CPSase, DHOase, and ATCase catalyze the first, third, and second steps in the *de novo* pyrimidine biosynthetic pathway. CPSase can be divided into domains with glutaminase (GLNase) and synthetase (SYNase) activities. In mammals and *Drosophila,* all activities are encoded by a single gene (CAD). The yeast URA2 gene codes for a similar length peptide, but the DHOase-like domain is inactive. In yeast, DHOase activity is encoded by the URA4 gene. All activities in *E. coli* are encoded by separate genes and include a regulatory submit for ATCase (pyrI). Sequential, structural, and functional similarities suggest that the CPSase and ATCase genes are evolutionarily related to other enzymes. GLNase is related to several enzymes that utilize cleavage of glutamine as the basis for catalysis. Acetyl-CoA carboxylase and pyruvate carboxylase contain domains with homology to the amino terminal half of SYNase, itself a product of gene duplication. This makes it likely that all three genes arose by domain recruitment from an ancestral proto-SYNase. Sequence similarities throughout the ornithine transcarbamylase and ATCase genes, both of which use carbamoyl phosphate as a substrate, suggest that they arose by duplication of an ancestral gene and divergent evolution.

enzymatic activities necessary for a sequential set of reactions (Srere, 1987; Hawkins and Lamb, 1995). Two examples of multidomain proteins are discussed: eukaryotic and prokaryotic enzymes in the *de novo* purine and pyrimidine biosynthetic pathways.

### *1. Purine Biosynthesis*

Purines are synthesized *de novo* from phosphoribosyl pyrophosphate. The first ten enzymatic steps lead to the formation of inosine monophosphate, from which adenosine monophosphate and guanosine monophosphate are then synthesized in two steps. The second, third, and fifth steps in this pathway are catalyzed by glycinamide ribonucleotide synthetase (GARS), glycinamide ribonucleotide formyltransferase (GART), and aminoimidazole ribonucleotide synthetase (AIRS), respectively. In *Escherichia coli* (Zalkin and Nygaard, 1999) and *Bacillus subtilis* (Ebbole and Zalkin, 1987) these enzymatic activities lie on separate proteins, whereas in human, chicken (Aimi *et al.,* 1990) and mouse (Kan *et al.,* 1993), a single gene locus codes for a trifunctional protein with the domain order GARS-AIRS-GART (Fig. 2). Sequence and structural data clearly indicate that the individual genes from bacteria and the gene fragments corresponding to the functional domains of mammals and chicken have a common evolutionary origin. Examination of the organization of the corresponding genes in *Drosophila* (Henikoff *et al.,* 1986) and yeast (Henikoff, 1986) suggests a two-step evolutionary pathway for the fusion of these genes. Initially, as seen in yeast, genes for GARS and AIRS activities were fused, with GART activity remaining on a single gene. Subsequently the GARS-AIRS and GART genes fused. In *Drosophila,* tandem duplication of the AIRS region has led to a protein with two tandem AIRS domains. How the genes for the GARS-AIRS-GART multienzymatic protein became fused is an open question. A reasonable model has been proposed that employed intron-intron mediated rearrangement as the evolutionary mechanism, based largely on the presence of introns at domain boundaries (Davidson and Peterson, 1997).

### *2. Pyrimidine Biosynthesis*

Carbamoyl-phosphate synthetase (CPSase), aspartate transcarbamoylase (ATCase), and dihydroorotase (DHOase) catalyze the first, second, and third steps in the pyrimidine *de novo* biosynthetic pathway, respectively. CPSase has two enzymatic functions that are found as separate units in bacteria: glutamine amidotransferase (GLNase) and synthetase (SYNase). GLNase transfers the amino group of glutamine to the catalytic site of SYNase, which in turn catalyzes the formation of carbamoyl phos-

phate in a complex reaction. GLNase- and SYNase-like domains are found in CPSase domains of higher species. CPSase, ATCase and CHOase are independent in bacteria but fused into multienzyme proteins in eukaryotes (CAD) in the order CPSase-DHOase-ATCase (Fig. 3). Unlike in the purine enzymes previously described, there is no intermediate level of gene fusions in species between bacteria and human to aid in the elucidation of the order of gene assembly. Interestingly, the yeast multifunctional enzyme URA2 contains an inactive DHOase-like domain between the CPSase and ATCase domains with an individual DHOase protein coded for on a separate operon (Souciet *et al.,* 1989). This observation is clearly suggestive of a gene duplication event.

Although channeling can be found in these multifunctional proteins, it seems that the creation of more sophisticated allosteric control was the driving force behind gene consolidation. Several lines of evidence indicate that URA2 channels carbamoyl phosphate from the catalytic site of CPSase to the catalytic site of ATCase (Belkaïd *et al.,* 1987; Hervé *et al.,* 1993) and partial channeling has been observed in mammalian CAD (Mally *et al.,* 1980). However, evidence for channeling has been observed in the hyperthermophilic archaeon *Pyrococcus abyssi* CPSase, ATCase, and DHOase enzymes, which are separate enzymes (Purcarea *et al.,* 1999), and in a truncated CAD with ATCase as a separate domain (Davidson *et al.,* 1993). Studies on independent domains of yeast and mammalian CAD and yeast-mammalian chimeric CAD indicate that catalytic activity resides entirely within the independent domains but that regulation requires interdomain interactions (Serre *et al.,* 1999). Thus, it seems that the evolutionary driving forces behind consolidation of these genes into a multifunctional enzyme were incorporating more sophisticated allosteric control and, presumably, coordinate expression.

### *B. Tandem Duplication*

Duplication of genes or gene fragments in tandem is one method observed in nature to evolve new function. Its occurrence in so many proteins clearly supports its advantages: (1) increased stability, (2) new cooperative functions, (3) formation of a binding site in a cleft, (4) production of multiple binding sites in series resulting in more efficient or specific binding, and (5) growth of long repetitive structures in modular proteins (McLachlan, 1987). The two possible evolutionary outcomes for such duplications are domains that function in isolation and domains that interact. The former are likely to exist initially after duplication, with interactions, if evolutionarily favorable, developing concomitant with genetic changes in the two domains. Tandem duplication is also a

proposed first step in a mechanism for circular permutation of proteins (Ponting and Russell, 1995).

*1. Pyrimidine Biosynthesis*

CPSase catalyzes the formation of carbamyl phosphate from glutamine, bicarbonate, and two equivalents of ATP. The biosynthesis involves four partial reactions. GLNase catalyzes the formation of ammonia from glutamine. The remaining three partial reactions are catalyzed by SYNase. Bicarbonate is activated by ATP to form carboxyphosphate, which reacts with ammonia to form carbamate. The ATP-dependent phosphorylation of carbamate results in the production of carbamyl phosphate.

Comparison of the N-terminal and C-terminal fragments of the *E. coli carB* gene, which codes for the large subunit of CPSase, suggests that it arose from tandem duplication of a smaller ancestral gene (Nyunoya and Lusty, 1983) (Fig. 3). The homology is especially strong between residues 1-400 and 553-933, which exhibit 39% identity and 64% homology with only four minor adjustments for insertions and deletions. Although these two domains are commonly referred to as CPS.A and CPS.B, they will be referred to as SYN.A and SYN.B, respectively, to avoid confusion with CPSase. Since the two ATP-dependent reactions of SYNase utilize similar substrates (bicarbonate and carbamate), one might assume that tandem duplication of an ancient SYNase half-domain occurred to catalyze these two reactions at different sites. However, the picture is more complicated than that.

Remarkably, the isolated GLNase-SYN.A domains and a GLNase-SYN.B fusion catalyze the entire series of reactions involved in glutamine-dependent carbamyl phosphate synthesis (Guy and Evans, 1996). Also, a truncated URA2 gene with the GLNase and SYN.A domains removed still codes for an active enzyme that retains the ability to channel carbamoyl phosphate and be allosterically regulated by UTP (Serre *et al.,* 1999). Furthermore, *Pyrococcus* archaebacteria posses a SYNase that is less than half the size of other SYNases (Purcarea *et al.,* 1996; Durbecq *et al.,* 1997). All this evidence suggests that the ability to synthesize carbamylphosphate predates tandem duplication.

This begs the question as to why the SYNase domain has evolved by tandem duplication, since both SYN.A and SYN.B seem to be functionally equivalent. However, evidence that the two ATP-dependent reactions occur at different sites in SYNase is very strong (Guy *et al.,* 1996), and it has been subsequently shown that the functional form of SYN.A and SYN.B domains in the absence of the other is a homodimer (Guy *et al.,* 1998). It appears that it is somehow advantageous to have separate, but

proximal, domains for the ATP-dependent partial reactions. Although this can occur intermolecularly, fusing the two domains together ensures this geometry and stoichiometry. That both SYN.A and SYN.B can catalyze the two reactions is probably an evolutionary leftover from an ancestral single domain SYNase, like that of *Pyrococcus,* which catalyzed both reactions. This view of the function of SYN.A and SYN.B supports an evolutionary pathway by tandem duplication, followed by optimization by point mutation.

### 2. *Proteases*

The pepsin and chymotrypsin families of proteases are believed to have evolved by tandem duplication. Members of the pepsin family of proteases are composed of two large $\beta$-sheet domains, with their active sites in between composed of catalytic aspartate residues from each domain. Aside from structural similarities, the domains have sequence homology, particularly around the active site aspartate residues. This clearly suggests that the two domains arose by tandem duplication (Tang *et al.,* 1978). It seems that tandem duplication in pepsin allowed the creation of a binding site cleft. Members of the chymotrypsin family of proteases are composed of two homologous $\beta$-barrel domains packed together asymmetrically. The important residues for catalysis are His57, Asp102, and Ser195. The domain interface forms the active site with the first two catalytic residues on the N-terminal barrel and the last one on the C-terminal barrel. Superposing the $\alpha$-carbon atoms of the two domains shows forty-six carbons fitting within a root-mean-square distance of 2.4 Å (McLachlan, 1979). This suggests that chymotrypsin evolved by tandem duplication, although no detectable sequence similarity remains.

### 3. *Modular Proteins*

Modular proteins are described as displaying a beads-on-a-string organization of domains or modules because the individual domains (beads) function independently but are connected via the peptide backbone (string). Such an organization has been found in many eukaryotic proteins such as fibronectin, collagen XII, factors involved in blood clotting and fibrinolysis, muscle associated proteins such as twitchin and titin, and cell surface receptors (Baron *et al.,* 1991; Doolittle and Bork, 1993; Hegyi and Bork, 1997). The demonstration that domain folding in prokaryotes is posttranslational and domain folding in eukaryotes is cotranslational and sequential may have been critical in the evolution of modular proteins by allowing tandem duplication events to create immediately foldable protein structures (Netzer and Hartl, 1997).

### C. *Domain Recruitment*

Domain recruitment is a mechanism by which functional units from one protein are ''recruited'' by another protein. This mixing and matching of existing domains constitute an efficient method to evolve proteins. An apt explanation of this is a variation on the saying that ''a thousand monkeys typing at a thousand typewriters would eventually reproduce the works of Shakespeare.'' The monkeys would obviously work much more efficiently if, once they managed to type a coherent word, sentence, paragraph, or chapter, they could reproduce these words, sentences, paragraphs, and chapters with a single keystroke.

#### 1. *Purine Biosynthesis*

The *E. coli* genes for glycinamide ribonucleotide transformylase (PurN), methionyl-tRNA formyltransferase (FMT), and formyltetrahydrofolate hydrolase (purU) catalyze the transfer of the formyl group from formyltetrahydrofolate to glycinamide ribonucleotide, methionyl-tRNA, and water, respectively. Sequence homology, conservation of catalytic residues, and structural similarities between PurN and FMT (see Fig. 4) suggest a common ancestor for the subdomain responsible for formyltetrahydrofolate binding and deformylation.

The N-terminal domain of PurN is structurally homologous to other N-terminal domains of GAR synthetase (PurD) and $N^5$-carboxyaminoimidazole ribonucleotide synthetase (PurK)—*E. coli* enzymes that are involved in purine biosynthesis (Wang *et al.,* 1998; Thoden *et al.,* 1999). This N-terminal domain is responsible for ribonucleotide binding and adopts a Rossman-fold (Rossmann *et al.,* 1974) that is common in nucleotide binding proteins. As a result of the structural similarity in the N-terminus of three enzymes in the *de novo* purine biosynthetic pathway, it is tempting to speculate that these enzymes have evolved by domain recruitment. However, the very low sequence homology makes this difficult to prove. If domain recruitment did occur, it must have occurred with an early ancestral nucleotide-binding motif for such a divergence in sequence to accumulate.

#### 2. *Pyrimidine Biosynthesis*

ATCase and ornithine carbamoyltransferase (OTCase) catalyze analogous reactions. ATCase transfers the carbamoyl moiety from carbamoyl phosphate to aspartate, and OTCase transfers the carbamoyl moiety from carbamoyl phosphate to ornithine. They both share a common N-terminal functional domain, which binds carbamoyl phosphate. The C-terminal domains of these enzymes are structurally similar but have

a

```
purU MHSLQRKVLRTICPDQKGLIARITNICYKHELNIVQNNEFVDHRTGRFFMRTELEGIFNDSTLLADLDSA
purN
FMT

purU LPEGSVRELNPAGRRRIVILVTKEAHCLGDLLMKANYGGLDVEIAAVIGNHDTLRSLVERFDIPFELVSH
purN            MNIVVLISGNGSNLQAIIDACKTNKIKGTVRAVFSNKADAFGLERARQAGIATHTLIAS
FMT   MSESLRIIFAGTPDFAARHLDALLSSGHNVVGVFTQPDRPAGRGKKLMPSPVKVLAEEKGLPVFQPVS-

                                                        *  *
purU EGLTRNEHDQKMADAIDAYQPDYVVLAKYMRVLTPEFVARFPNK-IINIHHSFLPAFIGARPYHQAYERG
purN AFDSREAYDRELIHEIDMYAPDVVVLAGFMRILSPAFVSHYAGR-LLNIHPSLLPKYPGLHTHRQALENG
FMT  ---LRPQENQQLVAEL---QADVMVVVAYGLIL-PKAVLEMPRLGCINVHGSLLPRWRGAAPIQRSLWAG

                     *
purU VKIIGATAHYVNDNLDEGPIIMQDVIHVDHTYTAEDMMRAGRDVEKNVLSRALYKVLAQRVFVYGNRTII
purN DEEHGTSVHFVTDELDGGPVILQAKVPVFAGDSEDDITARVQTQEHAIYPLVI----SWFADGRLKMHEN
FMT  DAETGVTIMQMDVGLDTGDMLYK----LSCPITAEDTSGTLYDKLAELGPQGLITTLKQLADGTAKPEVQ

purU L
purN AAWLDGQRLPPQGYAADE
FMT  DETLVTYAEKLSKEEARIDWSLSAAQLERCIRAFNPWPMSWLEIEGQPVKVWKASVIDTATNAAPGTILE

purU
purN
FMT  ANKQGIQVATGDGILNLLSLQPAGKKAMSAQDLLNSRREWFVPGNRLV
```

b

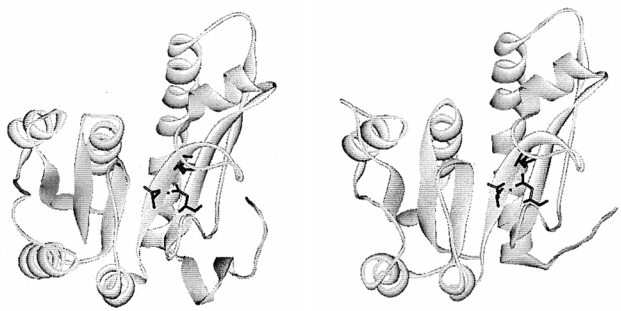



FIG. 4. (a) Sequence alignment of the *E. coli* genes for glycinamide ribonucleotide formyltransferase (*purN*), methionyl-tRNA formyltransferase (*FMT*), and formyltetrahydrofolate hydrolase (*purU*). Identical amino residues are shaded in gray. The key active site residues of PurN are indicated by (*). The N-termini of all three enzymes do not align well, but a sequence alignment of the N-termini of *purN* and *FMT* based on a structural alignment (not shown here) has been proposed (Schmitt *et al.,* 1996). (b) Crystal structures of PurN (Almassy *et al.,* 1992) and the domain of FMT (Schmitt *et al.,* 1996) homologous to PurN. The C-terminal domain of FMT is not shown. Side chains of the key active site residues of PurN (Asn106, His108 and Asp144) and FMT (Asn108, His110 and Asp 146) are shown in black.

very divergent sequences. Based on multiple sequence alignments, the simplest explanation is that the two carbamoyltransferases arose not by domain recruitment but by duplication of an ancestral gene followed by divergent evolution (Labedan *et al.,* 1999) (Fig. 3).

