

**Mechanisms of Protein Evolution and Their Application to Protein Engineering**

Margaret E. Glasner<sup>1</sup>, John A. Gerlt<sup>2</sup>, and Patricia C. Babbitt<sup>1</sup>

<sup>1</sup>Department of Biopharmaceutical Sciences  
University of California, San Francisco, California 94143

<sup>2</sup>Departments of Biochemistry and Chemistry,  
University of Illinois, Urbana, Illinois 61801

in

**Advances in Enzymology and Related Areas of Molecular Biology Volume 75:**

**Protein Evolution**

Wiley & Sons, 2007, Vol. 75, pp. 193-239

Edited by Eric Toone

## I. Introduction

## II. Conservation in Protein Evolution

- A. Definition of Terms
- B. Chemistry-Constrained Evolution in Mechanistically Diverse Superfamilies
- C. Exploiting Conservation of Catalysis in Protein Design: Template Selection
- D. Substrate-constrained Evolution in Suprafamilies
- E. Active Site Architecture-Constrained Evolution in Suprafamilies
- F. Exploiting Suprafamilies in Protein Design

## III. Intermediates in Evolutionary Pathways

- A. Promiscuous Enzymes
  - 1. Promiscuity of Natural Enzymes
  - 2. Promiscuity of Recently Evolved Enzymes
  - 3. Promiscuity of Engineered Enzymes
- B. Cryptic Genes
- C. Pseudogenes
- D. Exploiting Promiscuity in Protein Design

## IV. Perspective and Conclusions

Acknowledgements

References

## I. Introduction

Nature has evolved enzymes to catalyze an amazing array of reactions. Looking to this diversity, we see the possibility of redesigning proteins to meet new demands: biosensors, diagnostics, therapeutics, bioremediation, and other applications not yet imagined are driving protein engineering research. Given the wealth of enzymatic and binding activities of natural proteins, it is not unreasonable to believe that by developing powerful experimental methods in parallel with computational approaches we can design proteins to bind almost any compound or catalyze almost any reaction.

On the other hand, consider the complexity of the problem. A small protein of 100 amino acids has  $20^{100}$  possible sequences; as this exceeds the number of atoms in the universe, it is inconceivable that we can explore even a small portion of sequence space. Nevertheless, there have been some remarkable successes in protein engineering, including *de novo* design of a stable protein whose topology has not been observed in nature (1) and transfer of the catalytic motif of triose phosphate isomerase into the scaffold of the nonenzymatic ribose-binding protein (2). Despite these successes, it remains challenging to alter substrate specificity and catalysis of enzymes, and engineered enzymes rarely have the rates or efficiencies of natural proteins.

Computational and experimental methods for protein design routinely imitate some aspects of natural evolution, including mutation, recombination, and selection. Our inability to fully recapitulate the successes of natural evolution suggests that there are evolutionary principles which have not been fully exploited. Evolution proceeds by reusing structures and catalytic motifs, subtly altering proteins to achieve new functions. Learning the rules of this process could revolutionize protein design methods, allowing us to create efficient enzymes catalyzing any variety of reactions not found in nature.

In this review, we highlight two evolutionary concepts that have been underutilized in protein engineering: the conservation of catalytic mechanisms and functional promiscuity. In natural evolution, the amount of sequence space that must be explored to evolve new functions has been largely limited by reusing protein structures. Studies of protein superfamilies have demonstrated that catalytic motifs are often conserved. As a result, some aspect of catalysis, such as a partial chemical reaction, is also conserved between evolutionarily related, but functionally distinct proteins. Thus, knowledge of the structure-function relationships of conserved superfamily catalytic motifs could be used to identify the most promising scaffold, or template, for protein engineering (3-5). Second, considerable evidence has accumulated demonstrating that evolution often proceeds through promiscuous intermediates, suggesting that templates which are naturally promiscuous for a target reaction are ideal, and possibly requisite, for successful protein engineering (6-8). We develop these ideas below, discussing the evidence for and alternatives to these hypotheses, as well as how these concepts might be used in protein engineering.

## II. Conservation in Protein Evolution

Protein design and engineering are confounded by the impossibility of completely sampling sequence space. Much of the focus in protein engineering has been on developing efficient mutagenesis, recombination and screening protocols. Many

experiments to alter substrate specificity or the physico-chemical properties of proteins have been successful, but it has proven difficult to significantly alter the overall reaction or catalytic mechanism. It has been thought that, given the limited amount of sequence space that can be searched, it might be necessary to start with a template that already has some activity for the desired reaction (7). This has often been sufficient for altering substrate specificity, but it is a daunting task to screen potential template enzymes for promiscuous activities significantly different from known enzymatic reactions. Yet this is a primary goal of protein engineering: to redesign enzymes to catalyze novel reactions not observed in nature. Utilizing bioinformatic and experimental methods to understand the requirements for protein activity, such as identifying residues required for catalysis and binding, determining structurally appropriate break points for recombination, and determining the optimal degree, position, and type of mutagenesis, has the potential to revolutionize methods for designing novel catalysts (3, 5).

An aspect of protein engineering pertinent to this problem is how to determine the most promising template for enzyme design. Template selection has received little attention in the field of protein engineering. One exception is the use of family shuffling to create sequence diversity (9). Instead of using random mutagenesis, several homologous genes are fragmented and reassembled by PCR. This results in a library in which mutations tend to be conservative, decreasing the likelihood that the structure and function will be disrupted. As a result of combining conservative mutations, family-shuffled libraries have higher effective diversity and a greater number of improved variants than random mutagenesis libraries. This accelerates the evolutionary process and reduces the number of variants that must be screened (10). Family DNA shuffling is limited by the degree of homology required for effective recombination and is typically performed with sets of genes sharing greater than 60% identity at the protein level (9, 11). At this level of identity, most proteins share the same function (12). Thus, it may still be difficult to significantly alter protein function using this method. Careful consideration of template selection, however, might greatly expand the diversity of catalysts that can be successfully designed.

Understanding how proteins evolve—how nature has chosen templates for protein redesign—is vital if we are to determine how to choose appropriate templates for protein engineering. Two different models for how protein sequence and structure diverge during evolution have been discussed extensively (4, 12-22). The first, suggested by Horowitz in 1945, proposes that substrate binding is the primary evolutionary constraint. This means that the active site residues required to bind a specific substrate (or part of a substrate) are conserved during evolution, whereas active site elements required to catalyze different reactions change. The other model, primary elements of which were first suggested by Jensen in 1976, proposes that catalysis is the primary constraint. In this model, an aspect of catalysis, such as a partial chemical reaction, is conserved, and the protein scaffold evolves to bind sometimes very different substrates and to perform quite different overall reactions. The explosion in genomic information in recent years has allowed an analysis of these hypotheses. The conclusion is that the evolution of new protein activities is primarily constrained by the requirements of catalysis (4, 12, 16-20, 23). Thus, although there are many solutions to substrate binding, it is much more difficult to ensure the precise placement of catalytic residues to achieve an appropriate chemical environment. However, there are instances in which substrate binding or

another aspect of active site architecture are conserved. Several examples of these evolutionary modes are discussed in this section. From this discussion, we will derive principles to guide template selection for protein design and engineering.

### A. Definition of Terms

Before commencing a discussion of these evolutionary models and how they can inform protein design methodology, we define the following terms (4):

1. *Family*: Set of homologous enzymes that share the same function (both mechanism and substrate specificity). Family members usually share greater than 30% sequence identity, but there are examples in which the sequence identity of family members is much lower.
2. *Superfamily*: Sets of homologous enzymes that are unified by a chemical attribute of catalysis. Because both functional and structural information are considered, this definition differs from that commonly used in structural classification schemes such as Structural Classification of Proteins (SCOP, <http://scop.mrc-lmb.cam.ac.uk/scop/>) and Class-Architecture-Topology-Homologous superfamily (CATH, <http://cathwww.biochem.ucl.ac.uk/>) (24-27). Members of superfamilies generally share less than 50% identity, and frequently share less than 20% identity. Superfamilies can be classified into two types, specificity diverse and mechanistically diverse.
  - a. *Specificity diverse superfamily*: Sets of homologous enzymes that catalyze the same overall reaction with differing substrate specificities. The serine proteases are an example.
  - b. *Mechanistically diverse superfamily*: Sets of homologous enzymes that catalyze different overall reactions with differing substrate specificities, but which share a common mechanistic attribute (a specific partial reaction, intermediate, transition state, or the use of a fundamental chemical capability such as an oxyanion hole). The enolase superfamily, discussed below, is an example. Recently, a curated database called the Structure-Function Linkage Database (SFLD, <http://sfld.rbvi.ucsf.edu/>) was developed to capture information on shared mechanistic attributes (such as partial reactions), conserved catalytic residues, structures, and functions in mechanistically diverse superfamilies (28).
3. *Suprafamily*: Sets of homologous enzymes that catalyze different overall reactions which do not share a common mechanistic attribute. Active site residues might be conserved, but these perform different functions. Suprafamilies can also be classified into two types.
  - a. *Substrate-constrained suprafamily*: Sets of homologous enzymes in which binding of a substrate (or part of a substrate) has been conserved. For instance, in the histidine and tryptophan biosynthesis suprafamily discussed below, a phosphate binding site has been conserved. This type of suprafamily demonstrates conservation of substrate binding as proposed by Horowitz (13, 14).

- b. *Active site architecture-constrained suprafamily*: Sets of homologous enzymes in which some aspect of active site architecture has been conserved, such as the orientation and placement of catalytic residues.

TABLE 1  
Examples of Mechanistically Diverse Superfamilies

Superfamily	Fundamental Common Chemistry	Representative Superfamily Members	References
Enolase	Abstraction of a proton $\alpha$ to a carboxylate to form metal-ion stabilized enolate intermediates	Enolase Mandelate racemase <i>o</i> -succinylbenzoate synthase	(32)
Haloacid Dehalogenase	Formation of covalent enzyme-substrate intermediates via a conserved aspartate	$\beta$ -phosphoglucomutase Phosphonoacetaldehyde hydrolase Cu <sup>2+</sup> /H <sup>+</sup> -ATPase	(37)
Crotonase	Utilization of an oxyanion hole to stabilize enolate-anion intermediates generally derived from thioesters	4-Chlorobenzoyl-CoA dehalogenase Methylmalonyl-CoA decarboxylase ClpP Protease	(4, 36)
Amidohydrolase	Metal-assisted hydrolysis	Urease Phosphotriesterase Adenosine deaminase	(35, 38)
Vicinal Oxygen Chelate	Stabilization of oxyanion intermediates formed by metal-dependent catalysis	Dioxygenase Glyoxylase I Methylmalonyl CoA epimerase	(4, 33)
<i>N</i> -Acetylneuraminatase Lyase	Utilization of a protonated Schiff base as an electron sink	<i>N</i> -acetylneuraminatase lyase Dihydrodipicolinate synthase 2-keto-3-deoxygluconate aldolase	(23, 34, 120)

## B. Chemistry-Constrained Evolution in Mechanistically Diverse Superfamilies

There is overwhelming evidence that partial chemical reactions are frequently conserved in protein evolution, suggesting that consideration of partial reactions will be valuable for selecting templates for protein engineering (4, 12, 16-20, 23). For example, surveys of the *E. coli* genome found that most protein domains have homologs in more than one metabolic pathway and that some aspect of the catalytic mechanism is conserved much more frequently than substrate recognition (20-22, 29). In addition, analysis of structurally defined superfamilies in the CATH database reveal that chemistry is at least partially conserved in 22 out of 27 superfamilies which exhibit significant functional variation (12). Detailed studies of mechanistically diverse superfamilies revealed that the shared attribute is frequently the conservation of a partial reaction, intermediate, or transition state (4, 30). In mechanistically diverse superfamilies, some catalytic residues are well-conserved across the superfamily and are required for the conserved partial reaction (or other mechanistic attribute common to the superfamily), whereas the identities and sometimes the positions of additional catalytic residues can vary (30, 31). For instance, in the enolase superfamily which is discussed in more detail below, the three metal-binding residues are nearly universally conserved, but the identity and position of the general base differs in different subgroups of the superfamily (32). Several mechanistically diverse superfamilies have been extensively characterized (Table 1) (reviewed in ref. 4, 23, 32-38). For example, the haloacid dehalogenase superfamily uses a Mg<sup>2+</sup> cofactor bound to a conserved motif to form a covalent enzyme-substrate