In contrast, sequence relationships between CPSase and enzymes outside the pyrimidine pathway suggest domain recruitment as an evolutionary mechanism. Homology between the N-terminal half of SYNase and subdomains of acetyl-CoA carboxylase (Takai *et al.,* 1988) and pyruvate carboxylate (Lim *et al.,* 1988) suggest that these enzymes evolved by domain recruitment. Similarly, GLNase is sequentially and functionally related to domains in several enzymes that utilize cleavage of glutamine as the basis for catalysis. GLNase probably evolved from an ancestral glutaminase that was duplicated and inserted into other proteins by domain recruitment (Davidson *et al.,* 1993).

### *3. Insertion of Domains*

Identification of homologous domains can be confounded by irregularities in the relationships between linear sequences and three-dimensional structures (Russell and Ponting, 1998). This arises primarily from the assumption that domains form distinct compact structures that are coded for by a single, uninterrupted stretch of DNA. Although end-to-end fusion is the predominant manner of linking two domains, the increasing number of exceptions to this rule suggests an evolutionary tolerance or even advantage to domain insertions. Although a protein created by a domain insertion would be expected to incur protein folding problems for entropic reasons, a number of advantages for insertions have been proposed (Russell, 1994). Fixing both ends of an inserted domain offers the advantages of a more compact structure to avoid proteolysis and a more rigid structure to fix the relative spatial orientation of the two domains in order to make functional combinations more viable. Finally, a possible reason why one third of all proteins have their N- and C-termini proximal (Thornton and Sibanda, 1983)—a proximity that is required for a domain to be inserted—could be to allow for domain insertion as an evolutionary mechanism.

Protein structure determinations have identified several examples of one domain inserted within another. One example is the *E. coli* DsbA protein, which catalyzes the formation of disulfide bonds in the periplasm. The enzyme consists of two domains: a thioredoxin-like domain that contains the active site, and an inserted helical domain similar to the C-terminal domain of thermolysins (Martin *et al.,* 1993). The inserted domain forms a cap over the active site, suggesting that it plays a role in binding to partially folded polypeptide chains before oxidation of

cysteine residues. It has been proposed that DsbA's potent disulfide catalyzing ability results from the relative motions of the two domains, though recent evidence shows that the motions are independent of the redox state of the active site and thus may only be related to substrate binding (Guddat *et al.,* 1998). It seems likely that DsbA arose by the insertion of a domain into an ancestral thioredoxin-like domain in order to provide improved substrate binding, raising the catalytic power of the enzyme. Of course, an alternative explanation that must be considered is that thioredoxin is a result of the deletion of the helical domain from an ancestral DsbA protein (Russell, 1994). Other examples of domain insertion have been summarized elsewhere (Russell, 1994; Russell and Ponting, 1998).

### *D. Circular Permutations*

A circularly permuted protein has its original N- and C-termini fused and new N- and C-termini created by a break elsewhere in the sequence. Circular permutation can be thought of as a form of domain swapping in which the C-terminal fragment of a gene has been moved to the beginning of that gene. Circular permutation is believed to result when tandem duplication of a gene is followed by the introduction of a new open reading frame within the first repeat and a new stop codon within the second repeat (see Fig. 1). This model is supported by the observation of tandem duplication in prosaposins and adenine-$N^6$ DNA methyltransferase, genes for which circular permutated variants are known (Ponting and Russell, 1995; Jeltsch, 1999).

The first *in vitro* construction of a circular permuted protein was carried out on bovine pancreatic trypsin inhibitor by chemical means (Goldenberg and Creighton, 1983). Since then, a number of circular permuted proteins have been constructed, primarily by genetic methods (for a review, see Heinemann and Hahn, 1995). These studies have shown that circular permuted proteins very often fold up into stable, active proteins. Comparisons of primary and tertiary structures within several protein families have led to the conclusion that circular permutation occurs in natural protein sequences. A recent review noted evidence of naturally occurring circular permutations in lectins, bacterial $\beta$-gluconases, aspartic proteases, glucosyltransferases, $\beta$-glucosidases, surface layer homology domains, transaldolases, and C2 domains (Lindqvist and Schneider, 1997). Another recent example is adenine-$N^6$ DNA methyltransferases (Jeltsch, 1999).

An interesting case is a family of plant aspartic proteases in which sequence alignments have predicted evolution by insertion of a saposin-

like domain that itself has been circularly permuted. Circularly permuted variations of saposin-like domains have been termed ''swaposins'' and were first identified in plant aspartic proteases (Guruprasad *et al.,* 1994). The four $\alpha$-helices of saposin have been permuted in swaposin such that the order in swaposin is helices 3-4-1-2. The cDNA sequence that codes for saposins suggests a mechanism for the evolution of swaposin (see Fig. 1) (Ponting and Russell, 1995). Saposin domains are formed by cleavage of prosaposin, which consists of four tandem repeats of saposin domains. This is suggestive of an evolutionary event in which swaposin was created by the creation of a new ORF whose N-terminus is helixes 3 and 4 followed by a linker and helices 1 and 2.

The evolutionary advantages of circular permutation are not clear from the current examples of naturally occurring circularly permuted proteins, as most have conserved function. However, examples of engineered circularly permuted proteins have been shown to result in higher stability (Wieligmann *et al.,* 1998), effect quaternary domain assembly in multi-domain proteins (Wright *et al.,* 1998), and increase antitumor activity in fusions of interleukin 4 and *Pseudomonas* exotoxin by changing of their topology (Kreitman *et al.,* 1995).

## IV. Domain Swapping for Protein Engineering

The following sections describe examples of domain swapping performed to engineer proteins with desired function and properties. Studies are categorized under the intended goals of the protein engineering: engineering of allosteric regulation, creation of activators and inhibitors, improvement in stability or expression, modification of substrate specificity, improvement of catalytic efficiency, alteration of multimodular synthetases, improvement of therapeutic properties, creation of molecular biosensors, and creation of novel enzymes. Some of this work has also been explored in two recent reviews on hybrid enzymes (Nixon *et al.,* 1998; Beguin, 1999).

Bifunctional enzymes, generally end-to-end fusions of two domains or proteins in which the two proteins have distinct and independent functions, are mentioned only briefly in order to devote more space to domain-swapped enzymes that have a higher degree of interaction. Recent reviews that discuss the use of bifunctional enzymes in more detail include those on recombinant immunotoxins (Reiter and Pastan, 1998), signal sequences for cellular targeting (Martoglio and Dobberstein, 1998), gene fusions for phage display (Hoogenboom *et al.,* 1998; Rodi and Makowski, 1999) or bacterial display (Georgiou *et al.,* 1997), affinity purification (Nilsson *et al.,* 1997), improved recombinant protein pro-

duction (Baneyx, 1999), *in vivo* detection of proteins (Tsien, 1998), chimeric transcription activators (Clackson, 1997), artificial systems for assembling enzymes (Beguin, 1999), and two- and three-hybrid systems (Drees, 1999). Similarly, coverage will be minimal of hybrids between interspecies homologs that were constructed for purposes of understanding differences in their properties.

### A. *Engineering of Allosteric Regulation*

The engineering of new allosteric regulations by domain swapping or other methods has not been extensively explored, but two studies suggest its general feasibility. First, insertion of TEM $\beta$-lactamase into two different loops of the *E. coli* maltose binding protein (MalE) have been found to result in $\beta$-lactamase activity that is more stable to urea denaturation in the presence of maltose than in its absence (Betton *et al.,* 1997). Second, heterotetramers of L-lactate dehydrogenase (LDH), which naturally forms homotetramers, and an engineered form of LDH with the substrate specificity of malate dehydrogenase (MDH) exhibit allosteric properties (Fushinobu *et al.,* 1998). MDH activity in these LDH/MDH mixed tetramers was found to increase in the presence of oxamate, which was found to bind to LDH in the complex.

An enzymatic two-hybrid system, termed protein-fragment complementation assay (PCA) (Pelletier *et al.,* 1998) has potential for engineering of allosteric regulation. Developed to link protein interactions to enzymatic activity, PCA has elements of both domain swapping and regulation of enzyme activity. PCA consists of two designed fragments of murine dihydrofolate reductase (mDHFR), each of which is fused to a target domain (i.e., two domains that potentially interact). These two mDHFR domains were engineered by circular permutation followed by cleavage at a designed location. The two mDHFR domains without the target domains do not associate efficiently enough to produce DHFR activity. Association between the two target domains drives the association of the mDHFR fragments. Since mDHFR activity is essential for complementation of *E. coli* grown in the presence of the anti-folate drug trimethoprim, domains that interact can be selected by functional complementation. Allosteric regulation dependent on a small molecule can be created if the association between the two interacting domains is dependent on a small molecule. Such a case was demonstrated, though not for the purposes of regulating DHFR activity, with the FK506 binding protein (FKBP) and a domain of the FKBP-rapamycin binding protein (FRB) fused to the mDHFR fragments (Remy *et al.,* 1999). FKBP and FRB only associate in the presence of rapamycin. Hence, the mDHFR

PCA can be used for quantitative pharmocological analysis of protein–protein and protein–small molecule binding *in vivo*. However, if the mDHFR domains are swapped for complementary fragments of another enzyme, then perhaps a PCA with ligand-dependent interacting domains such as FKBP and FRB can be used to regulate the activity of this enzyme.

### *B. Creation of Activators and Inhibitors*

In *Drosophila,* three proteins with epidermal growth factor-like domains modulate the function of the epidermal growth factor (EGF) receptor. Spitz (Spi) is a potent activator, Vein (Vn) is a moderate activator, and Argos (Aos) is an inhibitor. Aside from the EGF-like domains, these three proteins are structurally unrelated, suggesting that they evolved by EGF-like domains being fused to ancestral Spi, Vn, and Aos. Swapping of the EGF-like domains from Spi and Aos into Vn resulted in hybrids that behaved *in vitro* and *in vivo* like Spi and Aos (Schnepp *et al.,* 1998). This work demonstrates the feasibility of swapping homologous structures between proteins containing other domains that are structurally unrelated.

### *C. Improvement in Stability or Expression*

#### *1. Pyrroloquinoline Quinone Glucose Dehydrogenase*

Pyrroloquinoline quinone glucose dehydrogenase (PQQGDH), pyrroloquinoline quinone ethanol dehydrogenase (PQQEDH), and pyrroloquinoline quinone methanol dehydrogenase (PQQMDH) use pyrroloquinoline quinone (PQQ) as a bound cofactor. These enzymes have potential for use in diagnostics and as biosensors for organic compounds, such as using PQQGDH as an enzyme sensor for glucose. Based on the crystal structure of PQQMDH (Ghosh *et al.,* 1995) and predicted structures of PQQGDH and PQQEDH (Cozier and Anthony, 1995a; Cozier *et al.,* 1995b), all are $\beta$-propeller proteins with eight W-motifs (propellers) of four $\beta$-sheets each. PQQGDH are found in a variety of bacteria and, despite their sequence homology, have a great deal of variety in PQQ binding stability, thermal stability, and substrate specificity (Yoshida *et al.,* 1999).

To delineate regions of the protein responsible for the observed differences and create PQQGDH enzymes with desired substrate activities and stability, Sode and co-workers have constructed a variety of domain-swapped hybrids between the *E. coli* and *Acinetobacter calcoaceticus* PQQGDHs. They identified the region between 32% and 59% of the

N-terminus as being responsible for EDTA tolerance (PQQ binding stability) in the *A. calcoaceticus* enzyme. Based on this data, a domain-swapped enzyme was constructed that exhibited *E. coli* enzyme specificity and *A. calcoaceticus* EDTA tolerance (Yoshida *et al.,* 1999). In addition, the C-terminal 3% of the *A. calcoaceticus* enzyme was found to confer thermal stability to the *E. coli* enzyme when swapped for the *E. coli* enzyme's C-terminal 3% (Sode *et al.,* 1995). Interestingly, the 32%–59% region also conferred thermal stability. A domain-swapped enzyme containing the 32%–59% and C-terminal 3% regions of *A. calcoaceticus* but elsewhere derived from *E. coli* had the highest thermal stability while maintaining an EDTA tolerance that was essentially that of *A. calcoaceticus* (Yoshida *et al.,* 1999).

This work utilizes a functional definition of domains rather than a topological one. Mapping the swapped regions onto the predicted structure of PQQGDH shows that the region between 32% and 59% correlates to the second, third, and fourth W-motifs of the eight W-motifs, and the C-terminal 3% correlates to part of the second-to-last $\beta$-sheet in the eight W-motif. It is important to note that although the engineered enzyme had the desired thermal stability, EDTA tolerance, and specificity, activity toward glucose had decreased to 6% of the level of the *E. coli* enzyme as judged by $V_{max}/K_m$. This correlated with an increase in random structure as judged by circular dichroism. Large reductions in activity are likely to be common in domain-swapped enzymes with lower levels of homology. However, directed evolution strategies to improve activity, which have been particularly successful in improving the activity of poor enzymes (Arnold and Volkov, 1999), will likely be successful in fine-tuning domain-swapped enzymes.

### 2. *Other Examples*

The three-dimensional structure of human extracellular superoxide dismutase (EC-SOD) is unknown. Studies of structure–function relationships have been severely limited by its poor production in mammalian cell lines and failure to be expressed in prokaryotic and yeast systems. In contrast, extra- and intracellular Cu- and Zn-containing superoxide dismutases (CuZn-SOD) are expressed very well in *E. coli* and yeast. CuZn-SOD is homologous to a large interior fragment of EC-SOD, but lacks its extra N-terminal and C-terminal domains. Fusions of either the N-terminal domain of EC-SOD or both the N- and C-terminal domains of EC-SOD to CuZn-SOD resulted in a domain-swapped enzyme that expressed well and whose characteristics resemble EC-SOD (Stenlund and Tibell, 1999).

Domain swapping can also result in the acquisition of proteolytic stability. Insertion of TEM $\beta$-lactamase into loops of the *E. coli* maltose binding protein (MalE) results in fusions that are less susceptible to proteolysis than simple end-to-end fusion (Betton *et al.,* 1997). The stability of the $\beta$-lactamase activity in the insertion proteins to urea induced denaturation was greater in the presence of maltose than in its absence. In addition, fusion partners, particularly MalE, can greatly improve the solubility and expression of aggregation-prone proteins (Kapust and Waugh., 1999). Engineered circular permutation almost always results in a slight to significant decrease in stability, but a recent exception to the rule has been found. Circular permutation of eye lens $\beta$B2-crystallin resulted in a modest improvement of stability to urea denaturation, with a midpoint of transition of 2.1 M compared to 1.9 M for the wild-type crystallin (Wieligmann *et al.,* 1998). The increased stability was attributed to a more compact structure of the individual domains of the protein.

### *D. Modification of Substrate Specificity*

#### *1. Chimeric Restriction Endonucleases*

One of the greatest successes in utilizing domain swapping for protein engineering has been the creation of chimeric restriction endonucleases by Chandrasegaran and co-workers (Fig. 5) (Chandrasegaran and Smith, 1999). Although most known restriction enzymes recognize a particular DNA sequence and then cleave the DNA within that sequence, type IIS restriction enzymes recognize a particular site and then cleave the DNA a certain number of bases away from that site, irrespective of the cleavage site's sequence. This property is suggestive of two domains: a DNA-binding domain responsible for recognition and specificity, and a non-specific endonuclease domain. The type IIS restriction enzyme *Fok*I has been shown by biochemical experiments (Li *et al.,* 1992; Li *et al.,* 1993) and confirmed by recently solved crystal structures (Wah *et al.,* 1997; Wah *et al.,* 1998) to display such a domain structure. This structural arrangement suggests that *Fok*I evolved by random fusion of a nonspecific endonuclease domain ($F_N$) and a DNA-binding domain.

A systematic method for constructing restriction enzymes that recognize a given desired sequence has obvious applications in molecular biology. More important, the construction of a restriction enzyme that specifically recognizes sequences of 16 to 18 bp, long enough to make it probable that such a sequence is unique in a genome, has important applications in gene therapy and the construction of transgenic animals.

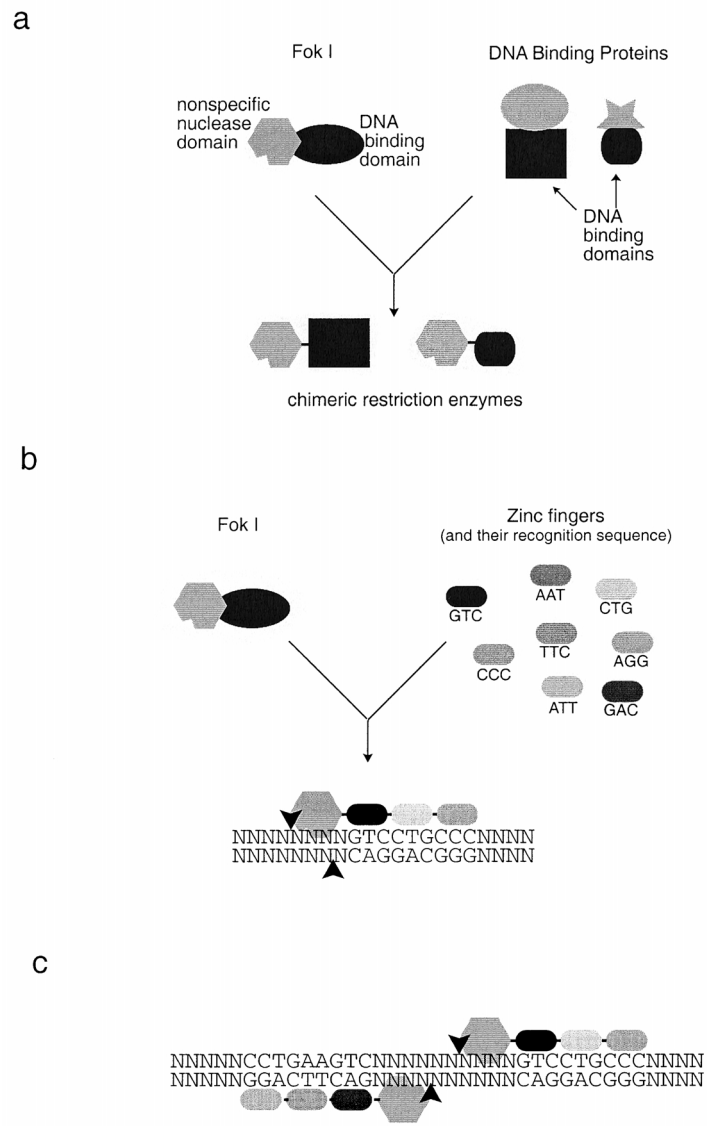



FIG. 5. Schematic of the construction and function of chimeric restriction endonucleases (Chandrasegaran and Smith, 1999). (a) The type IIS restriction enzyme *Fok*I has a nonspecific endonuclease domain ($F_N$) and a DNA-binding domain. Swapping *Fok*I's DNA binding-domain for that of another DNA binding-domain results in a chimeric restriction enzyme with new specificity. The flexibility resulting from less intimate contact between the $F_N$ domain and its new DNA-binding domain translates into DNA cleavage at several locations near the binding site. (b) The modular nature and tunable specificity of zinc

Chandrasegaran and co-workers have made fine progress toward this goal by creating chimeric restriction enzymes using *Fok*I in which the natural DNA-binding domain has been swapped for a variety of DNA-binding domains, particularly zinc fingers (Chandrasegaran and Smith, 1999).

Functional chimeras have been constructed between $F_N$ and the *Drosophila* Ubx homeodomain (Kim and Chandrasegaran, 1994), the DNA-binding domain of the transcription factor Sp1 (Huang *et al.,* 1996), yeast Gal4 (Kim *et al.,* 1998), and zinc-finger DNA-binding motifs (Kim *et al.,* 1996; Kim *et al.,* 1997). Zinc-finger motifs hold the most promise due to their modular nature (Fig. 5b). Crystal structures of zinc finger bound to DNA indicate that each zinc finger binds specifically to three bases (Pavletich and Pabo, 1991). This has stimulated efforts for creating a library of zinc fingers specific for each possible codon (Greisman *et al.,* 1997; Isalan *et al.,* 1997). Although fused zinc fingers do not bind totally independently, the linkage of several zinc fingers is believed to be an excellent framework to build DNA-binding proteins specific for sites up to 18 bp in length (Chandrasegaran and Smith, 1999). Sequence-specific cytosine methyltransferases were created by fusing zinc fingers and a CpG-specific DNA methyltransferase (Xu and Bestor, 1997).

A detailed study of a fusion of $F_N$ and the designed zinc finger ΔQNK (a fusion of three zinc fingers) has demonstrated that the fusion does not significantly change the sequence specificity or binding affinity of the zinc fingers (Smith *et al.,* 1999). The sequence specificity of the $F_N$-Δ*QNK* chimeric endonuclease is a reflection of that of the Δ*QNK*. $F_N$-Δ*QNK* had no activity on some 9-bp sites with only one base difference from its target site. However, other sites with one base change were cleaved 16% as efficiently as the target site, as judged by the initial rate of cleavage. All chimeric $F_N$ restriction enzymes described to date show nonspecific nucleotide activity, particularly at high $MgCl_2$ concentration and have multiple cleavage products (Kim *et al.,* 1998). This presumably results from the lack of interaction between the two domains in the chimera to ''lock'' the $F_N$ domain in a fixed orientation relative to the DNA-binding domain: an interaction that is present in the natural *Fok*I enzyme. Constructing $F_N$-fusions with six zinc fingers in order to recog-

←

fingers potentially enables the construction of restriction enzymes for any desired sequence by the fusion of multiple zinc fingers together in a specific order. An $F_N$-three zinc finger chimeric restriction enzyme recognizing a nine-bp sequence is shown. (c) $F_N$-three zinc finger chimeric restriction enzymes that nick DNA but do not make double stranded breaks can be used in pairs to specifically recognize a noncontinuous 18-bp sequence.

nize a 18-bp site may achieve very tight binding, but at the expense of specificity. Interestingly, a zinc finger-$F_N$ fusion with only DNA-nicking activity would allow circumvention of this problem. The use of two chimeras with nicking activity and different specificities potentially allows the specific cleavage within a noncontinuous 18-bp site (Fig. 5c) (Chandrasegaran and Smith, 1999).

### 2. *Proteases*

Members of the chymotrypsin family of serine proteases are composed of two homologous $\beta$-barrels that are believed to have evolved from an ancient gene duplication event (McLachlan, 1979). Numerous crystal structures of chymotrypsin-like proteases have shown that the two $\beta$-barrels stack together at roughly 90° to each other and are structurally conserved. The catalytic residues His57, Asp102, and Ser195 are at the interface of the two barrels, with His57 and Asp102 being from one domain and Ser195 being from the other. Variations between members of the chymotrypsin family primarily manifest themselves in surface loops near the active site that determine substrate specificity (Perona and Craik, 1997). Sites important for binding peptide substrates are distributed between the two domains.

Trypsin and factor Xa (fXa) are two members of the chymotrypsin family that have 38% sequence identity on the amino acid level and have distinguishable substrate specificities. Recently, the N-terminal $\beta$-barrel of fXa and the C-terminal $\beta$-barrel of trypsin were fused at a rationally designed site in the linker region between the two domains in order to create a hybrid fXa-trypsin protease (Hopfner *et al.,* 1998). The fXa-trypsin hybrid was highly active and more active than either parent on three of the ten substrates assayed, as determined by $k_{cat}/K_m$. For most substrates, the activity of fXa-trypsin was an admixture of the two parents, probably because trypsin had higher activity than fXa for all the substrates tested.

However, some of the most important results that impinge on domain swapping derive from the crystal structure of fXa-trypsin, which was determined to 2.15 Å. The general structural features of the chymotrypsin family are conserved between the respective parent molecules, including the core elements and the residues of the active site. Surface elements, particularly loops, exhibited much greater variability. Most important is the fact that although approximately 30% of the residues at the interface of the two domains differ between fXa and trypsin—presenting the possibility of unfavorable interactions that negatively effect folding and activity—these residues generally retained their structure with minor adjustments to accommodate mismatches. That plasticity

apparently did not grossly affect stability or activity bodes well for a strategy of domain swapping that exchanges homologous structures irrespective of sequence homology. These results suggest that such exchanged structures are flexible enough to accommodate mismatches while generally retaining their fold. This possibly illustrates how large changes in sequence space can occur in nature, enabling the escape from local energy minimum and furthering the evolution of novel function.

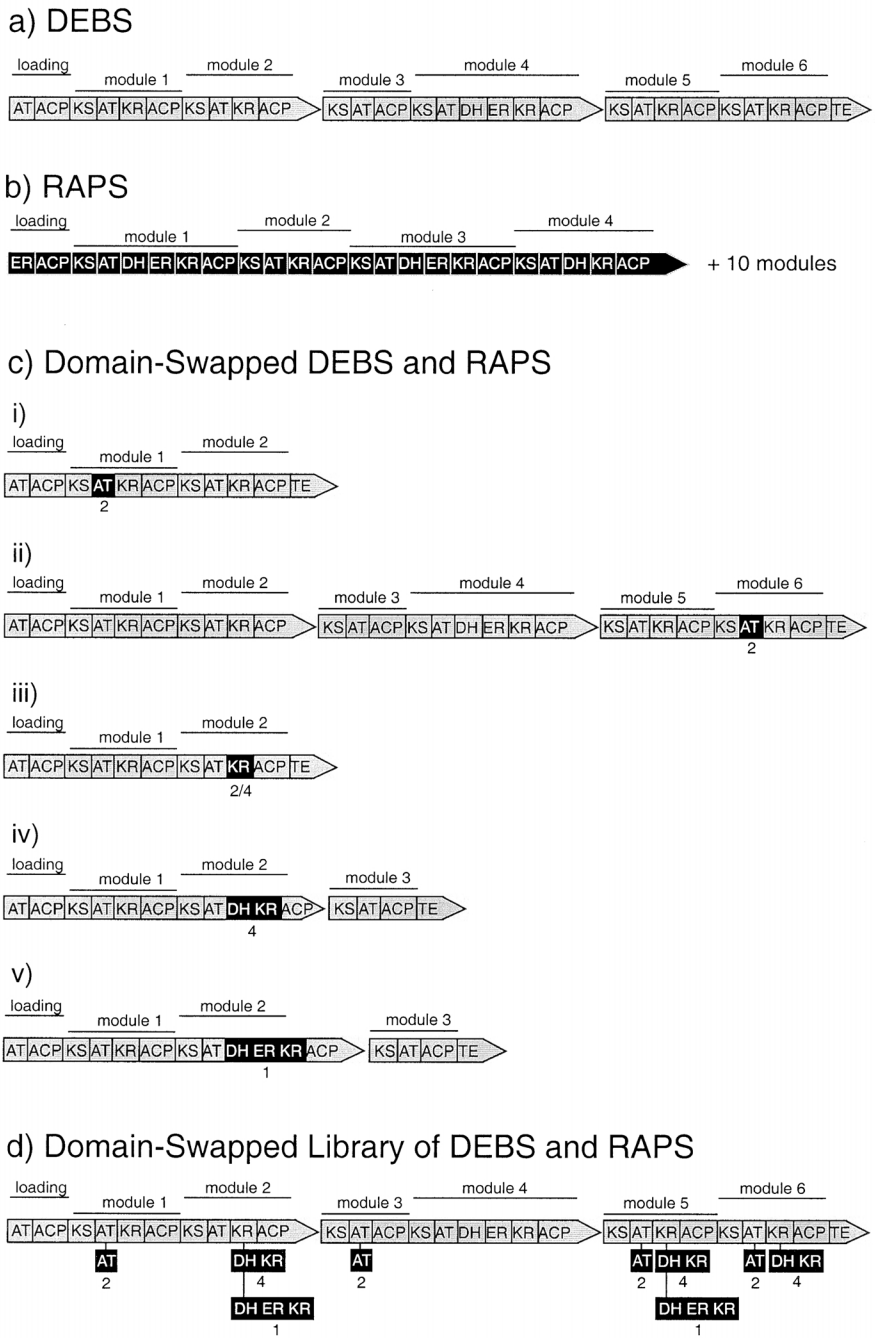
### *E. Improvement of Catalytic Efficiency*

Studies on in-frame fusions between enzymes have demonstrated several benefits, including improved overall activity, pre-steady-state lag reduction, and sequestering of intermediates, presumably through the advent of substrate channeling (Bulow, 1998). Hybrids of $\beta$-galactosidase and galactose dehydrogenase (Ljungcrantz *et al.,* 1989), galactose dehydrogenase and luciferase (Lindbladh *et al.,* 1992), citrate synthetase and malate dehydrogenase (Lindbladh *et al.,* 1994), and two unique $\beta$-glucanases (Olsen *et al.,* 1996) have exhibited evidence of substrate channeling. All these bifunctional enzymes were constructed by end-to-end fusion of their parental genes with an intervening linker sequence. Presumably, this could be improved on by better organization of the active sites through variation of the linker length or through a change in the topology from end-to-end fusion to insertion of one enzyme into another. As the active sites are brought closer together, specific point mutations, perhaps identified through random mutagenesis, could be useful in developing interactions between the two domains.