Figure 1A

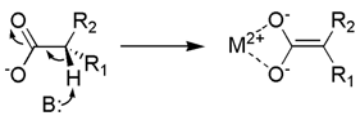


Figure 1C

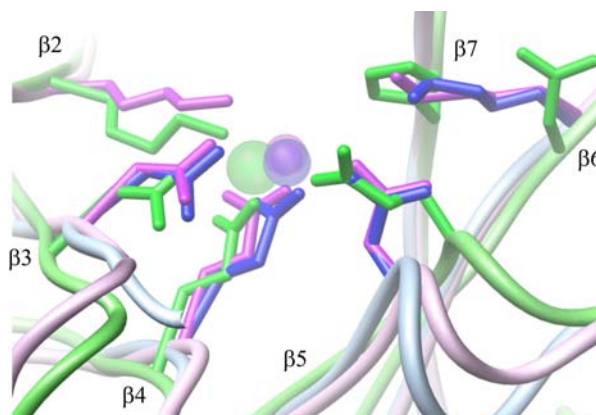


Figure 1B

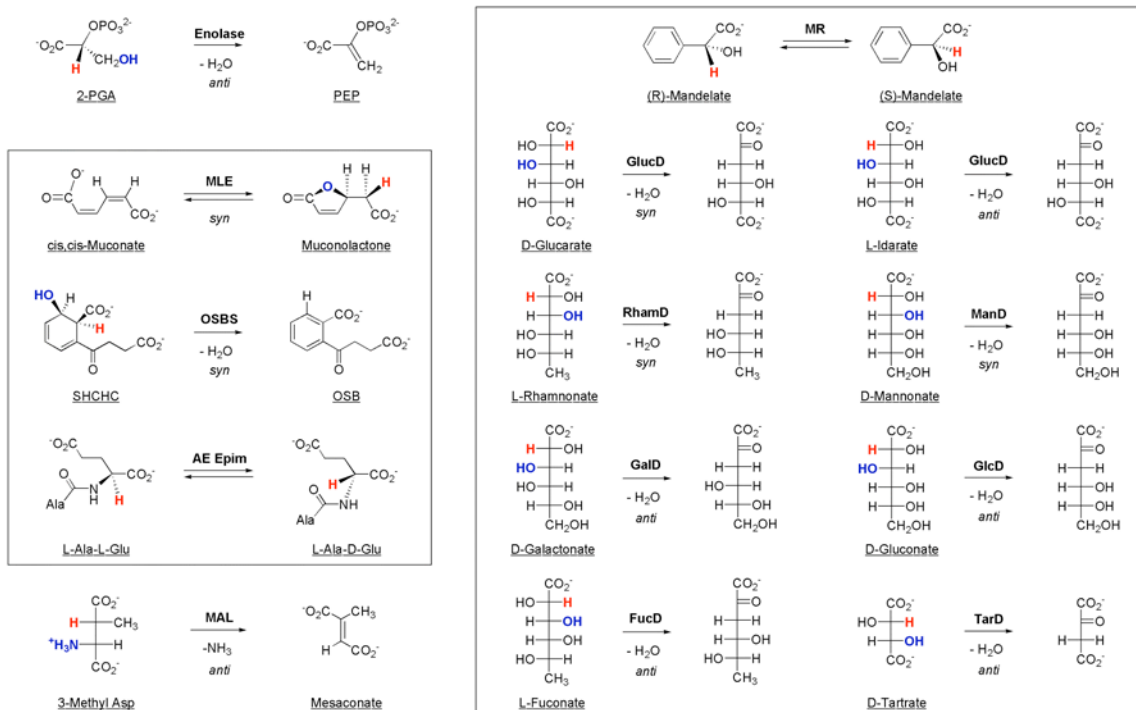


Figure 1. The enolase superfamily. A) The enolase superfamily partial reaction, in which the  $\alpha$  proton to a carboxylate is abstracted, leading to a metal-stabilized enolate anion intermediate. B) Reactions catalyzed by the enolase superfamily. The abstracted hydrogen is shown in red and the leaving group in blue. C) Structural superposition of Enolase (blue; pdb code 7ENL), MR (green; pdb code 2MNR), and MLE (purple; pdb code 1MUC) showing the positions of conserved catalytic residues in the active site. The  $\beta$ -strands are numbered, and the metal ions are shown as semi-transparent spheres.

intermediate; the crotonase superfamily forms an oxyanion hole using two structurally conserved peptide NH groups to stabilize an enolate anion intermediate derived from an acyl-CoA substrate; and the amidohydrolase superfamily relies on metal-assisted hydrolysis (4, 28, 36-38). Each of these superfamilies is known to catalyze at least eleven different reactions.

To describe in more detail the characteristics of mechanistically diverse superfamilies, we will focus on the enolase superfamily. The enolase superfamily is one of the most extensively characterized superfamilies, and it serves as an example for many of the concepts discussed in this review, including how to utilize knowledge of mechanistically diverse superfamilies to select templates for protein engineering. The discovery of the enolase superfamily, the first recognized mechanistically diverse superfamily, was unexpected. Structural superposition of mandelate racemase (MR) and muconate lactonizing enzyme (MLE) from *Pseudomonas putida* revealed that they have not only strikingly similar structures, but also the same placement of catalytic residues (39). These enzymes share only ~25% identity, and at the time these structures were solved, it was not conclusive from their sequence alignment that they were evolutionarily related. Further analysis demonstrated that these two enzymes, along with enolase and five other homologous enzymes which catalyze different reactions, have identically positioned catalytic residues and catalyze a common partial reaction leading to an enolate intermediate stabilized by a divalent metal ion (Figure 1A) (17, 18). Currently, the Structure-Function Linkage Database lists ~1000 enolase superfamily members which catalyze at least fourteen different reactions (Figure 1B) (28, 32).

Structurally, the enolase superfamily is defined by a common scaffold and conserved active site residues which catalyze the superfamily partial reaction. Members of this superfamily have two domains—a C-terminal modified  $(\beta/\alpha)_8$ -barrel domain [ $(\beta/\alpha)_7\beta$ ] and an  $\alpha + \beta$  domain comprised of elements from both the N- and C-terminals. Like other  $(\beta/\alpha)_8$ -barrel proteins, the active site lies in a depression formed by the C-terminal ends of the barrel strands; the N-terminal domain caps the barrel to close off the active site and appears to have a role in substrate specificity. Catalytic residues are arranged around the inside of the barrel at the ends of the  $\beta$ -strands (Figure 1C). All superfamily members have conserved residues (usually Glu or Asp) at the ends of the third, fourth, and fifth  $\beta$ -strand which coordinate the essential metal ion. In addition, a Lys at the end of the sixth  $\beta$ -strand or His at the end of the seventh  $\beta$ -strand acts as a general base. Using this catalytic machinery, the enolase superfamily catalyzes the abstraction of a proton  $\alpha$  to a carboxylate group to form a metal-stabilized enolate intermediate.

Aside from the canonical superfamily partial reaction, enolase superfamily members catalyze extremely divergent reactions, including racemization, cycloisomerization, and  $\beta$ -elimination (Figure 1B). The superfamily can be divided into four subgroups based on the identity and position of the catalytic residues (32). A survey of currently available sequences shows that two of these, the enolase and 3-methylaspartate ammonia lyase (MAL) subgroups, appear to be monofunctional, whereas the MR and MLE subgroups include proteins of several different functions. In addition to mandelate racemase, the MR subgroup includes at least seven different acid-sugar dehydratases (32, 40). The MLE subgroup is possibly the most diverse, as it includes MLE (cycloisomerization), *o*-succinylbenzoate synthase ( $\beta$ -elimination), and L-Ala-D/L-



Glu epimerase (racemization). Since the functions of only about half of the sequences in the MR and MLE subgroups can be assigned reliably, more enolase superfamily activities undoubtedly remain to be discovered.

### C. Exploiting Conservation of Catalysis in Protein Design: Template selection

Because conservation of catalysis appears to be a common theme in protein evolution, an approach for identifying suitable templates for protein engineering is to imitate nature and utilize mechanistically diverse superfamilies. By carefully considering the set of partial reactions required to perform a target reaction, an appropriate template could be chosen from a superfamily that relies on the most critical of those partial reactions. The hypothesis is that although the template might lack activity for the target reaction, the ability to catalyze a required partial reaction would be hard-wired into the active site scaffold. Consequently, achieving activity for the target reaction might require very few mutations.

As proof-of-principle for this methodology, Schmidt et al. assessed the difficulty of engineering an enzyme to perform a reaction catalyzed by another member of its superfamily (41). They redesigned two members of the enolase superfamily, muconate lactonizing enzyme II (MLE II) from *Pseudomonas* sp. P51 and L-Ala-D/L-Glu epimerase (AEE) from *E. coli*, to catalyze the reaction of a third enolase superfamily member, *o*-succinylbenzoate synthase (OSBS) (Figure 1B). Neither template exhibits detectable OSBS activity. DNA shuffling (gene fragmentation followed by mutagenic PCR to reassemble the gene) was performed using MLE II as a template, and OSBS activity was selected by complementation of an OSBS auxotroph. The selected mutant differed from wild-type MLE II by a single mutation. In parallel, the structures of AEE and OSBS from *E. coli* were compared, and a point mutant of AEE which also complemented the OSBS auxotroph was designed. Remarkably, this mutation was in the analogous position in both enzymes, exchanging an Asp (AEE) or Glu (MLE II) for Gly to make room for the succinyl moiety of OSB. The rate acceleration ( $k_{\text{cat}}/k_{\text{uncat}}$ ) of the MLE II mutant for OSBS activity ( $10^{10}$ ) is only 10-fold lower than the rate acceleration of wild-type OSBS from *E. coli* ( $10^{11}$ ), although its efficiency ( $k_{\text{cat}}/K_{\text{M}}$ ) is 1000-fold less. In addition, the MLE II mutant retained cycloisomerization activity at a rate only ~10-fold slower than wild-type MLE II and at an efficiency comparable to its OSBS activity. The AEE mutant is a much poorer catalyst, achieving a rate acceleration of  $10^7$  and an efficiency ~ $10^5$ -fold less than wildtype OSBS. It also retained its native AEE activity, although at a 1000-fold decrease in efficiency. These results demonstrate that it is not necessary to have pre-existing activity for the target reaction if starting from a suitable template. However, OSBS activity is very exergonic, and the OSBS family of enzymes is highly divergent—proteins with as little as 15% identity catalyze the same reaction. This implies that constraints on the active site geometry of OSBS are fairly relaxed and that it might be relatively easy to evolve OSBS activity (42, 43). As a further test of the effectiveness of basing template selection on mechanistically diverse superfamilies, it will be necessary to determine whether other superfamily reactions can be as easily engineered onto templates within the enolase superfamily.

Although more experiments need to be performed to demonstrate the general utility of basing template selection on knowledge of superfamily partial reactions, the

preceding example begins to demonstrate the feasibility of this approach. The ultimate test will be to engineer a protein to perform a reaction which is not catalyzed by any other member of its superfamily but which requires the superfamily partial reaction. The success of this experiment would suggest a recipe for protein engineering in which a target reaction would be broken down into potential partial reactions, and a template catalyzing one of the partial reactions would be selected from a catalog of superfamily partial reactions, such as provided in the SFLD (28).

#### D. Substrate-Constrained Evolution in Suprafamilies

Understanding that proteins frequently evolve through conservation of chemistry suggests that selecting templates based on the ability to catalyze a required partial reaction could improve protein engineering methodology. However, deciphering the common attributes of mechanistically diverse superfamilies and analyzing the critical aspects of a target reaction to select the most promising superfamily template is not trivial. Do other modes of protein evolution offer simpler, effective alternatives? As discussed below, conservation of substrate binding is relatively rare in nature and the examples of this type of evolution are not simple.

The concept of substrate-constrained evolution was developed from a hypothesis suggesting that metabolic pathways evolved backwards in a process often called retrograde evolution (13, 14). According to this hypothesis, as prebiotically synthesized nutrients became limiting, there was selective pressure to add steps to metabolic pathways to utilize available precursors. The evolution of these new activities would have followed gene duplication and divergence, giving rise to operons consisting of homologous enzymes with similar binding specificities but catalyzing different chemical reactions. In an exhaustive survey of the protein domains involved in small molecule metabolism in *E. coli*, Teichmann et al. demonstrated that although 56 protein domains have homologs within the same pathway, only seven of these appear to retain substrate binding and alter the catalytic mechanism, demonstrating that substrate-constrained evolution is uncommon (20). More recently, however, Nobeli, et al. determined that for the majority of structurally defined superfamilies from the CATH database, some aspect of the substrate structure was conserved (44). Conservation of ligand substructures might represent substrate-constrained evolution, but it could also reflect structural requirements necessary for the common partial reactions or other mechanistic attributes of mechanistically diverse superfamilies.