### *F. Alteration of Multimodular Synthetases*

#### *1. Polyketide Synthases*

Polyketides are made by the sequential activity of domains of large, multifunctional enzymes called polyketide synthases (PKSs) (Fig. 6a and b). Polyketides are formed by the condensation and modification of acyl units derived from acyl-CoA precursors. Domains are organized in modules and each module carries out the series of steps necessary for one cycle of polyketide chain elongation. A single protein can have more than one module, and several different proteins together can make up a PKS. The number of modules determines the size of the polyketide. A growing polyketide chain is tethered to the enzyme as a thiol ester and moves sequentially from the N- to the C-terminus of a module, lengthened by two carbon units per module. The first module in a PKS


a) DEBS
loading
module 1
module 2
module 3
module 4
module 5
module 6
AT ACP KS AT KR ACP KS AT KR ACP
KS AT ACP KS AT DH ER KR ACP
KS AT KR ACP KS AT KR ACP TE
b) RAPS
loading
module 1
module 2
module 3
module 4
ER ACP KS AT DH ER KR ACP KS AT KR ACP KS AT DH ER KR ACP KS AT DH KR ACP
+ 10 modules
c) Domain-Swapped DEBS and RAPS
i)
loading
module 1
module 2
AT ACP KS AT KR ACP KS AT KR ACP TE
2
ii)
loading
module 1
module 2
module 3
module 4
module 5
module 6
AT ACP KS AT KR ACP KS AT KR ACP
KS AT ACP KS AT DH ER KR ACP
KS AT KR ACP KS AT KR ACP TE
2
iii)
loading
module 1
module 2
AT ACP KS AT KR ACP KS AT KR ACP TE
2/4
iv)
loading
module 1
module 2
module 3
AT ACP KS AT KR ACP KS AT DH KR ACP
4
KS AT ACP TE
v)
loading
module 1
module 2
module 3
AT ACP KS AT KR ACP KS AT DH ER KR ACP
1
KS AT ACP TE
d) Domain-Swapped Library of DEBS and RAPS
loading
module 1
module 2
module 3
module 4
module 5
module 6
AT ACP KS AT KR ACP KS AT KR ACP
KS AT ACP KS AT DH ER KR ACP
KS AT KR ACP KS AT KR ACP TE
AT
2
DH KR
4
DH ER KR
1
AT
2
AT
2
DH KR
4
DH ER KR
1
AT
2
DH KR
4

has a loading domain that recognizes a particular acyl-CoA to start the chain. The mimimal module consists of ketosynthase (KS), acyltransferase (AT), and acyl carrier protein (ACP) domains. The specificity of the AT domain determines the $\alpha$-alkyl moiety of the two-carbon unit. Each module can have between zero and three domains that are responsible for modifying the $\beta$-carbon of the polyketide chain. Those modules with modification domains can have a ketoreductase (KR), a KR and a dehydratase (DH), or a KR, a DH, and an enoylreductase (ER) domain. In the case when all three are present, the KR reduces the ketone to an alcohol, the DH dehydrates the alcohol to a double bond, and the ER converts the double bond to a saturate single bond.

Numerous studies have shown that the module nature of PKSs is very amenable for engineering new PKSs by domain swapping. Several recent reviews offer a detailed description of this work (Carreras and Santi, 1998; Hutchinson, 1999; Keating and Walsh, 1999). Much of the work has been with two PKSs: 6-deoxyerythronolide B synthase (DEBS; Fig. 6a) and rapamycin PKS (RAPS; Fig. 6b). The chain initiation is one step in the synthesis that can be altered by swapping loading domains. This was first demonstrated by replacing the loading domain of the platenolide PKS with that of the tylactone PKS, switching the specificity from acetate to propionate (Kuhstoss *et al.,* 1996). When the loading domain of DEBS was replaced with one from avermectin PKS, polyketides were obtained with side chains derived from a broader distribution of starter units, consistant with the avermectin PKS-loading unit's broader specificity for branched-chain carboxylic acids (Marsden *et al.,* 1998).

←

FIG. 6. Schematic of engineered polyketide synthases. Polyketide synthases (PKS) are made up of acyltransferase (AT), ketosynthase (KS), acyl carrier protein (ACP), ketoreductase (KR), dehydratase (DH), enoylreductase (ER), and chain termination (TE) domains that are grouped into modules. A single gene (indicated by the pointed wedges) can encode for more than one module. (a) The PKS 6-deoxyerythronolide B synthase (DEBS) has seven modules, including a loading module encoded by three genes (Cortes *et al.,* 1990; Caffrey *et al.,* 1992). (b) Rapamycin PKS (RAPS) has fifteen modules including a loading module encoded by three genes (Schwecke *et al.,* 1995). Only the first gene of RAPS is shown. (c) Many hybrid PKS have been engineered by replacing a domain from DEBS with domain(s) from RAPS. The number directly below the inserted domain(s) indicates the RAPS module from which the domain came. In (i), the AT domain from RAPS module 2 has replaced the AT domain of DEBS module 1 (Oliynyk *et al.,* 1996). Other functional hybrids include (ii) (Liu *et al.,* 1997), (iii) (Kao *et al.,* 1998), (iv) (McDaniel *et al.,* 1997) and (v) (Kao *et al.,* 1997). (d) A combinatorial library using three separate plasmids for the three genes of DEBS and the indicated array of domain swaps has been constructed (Xue *et al.,* 1999).

The AT domain is one logical place to alter PKS since it determines the $\alpha$-alkyl moiety of the two-carbon unit that is added. The first demonstration of the feasibility of swapping AT domains came from the successful replacement of the AT domain of DEBS module 1 with the AT domain from RAPS module 2 that resulted in a polyketide with the intended substitution (Fig. 6c,i) (Oliynyk *et al.,* 1996). Replacements of the AT domain from DEBS module 6 with the AT domain from RAPS module 2 (Fig. 6c,ii) (Liu *et al.,* 1997), and AT domains of DEBS modules 1 and 2 with three different heterologous AT domains (Ruan *et al.,* 1997) resulted in novel polyketides being synthesized.

$\beta$-carbon-modifying domains (KR, DH, and ER) can also be swapped to produce novel polyketides. This can result in polyketides with stereochemical changes, such as changing an alcohol configuration from S to R by swapping the KR domain of DEBS module 2 for that of RAPS module 2 or 4 (Fig. 6c,iii) (Kao *et al.,* 1998). Modules can also be extended in length by, in effect, adding DH or DH and ER domains to modules that have only a KR domain such as DEBS module 2. Replacement with a KR and DH pair (Fig. 6c,iv) or KR, DH, and ER trio (Fig. 6c,v) from RAPS resulted in the expected products and thereby increased the functionality of the DEBS (McDaniel *et al.,* 1997; Kao *et al.,* 1997).

Recent publications in the field of engineered modular PKSs demonstrate the combinatorial potential of domain swapping to create novel PKSs to synthesize novel polyketides. Systematic replacement of AT and KR domains in modules 2, 5, and 6 of DEBS with AT and (KR or KR/DH or KR/DH/ER) domains from RAPS, respectively, resulted in a library of PKSs capable of synthesizing more than fifty different polyketides (McDaniel *et al.,* 1999). An approach for generating much larger libraries takes advantage of the fact that DEBS is encoded by three distinct genes whose proteins interact to form the PKS. Variants of individual genes are cloned into separate compatible plasmids, all of which are transformed into the same *Streptomysces lividans* strain. In this way, once the original set of variant genes are constructed, a library of all combinations of these variant proteins can be obtained in a relatively short amount of time by transforming cells with all three plasmid types (Fig. 6d). To demonstrate this system, fourteen plasmids were constructed and combinatorialized. Sixty-four cells received all three plasmids, of which forty-six produced detectable levels of forty-three different polyketides (Xue *et al.,* 1999). A combinatorial method of generating the original variants would significantly expand this combinatorial approach.

### 2. *Peptide Synthetases*

Nonribosomal peptide synthetases have a modular structure and carry out the syntheses of peptides about two to fifteen amino acids in length

through the sequential action of several semi-independent domains, each of which carries out a specific reaction on the growing polypeptide chain. Distinct domains have been shown to be involved in amino acid recognition and adenylation (A-domain), covalent linkage of the substrate to the cofactor 4′-phosphopantetheine (T-domain), condensation of two activated petidyl or aminoacyl moieties (C-domain), and release of the product by cyclization or hydrolysis (Te-domain) (Mootz and Marahiel, 1999). The nature, number, and order of these domains determine the structure of the product.

Although not as extensively explored as recombining polyketide synthases has been, redesign of peptide synthetases by recombining their modules has been successful in a handful of studies (Stachelhaus *et al.,* 1995; de Ferra *et al.,* 1997; Schneider *et al.,* 1998) and was the subject of a recent review (Mootz and Marahiel, 1999). The first example of successful engineering of an active peptide synthetase to produce a new product was engineered by swapping A-domain and T-domains (AT-domain) from different species for the AT-domain of the seventh module of the seven module SrfA complex in *Bacillus subtilis* (Stachelhaus *et al.,* 1995). Phe-, Orn- and Leu-AT-domains from *Bacillus brevis* and Cys- and Val- AT-domains from *Penicillium chrysogenum* were swapped for the Leu-AT-domain. These five domain-swapped enzymes encoded peptide synthetases with the desired amino-acid specificities and produced a surfactin (containing seven amino acid moieties including two D-amino acids) with the desired amino acid substitutions, albeit at lower levels. Subsequent experiments in which AT-domains were swapped for the second and fifth position of the SrfA complex also resulted in an altered, but not always the expected, product (Schneider *et al.,* 1998). Swapping the Te-domain (responsible for release of the product from the seventh module) for the fourth and the fifth modules resulted in the expected shortened peptide products (de Ferra *et al.,* 1997). Fusion points chosen at estimated domain boundaries (Stachelhaus *et al.,* 1995; Schneider *et al.,* 1998), within highly conserved motifs (Ritsema *et al.,* 1998), and at fortuitously located restriction sites (de Ferra *et al.,* 1997) have been successful for creating peptide synthetases with some activity. For creating improved domain-swapped peptide synthetase, identifying the optimal choice of fusion points is likely to be crucial (Mootz and Marahiel, 1999) and may be well suited for the combinatorial methods described later.

### *G. Improvement of Therapeutic Properties*

The spatial orientation of fused domains can affect important properties of the fusions. The use of circularly permuted domains overcomes

the limited number of spatial orientations available by fusion at natural N- and C- termini. For example, fusion proteins of interleukin-4 (IL-4) and a *Pseudomonas* exotoxin (PE) have antitumor activity. However, the fusion point of the two domains is near the site of IL-4 that binds its receptor, limiting its effectiveness. However, circular permutation of Il-4 at residues 37 and 38 and fusion of PE to the new C-terminus resulted in a chimera that bound its receptor with ten-fold higher affinity (Kreitman *et al.,* 1994). It possessed improved, specific antitumor activity in a variety of different types of tumor cell cultures and increased efficacy in mice (Kreitman *et al.,* 1995). Similarly, granulocyte stimulating factor (GCSF) was circularly permuted at several locations and fused to interleukin-3 (IL-3) in an attempt to make a set of spatially permuted GCSF-IL-3 fusions (called myelopoietins (MPOs)) (McWherter *et al.,* 1999; Feng *et al.,* 1999). MPOs display enhanced activity to promote cell growth and maturation of hematopoietic cells. The activity of the circularly permuted GCSF ranged from 10% to greater than 100% of that of wild-type GCSF. Some of the fusions had altered ratios of GCSF to Il-3 activity, expanding the range of MPOs with therapeutic potential.

### *H. Creation of Molecular Biosensors*

The green fluorescent protein (GFP) holds much potential for generating genetically encoded indicators for biochemical and physiological signals (Tsien, 1998). One possible arrangement to generate a signal is to fuse two variants of GFP, a donor and an acceptor with overlapping absorption and emission spectra, to the same or to interacting proteins. In this way, fluorescence resonance energy transfer (FRET) changes can be linked to molecular events that alter the distance ($d$) between the two GFPs. For example, variants of GFP have been fused to both ends of a calmodulin-binding domain (Romoser *et al.,* 1997) or to both ends of a calmodulin-(calmodulin-binding peptide M13) fusion (Miyawaki *et al.,* 1997) to generate a molecular sensor for $Ca^{2+}$. This domain undergoes large conformational changes on binding $Ca^{2+}$, effectively changing the distance between the two GFP variants such that a FRET signal can be linked to $Ca^{2+}$ levels *in vitro* or *in vivo.*

Because FRET requires the two domains to be very close, proteins with their N- and C-termini distal may have difficulty in generating a FRET signal when the GFP variants are fused at both ends, since FRET efficiency decreases by a factor of $d^6$. Using such a system to monitor the interaction of two proteins, where each of the two variants of GFP is fused to different proteins, will have similar problems (Fig. 7a and b). Moving beyond a simple end-to-end fusion scheme can possibly bring

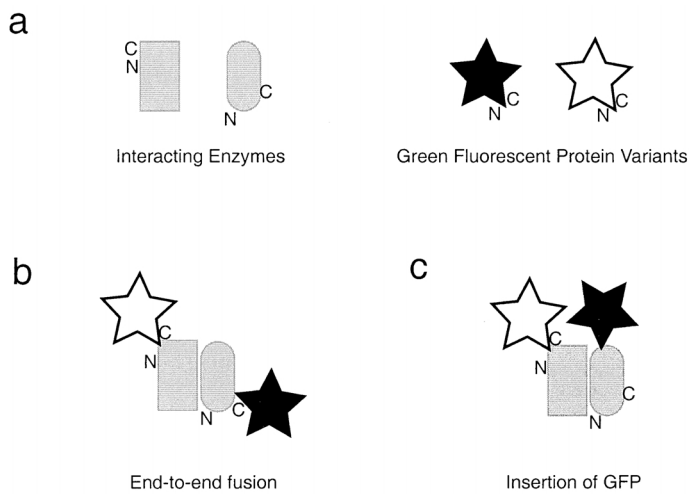



FIG. 7. Topological modifications to increase a FRET signal between two variants of the green fluorescent protein (GFP) fused to interacting proteins. (a) Schematic representation of interacting proteins and GFP variants with the location of the N- and C-termini indicated. (b) The GFP variants fused to the ends of the interacting proteins are too far apart to generate a FRET signal. (c) Insertion of one of the GFP variants into an interacting protein brings the two GFP variants close enough for FRET to be detected.

the two GFPs closer to allow the observance of FRET (Fig. 7c). To this end, random insertion of GFP into the cAMP-dependent protein kinase regulatory subunit from *Dictostelium discoideum* has been performed to increase the probability of observing FRET between the regulatory and catalytic subunits of this enzyme (Biondi *et al.,* 1998). However, this strategy did not result in any GFP insertion proteins with high fluorescence that maintained high affinity for cAMP. The binding of cAMP was still significant in many cases and 20% of the insertion proteins still interacted with the catalytic unit *in vitro.*

Even greater topological changes in GFP have been performed by random circular permutation (Graf and Schachman, 1996) in order to create new N- and C- termini for end-to-end fusion with other genes (Baird *et al.,* 1999). Ten nontrivial fluorescent circular permutations of GFP were found that had altered p$K_a$ values and orientation of the chromophore with respective to its N- and C-termini. The systematic identification of sites for circular permutation in GFP also identifies plausible sites for insertions of other proteins into GFP. This work speaks strongly about the potential of random circular permutation for protein

engineering and suggests a variety of protein topologies as targets for protein engineering (Baird *et al.,* 1999).