The canonical examples of substrate-constrained evolution are consecutive enzymes in the histidine and tryptophan biosynthesis pathways (4). In histidine biosynthesis, phosphoribosyl-formimino-5-aminoimidazole carboxamide ribonucleotide isomerase (HisA; ProFAR) and imidazole-3-glycerol phosphate synthase (HisF; ImGPS) are not only consecutive in the pathway, but are also adjacent to each other on the *E. coli* chromosome. In tryptophan biosynthesis, phosphoribosylanthranilate isomerase (TrpF; PRAI) and indoleglycerol phosphate synthase (TrpC; InGPS) are fused to form a single, two-domain protein in *E. coli* (often designated together as TrpC) but are separate genes in other organisms. Both pairs of enzymes are  $(\beta/\alpha)_8$ -barrels, and both are quite divergent: HisA and HisF share ~25% identity, and TrpF and TrpC share 22% identity (4, 45, 46). Both HisA and TrpF, which share ~10% identity, catalyze Amadori

Figure 2A

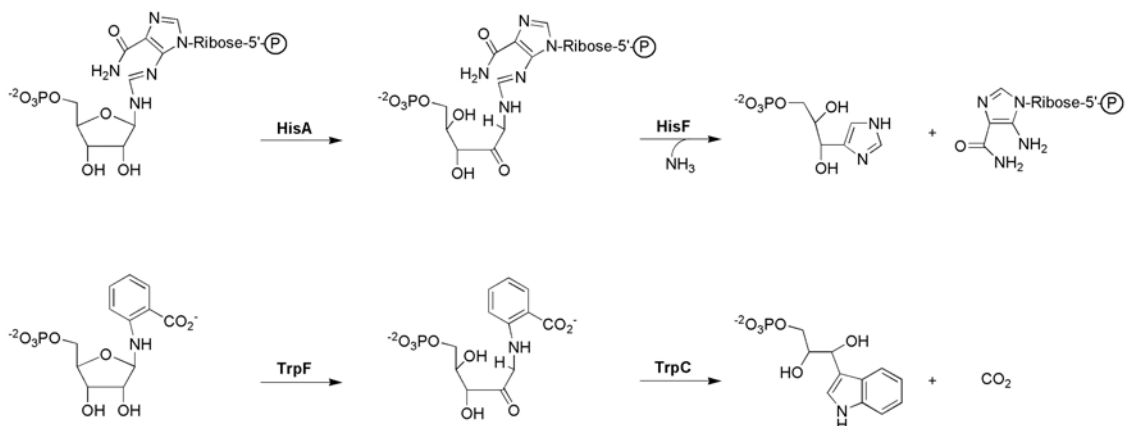


Figure 2B

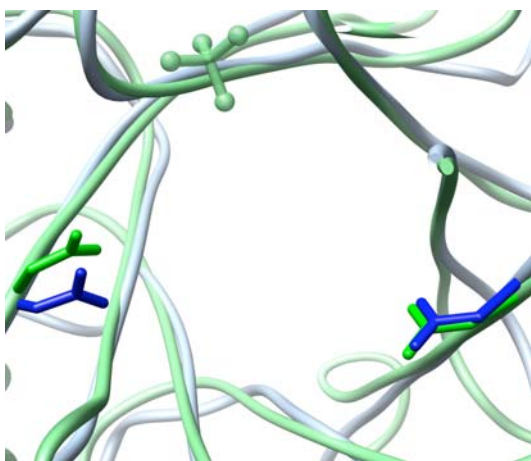


Figure 2C

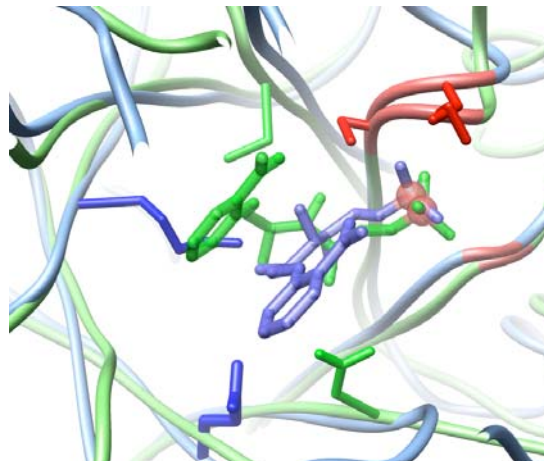


Figure 2. Histidine and tryptophan biosynthesis enzymes. A) Reactions catalyzed by HisA, HisF, TrpF, and TrpC. B) Structural superposition of HisA (blue; pdb code 1QO2) and HisF (green; pdb code 1THF) showing that the catalytic residues are positioned nearly identically in the active site. Phosphate (green ball and stick) is shown bound to HisF. C) Active site superposition of TrpF (green; pdb code 1LBM) and TrpC (blue; pdb code 1LBF) bound to 1-(*o*-carboxyphenylamino)-1-deoxyribulose 5-phosphate, a product analog of TrpF and substrate analog of TrpC. Catalytic residues of TrpF (green) and TrpC (blue) are shown and do not superimpose. However, the phosphate binding site (red) and phosphorous (red sphere) superimpose well.

rearrangements on structurally similar substrates (substituted 5'-phosphoribosylamines), while HisF and TrpC catalyze cleavage of the substrate to form imidazole glycerol phosphate or ring closure to form indoleglycerol phosphate, respectively (Figure 2A) (47, 48).

Recent evidence has called into question the hypothesis that these enzyme pairs represent purely substrate-constrained, retrograde evolution. Structure and mutagenesis experiments demonstrated that HisA and HisF utilize general acid/base mechanisms requiring aspartate residues at analogous positions (Figure 2B) (47-49). Although the specific enzymatic mechanisms of the two proteins are quite different, the striking conservation and superposition of catalytic residues indicates that both substrate binding and some aspects of the catalytic machinery have been conserved during the evolution of HisA and HisF.

The case for retrograde, substrate-constrained evolution of TrpF and TrpC is tenuous for different reasons. TrpF and TrpC also use general acid/base mechanisms, but their catalytic residues do not align in the sequence or structure (Figure 2C) (47, 50). This would suggest that this pair has evolved by substrate-constrained evolution, except for the fact that the ligand is bound in different orientations (Figure 2C). In two structures of TrpC bound to a substrate analog or the product, the ring portion of the ligand is in different orientations, neither of which matches the orientation of the ligand in TrpF. The main sequence and structural similarity between TrpF and TrpC is the phosphate binding pocket, a motif which is also found in many other  $(\beta/\alpha)_8$ -barrel proteins, including the histidine biosynthesis enzymes (19, 51, 52). Thus, only the binding of the chemically-inert portion of the ligand is conserved.

In addition to the common phosphate-binding motif, there is some functional evidence suggesting a relationship between TrpF and the histidine biosynthesis enzymes. TrpF and HisA catalyze analogous reactions in which a side group of the substrates differ. Interestingly, point mutagenesis at analogous positions in HisA and HisF abolishes their native activity but endows them with a low level of TrpF activity by removing a charged residue which prevents the TrpF substrate from binding (49, 53). In addition, *Streptomyces coelicolor* A3(2) and *Mycobacterium tuberculosis* HR37Rv encode a single domain, bifunctional HisA/TrpF instead of two separate enzymes (54). Although this functional evidence linking TrpF, HisA, and HisF is suggestive, the current data cannot establish a sequence- or structure-based evolutionary link because the catalytic residues of TrpF do not structurally superimpose on those of HisA or HisF, although they are located at similar positions in the active site (47). A rigorous analysis of all histidine and tryptophan biosynthesis enzymes will be required to determine if there are conserved sequence and structural elements beyond the phosphate-binding motif.

Although the studies of tryptophan and histidine biosynthesis enzymes discussed above have revealed considerable insight into the function and structure of these enzymes, the results suggest that our understanding of these enzymes' evolution is too simplistic. Thus, HisA and HisF may have evolved by retrograde evolution, but both substrate binding and catalytic residues have been conserved. Likewise, TrpF and TrpC might represent another example of retrograde evolution, or they may be part of a larger suprafamily of phosphate-binding  $(\beta/\alpha)_8$ -barrel proteins that coincidentally catalyze sequential reactions. Considering the evidence presented here, it would be worthwhile to reevaluate whether these enzymes or the other examples of substrate-constrained

evolution identified in *E. coli* are most closely related to their homologs within their metabolic pathways (as would be predicted by the retrograde evolution hypothesis) and if they share any catalytic similarities.

Although the evolutionary relationships of the histidine and tryptophan biosynthesis enzymes are not completely understood, protein engineering experiments of these enzymes offer another example of basing template selection on conserved catalytic capabilities. According to this hypothesis, HisA should be a reasonable template for engineering the TrpF reaction because HisA catalyzes a reaction analogous to that of TrpF and its catalytic residues are in similar positions. Jurgens et al. performed random mutagenesis on HisA from *Thermotoga maritima* and selected TrpF activity by complementation of a TrpF deficient *E. coli* strain (49). They isolated several variants of HisA that had TrpF activity. One of these had three mutations and was bifunctional for HisA and TrpF activities, reminiscent of the natural HisA/TrpF enzyme found in *S. coelicolor* and *M. tuberculosis* (54). The TrpF activity of another variant, however, was due to a single mutation which abolished HisA activity. Surprisingly, introducing the same mutation into HisF, which performs a reaction that bears much less similarity to the TrpF reaction, also endowed HisF with TrpF activity (53).

#### E. Active Site Architecture-Constrained Evolution in Suprafamilies

In the previous sections, we discussed two mechanisms by which proteins could evolve. As nature rarely follows such simple dichotomies, we must now consider a third alternative. In active site architecture-constrained suprafamilies, overall structure and placement of some catalytic residues are conserved; however, the catalytic residues serve different functions, and there is no common mechanistic attribute (4). In a study of 24 pairs of homologous enzymes which catalyze completely different reactions, Bartlett, et al. identified six pairs of enzymes that appear to share no mechanistic attributes but which have common active site architectures, suggesting that this mode of evolution is not uncommon (30). The dominance of active site architecture in protein evolution has been analyzed in two suprafamilies—the orotidine 5'-monophosphate decarboxylase (OMPDC) suprafamily and the thioredoxin suprafamily (55, 56).

The OMPDC suprafamily is comprised of two mechanistically distinct groups (55). OMPDC is the most proficient enzyme known and catalyzes the metal ion-independent decarboxylation of orotidine monophosphate to form uridine monophosphate (Figure 3A) (57). In contrast, several other members of the suprafamily catalyze  $Mg^{2+}$ -dependent reactions that are likely to proceed through an enediolate intermediate (58). Although the sequence identity between the two groups is less than 25%, there are several conserved residues which are well-aligned in the structures of OMPDC and the metal-dependent 3-keto-L-gulonate 6-phosphate decarboxylase (KGPDC) (Figure 3B,C). One of the most remarkable aspects of these two structures is that they form nearly identical homodimers in which the active site is formed at the dimer interface, with both monomers contributing residues to the active site. However, these residues serve different functions. For instance, K62 in OMPDC probably serves as the proton donor, while the homologous K64 of KGPDC stabilizes the anion intermediate (58-60).