$\beta$-Lactamase has been inserted into a rationally chosen loop of GFP (a loop that, incidentally, was found to be a permissive site for circular permutation in the study described in the previous paragraph) in order to create a fusion protein whose fluorescence changes in the presence of $\beta$-lactamase-inhibitory protein (Doi and Yanagawa, 1999). Although in the initial construct no fluorescence change was observed on addition of the inhibitory protein, two rounds of directed evolution resulted in two mutations in the $\beta$-lactamase gene that conferred an inhibitory-protein concentration dependence on the fluorescence of the fusion. The mutated fusion had approximately 70% greater emission intensity in the presence of a tenfold molar excess of the inhibitory protein than in its absence. This is an excellent example of using domain swapping (by insertion) to move large distances in sequence space, followed by error-prone PCR to nudge the construct into a functional area of sequence space very nearby.

### *I. Creation of Novel Enzymes*

Although all the examples of protein engineering previously discussed can be considered novel in some respects, the creation of an entirely novel enzyme by domain swapping has not been demonstrated. The phrase ''entirely novel enzyme'' is defined as an enzyme that displays catalytic activity that is neither present in nor meaningfully related to either parent. A theoretical example of creating an entirely novel enzyme by domain swapping would be to take a domain from protein A that is responsible for binding substrate 1 and fuse it to a domain from enzyme B to create a hybrid that could perform enzyme B's chemistry on substrate 1. The natural substrates for enzyme B would bear little resemblance to substrate 1, and protein A would not have catalytic activity similar to enzyme B. Examples of engineered entirely novel enzymes created by methods other than domain swapping include the creation of catalytic antibodies (Smithrud and Benkovic, 1997) and the introduction of a protease catalytic triad into an *E. coli* cyclophilin to create a novel endopeptidase (Quemeneur *et al.,* 1998).

A step toward using domain swapping to create an entirely novel enzyme was taken by successfully creating a domain-swapped enzyme with glycinamide ribonucleotide (GAR) formyltransferase by fusing a ribonucleotide binding domain from PurN and a $N^{10}$-tetrahydrofolate cofactor hydrolase domain from PurU (Nixon *et al.,* 1997). Both PurN and PurU are formyltransferase enzymes (see Fig. 4). PurN transfers the

formyl group of $N^{10}$-tetrahydrofolate to GAR to synthesize formyl-GAR and PurU transfers the formyl group to water to synthesize formate. Crystallographic, biochemical, and sequence comparisons were used to design several PurN–PurU hybrids. A PurN–PurU hybrid with GAR transformylase activity was characterized *in vitro* and found to have activity $\sim 10^4$ lower than PurN and to favor hydrolysis over transfer to GAR by a ratio of 40 : 1. Nevertheless, swapping a GAR binding domain into PurU and creating an active GAR transformylase is proof of principle that domain swapping to create novel activities is attainable. However, this work technically falls short of creating an entirely novel enzyme. This can be understood by instead viewing the active hybrid as derived from a swap of a formyltransferase domain of PurN for a formyltransferase domain of PurU.

## V. Methodologies

Non-combinatorial methods for the engineering of domain-swapped enzymes have primarily been employed in the rational design of proteins. These methods achieve the fusion of one or more fragments at precise positions that are predicted through rational means to achieve a desired function. The probability of success for such an approach is directly proportional to the completeness of our understanding of protein folding, structure, function, and enzymatic mechanism in general and of the particular protein or protein fragments to be fused. Given that our knowledge is not complete, non-combinatorial methods are most likely to be successful in those areas in which structure and function of the fragments are not coupled and where desired function is objectively less complicated, as, for example, in the creation of bifunctional proteins for cellular targeting, phage or cell surface display, affinity purification, and increased expression/stability. Although combinatorial methods for optimizing linker length can be useful and have been employed, combinatorial methods seem more suited for achieving domain swapped enzymes in which the fragments are functionally coupled, such as in altering catalytic activity, modifying substrate specificity, engineering allosteric regulation and the creation of entirely novel enzymes.

### A. *Non-Combinatorial*

For rationally constructing domain-swapped proteins, one consideration is whether to include a linker region between the two domains. In general, hybrids with two autonomously functioning domains require significant distance between the two domains so that they can fold and

achieve a functional tertiary structure without interference from each other. The most common method for creating such domain-swapped enzymes is to clone the desired genes or gene fragments between properly arranged restriction sites such that the two fragments are separated by a piece of DNA that codes for a sequence of amino acids that can serve as a flexible linker. A consequence of this is that the DNA of the restriction sites will code for amino acids in the protein. However, appropriate choice of restriction sites that code for acceptable amino acids (e.g., *Bam*HI that codes for Gly-Ser, which are common linker amino acids) can go a long way to avoiding potential problems.

When the domains require more intimate contact to achieve the desired function, the presence of a linker and extra residues coded for by amino acids is not desirable. Although domain-swapped enzymes can be created by taking advantage of fortuitously located restriction sites or sites artificially introduced by silent mutagenesis, the simplest method for creating ''seamless'' fusions is by overlap extension (Horton *et al.,* 1989). Overlap extension can make end-to-end fusions in a single cloning step, but requires more than one cloning step to perform swapping of an internal domain. A variation of overlap extension for swapping internal domains in one cloning step has been described (Grandori *et al.,* 1997). Although other methods for making seamless fusions exist, such as vectors designed to allow the construction of precise fusions by creating unidirectional deletions between the two sequences to be fused (Kim *et al.,* 1991), they have not gained widespread use.

### *B. Combinatorial*

The development of combinatorial processes for protein engineering that mirror natural evolution is a rational approach to achieving desired protein function. Homologous recombination is one method to create hybrids between highly homologous enzymes *in vivo* (Schneider *et al.,* 1981). *In vitro* methods of gene recombination, including DNA shuffling (Stemmer, 1994), staggered extension process (Zhao *et al.,* 1998), random priming recombination (Shao *et al.,* 1998), DNA reassembly by interrupting synthesis (Short, 1999), and restriction enzyme based shuffling (Kikuchi *et al.,* 1999), all differ in methods used to generate fragments to reassemble by sexual PCR. These methods depend on a stochastic process that reassembles genes based on homology and hence cannot combinatorialize genes with low DNA homology, cannot avoid biases due to homology, and cannot control maintenance of discrete domains or subdomains throughout the process. However, these methods clearly

have utility in domain swapping, and are discussed in detail elsewhere in this volume.

*1. Random Insertions*

The observation of insertions in naturally occurring proteins suggests that such a route can be viable to construct proteins with desired properties and functions. Furthermore, studies on insertions of DNA coding for less than five amino acids (Zebala and Barany, 1991; Hallet *et al.,* 1997), up to sixteen amino acids (Starzyk *et al.,* 1989; Ladant *et al.,* 1992), a randomized 120 amino acids library (Doi *et al.,* 1997), and even entire proteins (Betton *et al.,* 1997; Biondi *et al.,* 1998) have all succeeded in creating active, often fully active, hybrids.

A number of methods have been described for random insertion of short sequences into a target gene, a method commonly referred to as linker scanning mutagenesis. Combinatorial methods for insertion of one domain into another are logical extensions of linker scanning mutagenesis (Fig. 8). Common procedures involve either (1) limited digestion of a target plasmid with DNase I (Heffron *et al.,* 1978) or similar procedures (Luckow *et al.,* 1987) in which the plasmid molecule is randomly linearized at one position followed by ligation of a short linker sequence that often includes a restriction site, or (2) random insertion by transposons (Hallet *et al.,* 1997; Manoil and Bailey, 1997). If the inserted sequence has a restriction enzyme site, a cassette containing the desired gene insert can be cloned into the library of linker insertions. Alternatively, using the DNase I procedure, linkers containing a restriction site can be ligated to the ends of the randomly linearized vector, the linkers digested, and the desired insert cassette ligated to covalently close the vector. This procedure has been employed to randomly insert the green fluorescent protein (GFP) into cAMP-dependent protein kinase regulatory subunit (PKA) in an attempt to develop a viable two-component fluorescence resonance energy transfer (FRET) system for monitoring intracellular cAMP levels (Biondi *et al.,* 1998).

Scanning linker mutagenesis can be used to identify sites amenable to functional insertion of another protein. TEM $\beta$-lactamase was inserted into three sites of the *E. coli* maltose binding protein (MalE) (Betton *et al.,* 1997), previously identified by random insertion of a *Bam*H I linker to tolerate small insertions (Duplay *et al.,* 1987). In all three hybrids, both activities were found to be essentially that of the individual wild-type enzymes.

*2. Circular Permutations*

A genetic method for random circular permutation of any gene was first described by Graf and Schachmann (Graf *et al.,* 1996). The method

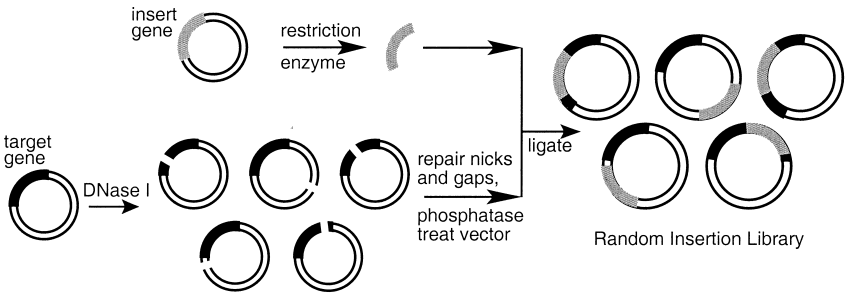



FIG. 8. Random insertion utilizing DNase I treatment (Heffron *et al.,* 1978; Biondi *et al.,* 1998). A plasmid containing the target gene is subjected to DNase I digestion under such conditions that, on average, one double strand break is randomly introduced per molecule. Because DNase I introduces nicks more frequently than double strand breaks and the double strand break does not generally produce blunt ends, repair of the plasmid DNA using a DNA polymerase and DNA ligase is required. The plasmid DNA is then dephosphorylated so that, in the subsequent ligation step, the vector does not religate without insertion of the desired insert fragment. The desired insert fragment is prepared, for example, by suitable restriction digest of the fragment from a second vector. Requirements for this fragment include blunt ends, and removal of the 3′ stop codon. For linker scanning mutagenesis, the insert fragment is a pair of designed oligonucleotides. Due to the random nature of the process, the fragments or oligonucleotide pairs can be inserted in either direction, in any of the three reading frames and in other locations of the plasmid. Thus, the majority of the library members do not have the desired insertions. However, a significant fraction will have an insert within the target gene and be in frame.

is composed of the following steps (Fig. 9): (1) isolation of a linear fragment of double stranded DNA with flanking compatible ends, (2) cyclization of this DNA fragment by ligation under dilute conditions, (3) random linearization of the cyclized gene using DNase I digestion in the presence of $Mn^{2+}$ at dilute concentrations of the enzyme such that the DNase I, on average, makes one double strand break, (4) repair of nicks and gaps using a DNA polymerase and a DNA ligase, and (5) ligation of the fragment into a desired vector by blunt end ligation to create the plasmid library of randomly circularly permuted genes. The vector must be designed so that the 5′ end of the cyclized gene is fused to a start codon and the 3′ is fused to a series of stop codons in all three frames. Using this procedure, one of six members of the library will be both in the correct orientation and the correct frame. Additionally, members of the library can have altered amino acids at either the N- or C-terminal and C-terminal extensions, depending on which stop codon is in frame with the circularly permuted gene. Portions of the experimental protocol described in the original paper (Graf *et*

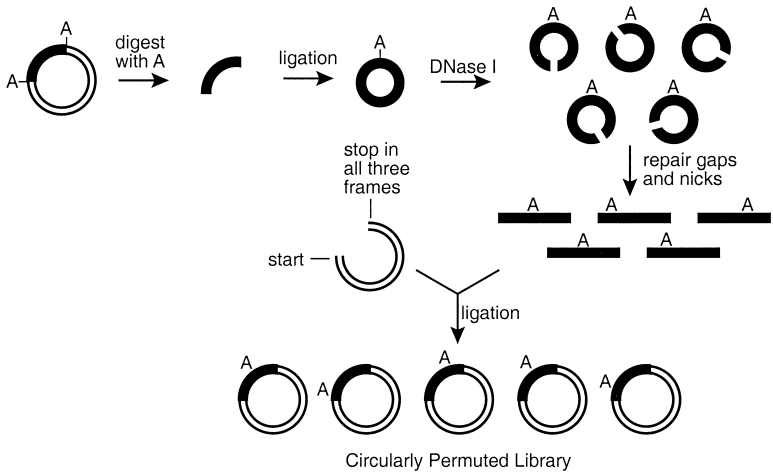



FIG. 9. Random circular permutation (Graf and Schachman, 1996). The gene to be randomly circularly permuted is excised from a suitable vector such that compatible restriction sites exist at the ends. In addition, the ligation of these compatible ends must produce an in-frame fusion of the 5′ and 3′ ends of the gene such that they code for amino acids that will form a suitable ''linker'' between the original N- and C-terminus of the protein. Ligation under dilute DNA concentrations results in cyclization of the DNA. The cyclized gene is subjected to DNase I digestion under such conditions that, on average, one double strand break is randomly introduced per molecule. As DNase I introduces nicks more frequently than double strand breaks and the double strand break does not generally produce blunt ends, repair of the plasmid DNA using a DNA polymerase and DNA ligase is required. Subsequent ligation into a suitably prepared vector that has dephosphorylated, blunt ends, a start codon, and stop codons in all three frames creates the circularly permuted library in which one out of six members will be both in-frame and in the correct orientation.

*al.,* 1996), particularly the concentration of DNA for circular ligation, do not appear to be correct or are at least suboptimal. The experimental conditions described in a recent paper appear more reasonable (Baird *et al.,* 1999). Thus far, the procedure has been used to systematically identify permissive sites for circular permutation in aspartate transcarbamoylase (Graf *et al.,* 1996), DsbA (Hennecke *et al.,* 1999), and GFP (Baird *et al.,* 1999). Aspartate transcarbamoylase and DsbA tolerated new termini in a variety of locations, whereas GFP was much less permissive.

### *3. Incremental Truncation*

Incremental truncation, in its simplest form, allows the creation of a library of every one bp deletion of a gene or gene fragment (Fig. 10). Despite its counterintuitive nature (i.e., that by deleting amino acids

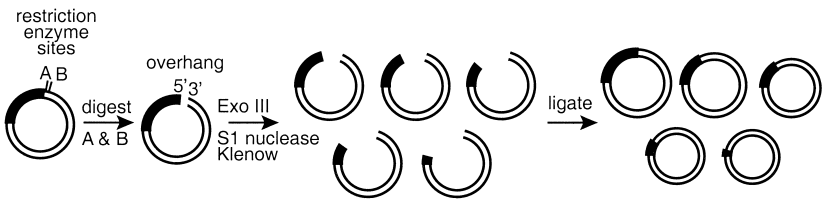



FIG. 10. Incremental truncation libraries (Ostermeier *et al.,* 1999b). Plasmid DNA is digested with two restriction enzymes: one that produces a 3′ recessed end (A; which is susceptible to Exo III digestion) and the other that produces a 5′ recessed end (B; which is resistant to Exo III digestion). Digestion with Exonuclease III proceeds under conditions in which the digestion rate is slow enough so that the removal of aliquots at frequent intervals results in a library of deletions of all possible lengths from one end of the fragment. The ends of the DNA can be blunted by treatment with S1 nuclease and Klenow so that unimolecular ligation results in the desired incremental truncation library.

one can arrive at new function), incremental truncation allows one to explore a number of novel protein engineering strategies (Ostermeier *et al.,* 1999a). A key step in the creation of these libraries is the digestion of the gene fragments with a 3′ to 5′ exonuclease such as Exonuclease III (Exo III) under conditions (e.g., low temperature or in the presence of NaCl) such that the rate is controlled to ~10 bases/min or less. During Exo III digestion, small aliquots are removed frequently and quenched by addition to a low pH, high salt buffer. Blunt ends are prepared by treatment with a single-strand nuclease and a DNA polymerase followed by unimolecular ligation to recyclize the vector. As Exo III digests DNA at a substantially uniform and synchronous rate (Wu *et al.,* 1976), this allows the creation of a library of every one bp deletion of a gene or gene fragment.