Figure 3A

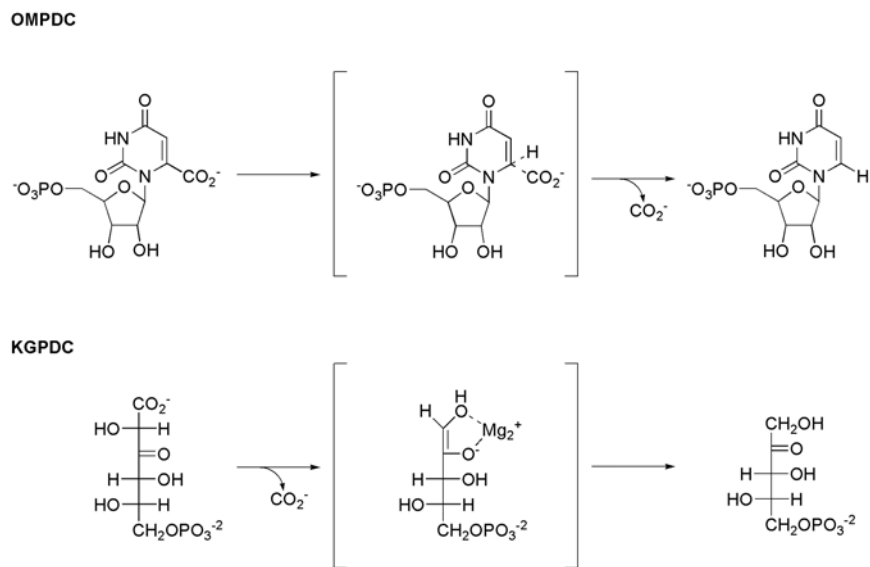


Figure 3B

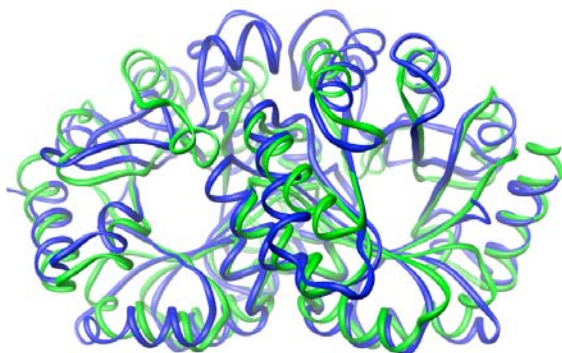


Figure 3C

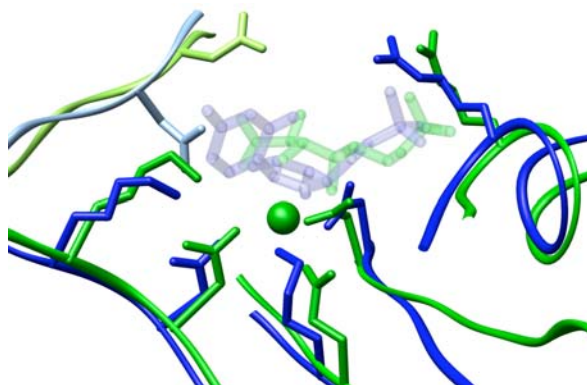


Figure 3. The OMPDC/KGPDC suprafamily. A) Reactions catalyzed by OMPDC and KGPDC. Several hypotheses of OMPDC's reaction mechanism have been proposed; a possible intermediate is shown (59, 121). B) Structural superposition of OMPDC (blue; pdb code 1DBT) and KGPDC (green; pdb code 1KW1) showing their unusual dimeric structure. C) Structural superposition of the active site showing conserved residues. Residues in light blue (OMPDC) and light green (KGPDC) are from the second monomer of the dimer. The ligands, shown as semitransparent structures, are uridine 5'-monophosphate (blue) and L-gulonate 6-phosphate (green)

Likewise, other residues which interact with the substrate directly to mediate catalysis in OMPDC also coordinate the  $Mg^{2+}$  ion in KGPDC (55, 61).

Recently, Copley et al. defined a second suprafamily by detecting sequence and structural relationships between the thioredoxin and peroxiredoxin superfamilies (56). The thioredoxins and peroxiredoxins, as well as glutathione S-transferases, protein disulfide isomerases, and several other superfamilies, are members of the thioredoxin fold class and carry out oxidation-reduction reactions (Figure 4A). Although this suggests an evolutionary relationship among these proteins, their sequence identity is nearly insignificant, and insertions and extensions of the thioredoxin fold further complicate the analysis of their relationships. Copley, et al. discovered that thioredoxins and peroxiredoxins are both related to a group of cytochrome maturation proteins (CMPs), which catalyze thiol oxidoreductase reactions, like thioredoxins. Like peroxiredoxins, however, CMPs have an insertion after the second  $\beta$ -strand of the canonical thioredoxin fold. Motif analysis identified one highly significant motif between CMPs and thioredoxin and several motifs that CMPs share with peroxiredoxins. Analysis of the active site revealed that thioredoxins, CMPs, and peroxiredoxins have an identically placed cysteine residue (Figure 4B). The positions of other important active site residues are also conserved, but their identities and function in catalysis vary. In peroxiredoxins, the conserved Cys, which aligns with the C-terminal Cys of a CXXC motif found in thioredoxin and CMPs, attacks the substrate; in thioredoxin and CMPs, however, the N-terminal Cys of the CXXC motif, which is replaced by threonine in peroxiredoxin, attacks the substrate. Thus, although thioredoxins, CMPs, and peroxiredoxins appear to have diverged from a common ancestor and share the same active site architecture, they catalyze different redox reactions, and their conserved residues function differently.

The characterization of two suprafamilies with divergent catalytic mechanisms (55, 56) and the identification of several others that also appear to have evolved by the same mechanism (30) raises the possibility that conservation of active site architecture is a widespread phenomenon. Indeed, it has been suggested that a majority of  $(\beta/\alpha)_8$ -barrel proteins are evolutionarily related (19, 52, 62). An intriguing possibility that could potentially be applied to protein engineering is that  $(\beta/\alpha)_8$ -barrel fold proteins have arisen by recombination of half barrels, thus introducing different catalytic combinations into new contexts (19, 63-69). Without much more detailed scrutiny of suprafamily relationships, however, the occurrence of common fold classes by convergent evolution cannot be ruled out. Further research into potential suprafamily relationships will elucidate whether evolution by conservation of active site architecture, convergent evolution, or other mechanisms such as recombination of subdomains has been predominant.

Figure 4A

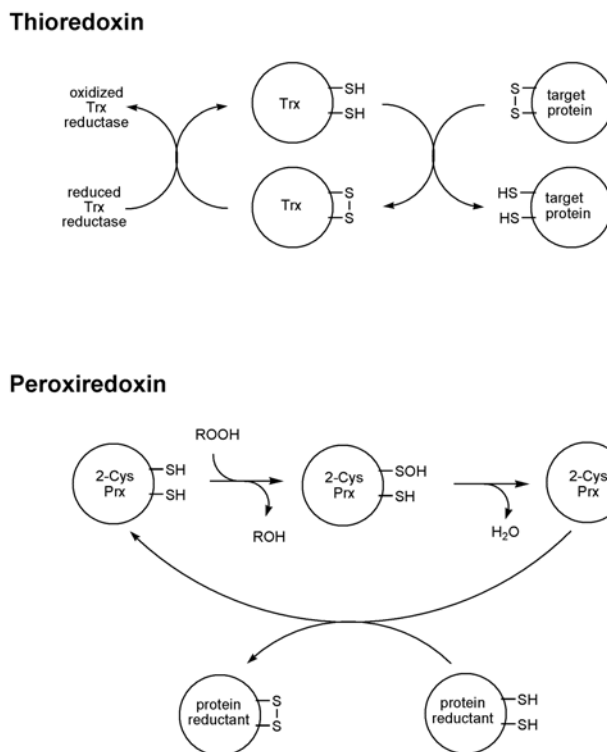


Figure 4B

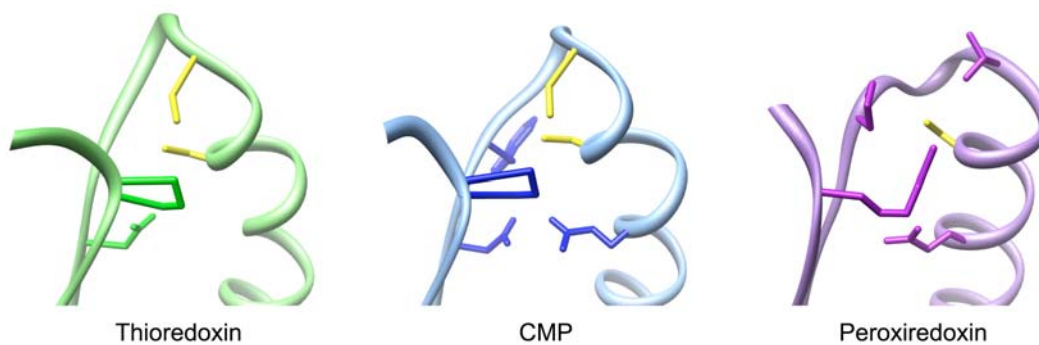


Figure 4. The thioredoxin/perioredoxin suprafamily. A) Reactions of thioredoxins and peroxiredoxins which use two Cys residues. The protein reductant used for 2-Cys peroxiredoxins include thioredoxin and glutaredoxin, among others. B) Comparison of the active sites of thioredoxin (green; pdb code 1XOA), a CMP (blue; pdb code 1KNG), and a peroxiredoxin (purple; pdb code 1PSQ). Positions of important active site residues conserved between at least two proteins are shown. Cysteine residues are shown in yellow. Figure 4A reproduced with permission from Copley et al. (2004), *Biochemistry*, 43:13981-13995. Copyright 2004 Am. Chem. Soc.



## F. Exploiting Suprafamilies in Protein Design

Can the suprafamily modes of protein evolution guide protein design and engineering? Substrate-constrained evolution is rare, and active site architecture-constrained evolution is difficult to decipher. This suggests that basing template selection on these criteria would rarely succeed. However, the examples discussed above do suggest two principles that might be useful for protein engineering. First, certain binding motifs, such as the phosphate binding motif, are well-conserved. Incorporating such motifs into template scaffolds so that a substrate or cofactor is oriented correctly for catalysis could be a useful strategy. Second, suprafamily analysis indicates that evolutionary conservation exceeds what is easily visible in sequence conservation. Some common folds such as  $(\beta/\alpha)_8$ -barrels appear to be especially plastic, in that catalytic residues can approach the ligand from any of the eight strands (23). The apparent modularity of  $(\beta/\alpha)_8$ -barrels suggests that using recombination to reassort barrel fragments might be a promising approach (66, 68). Undoubtedly, further research into superfamily and suprafamily relationships will suggest other possibilities for improving protein engineering methods.

## III. Intermediates in Evolutionary Pathways

The preceding section discussed elements of protein structure and function that are conserved in evolution. Here, we turn to pathways by which proteins evolve. For protein engineering and perhaps natural evolution, the principle concept is that protein evolution often proceeds through promiscuous intermediates, although evolution is opportunistic and can proceed by other mechanisms as well. Three genetic sources for new protein activities have been proposed: genes encoding functional proteins that have promiscuous side reactions; cryptic genes, which are functional but are expressed only under rare circumstances; and pseudogenes, which are nonfunctional and not expressed. (Although cryptic genes may also encode promiscuous proteins, they are differentiated by the fact that their expression is induced by unusual mechanisms such as frameshifts or point mutations.) The primary difference among these hypotheses is how they balance the accumulation of mutations with functional constraints. Pseudogenes are not under selective pressure to maintain a functional product and can accumulate more mutations, but there is no way to select against deleterious mutations. This severely limits the possibilities for resurrecting pseudogenes to perform new functions. Cryptic genes can accumulate mutations like pseudogenes if they are not expressed for generations, but deleterious mutations can be periodically purged from the population when expression of the cryptic gene is required (70). Because promiscuous proteins must maintain their primary function, they are the least likely to accumulate deleterious mutations, but they also have the most constraints on their evolution. As we discuss here, there is considerable evidence supporting the use of promiscuous genes as scaffolds for the evolution of new activities, suggesting that templates which are naturally promiscuous for a target reaction are ideal, and possibly requisite, for successful protein engineering (6-8).

## A. Promiscuous Enzymes

Gene duplication followed by specialization of the two copies to perform different functions has been the main paradigm advanced to explain protein evolution for decades (71). In recent years, however, this idea has been challenged by the converse hypothesis that evolution of new functions can precede gene duplication, and that gene duplication is positively selected because it allows the activities of multi-functional proteins to be optimized individually (72-74). In the context of enzyme evolution, this hypothesis proposes that new enzymes evolve from catalytically promiscuous intermediates (4, 6, 7, 15, 41, 75, 76). The promiscuity hypothesis differs from the concept of moonlighting in that all enzymatic activities are assumed to be catalyzed by the same active site (76, 77). Evidence from both functional characterization of natural enzymes and protein engineering experiments support the idea that evolution proceeds through promiscuous intermediates.