Incremental truncation libraries can be used to examine all possible bisection points within a given region of an enzyme that will allow the conversion of a monomeric enzyme into its functional heterodimer (Ostermeier *et al.,* 1999b). This strategy is described in the lower left panel of Figure 11 when the two starting fragments are overlapping fragments of the same gene. Conversion of a monomeric enzyme to a heterodimer by breaking a link in the peptide backbone can be considered ''reverse evolution'' since such a process is the reverse of domain recruitment. As such, it may identify ancestral fusion points and be an experimental approach to functionally defining domain boundaries.

The *E. coli* GARTase PurN has been systematically bisected by this method and found to tolerate bisection in two regions: the first in a nonconserved region in the vicinity of a domain boundary suggested by

sequence alignments, and the second in a highly conserved region three residues away from an active site residue (Ostermeier *et al.,* 1999b). Bisection in the second region indicated that it was possible to divide the active site residues onto two separate polypeptides and that two residues conserved across all GARTases and other formyltransferases are non-essential in the context of a heterodimer. Subsequent work on the creation of fusion libraries between *E. coli* and human GARTases confirmed the second region, but not the first, as a suitable junction for domain swapping, at least within GARTases (Ostermeier *et al.,* 1999c). However, the suitability of regions in the human enzyme for functional bisection has not been examined. It would be interesting to see if the intersection of the sets of functional bisection sites in the two enzymes would predict locations for successful domain swapping between them.

Incremental truncation can also be used to create a library of fusions between all possible fragments of two genes using a method called incremental truncation for the creation of hybrid enzymes (ITCHY) (Fig. 11). Overlapping fragments of two genes are cloned into suitable vectors for incremental truncation. As shown in the lower right panel of Figure 11, incremental truncation is performed as before, but the judicious location of a restriction site after the 3′ end of the C-terminal gene allows the fusion of one incremental truncation library to the other by ligation. Since this ligation is a random process, all possible fragments of one gene will be fused to all possible fragments of the other gene, provided a sufficient library size is obtained (which depends on the product of the maximum number of bases truncated in each fragment). The ITCHY library contains fusions between genes where they align, as well as internal ''duplications'' and ''deletions.'' The size range of ''duplications'' and ''deletions'' can be selected, for example, by digesting the library with 5′ and 3′ restriction sites, by separation by gel electrophoresis, and by subcloning the desired size. In this manner, the library can be selected to have fusions near the locations where the sequences or structures align. ITCHY performed on a single gene will create a library of internal tandem duplications and deletions of that gene.

ITCHY has been used to create functional interspecies hybrids between the *E. coli* and human genes for GARTase, genes that have low DNA sequence homology (50% identity on the DNA level) (Ostermeier *et al.,* 1999c). ITCHY was found to identify a more diverse set of active hybrids than DNA shuffling. Furthermore, the most active hybrid, fused at a region of non-DNA homology, was solely identified by ITCHY. All fusion points of active enzymes except one were exactly at the alignment of the two genes. This was somewhat surprising, as it was reasonable to

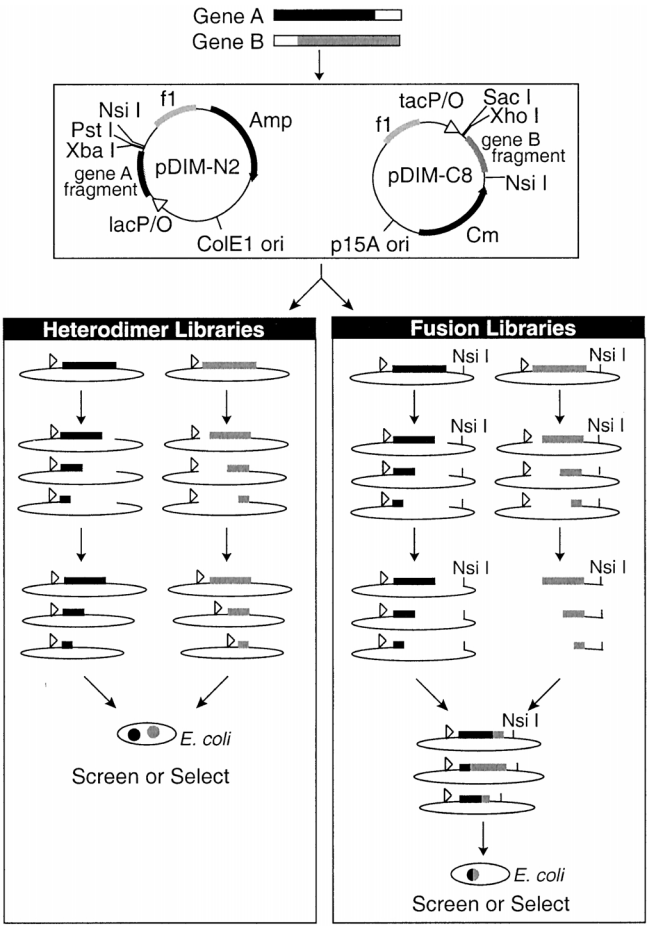



FIG. 11. Incremental truncation for the creation of hybrid enzymes (ITCHY) (Ostermeier *et al.,* 1999a; Ostermeier *et al.,* 1999c). A large 5′ fragment of gene A is cloned into phagemid pDIM-N2 and a large 3′ fragment of gene B is cloned into phagemid pDIM-C8. Phagemids pDIM-N2 and pDIM-C8 contain different antibiotic resistance genes (Amp, Cm) and different origins of replication (ColE1 and p15A) so that both can be maintained in the same cell, if desired. The phagemids also have restriction sites designed for creating incremental truncation libraries from the 3′ end of the gene fragment in pDIM-N2 and the 5′ end of the gene fragment in pDIM-C8. Noncovalent ITCHY is described in the lower left panel. Incremental truncation is performed on each plasmid and the vectors circularized as in Figure 9 such that a series of stop codons in all three frames is fused to the 3′ end of truncations of gene A and a start codon is fused to the 5′ end of truncations of gene B. Both vectors can be transformed into the same cell to create a library of (potentially) heterodimeric enzymes. If gene A and gene B are the same gene, this method allows one to search for loci that allow functional bisection

expect that fusions would tolerate a few extra amino acids or a few deleted amino acids, particularly in loops. However, the linear distances between conserved residues may have some importance for structure and/or function. Alternatively, the decrease in activity caused by extra or deleted amino acids may be small, but significant enough to inactivate a fusion protein that has an activity (as judged by $k_{cat}/K_m$) 500 to 10,000 less than wild type. Regardless, this work shows that ITCHY is a combinatorial solution to generating active fusion proteins that would be difficult to predict *a priori,* as most active hybrids found were fused in the proximity of the active site.

### *4. Internal Tandem Duplications or Deletions*

A strategy utilizing *Bal1*31 nuclease to create deletions, originally developed to investigate the arabinose promoter region (*araBAD*) (Reeder and Schleif, 1993), has been used to examine the tolerance of the linker region of the AraC protein for insertions and deletions (Eustance *et al.,* 1994) and of staphylococcal nuclease for internal tandem duplications (Nguyen and Schleif, 1998). *Bal*31 nuclease is a double-stranded nuclease that has both 3′ to 5′ and 5′ to 3′ exonuclease activities. A plasmid constructed of overlapping fragments of the same gene separated by a unique restriction site was linearized at this site, digested for various lengths of time with *Bal*31 nuclease, and recircularized. Although such libraries have not been characterized in great detail, they should be heavily biased toward having fusions between fragments that have had the same amount of bp deleted from them. Deviation from the bias would only result from variations within the sample between the amounts of DNA digested from each end (Eustance *et al.,* 1994). However, the inability to protect one end of DNA from *Bal*31 nuclease digestion precludes using it for creating incremental truncation-like libraries that are fused to defined sequences, such as start codons and stop codons, or for use in creating ITCHY libraries, without the use of a very long spacer.

## VI. Perspective

Progress in utilizing domain swapping for protein engineering has been partially hindered by the lack of combinatorial methods for per-

---

of a protein (Ostermeier *et al.,* 1999b) Covalent ITCHY is described in the lower right panel. Incremental truncation is followed by digestion with restriction enzyme *Nsi*I, isolation of the indicated DNA fragments, and ligation to form a library of gene fusions. The resulting fusion library has random fragments of the N terminus of protein A fused to random fragments of the C terminus of protein B. If gene A and gene B are the same gene, this method allows one to create a library of internal deletions and tandem duplications of a gene.

forming domain swapping on any two proteins independent of DNA homology. Often decisions made on where to fuse the enzymes are made irrespective of domain boundary considerations in lieu of the location of natural restrictions sites in the two genes that allow the construction of a set of hybrid genes. Recently developed combinatorial methods such as random insertion of large domains (Biondi *et al.,* 1998), random circular permutation (Graf and Schachman, 1996), and incremental truncation (Ostermeier *et al.,* 1999a; Ostermeier *et al.,* 1999c) should stimulate progress in the field.

DNA shuffling can combinatorialize related genes based on DNA homology. However, DNA shuffling's success in the evolution of novel function derives from the combinatorializing of homologous structures. Molecular breeding (i.e., DNA shuffling of families of proteins) can evolve new function more efficiently by accessing a more diverse sequence space, but in a rational manner: staying within homologous structures. This bodes well for a strategy of protein engineering that incorporates domain swapping of homologous structures. The high activity and crystal structure of domain-swapped fusions between trypsin and fXa, which have 38% sequence identity on the amino-acid level, illustrates how tolerant homologous structures can be in accommodating mismatches of proteins with low sequence homology (Hopfner *et al.,* 1998). Indeed, it is likely that such plasticity has been evolutionarily programmed into proteins to enable the natural evolution of new function. Continuing to explore the natural evolution of proteins is likely to be important for advances in protein engineering.

## Acknowledgments

We thank Dr. Stefan Lutz for many useful comments in the preparation of this manuscript.

## References

Aimi, J., Qiu, H., Williams, J., Zalkin, H., and Dixon, J. E. (1990). *De novo* purine nucleotide biosynthesis: cloning of human and avian cDNAs encoding the trifunctional glycinamide ribonucleotide synthetase-aminoimidazole ribonucleotide synthetase-glycinamide ribonucleotide transformylase by functional complementation in *E. coli. Nucleic Acids Res.,* **18,** 6665–6672.

Almassy, R. J., Janson, C. A., Kan, C.-C., and Hostomska, Z. (1992). Structures of apo and complexed *Escherichia coli* glycinamide ribonucleotide transformylase. *Proc. Natl. Acad. Sci. USA,* **89,** 6114–6118.

Arnold, F. H., and Volkov, A. A. (1999). Directed evolution of biocatalysts. *Curr. Opin. Chem. Biol,* **3,** 54–59.

Baird, G. S., Zacharias, D. A., and Tsien, R. Y. (1999). Circular permutation and receptor insertion within green fluorescent proteins. *Proc. Natl. Acad. Sci. USA,* **96,** 11241–11246.

Baneyx, F. (1999). Recombinant protein expression in *Escherichia coli. Curr. Opin. Biotechnol.,* **10,** 411–421.

Baron, M., Norman, D. G., and Campbell, I. D. (1991). Protein modules. *Trends Biochem. Sci.,* **16,** 13–17.

Beguin, P. (1999). Hybrid enzymes. *Curr. Opin. Biotechnol.,* **10,** 336–340.

Belkaïd, M., Penverne, B., Denis, M., and Hervé, G. (1987). *In situ* behavior of the pyrimidine pathway enzymes in *Saccharomyces cerevisiae.* 2. Reaction mechanism of aspartate transcarbamylase dissociated from carbamylphosphate synthetase by genetic alteration. *Arch. Biochem. Biophys.,* **254,** 568–578.

Betton, J.-M., Jacob, J. P., Hofnung, M., and Broome-Smith, J. K. (1997). Creating a bifunctional protein by insertion of $\beta$-lactamase into the maltodextrin-binding protein. *Nat. Biotechnology,* **15,** 1276–1279.

Biondi, R. M., Baehler, P. J., Reymond, C. D., and Véron, M. (1998). Random insertion of GFP into the cAMP-dependent protein kinase regulatory subunit from *Dictyostelium discoideum. Nucleic Acids Res.,* **26,** 4946–4952.

Bulow, L. (1998). Preparation of artificial bifunctional enzymes by gene fusion. *Biochem. Soc. Symp.,* **57,** 123–133.

Caffrey, P., Bevitt, D. J., Staunton, J., and Leadlay, P. F. (1992). Identification of DEBS 1, DEBS 2 and DEBS 3, the multienzyme polypeptides of the erythromycin-producing polyketide synthase from *Saccharopolyspora erythraea. FEBS Lett.,* **304,** 225–228.

Carreras, C. W., and Santi, D. V. (1998). Engineering of modular polyketide synthases to produce novel polyketides. *Curr. Opin. Biotechnol.,* **9,** 403–411.

Chandrasegaran, S., and Smith, J. (1999). Chimeric restriction enzymes: what is next? *Biol. Chem.,* **380,** 841–848.

Clackson, T. (1997). Controlling mammalian gene expression with small molecules. *Curr. Opin. Chem. Biol,* **1,** 210–218.

Cortes, J., Haydock, S. F., Roberts, G. A., Bevitt, D. J., and Leadlay, P. F. (1990). An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of *Saccharopolyspora erythraea. Nature,* **348,** 176–178.

Cozier, G. E., and Anthony, C. (1995a). Structure of the quinoprotein glucose dehydrogenase of *Escherichia coli* modelled on that of methanol dehydrogenase from *Methylobacterium extorquens. Biochem. J.,* **312,** 679–685.

Cozier, G. E., Giles, I. G., and Anthony, C. (1995b). The structure of the quinoprotein alcohol dehydrogenase of Acetobacter aceti modelled on that of methanol dehydrogenase from Methylobacterium extorquens. *Biochem. J.,* **308,** 375–379.

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

Davidson, J. N., Chen, K. C., Jamison, R. S., Musmanno, L. A., and Kern, C. B. (1993). The evolutionary history of the first three enzymes in pyrimidine biosynthesis. *BioEssays,* **15,** 157–164.

Davidson, J. N., and Peterson, M. L. (1997). Origin of genes encoding multi-enzymatic proteins in eukaryotes. *Trends Genet.,* **13,** 281–285.

de Ferra, F., Rodriguez, F., Tortora, O., Tosi, C., and Grandi, G. (1997). Engineering of peptide synthetases: Key role of the thioesterase-like domain for efficient production of recombinant peptides. *J. Biol. Chem.,* **272,** 25304–25309.

Doi, N., Itaya, M., Yomo, T., Tokura, S., and Yanagawa, H. (1997). Insertion of foreign random sequences of 120 amino acid residues into an active enzyme. *FEBS Lett.,* **402,** 177–180.