### 1. Promiscuity of Natural Enzymes

A number of proteins are known to have promiscuous activities (reviewed in refs. 6, 8, 76). Typically, the catalytic proficiency of the secondary reaction is substantially lower than that of the primary reaction, but these levels might be sufficient to provide a selective advantage under the right circumstances (6). In addition, there are different degrees of promiscuity, ranging from catalyzing essentially the same reaction on substrates differing only by the stereochemistry of a single atom, to catalyzing substantially different reactions on substrates with little chemical similarity (8, 76). The fact that the promiscuous activities of several enzymes are the physiological activities of evolutionarily related proteins is compelling evidence for the role of promiscuity in enzyme evolution (6). Below, we discuss several promiscuous enzymes which vary greatly in the similarity of their primary and promiscuous reactions and illustrate how this promiscuity could lead to the evolution of new enzymatic activities and, thus, to useful strategies for protein engineering in the laboratory (Table 2).

There are many examples of enzymes that utilize alternative substrates analogous to their natural substrates. The bifunctional HisA/TrpF enzyme from *S. coelicolor* is a prime example (54). In the enolase superfamily, there are at least seven families of acid-sugar dehydratases which perform the same overall reaction on different substrates (32, 40). Although specificity of the characterized members in each family are generally thought to be limited to a single substrate, there is a striking example in which dual substrate specificity has been selected by evolution. Not only does the gluconate dehydratase of *Sulfolobus solfataricus* utilize both gluconate and galactonate, but the other two enzymes in the pathway (glucose dehydrogenase and 2-keto-3-deoxygluconate aldolase) share the same dual specificity (78, 79). By sequence similarity, the *Sulfolobus* gluconate/galactonate dehydratase is more closely related to members of the gluconate dehydratase family than the galactonate dehydratase family, but further study is required to determine whether the *Sulfolobus* dehydratase is intermediate between the two families. Such data would add considerable support to the hypothesis that enzyme evolution proceeds through promiscuous intermediates.

In the second example, the differences among the native and promiscuous substrates are located at the site of chemistry, rather than at a distance from the reacting atoms. Alkaline phosphatase is a member of a superfamily which includes phosphate mono- and diesterases and sulfate esterases (80). Members of this superfamily utilize a common catalytic mechanism requiring divalent metal ions, but the sequence identity between alkaline phosphatase and the superfamily members phosphodiesterase and arylsulfatase is very low (<20%) (80-82). In spite of this, alkaline phosphatase has low levels of both phosphodiesterase and sulfatase activity, demonstrating the potential for the catalysis of similar reactions to evolve by catalytic promiscuity (83, 84).

In the last two examples, the native and promiscuous substrates bear little resemblance to each other, and the two reactions are quite dissimilar. First, human maleylacetoacetate isomerase (MAAI) has a central role in the catabolism of phenylalanine and tyrosine; unlike other enzymes in this pathway, which are expressed only in the liver and kidney, MAAI is expressed in several other tissues, suggesting that it has an additional function, possibly detoxification (85, 86). In fact, MAAI was first identified as the zeta-class glutathione-S-transferase GST Z1-1, which has low levels of peroxidase activity and higher levels of dehalogenation activity (87, 88). It is unknown whether these last two activities are physiologically relevant.

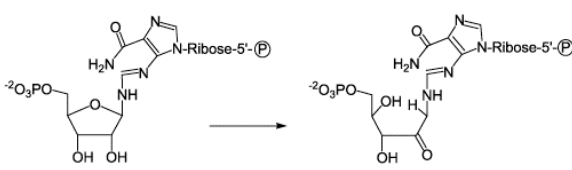
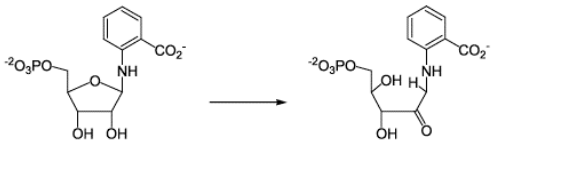
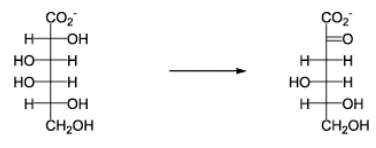
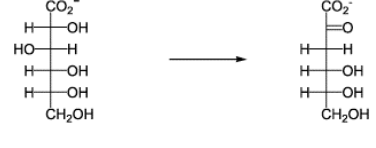
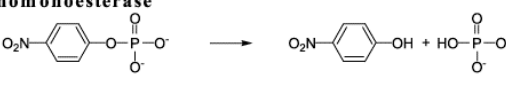
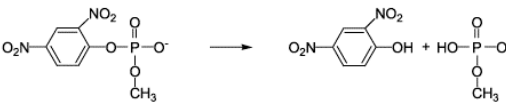
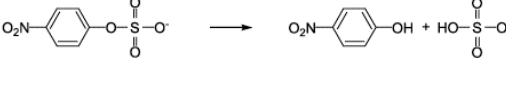
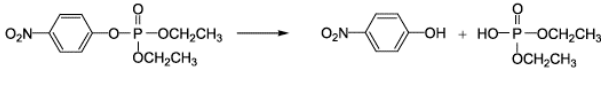
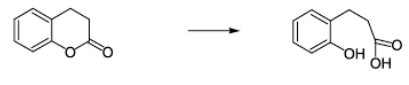
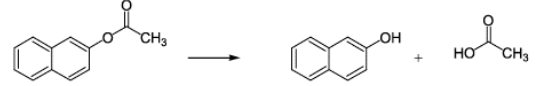
Finally, the enolase superfamily member *N*-acylamino acid racemase (NAAAR) from *Amycolatopsis* sp. T-1-60 was found to have *o*-succinylbenzoate synthase (OSBS) activity (42). NAAAR can complement OSBS-deficient strains of *E. coli*, and its activity for the OSBS reaction is three orders of magnitude higher than for racemization of *N*-acetylmethionine, the activity for which it was originally isolated (42, 89). It was recently discovered, however, that NAAAR catalyzes the racemization of *N*-succinylphenylglycine, which resembles OSB, at rates comparable to the OSBS reaction (90).

## 2. Promiscuity of Recently Evolved Enzymes

Evidence for the role of promiscuity in protein evolution also comes from examining contemporary evolutionary intermediates—enzymes which have evolved recently to degrade synthetic substrates (91). The fact that natural evolution has apparently optimized these extremely efficient enzymes on a very short time scale suggests that efficient methods of protein engineering based on evolutionary principles should be feasible. Phosphotriesterase, for example, hydrolyzes synthetic organophosphates such as paraoxon, which only became common after World War II (92). Remarkably, the efficiency of the enzyme ( $k_{\text{cat}}/K_M = 4 \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$ ) (93) approaches the diffusion-controlled limit (94). Since it would be surprising if a secondary promiscuous activity could be catalyzed at such a high rate, it appears that natural selection has operated over a very short time scale to generate this nearly catalytically-perfect enzyme (95). Although no known natural substrates have been discovered, phosphotriesterase does have a low level of lactonase and esterase activities, which may provide a clue as to its ancestral (and perhaps physiological) function (74).


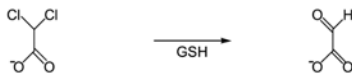
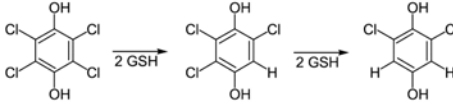
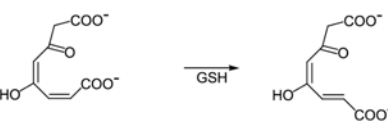


The second example provides the most compelling case for the role of promiscuity in the evolution of new functions. Pentachlorophenol, a pesticide introduced about 70 years ago, is degraded by a three step pathway in *Sphingobium*

TABLE 2  
Examples of Catalytic Promiscuity

Enzyme	Type of Promiscuity	Activities <sup>a</sup>
<i>S. coelicolor</i> HisA/TrpF	Similar substrates; Same overall reaction	<p><b>HisA</b></p> 
		<p><b>TrpF</b></p> 
<i>S. solfataricus</i> Gluconate/ Galactonate Dehydratase	Similar substrates; Same overall reaction	<p><b>Galactonate Dehydratase</b></p> 
		<p><b>Gluconate Dehydratase</b></p> 
Alkaline Phosphatase	Similar reaction mechanism, but reacting atoms differ	<p><b>Phosphomonoesterase</b></p> 
		<p>Phosphodiesterase</p> 
		<p>Sulfatase</p> 
Phosphotriesterase	Somewhat similar substrates and reaction mechanisms	<p><b>Phosphotriesterase</b></p> 
		<p>Lactonase</p> 
		<p>Esterase</p> 

<sup>a</sup>The primary activity is shown in bold.

TABLE 2 (cont.)  
 Examples of Catalytic Promiscuity

Enzyme	Type of Promiscuity	Activities <sup>a</sup>
Human Maleylacetoacetate Isomerase	Different substrates; Different overall reactions	<b>MAAI</b> 
		<b>Dehalogenase</b> 
<i>S. chlorophenolicum</i> TCHQD	Different substrates; Different overall reactions	<b>TCHQD</b> 
		<b>MAAI</b> 
<i>Amycolatopsis</i> N-Acylamino Acid Racemase	Different substrates; Different overall reactions	<b>OSBS</b> 
		<b>NAAAR</b> 

*chlorophenolicum* ATCC 39723 (91). The second enzyme in the pathway, tetrachlorohydroquinone dehalogenase (TCHQD), is distantly related to MAAI but shares a common sequence motif in the active site. Remarkably, TCHQD was found to catalyze the isomerization of maleylacetone nearly as well as a bona fide MAAI and at comparable levels to the dehalogenation reaction (96). It is thought that TCHQD evolved from an MAAI or related enzyme fairly recently. Unlike phosphotriesterase, TCHQD from *S. chlorophenolicum* has not approached catalytic perfection—it suffers from substrate inhibition, suggesting that either there has not been enough time or selection pressure to optimize the activity, or that dehalogenation of TCHQ is promiscuous, and the primary activity of the enzyme has not been discovered (96). It is tempting to speculate that there has not been adequate selective pressure on the *S. chlorophenolicum* TCHQD, since the TCHQD from *Sphingomonas* sp. UG30, which shares 94% identity with the *S. chlorophenolicum* enzyme, does not experience substrate inhibition and has an efficiency for TCHQ that is 35 times higher (97). Given the differences in TCHQD activity in these strains and the potential role for promiscuity in protein evolution both *in*

*vivo* and *in vitro*, it would be interesting to determine whether the UG30 TCHQD has lower levels of MAAI activity than the *S. chlorophenicum* TCHQD.

### 3. Promiscuity of Engineered Enzymes

It has proven difficult to engineer unique specificity into proteins. Instead, engineered proteins are often promiscuous, catalyzing both their ancestral reaction and the reaction for which they were designed or selected (7, 41, 74, 98-100). To investigate whether this observation is general and evolutionarily meaningful, several groups have studied the issue of promiscuity in evolution using directed evolution and rational design (Table 3).

TABLE 3  
Promiscuity in Engineered Enzymes

Engineered Enzyme	Engineered Activity	Relative Rates <sup>a</sup> (Engineered vs. Parental Enzyme)	
		Parental activity	Engineered activity
Ligase/HDV Ribozyme intersection sequence	Ligase	~10 <sup>-4</sup> (Ligation)	~10 <sup>3</sup> (Ligation)
	Nuclease	~10 <sup>-3</sup> (Cleavage)	~10 <sup>2</sup> (Cleavage)
β-glucuronidase 2 <sup>nd</sup> round isolate	β-galactosidase	~0.1	~10
MLE II E323G mutant	OSBS	~0.1	≥10 <sup>6</sup>
AEE D297G mutant D297G/I19F mutant	OSBS	~10 <sup>-4</sup>	≥10 <sup>3</sup>
	OSBS	~10 <sup>-6</sup>	≥10 <sup>4</sup>
Carbonic anhydrase 3 <sup>rd</sup> round isolate	Esterase	0.46	40
Phosphotriesterase 2 <sup>nd</sup> round isolate	Esterase	0.3	10.4
PON1 2.1 HT variant	Thiolactonase	0.25	72
2.1 HY variant	Esterase	0.23	31
2.2AC variant	Esterase	1.63	62
3.2PC variant	Phosphotriesterase	1.4	155

<sup>a</sup>Reported as  $k_{cat}/K_M$  ( $M^{-1} s^{-1}$ ) of the engineered enzyme divided by  $k_{cat}/K_M$  of the parental enzyme for either the parental activity or the engineered activity. The rates for the MLE II and AEE variants for the engineered activity are relative to the lower limit of detection because the activity of the parental enzymes for the engineered activity was undetectable.