Doi, N., and Yanagawa, H. (1999). Design of generic biosensors based on green fluorescent proteins with allosteric sites by directed evolution. *FEBS Lett.,* **453,** 305–307.

Doolittle, R. F., and Bork, P. (1993). Evolutionarily mobile modules in proteins. *Sci. Am.,* **269,** 50–56.

Drees, B. L. (1999). Progress and variations in two-hybrid and three-hybrid technologies. *Curr. Opin. Chem. Biol,* **3,** 64–70.

Duplay, P., Szmelcman, S., Bedouelle, H., and Hofnung, M. (1987). Silent and functional changes in the periplasmic maltose-binding protein of *Escherichia coli* K12. I. Transport of maltose. *J. Mol. Biol.,* **194,** 663–673.

Durbecq, V., Legrain, C., Roovers, M., Pierard, A., and Glansdorff, N. (1997). The carbamate kinase-like carbamoyl phosphate synthetase of the hyperthermophilic archaeon *Pyrococcus furiosus,* a missing link in the evolution of carbamoyl phosphate biosynthesis. *Proc. Natl. Acad. Sci. USA,* **94,** 12803–12808.

Ebbole, D. J., and Zalkin, H. (1987). Cloning and characterization of a 12-gene cluster from *Bacillus subtilis* encoding nine enzymes for de novo purine nucleotide synthesis. *J. Biol. Chem.,* **262,** 8274–8287.

Eustance, R. J., Bustos, S. A., and Schleif, R. F. (1994). Reaching out. Locating and lengthening the interdomain linker in AraC protein. *J. Mol. Biol.,* **242,** 330–338.

Feng, Y., Minnerly, J. C., Zurfluh, L. L., Joy, W. D., Hood, W. F., Abegg, A. L., Grabbe, E. S., Shieh, J. J., Thurman, T. L., McKearn, J. P., and McWherter, C. A. (1999). Circular permutation of granulocyte colony-stimulating factor. *Biochemistry,* **38,** 4553–4563.

Fushinobu, S., Ohta, T., and Matsuzawa, H. (1998). Homotropic activation via the subunit interaction and allosteric symmetry revealed on analysis of hybrid enzymes of L-lactate dehydrogenase. *J. Biol. Chem.,* **273,** 2971–2976.

Georgiou, G., Stathopoulus, C., Daugherty, P. S., Nayak, A. R., Iverson, B. L., and Curtiss III, R. (1997). Display of heterologous proteins on the surface of microorganisms: from the screening of combinatorial libraries to live recombinant vaccines. *Nature Biotechnology,* **15,** 29–34.

Ghosh, M., Anthony, C., Harlos, K., Goodwin, M. G., and Blake, C. (1995). The refined structure of the quinoprotein methanol dehydrogenase from *Methylobacterium extorquens* at 1.94 A. *Structure,* **3,** 177–187.

Goldenberg, D. P., and Creighton, T. E. (1983). Circular and circularly permuted forms of bovine pancreatic trypsin inhibitor. *J. Mol. Biol.,* **165,** 407–413.

Graf, R., and Schachman, H. K. (1996). Random circular permutation of genes and expressed polypeptide chains: Application of the method to the catalytic chains of aspartate transcarbamoylase. *Proc. Natl. Acad. Sci. USA,* **93,** 11591–11596.

Grandori, R., Struck, K., Giovanielli, K., and Carey, J. (1997). A three-step PCR protocol for construction of chimeric proteins. *Protein Eng.,* **10,** 1099–1100.

Greisman, H. A., and Pabo, C. O. (1997). A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science,* **275,** 657–661.

Guddat, L. W., Bardwell, J. C., and Martin, J. L. (1998). Crystal structures of reduced and oxidized DsbA: investigation of domain motion and thiolate stabilization. *Structure,* **6,** 757–767.

Guruprasad, K., Tormakangas, K., Kervinen, J., and Blundell, T. L. (1994). Comparative modelling of barley-grain aspartic proteinase: a structural rationale for observed hydrolytic specificity. *FEBS Lett.,* **352,** 131–136.

Guy, H. I., and Evans, D. R. (1996). Function of the major synthetase subdomains of carbamyl-phosphate synthetase. *J. Biol. Chem.,* **271,** 13762–13769.

Guy, H. I., Schmidt, B., Hervé, G., and Evans, D. R. (1998). Pressure-induced dissociation of carbamoyl-phosphate synthetase domains. The catalytically active form is dimeric. *J. Biol. Chem.,* **273,** 14172–14178.

Hallet, B., Sherratt, D. J., and Hayes, F. (1997). Pentapeptide scanning mutagenesis: random insertion of a variable five amino acid cassette in a target protein. *Nucleic Acids Res.,* **25,** 1866–1867.

Hawkins, A. R., and Lamb, H. K. (1995). The molecular biology of multidomain proteins: selected examples. *Eur. J. Biochem.,* **232,** 7–18.

Heffron, F., So, M., and McCarthy, B. J. (1978). In vitro mutagenesis of a circular DNA molecule by using synthetic restriction sites. *Proc. Natl. Acad. Sci. USA,* **75,** 6012–6016.

Hegyi, H., and Bork, P. (1997). On the classification and evolution of protein modules. *J. Protein Chem.,* **16,** 545–551.

Heinemann, U., and Hahn, M. (1995). Circular permutation of polypeptide chains: implications for protein folding and stability. *Prog. Biophys. Mol. Biol,* **64,** 121–143.

Henikoff, S. (1986). The *Saccharomyces cerevisiae* ADE5,7 protein is homologous to overlapping *Drosophila melanogaster* Gart polypeptides. *J. Mol. Biol.,* **190,** 519–528.

Henikoff, S., Keene, M. A., Sloan, J. S., Bleskan, J., Hards, R., and Patterson, D. (1986). Multiple purine pathway enzyme activities are encoded at a single genetic locus in *Drosophila. Proc. Natl. Acad. Sci. USA,* **83,** 720–724.

Hennecke, J., Sebbel, P., and Glockshuber, R. (1999). Random circular permutation of DsbA reveals segments that are essential for protein folding and stability. *J. Mol. Biol.,* **286,** 1197–1215.

Hervé, G., Nagy, M., Le Gouar, M., Penverne, B., and Ladjimi, M. (1993). The carbamoyl phosphate synthetase–aspartate transcarbamoylase complex of *Saccharomyces cerevisiae:* molecular and cellular aspects. *Biochem. Soc. Trans.,* **21,** 195–198.

Hoogenboom, H. R., de Bruine, A. P., Hufton, S. E., Hoet, R. M., Arends, J. W., and Roovers, R. C. (1998). Antibody phage display technology and its applications. *Immunotechnology,* **4,** 1–20.

Hopfner, K.-P., Kopetzki, E., Kreb(German B)e, G.-B., Bode, W., Huber, R., and Engh, R. A. (1998). New enzyme lineages by subdomain shuffling. *Proc. Natl. Acad. Sci. USA,* **95,** 9813–9818.

Horton, R. M., Hunt, H. D., Ho, S. N., Pullen, J. K., and Pease, L. R. (1989). Engineered hybrid genes without the use of restriction enzymes: gene splicing by overlap extension. *Gene,* **77,** 61–68.

Huang, B., Schaeffer, C. J., Li, Q., and Tsai, M. D. (1996). Splase: a new class IIS zinc-finger restriction endonuclease with specificity for Sp1 binding sites. *J. Protein Chem.,* **15,** 481–489.

Hutchinson, C. R. (1999). Microbial polyketide synthases: more and more prolific. *Proc. Natl. Acad. Sci. USA,* **96,** 3336–3338.

Isalan, M., Choo, Y., and Klug, A. (1997). Synergy between adjacent zinc fingers in sequence-specific DNA recognition. *Proc. Natl. Acad. Sci. USA,* **94,** 5617–5621.

Jeltsch, A. (1999). Circular permutations in the molecular evolution of DNA methyltransferases. *J. Mol. Evol.,* **49,** 161–164.

Kan, J. L., Jannatipour, M., Taylor, S. M., and Moran, R. G. (1993). Mouse cDNAs encoding a trifunctional protein of *de novo* purine synthesis and a related single-domain glycinamide ribonucleotide synthetase. *Gene,* **137,** 195–202.

Kao, C. M., Mcpherson, M., McDaniel, R. N., Fu, H., Cane, D. E., and Khosla, C. (1997). Gain of function mutagenesis of the erythromycin polyketide synthase. 2. Engineered

biosynthesis of an eight-membered ring tetraketide lactone. *J. Am. Chem. Soc.,* **119,** 11339–11340.

Kao, C. M., Mcpherson, M., McDaniel, R. N., Fu, H., Cane, D. E., and Khosla, C. (1998). Alcohol stereochemistry in polyketide backbones is controlled by the $\beta$-ketoreductase domains of modular polyketide synthases. *J. Am. Chem. Soc.,* **120,** 2478–2479.

Kapust, R. B., and Waugh, D. S. (1999). *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.,* **8,** 1668–1674.

Keating, T. A., and Walsh, C. T. (1999). Initiation, elongation, and termination strategies in polyketide and polypeptide antibiotic biosynthesis. *Curr. Opin. Chem. Biol.,* **3,** 598–606.

Kikuchi, M., Ohnishi, K., and Harayama, S. (1999). Novel family shuffling methods for the *in vitro* evolution of enzymes. *Gene,* **236,** 159–167.

Kim, S. C., Posfai, G., and Szybalski, W. (1991). A novel gene-fusing vector: construction of a 5′-GGmCC-specific chimeric methyltransferase, M-*Bsp*RI/M-*Bsu*RI. *Gene,* **100,** 45–50.

Kim, Y. G., and Chandrasegaran, S. (1994). Chimeric restriction endonuclease. *Proc. Natl. Acad. Sci. USA,* **91,** 883–887.

Kim, Y. G., Cha, J., and Chandrasegaran, S. (1996). Hybrid restriction enzymes: zinc finger fusions to *Fok*I cleavage domain. *Proc. Natl. Acad. Sci. USA,* **93,** 1156–1160.

Kim, Y. G., Shi, Y., Berg, J. M., and Chandrasegaran, S. (1997). Site-specific cleavage of DNA-RNA hybrids by zinc finger/*Fok*I cleavage domain fusions. *Gene,* **203,** 43–49.

Kim, Y. G., Smith, J., Durgesha, M., and Chandrasegaran, S. (1998). Chimeric restriction enzyme: Gal4 fusion to *Fok*I cleavage domain. *Biol. Chem.,* **379,** 489–495.

Kreitman, R. J., Puri, R. K., and Pastan, I. (1994). A circularly permuted recombinant interleukin 4 toxin with increased activity. *Proc. Natl. Acad. Sci. USA,* **91,** 6889–6893.

Kreitman, R. J., Puri, R. K., and Pastan, I. (1995). Increased antitumor activity of a circularly permuted interleukin 4- toxin in mice with interleukin 4 receptor-bearing human carcinoma. *Cancer Res.,* **55,** 3357–3363.

Kuhstoss, S., Huber, M., Turner, J. R., Paschal, J. W., and Rao, R. N. (1996). Production of a novel polyketide through the construction of a hybrid polyketide synthase. *Gene,* **183,** 231–236.

Labedan, B., Boyen, A., Baetens, M., Charlier, D., Chen, P., Cunin, R., Durbeco, V., Glansdorff, N., Herve, G., Legrain, C., Liang, Z., Purcarea, C., Roovers, M., Sanchez, R., Toong, T. L., Van de, C. M., van Vliet, F., Xu, Y., and Zhang, Y. F. (1999). The evolutionary history of carbamoyltransferases: A complex set of paralogous genes was already present in the last universal common ancestor. *J. Mol. Evol.,* **49,** 461–473.

Ladant, D., Glaser, P., and Ullmann, A. (1992). Insertional mutagenesis of *Bordetella pertussis* adenylate cyclase. *J. Biol. Chem.,* **267,** 2244–2250.

Lawrence, J. G. (1997). Selfish operons and speciation by gene transfer. *Trends Microbiol.,* **5,** 355–359.

Li, L., Wu, L. P., and Chandrasegaran, S. (1992). Functional domains in *Fok*I restriction endonuclease. *Proc. Natl. Acad. Sci. USA,* **89,** 4275–4279.

Li, L., Wu, L. P., Clarke, R., and Chandrasegaran, S. (1993). C-terminal deletion mutants of the *Fok*I restriction endonuclease. *Gene,* **133,** 79–84.

Lim, F., Morris, C. P., Occhiodoro, F., and Wallace, J. C. (1988). Sequence and domain structure of yeast pyruvate carboxylase. *J. Biol. Chem.,* **263,** 11493–11497.

Lindbladh, C., Persson, M., Bulow, L., and Mosbach, K. (1992). Characterization of a recombinant bifunctional enzyme, galactose dehydrogenase/bacterial luciferase,

displaying an improved bioluminescence in a three-enzyme system. *Eur. J. Biochem.,* **204,** 241–247.

Lindbladh, C., Rault, M., Hagglund, C., Small, W. C., Mosbach, K., Bulow, L., Evans, C., and Srere, P. A. (1994). Preparation and kinetic characterization of a fusion protein of yeast mitochondrial citrate synthase and malate dehydrogenase. *Biochemistry,* **33,** 11692–11698.

Lindqvist, Y., and Schneider, G. (1997). Circular permutations of natural protein sequences: structural evidence. *Curr. Opin. Struct. Biol,* **7,** 422–427.

Liu, L., Thanchaipenet, A., Fu, H., Betlach, M., and Ashley, G. (1997). Biosynthesis of 2-nor-6-deoxyerythronolide B by rationally designed domain substitutions. *J. Am. Chem. Soc.,* **119,** 10553–10554.

Ljungcrantz, P., Carlsson, H., Mansson, M. O., Buckel, P., Mosbach, K., and Bulow, L. (1989). Construction of an artificial bifunctional enzyme, beta-galactosidase/galactose dehydrogenase, exhibiting efficient galactose channeling. *Biochemistry,* **28,** 8786–8792.

Luckow, B., Renkawitz, R., and Schutz, G. (1987). A new method for constructing linker scanning mutants. *Nucleic Acids Res.,* **15,** 417–429.

Mally, M. I., Grayson, D. R., and Evans, D. R. (1980). Catalytic synergy in the multifunctional protein that initiates pyrimidine biosynthesis in Syrian hamster cells. *J. Biol Chem.,* **255,** 11372–11380.

Manoil, C., and Bailey, J. (1997). A simple screen for permissive sites in proteins: analysis of *Escherichia coli* lac permease. *J. Mol. Biol.,* **267,** 250–263.

Marsden, A. F., Wilkinson, B., Cortes, J., Dunster, N. J., Staunton, J., and Leadlay, P. F. (1998). Engineering broader specificity into an antibiotic-producing polyketide synthase. *Science,* **279,** 199–202.

Martin, J. L., Bardwell, J. C., and Kuriyan, J. (1993). Crystal structure of the DsbA protein required for disulphide bond formation *in vivo. Nature,* **365,** 464–468.

Martoglio, B., and Dobberstein, B. (1998). Signal sequences: more than just greasy peptides. *Trends Cell. Biol.,* **8,** 410–415.

McDaniel, R., Kao, C. M., Fu, H., Hevezi, P., Gustafsson, C., Betlach, M., Ashley, G., Cane, D. E., and Khosla, C. (1997). Gain-of-function mutagenesis of a modular polyketide synthase. *J. Am. Chem. Soc.,* **119,** 4309–4310.