A ribozyme-design experiment by Schultes and Bartel demonstrated that enzymes can evolve along neutral paths through promiscuous intermediates (73). They started with two ribozymes: the naturally occurring self-cleaving hepatitis delta virus (HDV) ribozyme, which cleaves a phosphodiester bond to produce a cyclic 2'-3'-phosphate and a 5'-hydroxyl, and the *in vitro*-selected Class III ligase ribozyme, which forms a 2'-5'-phosphodiester bond with the 5'-terminal triphosphate, releasing pyrophosphate. These two ribozymes not only catalyze different reactions, but also have entirely different secondary (and presumably tertiary structures). They cannot be related, since one of them does not occur naturally. Schultes and Bartel designed a series of intermediate sequences linking the two ribozymes that differed by no more than two mutations (Figure 5). The activity of the intermediates along the evolutionary path connecting the two ribozymes remained fairly constant (within 10-fold of the parental reaction), except for

Figure 5A

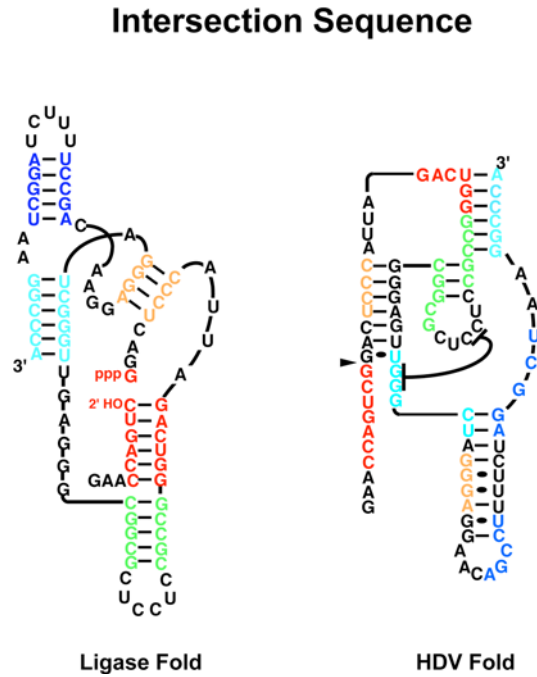


Figure 5B

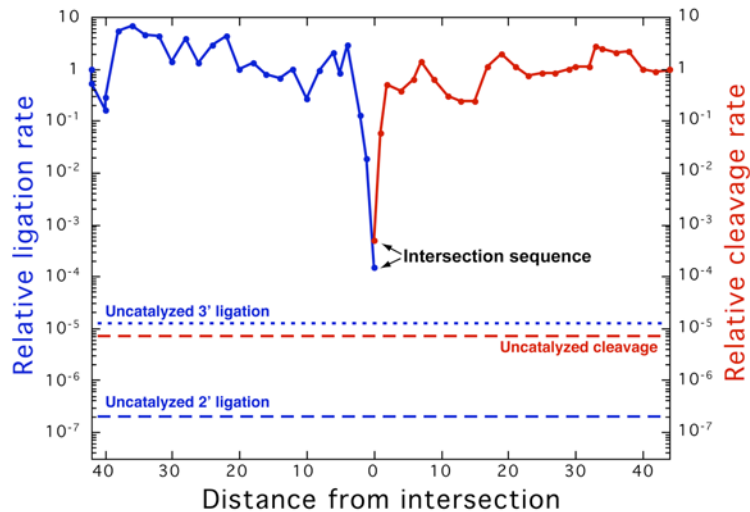


Figure 5. Design of an evolutionary pathway between the Class III ligase ribozyme and the HDV self-cleaving ribozyme (73). A) Secondary structure of the intersection sequence folded into the ligase or HDV fold. Colored segments are base-paired in the ligase fold. The arrowhead indicates the site of cleavage in the HDV ribozyme. B) Relative reaction rates of constructs along an evolutionary path from the ligase (left axis) or HDV (right axis) ribozymes to the intersection sequence. Each point represents a designed construct differing from the intersection sequence by the number of substitutions indicated on the horizontal axis, while the vertical axes indicate the reaction rate of each construct relative to the parental sequence. Self-ligation rates are in blue; Self-cleavage rates are in red. Both ligation and cleavage rates are shown for the intersection sequence. Abstracted with permission from Schultes, E. A., and Bartel, D. P. (2000), *Science* 289, 448-52. Copyright 2000 AAAS.

sequences neighboring the bifunctional intersection sequence. Although the activity of the intersection sequence was several orders of magnitude lower than either parental reaction, it is remarkable that this intermediate must adopt entirely different folds to catalyze the two reactions. Although protein evolution as discussed here is not strictly comparable—it is less extreme in that evolving a new fold is not necessary but more extreme when the chemistry catalyzed is more diverse than the two phosphoribosyl-transfer reactions—this experiment demonstrated how a new activity can evolve prior to gene duplication.

The remaining examples illustrate how proteins might evolve through promiscuous intermediates along neutral paths. Matsumura and Ellington investigated whether proteins proceed through broadly specific intermediates by directed evolution of a  $\beta$ -glucuronidase (7).  $\beta$ -glucuronidase possesses weak  $\beta$ -galactosidase activity ( $k_{\text{cat}}/K_M$  is 6 orders of magnitude lower than its  $\beta$ -glucuronidase activity). Matsumura and Ellington performed three rounds of DNA shuffling to improve this activity. They sequenced isolates with increased  $\beta$ -galactosidase activity after each round and discovered that a second-round isolate had the broadest specificity, utilizing a novel substrate not used by the wild-type or evolved isolates but retaining a preference for  $\beta$ -glucuronide over  $\beta$ -galactoside. The third round isolate, in contrast, had an 18-fold preference for  $\beta$ -galactoside. They concluded that evolving new substrate specificity through intermediates with broader specificity might be a general phenomenon.

In contrast to the previous example, Schmidt et al. began their experiments with two enzymes whose overall reactions are quite different from the target reaction and which possess no detectable activity for the target reaction (41). The goal of the experiment, as discussed in section IIC, was to determine whether other enolase superfamily members could be engineered to catalyze the OSBS reaction, like the bifunctional *Amycolatopsis* NAAAR. Directed evolution of muconate lactonizing enzyme II (MLEII) from *Pseudomonas* sp. P51 and rational design of L-Ala-D/L-Glu epimerase (AEE) from *E. coli* resulted in the isolation of single point mutants of each protein that could catalyze the OSBS reaction. Both point mutants retained their native activity as well as gaining OSBS activity. Thus, these two engineered, promiscuous enzymes plus the promiscuous *Amycolatopsis* enzyme link three different activities to the OSBS reaction, illustrating how evolution could proceed through promiscuous intermediates.

Having observed that directed evolution often produces promiscuous enzymes, Aharoni et al. studied the effect of directed evolution on the parental activity of the template (74). After assessing the promiscuous activities of carbonic anhydrase, phosphotriesterase, and paraoxonase (PON1), they performed six directed evolution experiments to select for the promiscuous activities (one activity each for carbonic anhydrase and phosphotriesterase, and four different activities for PON1). They characterized improved isolates from early points in the selection (the first and second rounds of DNA shuffling). In all six experiments, the activity of the promiscuous reaction improved at least ten-fold, while the parental activity changed less than three-fold and even increased slightly in some cases (Table 3). To see if this result was general, they reviewed the literature and found eighteen other experiments in which the promiscuous activity increased by more than 1000-fold on average, while the parental



activity decreased by an average factor of three. In both Aharoni et al.'s experiments and those they review, this broadening of substrate specificity required very few mutations—often a single mutation was sufficient. Typically, these mutated residues were located on surface loops that form the walls or perimeter of the active site and were likely to have greater conformational flexibility than catalytic residues or those that define the structural scaffold of the protein. Further, these authors suggest that there is a robustness to the parental activity that is resistant to mutation. While the parental activity is robust, a certain conformational flexibility would increase the likelihood of promiscuous activities (101). This supports a scenario in which promiscuous activities arising through mutations that have minor effects on the parental activity provide a selective advantage in the appropriate environment, leading to optimization of the promiscuous activity.

Although the experiments performed and cited by Aharoni et al. support the hypothesis that the native activity of a protein is robust to mutation, a number of other experiments do not. For instance, although the single point mutant of MLE II retained considerable MLE activity as well as high levels of OSBS activity, the AEE point mutant was a poor catalyst for both its parental activity and the OSBS reaction (41). In addition, more recent directed evolution to improve the AEE point mutant resulted in the identification of an additional mutation which increased the enzyme's OSBS activity tenfold but decreased its parental activity an additional 100-fold (102). Likewise, a single mutation in HisA endowed it with TrpF activity but abolished its native activity, while three mutations acting synergistically were required for a bifunctional HisA/TrpF variant isolated in the same directed evolution experiment (49). In the absence of extensive structural information, it is not always clear why some enzymes are more robust to mutation than others. In the HisA single point mutant, the mutation removed an aspartate which was catalytically essential for the HisA reaction but likely to interfere electrostatically with substrate binding in the TrpF reaction (49, 53). In contrast, directed evolution of carbonic anhydrase improved its esterase activity using a substrate considerably larger than its native substrate without severely affecting its native activity (74). However, the two reactions are mechanistically similar, involving nucleophilic attack followed by leaving group departure, so the fact that the selected variant retained substantial activity with its smaller, native substrate is not surprising. Thus, the robustness of an enzyme to mutation depends on how the mutations necessary for accommodating binding and catalysis of the promiscuous substrate affect the native activity (102).

## B. Cryptic Genes

The second potential source material for the evolution of new protein functions is cryptic genes. Cryptic genes, like pseudogenes, are not normally expressed, but they can be activated to express functional proteins by uncommon environmental conditions or mutations, such as reversion of frameshifts, transposition of insertion sequences, or other mechanisms (70, 103). Numerous cryptic genes have been identified in microorganisms, but in many cases it is unclear whether an unknown inducer is capable of promoting gene expression in the canonical fashion or whether a mutation or other uncommon event is actually required (70, 103-107). In either case, cryptic genes which lie dormant for generations are able to accumulate multiple mutations, which are purged if deleterious or

selected if advantageous when the cryptic gene is activated and becomes required for growth. As an example, the *cel* operon in *E. coli*, one of at least four cryptic gene systems that allow utilization of  $\beta$ -glucosides as the sole carbon source, can be activated by transposition or point mutations to allow growth on cellobiose (108, 109). More recently, it was demonstrated that this operon, while cryptic for cellobiose utilization, is actually induced by and encodes genes for the utilization of N,N'-diacetylchitobiose (Figure 6) (103, 110). In order to metabolize cellobiose, mutations in the permease and putative phospho- $\beta$ -glucosidase are required in addition to the mutation required to activate transcription of the operon.

Figure 6

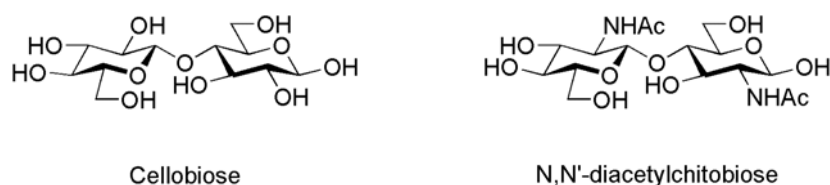


Figure 6. Structures of cellobiose and N,N'-diacetylchitobiose.

Eukaryotes appear to activate cryptic genes by a different mechanism. In *Saccharomyces cerevisiae*, the prion [PSI<sup>+</sup>] epigenetically modifies the fidelity of translation termination, resulting in read-through of stop codons. Observing the effects of the [PSI<sup>+</sup>] prion under many different conditions in seven different yeast strains revealed a variety of phenotypes, indicating that there is a great deal of genetic variation in yeast that is masked (111). In animals and plants, Hsp90 plays a similar role by acting as a buffering system that masks genetic variation (112, 113). Hsp90 stabilizes many partially folded proteins required for development until they are activated and achieve stable conformations. Under high-stress environmental conditions such as elevated temperature, Hsp90 is diverted from its normal targets to the general population of stress-damaged proteins. Dilution of Hsp90 function results in the phenotypic expression of hidden genetic variation often involving multiple genes (112). Although specific examples of protein evolution by this mechanism remain to be demonstrated, the idea that hidden genetic variation can be selected for when unmasked by epigenetic events is compelling.

These examples demonstrate how cryptic genes and pathways can be a source of new protein function in both prokaryotes and eukaryotes. Although the mechanisms are different, mutations are allowed to accumulate by neutral drift during the generations when cryptic genes are unexpressed or epigenetically masked. Under conditions in which these mutations become advantageous, the expression of these cryptic genes is positively selected, allowing the evolution of new functions that require multiple mutations without the expression of nonfunctional intermediates (110-112). For protein engineering, exploiting cryptic genes by utilizing *in vivo* selections of bacteria or yeast has the advantage that no prior, detailed knowledge of the template is required.

However, *in vivo* selections are somewhat limited by a lower level of mutagenesis and smaller library size than is achievable by some *in vitro* directed evolution methods.

### C. Pseudogenes

The idea that new enzymes arise from pseudogenes was originated by Koch, who postulated that the fastest way to evolve a new enzymatic activity, which generally requires multiple mutations, would be for mutations to accumulate in untranslatable intermediates (114). The main problem with this hypothesis is that deleterious mutations that destabilize the protein and remove critical catalytic residues are as free to accumulate as mutations that alter specificity, and they are undoubtedly more numerous. On occasion, however, “resurrected” pseudogenes might provide a shorter evolutionary route to a new function (115).

In contrast to Koch’s hypothesis that pseudogenes can give rise to new functions without the benefit of recombination, there are several examples in which portions of pseudogenes have been resurrected. For instance, an alternatively spliced variant of the nuclear protein SP100 is terminated by an exon encoding an HMG1-DNA binding domain that appears to have originated from a processed pseudogene (i.e., a pseudogene derived by reverse transcription of an mRNA) (116). In this case, the new (or additional) function arose in a process analogous to domain shuffling. Second, pseudogenes are used to generate antibody diversity by gene conversion in birds and some mammals (117). In another case of gene conversion, bovine seminal ribonuclease evolved by the repair of a pseudogene using the functional pancreatic ribonuclease (118). As in the first two examples, the function of the resurrected pseudogene is not drastically altered: it is still a ribonuclease even though its physiological function has been modified.

Although these examples do not demonstrate the evolution of new catalytic specificities of the sort observed in mechanistically diverse superfamilies, it is conceivable that gene conversion could be used to incorporate faster-evolving genetic material from pseudogenes to generate enzymes with new substrate or catalytic specificities. How common might this mechanism be in nature? Systematic analyses to discover resurrected pseudogenes have not been performed. However, pseudogenes comprise a much larger proportion of the genome in eukaryotes compared to prokaryotes, and prokaryotes are generally under selective pressure to delete them (115). Although prokaryotes are structurally and organizationally simpler than eukaryotes, prokaryotic metabolism is as complex and sometimes more varied—a higher fraction of the genes in prokaryotic genomes encodes enzymes, and the functional repertoire of enzymes in prokaryotes is much less redundant compared to that of eukaryotes (119). Thus, gene conversion of pseudogenes does not adequately explain the evolution of new enzymatic activities in prokaryotes, which have large metabolic repertoires but small pseudogene pools.

### D. Exploiting Promiscuity in Protein Design

Protein evolution can undoubtedly proceed by any of the mechanisms discussed here, and there is strong evidence that promiscuity plays a major role. How can these evolutionary principles be applied to protein design? The use of cryptic genes and

pseudogenes are mimicked implicitly in mutagenesis protocols that incorporate multiple mutations per gene. Recombination in single gene and family DNA shuffling also imitates gene conversion of pseudogenes. Promiscuity, however, has not only been looked upon as a by-product of protein engineering, but as a potential source of activities to be optimized by directed evolution (8). The success of many protein engineering experiments have hinged on the use of promiscuous templates that already exhibited low levels of the target activity (7, 74). The main challenge is to harness the potential of promiscuity and identify proteins that are promiscuous for a given reaction without examining the entire protein universe. Studying mechanistically diverse superfamilies can aid in this process by narrowing the field to candidates that can catalyze a required partial reaction. In addition, experimental characterization of superfamilies might identify other proteins like NAAAR that catalyze multiple superfamily reactions, which might prove to be ideal templates for protein engineering because of their evolutionary plasticity. Little research has been conducted to investigate these possibilities, but they have the potential to greatly improve protein design methodologies.

#### **IV. Perspective and Conclusions**

Great strides have been made in protein engineering. Proteins catalyzing a wide variety of chemical reactions have been designed by both experimental directed evolution and computational design. These methods utilize a number of evolutionary concepts, including mutation, recombination, and selection. However, engineered proteins rarely achieve the efficiency or specificity of natural enzymes. This suggests that current protein design methods are missing some critical components, preventing us from being able to recapitulate natural evolution.

Highlighted in this review are two evolutionary principles that promise to improve protein design methodology. First, protein evolution often proceeds by conserving an aspect of catalysis such as a partial chemical reaction. Selecting templates for protein engineering based on their ability to catalyze a required partial reaction could expand the diversity of enzymes that are successfully designed. Second, promiscuity has played a major role in protein evolution and has been seen as a potential source of activities to be optimized by directed evolution (8). The evolutionary plasticity of promiscuous proteins suggests that they might be particularly suitable as templates for protein engineering. Applying both of these principles requires a comprehensive database of mechanistically diverse superfamilies such as that initiated in the first release of the Structure-Function Linkage Database, which could be used as a catalog to identify partial chemical reactions, promiscuous enzymes, and the structural requirements for catalysis (28). Future research will determine if application of these principles will lead to a protein engineering methodology governed by predictable rules for designing efficient, novel catalysts.

#### **Acknowledgements**

This work was supported by National Institutes of Health Grant GM60595 to P.C.B. and Grants GM52594 and GM65155 to J.A.G. M.E.G. is supported by a Postdoctoral Fellowship in Informatics from the Pharmaceutical Researchers and

Manufacturers of America. We thank Erik Schultes and David Bartel for providing figures.

## References

1. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy, *Science* 302, 1364-8.
2. Dwyer, M. A., Looger, L. L., and Hellinga, H. W. (2004) Computational design of a biologically active enzyme, *Science* 304, 1967-71.
3. Babbitt, P. C., and Gerlt, J. A. (2000) New functions from old scaffolds: how nature reengineers enzymes for new functions, *Adv Protein Chem* 55, 1-28.
4. Gerlt, J. A., and Babbitt, P. C. (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies, *Annu Rev Biochem* 70, 209-46.
5. Minshull, J., Ness, J. E., Gustafsson, C., and Govindarajan, S. (2005) Predicting enzyme function from protein sequence, *Curr Opin Chem Biol* 9, 202-9.
6. O'Brien, P. J., and Herschlag, D. (1999) Catalytic promiscuity and the evolution of new enzymatic activities, *Chem Biol* 6, R91-R105.
7. Matsumura, I., and Ellington, A. D. (2001) In vitro evolution of beta-glucuronidase into a beta-galactosidase proceeds through non-specific intermediates, *J Mol Biol* 305, 331-9.
8. Bornscheuer, U. T., and Kazlauskas, R. J. (2004) Catalytic promiscuity in biocatalysis: using old enzymes to form new bonds and follow new pathways, *Angew Chem Int Ed Engl* 43, 6032-40.
9. Cramer, A., Raillard, S. A., Bermudez, E., and Stemmer, W. P. (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution, *Nature* 391, 288-91.
10. Kurtzman, A. L., Govindarajan, S., Vahle, K., Jones, J. T., Heinrichs, V., and Patten, P. A. (2001) Advances in directed protein evolution by recursive genetic recombination: applications to therapeutic proteins, *Curr Opin Biotechnol* 12, 361-70.
11. Ness, J. E., Welch, M., Giver, L., Bueno, M., Cherry, J. R., Borchert, T. V., Stemmer, W. P., and Minshull, J. (1999) DNA shuffling of subgenomic sequences of subtilisin, *Nat Biotechnol* 17, 893-6.
12. Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001) Evolution of function in protein superfamilies, from a structural perspective, *J Mol Biol* 307, 1113-43.
13. Horowitz, N. H. (1945) On the Evolution of Biochemical Syntheses, *Proc Natl Acad Sci U S A* 31, 153-157.
14. Horowitz, N. H. (1965) in *Evolving Genes and Proteins* (Bryson, V., and Vogel, H. J., Eds.) pp 15-23, Academic Press, New York.
15. Jensen, R. A. (1976) Enzyme recruitment in evolution of new function, *Annu Rev Microbiol* 30, 409-25.
16. Petsko, G. A., Kenyon, G. L., Gerlt, J. A., Ringe, D., and Kozarich, J. W. (1993) On the origin of enzymatic species, *Trends Biochem Sci* 18, 372-6.

17. Babbitt, P. C., Mrachko, G. T., Hasson, M. S., Huisman, G. W., Kolter, R., Ringe, D., Petsko, G. A., Kenyon, G. L., and Gerlt, J. A. (1995) A functionally diverse enzyme superfamily that abstracts the alpha protons of carboxylic acids, *Science* 267, 1159-61.
18. Babbitt, P. C., Hasson, M. S., Wedekind, J. E., Palmer, D. R., Barrett, W. C., Reed, G. H., Rayment, I., Ringe, D., Kenyon, G. L., and Gerlt, J. A. (1996) The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids, *Biochemistry* 35, 16489-501.
19. Copley, R. R., and Bork, P. (2000) Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways, *J Mol Biol* 303, 627-41.
20. Teichmann, S. A., Rison, S. C., Thornton, J. M., Riley, M., Gough, J., and Chothia, C. (2001) The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*, *J Mol Biol* 311, 693-708.
21. Teichmann, S. A., Rison, S. C., Thornton, J. M., Riley, M., Gough, J., and Chothia, C. (2001) Small-molecule metabolism: an enzyme mosaic, *Trends Biotechnol* 19, 482-6.
22. Rison, S. C., Teichmann, S. A., and Thornton, J. M. (2002) Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in *Escherichia coli*, *J Mol Biol* 318, 911-32.
23. Babbitt, P. C., and Gerlt, J. A. (1997) Understanding enzyme superfamilies. Chemistry As the fundamental determinant in the evolution of new catalytic activities, *J Biol Chem* 272, 30591-4.
24. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol* 247, 536-40.
25. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997) CATH--a hierarchic classification of protein domain structures, *Structure* 5, 1093-108.
26. Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., Akpor, A., Maibaum, M., Harrison, A., Dallman, T., Reeves, G., Diboun, I., Addou, S., Lise, S., Johnston, C., Sillero, A., Thornton, J., and Orengo, C. (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis, *Nucleic Acids Res* 33, D247-51.
27. Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data, *Nucleic Acids Res* 32, D226-9.
28. Pegg, S. C., Brown, S., Ojha, S., Huang, C. C., Ferrin, T. E., and Babbitt, P. C. (2005) Representing structure-function relationships in mechanistically diverse enzyme superfamilies, *Pac Symp Biocomput* 10, 358-69.
29. Orengo, C. A., and Thornton, J. M. (2005) Protein Families and Their Evolution--a Structural Perspective, *Annu Rev Biochem* 74, 867-900.
30. Bartlett, G. J., Borkakoti, N., and Thornton, J. M. (2003) Catalysing new reactions during evolution: economy of residues and mechanism, *J Mol Biol* 331, 829-60.
31. Todd, A. E., Orengo, C. A., and Thornton, J. M. (2002) Plasticity of enzyme active sites, *Trends Biochem Sci* 27, 419-26.