McDaniel, R., Thamchaipenet, A., Gustafsson, C., Fu, H., Betlach, M., and Ashley, G. (1999). Multiple genetic modifications of the erythromycin polyketide synthase to produce a library of novel ''unnatural'' natural products. *Proc. Natl. Acad. Sci. USA,* **96,** 1846–1851.

McLachlan, A. D. (1979). Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.,* **128,** 49–79.

McLachlan, A. D. (1987). Gene duplication and the origin of repetitive protein structures. *Cold Spring Harbor Symp. Quant. Biol.,* **52,** 411–420.

McWherter, C. A., Feng, Y., Zurfluh, L. L., Klein, B. K., Baganoff, M. P., Polazzi, J. O., Hood, W. F., Paik, K., Abegg, A. L., Grabbe, E. S., Shieh, J. J., Donnelly, A. M., and McKearn, J. P. (1999). Circular permutation of the granulocyte colony-stimulating factor receptor agonist domain of myelopoietin. *Biochemistry,* **38,** 4564–4571.

Miyawaki, A., Llopis, J., Heim, R., McCaffery, J. M., Adams, J. A., Ikura, M., and Tsien, R. Y. (1997). Fluorescent indicators for Ca2+ based on green fluorescent proteins and calmodulin. *Nature,* **388,** 882–887.

Mootz, H. D., and Marahiel, M. A. (1999). Design and application of multimodular peptide synthetases. *Curr. Opin. Biotechnol.,* **10,** 341–348.

Netzer, W. J., and Hartl, F. U. (1997). Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature,* **388,** 343–349.

Nguyen, D. M., and Schleif, R. F. (1998). Isolation and physical characterization of random insertions in *Staphylococcal nuclease. J. Mol. Biol.,* **282,** 751–759.

Nilsson, J., Stahl, S., Lundeberg, J., Uhlen, M., and Nygren, P.-A. (1997). Affinity fusion strategies for detection, purification, and immobilization of recombinant proteins. *Protein Expression Purif.,* **11,** 1–16.

Nixon, A. E., Warren, M. S., and Benkovic, S. J. (1997). Assembly of an active enzyme by the linkage of two protein modules. *Proc. Natl. Acad. Sci. USA,* **94,** 1069–1073.

Nixon, A. E., Ostermeier, M., and Benkovic, S. J. (1998). Hybrid enzymes: manipulating enzyme design. *Trends Biotechnol.,* **16,** 258–264.

Nyunoya, H., and Lusty, C. J. (1983). The carB gene of *Escherichia coli:* a duplicated gene coding for the large subunit of carbamoyl-phosphate synthetase. *Proc. Natl. Acad. Sci. USA,* **80,** 4629–4633.

Oliynyk, M., Brown, M. J. B., Cortes, J., Staunton, J., and Leadlay, P. F. (1996). A hybrid modular polyketide synthase obtained by domain swapping. *Chem. Biol.,* **3,** 833–839.

Olsen, O., Thomsen, K. K., Weber, J., Duus, J. O., Svendsen, I., Wegener, C., and von Wettstein, D. (1996). Transplanting two unique beta-glucanase catalytic activities into one multienzyme, which forms glucose. *Biotechnology (NY),* **14,** 71–76.

Ostermeier, M., Nixon, A. E., and Benkovic, S. J. (1999a). Incremental truncation as a strategy in the engineering of novel biocatalysts. *Bioorg. Med. Chem.,* **7,** 2139–2144.

Ostermeier, M., Nixon, A. E., Shim, J. H., and Benkovic, S. J. (1999b). Combinatorial protein engineering by incremental truncation. *Proc. Natl. Acad. Sci. USA,* **96,** 3562–3567.

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

Pavletich, N. P., and Pabo, C. O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science,* **252,** 809–817.

Pelletier, J. N., Campbell-Valois, F. X., and Michnick, S. W. (1998). Oligomerization domain-directed reassembly of active dihydrofolate reductase from rationally designed fragments. *Proc. Natl. Acad. Sci. USA,* **95,** 12141–12146.

Perona, J. J., and Craik, C. S. (1997). Evolutionary divergence of substrate specificity within the chymotrypsin-like serine protease fold. *J. Biol. Chem.,* **272,** 29987–29990.

Ponting, C. P., and Russell, R. B. (1995). Swaposins: circular permutations within genes encoding saposin homologues. *Trends Biochem. Sci.,* **20,** 179–180.

Purcarea, C., Simon, V., Prieur, D., and Hervé, G. (1996). Purification and characterization of carbamoyl-phosphate synthetase from the deep-sea hyperthermophilic archaebacterium *Pyrococcus abyssi. Eur. J. Biochem.,* **236,** 189–199.

Purcarea, C., Evans, D. R., and Hervé, G. (1999). Channeling of carbamoyl phosphate to the pyrimidine and arginine biosynthetic pathways in the deep sea hyperthermophilic archaeon *Pyrococcus abyssi. J. Biol. Chem.,* **274,** 6122–6129.

Quemeneur, E., Moutiez, M., Charbonnier, J. B., and Menez, A. (1998). Engineering cyclophilin into a proline-specific endopeptidase. *Nature,* **391,** 301–304.

Reeder, T., and Schleif, R. (1993). AraC protein can activate transcription from only one position and when pointed in only one direction. *J. Mol. Biol.,* **231,** 205–218.

Reiter, Y., and Pastan, I. (1998). Recombinant Fv immunotoxins and Fv fragments as novel agents for cancer therapy and diagnosis. *Trends Biotechnol.,* **16,** 513–520.

Remy, I., and Michnick, S. W. (1999). Clonal selection and *in vivo* quantitation of protein interactions with protein-fragment complementation assays. *Proc. Natl. Acad. Sci. USA,* **96,** 5394–5399.

Ritsema, T., Gehring, A. M., Stuitje, A. R., van der Drift, K. M., Dandal, I., Lambalot, R. H., Walsh, C. T., Thomas-Oates, J. E., Lugtenberg, B. J., and Spaink, H. P. (1998). Functional analysis of an interspecies chimera of acyl carrier proteins indicates a specialized domain for protein recognition. *Mol. Gen. Genet.,* **257,** 641–648.

Rodi, D. J., and Makowski, L. (1999). Phage-display technology—finding a needle in a vast molecular haystack. *Curr. Opin. Biotechnol.,* **10,** 87–93.

Romoser, V. A., Hinkle, P. M., and Persechini, A. (1997). Detection in living cells of Ca2+-dependent changes in the fluorescence emission of an indicator composed of two green fluorescent protein variants linked by a calmodulin-binding sequence. A new class of fluorescent indicators. *J. Biol. Chem.,* **272,** 13270–13274.

Rossmann, M. G., Moras, D., and Olsen, K. W. (1974). Chemical and biological evolution of nucleotide-binding protein. *Nature,* **250,** 194–199.

Ruan, X., Pereda, A., Stassi, D. L., Zeidner, D., Summers, R. G., Jackson, M., Shivakumar, A., Kakavas, S., Staver, M. J., Donadio, S., and Katz, L. (1997). Acyltransferase domain substitutions in erythromycin polyketide synthase yield novel erythromycin derivatives. *J. Bacteriol.,* **179,** 6416–6425.

Russell, R. B. (1994). Domain insertion. *Protein Eng.,* **7,** 1407–1410.

Russell, R. B., and Ponting, C. P. (1998). Protein fold irregularities that hinder sequence analysis. *Curr. Opin. Struct. Biol.,* **8,** 364–371.

Schlunegger, M. P., Bennett, M. J., and Eisenberg, D. (1997). Oligomer formation by 3D domain swapping: a model for protein assembly and misassembly. *Adv. Protein Chem.,* **50,** 61–122.

Schmitt, E., Blanquet, S., and Mechulam, Y. (1996). Structure of crystalline *Escherichia coli* methionyl-tRNA(f)Met formyltransferase: comparison with glycinamide ribonucleotide formyltransferase. *EMBO J.,* **15,** 4749–4758.

Schneider, A., Stachelhaus, T., and Marahiel, M. A. (1998). Targeted alteration of the substrate specificity of peptide synthetases by rational module swapping. *Mol. Gen. Genet.,* **257,** 308–318.

Schneider, W. P., Nichols, B. P., and Yanofsky, C. (1981). Procedure for production of hybrid genes and proteins and its use in assessing significance of amino acid differences in homologous tryptophan synthetase alpha polypeptides. *Proc. Natl. Acad. Sci. USA,* **78,** 2169–2174.

Schnepp, B., Donaldson, T., Grumbling, G., Ostrowski, S., Schweitzer, R., Shilo, B. Z., and Simcox, A. (1998). EGF domain swap converts a *Drosophila* EGF receptor activator into an inhibitor. *Genes Dev.,* **12,** 908–913.

Schultz, G. E., and Schirmer, R. H. (1979). Principles of Protein Structure. Springer-Verlag, New York.

Schwecke, T., Aparicio, J. F., Molnár, I., König, A., Khaw, L. E., Haydock, S. F., Oliynyk, M., Caffrey, P., Cortés, J., Lester, J. B., Böhn, G. A., Staunton, J., and Leadlay, P. F. (1995). The biosynthetic gene cluster for the polyketide immunosuppressant rapamycin. *Proc. Natl. Acad. Sci. USA,* **92,** 7839–7843.

Serre, V., Guy, H., Penverne, B., Lux, M., Rotgeri, A., Evans, D., and Hervé, G. (1999). Half of *Saccharomyces cerevisiae* carbamoyl phosphate synthetase produces and channels carbamoyl phosphate to the fused aspartate transcarbamoylase domain. *J. Biol. Chem.,* **274,** 23794–23801.

Shao, Z., Zhao, H., Giver, L., and Arnold, F. H. (1998). Random-priming *in vitro* recombination: an effective tool for directed evolution. *Nucleic Acids Res.,* **26,** 681–683.

Short, J. M. (1997). Recombinant approaches for accessing biodiversity. *Nat. Biotechnol.,* **15,** 1322–1323.

Short, J. M. (1999). Method of DNA reassembly by interrupting synthesis. U.S. Patent, no. 5,965,408.

Smith, J., Berg, J. M., and Chandrasegaran, S. (1999). A detailed study of the substrate specificity of a chimeric restriction enzyme. *Nucleic Acids Res.,* **27,** 674–681.

Smithrud, D. B. and Benkovic, S. J. (1997). The state of antibody catalysis. *Curr. Opin. Biotechnol.,* **8,** 459–466.

Sode, K., Watanabe, K., Ito, S., Matsumura, K., and Kikuchi, T. (1995). Thermostable chimeric PQQ glucose dehydrogenase. *FEBS Lett.,* **364,** 325–327.

Souciet, J. L., Nagy, M., Le Gouar, M., Lacroute, F., and Potier, S. (1989). Organization of the yeast URA2 gene: identification of a defective dihydroorotase-like domain in the multifunctional carbamoylphosphate synthetase-aspartate transcarbamylase complex. *Gene,* **79,** 59–70.

Srere, P. A. (1987). Complexes of metabolic enzymes. *Ann. Rev. Biochem.,* **56,** 89–124.

Stachelhaus, T., Schneider, A., and Marahiel, M. A. (1995). Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains. *Science,* **269,** 69–72.

Starzyk, R. M., Burbaum, J. J., and Schimmel, P. (1989). Insertion of new sequences into the catalytic domain of an enzyme. *Biochemistry,* **28,** 8479–8484.

Stemmer, W. P. C. (1994). DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution. *Proc. Natl. Acad. Sci. USA,* **91,** 10747–10751.

Stenlund, P., and Tibell, L. A. (1999). Chimeras of human extracellular and intracellular superoxide dismutases. Analysis of structure and function of the individual domains. *Protein Eng.,* **12,** 319–325.

Takai, T., Yokoyama, C., Wada, K., and Tanabe, T. (1988). Primary structure of chicken liver acetyl-CoA carboxylase deduced from cDNA sequence. *J. Biol. Chem.,* **263,** 2651–2657.

Tang, J., James, M. N., Hsu, I. N., Jenkins, J. A., and Blundell, T. L. (1978). Structural evidence for gene duplication in the evolution of the acid proteases. *Nature,* **271,** 618–621.

Thoden, J. B., Kappock, T. J., Stubbe, J., and Holden, H. M. (1999). Three-dimensional structure of N(5)-carboxyaminoimidazole ribonucleotide synthetase: A member of the ATP grasp protein superfamily. *Biochemistry,* **38,** 15480–15492.

Thornton, J. M., and Sibanda, B. L. (1983). Amino and carboxy-terminal regions in globular proteins. *J. Mol. Biol.,* **167,** 443–460.

Tsien, R. Y. (1998). The green fluorescent protein. *Ann. Rev. Biochem.,* **67,** 509–544.

Wah, D. A., Hirsch, J. A., Dorner, L. F., Schildkraut, I., and Aggarwal, A. K. (1997). Structure of the multimodular endonuclease *Fok*I bound to DNA. *Nature,* **388,** 97–100.

Wah, D. A., Bitinaite, J., Schildkraut, I., and Aggarwal, A. K. (1998). Structure of *Fok*I has implications for DNA cleavage. *Proc. Natl. Acad. Sci. USA,* **95,** 10564–10569.

Wang, W., Kappock, T. J., Stubbe, J., and Ealick, S. E. (1998). X-ray crystal structure of glycinamide ribonucleotide synthetase from *Escherichia coli. Biochemistry,* **37,** 15647–15662.

Wieligmann, K., Norledge, B., Jaenicke, R., and Mayr, E. M. (1998). Eye lens betaB2-crystallin: circular permutation does not influence the oligomerization state but enhances the conformational stability. *J. Mol. Biol.,* **280,** 721–729.

Wright, G., Basak, A. K., Wieligmann, K., Mayr, E. M., and Slingsby, C. (1998). Circular permutation of $\beta$B2-crystallin changes the hierarchy of domain assembly. *Protein Sci.,* **7,** 1280–1285.

Wu, R., Ruben, G., Siegel, B., Jay, E., Spielman P., and Tu, C. D. (1976). Synchronous digestion of SV40 DNA by exonuclease III. *Biochemistry,* **15,** 734–740.

Xu, G. L., and Bestor, T. H. (1997). Cytosine methylation targetted to pre-determined sequences. *Nat. Genet.,* **17,** 376–378.

Xue, Q., Ashley, G., Hutchinson, C. R., and Santi, D. V. (1999). A multiplasmid approach to preparing large libraries of polyketides. *Proc. Natl. Acad. Sci. USA,* **96,** 11740–11745.

Yoshida, H., Kojima, K., Witarto, A. B., and Sode, K. (1999). Engineering a chimeric pyrroloquinoline quinone glucose dehydrogenase: improvement of EDTA tolerance, thermal stability and substrate specificity. *Protein Eng.,* **12,** 63–70.

Zalkin, H., and Nygaard, P. (1999). Biosynthesis of purine nucloetides. In: *"Escherichia coli and Samonella: Cellular and molecular biology"* (F. C. Neidhardt, ed.), pp. 561–579. ASM Press, Washington, D.C.

Zebala, J., and Barany, F. (1991). Mapping catalytically important regions of an enzyme using two-codon insertion mutagenesis: a case study correlating $\beta$-lactamase mutants with the three-dimensional structure. *Gene,* **100,** 51–57.

Zhao, H., Giver, L., Shao, Z., and Arnold, F. H. (1998). Molecular evolution by staggered extension process (StEP) *in vitro* recombination. *Nat. Biotechnology,* **16,** 258–261.