32. Gerlt, J. A., Babbitt, P. C., and Rayment, I. (2005) Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity, *Arch Biochem Biophys* 433, 59-70.
33. Armstrong, R. N. (2000) Mechanistic diversity in a metalloenzyme superfamily, *Biochemistry* 39, 13625-32.
34. Barbosa, J. A., Smith, B. J., DeGori, R., Ooi, H. C., Marcuccio, S. M., Campi, E. M., Jackson, W. R., Brossmer, R., Sommer, M., and Lawrence, M. C. (2000) Active site modulation in the N-acetylneuraminase lyase sub-family as revealed by the structure of the inhibitor-complexed *Haemophilus influenzae* enzyme, *J Mol Biol* 303, 405-21.
35. Holm, L., and Sander, C. (1997) An evolutionary treasure: unification of a broad set of amidohydrolases related to urease, *Proteins* 28, 72-82.
36. Holden, H. M., Benning, M. M., Haller, T., and Gerlt, J. A. (2001) The crotonase superfamily: divergently related enzymes that catalyze different reactions involving acyl coenzyme a thioesters, *Acc Chem Res* 34, 145-57.
37. Allen, K. N., and Dunaway-Mariano, D. (2004) Phosphoryl group transfer: evolution of a catalytic scaffold, *Trends Biochem Sci* 29, 495-503.
38. Seibert, C. M., and Raushel, F. M. (2005) Structural and catalytic diversity within the amidohydrolase superfamily, *Biochemistry* 44, 6383-91.
39. Neidhart, D. J., Kenyon, G. L., Gerlt, J. A., and Petsko, G. A. (1990) Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous, *Nature* 347, 692-4.
40. Yew, W. S. and Gerlt, J. A., unpublished work, 2005.
41. Schmidt, D. M. Z., Mundorff, E. C., Dojka, M., Bermudez, E., Ness, J. E., Govindarajan, S., Babbitt, P. C., Minshull, J., and Gerlt, J. A. (2003) Evolutionary potential of  $(\beta/\alpha)_8$ -barrels: functional promiscuity produced by single substitutions in the enolase superfamily, *Biochemistry* 42, 8387-8393.
42. Palmer, D. R., Garrett, J. B., Sharma, V., Meganathan, R., Babbitt, P. C., and Gerlt, J. A. (1999) Unexpected divergence of enzyme function and sequence: "N-acylamino acid racemase" is o-succinylbenzoate synthase, *Biochemistry* 38, 4252-8.
43. Taylor, E. A., Palmer, D. R., and Gerlt, J. A. (2001) The lesser "burden borne" by o-succinylbenzoate synthase: an "easy" reaction involving a carboxylate carbon acid, *J Am Chem Soc* 123, 5824-5.
44. Nobeli, I., Spriggs, R. V., George, R. A., and Thornton, J. M. (2005) A ligand-centric analysis of the diversity and evolution of protein-ligand relationships in *E. coli*, *J Mol Biol* 347, 415-36.
45. Fani, R., Lio, P., and Lazcano, A. (1995) Molecular evolution of the histidine biosynthetic pathway, *J Mol Evol* 41, 760-74.
46. Fani, R., Tamburini, E., Mori, E., Lazcano, A., Lio, P., Barberio, C., Casalone, E., Cavalieri, D., Perito, B., and Polsinelli, M. (1997) Paralogous histidine biosynthetic genes: evolutionary analysis of the *Saccharomyces cerevisiae* HIS6 and HIS7 genes, *Gene* 197, 9-17.
47. Henn-Sax, M., Thoma, R., Schmidt, S., Hennig, M., Kirschner, K., and Sterner, R. (2002) Two (betaalpha)(8)-barrel enzymes of histidine and tryptophan

- biosynthesis have similar reaction mechanisms and common strategies for protecting their labile substrates, *Biochemistry* 41, 12032-42.
48. Beismann-Driemeyer, S., and Sterner, R. (2001) Imidazole glycerol phosphate synthase from *Thermotoga maritima*. Quaternary structure, steady-state kinetics, and reaction mechanism of the hienzyme complex, *J Biol Chem* 276, 20387-96. Epub 2001 Mar 22.
  49. Jurgens, C., Strom, A., Wegener, D., Hettwer, S., Wilmanns, M., and Sterner, R. (2000) Directed evolution of a (beta alpha)8-barrel enzyme to catalyze related reactions in two different metabolic pathways, *Proc Natl Acad Sci U S A* 97, 9925-30.
  50. Hennig, M., Darimont, B. D., Jansonius, J. N., and Kirschner, K. (2002) The catalytic mechanism of indole-3-glycerol phosphate synthase: crystal structures of complexes of the enzyme from *Sulfolobus solfataricus* with substrate analogue, substrate, and product, *J Mol Biol* 319, 757-66.
  51. Wilmanns, M., Hyde, C. C., Davies, D. R., Kirschner, K., and Jansonius, J. N. (1991) Structural conservation in parallel beta/alpha-barrel enzymes that catalyze three sequential reactions in the pathway of tryptophan biosynthesis, *Biochemistry* 30, 9161-9.
  52. Nagano, N., Orengo, C. A., and Thornton, J. M. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions, *J Mol Biol* 321, 741-65.
  53. Leopoldseder, S., Claren, J., Jurgens, C., and Sterner, R. (2004) Interconverting the catalytic activities of (betaalpha)(8)-barrel enzymes from different metabolic pathways: sequence requirements and molecular analysis, *J Mol Biol* 337, 871-9.
  54. Barona-Gomez, F., and Hodgson, D. A. (2003) Occurrence of a putative ancient-like isomerase involved in histidine and tryptophan biosynthesis, *EMBO Rep* 4, 296-300.
  55. Wise, E., Yew, W. S., Babbitt, P. C., Gerlt, J. A., and Rayment, I. (2002) Homologous (beta/alpha)8-barrel enzymes that catalyze unrelated reactions: orotidine 5'-monophosphate decarboxylase and 3-keto-L-gulonate 6-phosphate decarboxylase, *Biochemistry* 41, 3861-9.
  56. Copley, S. D., Novak, W. R., and Babbitt, P. C. (2004) Divergence of function in the thioredoxin fold suprafamily: evidence for evolution of peroxiredoxins from a thioredoxin-like ancestor, *Biochemistry* 43, 13981-95.
  57. Radzicka, A., and Wolfenden, R. (1995) A proficient enzyme, *Science* 267, 90-3.
  58. Wise, E. L., Yew, W. S., Gerlt, J. A., and Rayment, I. (2003) Structural evidence for a 1,2-enediolate intermediate in the reaction catalyzed by 3-keto-L-gulonate 6-phosphate decarboxylase, a member of the orotidine 5'-monophosphate decarboxylase suprafamily, *Biochemistry* 42, 12133-42.
  59. Appleby, T. C., Kinsland, C., Begley, T. P., and Ealick, S. E. (2000) The crystal structure and mechanism of orotidine 5'-monophosphate decarboxylase, *Proc Natl Acad Sci U S A* 97, 2005-10.
  60. Yew, W. S., Wise, E. L., Rayment, I., and Gerlt, J. A. (2004) Evolution of enzymatic activities in the orotidine 5'-monophosphate decarboxylase suprafamily: mechanistic evidence for a proton relay system in the active site of 3-keto-L-gulonate 6-phosphate decarboxylase, *Biochemistry* 43, 6427-37.



61. Wu, N., Gillon, W., and Pai, E. F. (2002) Mapping the active site-ligand interactions of orotidine 5'-monophosphate decarboxylase by crystallography, *Biochemistry* 41, 4002-11.
62. Farber, G. K., and Petsko, G. A. (1990) The evolution of alpha/beta barrel enzymes, *Trends Biochem Sci* 15, 228-34.
63. Lang, D., Thoma, R., Henn-Sax, M., Sterner, R., and Wilmanns, M. (2000) Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion, *Science* 289, 1546-50.
64. Henn-Sax, M., Hocker, B., Wilmanns, M., and Sterner, R. (2001) Divergent evolution of (beta/alpha)<sub>8</sub>-barrel enzymes, *Biol Chem* 382, 1315-20.
65. Hocker, B., Jurgens, C., Wilmanns, M., and Sterner, R. (2001) Stability, catalytic versatility and evolution of the (beta/alpha)<sub>8</sub>-barrel fold, *Curr Opin Biotechnol* 12, 376-81.
66. Gerlt, J. A., and Babbitt, P. C. (2001) Barrels in pieces?, *Nat Struct Biol* 8, 5-7.
67. Hocker, B., Schmidt, S., and Sterner, R. (2002) A common evolutionary origin of two elementary enzyme folds, *FEBS Lett* 510, 133-5.
68. Hocker, B., Claren, J., and Sterner, R. (2004) Mimicking enzyme evolution by generating new (beta/alpha)<sub>8</sub>-barrels from (beta/alpha)<sub>4</sub>-half-barrels, *Proc Natl Acad Sci U S A* 101, 16448-53. Epub 2004 Nov 11.
69. Soberon, X., Fuentes-Gallego, P., and Saab-Rincon, G. (2004) In vivo fragment complementation of a (beta/alpha)<sub>8</sub> barrel protein: generation of variability by recombination, *FEBS Lett* 560, 167-72.
70. Hall, B. G., Yokoyama, S., and Calhoun, D. H. (1983) Role of cryptic genes in microbial evolution, *Mol Biol Evol* 1, 109-24.
71. Ohno, S. (1970) *Evolution by Gene Duplication*, Springer-Verlag, New York.
72. Hughes, A. L. (1994) The evolution of functionally novel proteins after gene duplication, *Proc Biol Sci* 256, 119-24.
73. Schultes, E. A., and Bartel, D. P. (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds, *Science* 289, 448-52.
74. Aharoni, A., Gaidukov, L., Khersonsky, O., Gould, S. M., Roodveldt, C., and Tawfik, D. S. (2005) The 'evolvability' of promiscuous protein functions, *Nat Genet* 37, 73-6. Epub 2004 Nov 28.
75. Ycas, M. (1974) On earlier states of the biochemical system, *J Theor Biol* 44, 145-60.
76. Copley, S. D. (2003) Enzymes with extra talents: moonlighting functions and catalytic promiscuity, *Curr Opin Chem Biol* 7, 265-72.
77. Jeffery, C. J. (1999) Moonlighting proteins, *Trends Biochem Sci* 24, 8-11.
78. Lambie, H. J., Heyer, N. I., Bull, S. D., Hough, D. W., and Danson, M. J. (2003) Metabolic pathway promiscuity in the archaeon *Sulfolobus solfataricus* revealed by studies on glucose dehydrogenase and 2-keto-3-deoxygluconate aldolase, *J Biol Chem* 278, 34066-72. Epub 2003 Jun 24.
79. Lambie, H. J., Milburn, C. C., Taylor, G. L., Hough, D. W., and Danson, M. J. (2004) Gluconate dehydratase from the promiscuous Entner-Doudoroff pathway in *Sulfolobus solfataricus*, *FEBS Lett* 576, 133-6.
80. Galperin, M. Y., Bairoch, A., and Koonin, E. V. (1998) A superfamily of metalloenzymes unifies phosphopentomutase and cofactor-independent

- phosphoglycerate mutase with alkaline phosphatases and sulfatases, *Protein Sci* 7, 1829-35.
81. Bond, C. S., Clements, P. R., Ashby, S. J., Collyer, C. A., Harrop, S. J., Hopwood, J. J., and Guss, J. M. (1997) Structure of a human lysosomal sulfatase, *Structure* 5, 277-89.
  82. Lukatela, G., Krauss, N., Theis, K., Selmer, T., Gieselmann, V., von Figura, K., and Saenger, W. (1998) Crystal structure of human arylsulfatase A: the aldehyde function and the metal ion at the active site suggest a novel mechanism for sulfate ester hydrolysis, *Biochemistry* 37, 3654-64.
  83. O'Brien, P. J., and Herschlag, D. (1998) Sulfatase activity of *E. coli* alkaline phosphatase demonstrates a functional link to arylsulfatases, an evolutionarily related enzyme family, *J. Am. Chem. Soc.* 120, 12369-12370.
  84. O'Brien, P. J., and Herschlag, D. (2001) Functional interrelationships in the alkaline phosphatase superfamily: phosphodiesterase activity of Escherichia coli alkaline phosphatase, *Biochemistry* 40, 5691-9.
  85. Fernandez-Canon, J. M., and Penalva, M. A. (1998) Characterization of a fungal maleylacetoacetate isomerase gene and identification of its human homologue, *J Biol Chem* 273, 329-37.
  86. Fernandez-Canon, J. M., Hejna, J., Reifsteck, C., Olson, S., and Grompe, M. (1999) Gene structure, chromosomal location, and expression pattern of maleylacetoacetate isomerase, *Genomics* 58, 263-9.
  87. Board, P. G., Baker, R. T., Chelvanayagam, G., and Jermiin, L. S. (1997) Zeta, a novel class of glutathione transferases in a range of species from plants to humans, *Biochem J* 328, 929-35.
  88. Tong, Z., Board, P. G., and Anders, M. W. (1998) Glutathione transferase zeta catalyses the oxygenation of the carcinogen dichloroacetic acid to glyoxylic acid, *Biochem J* 331, 371-4.
  89. Tokuyama, S., and Hatano, K. (1995) Purification and properties of thermostable N-acylamino acid racemase from *Amycolatopsis* sp. TS-1-60, *Appl Microbiol Biotechnol* 42, 853-9.
  90. Taylor Ringia, E. A., Garrett, J. B., Thoden, J. B., Holden, H. M., Rayment, I., and Gerlt, J. A. (2004) Evolution of enzymatic activity in the enolase superfamily: functional studies of the promiscuous o-succinylbenzoate synthase from *Amycolatopsis*, *Biochemistry* 43, 224-9.
  91. Copley, S. D. (2000) Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach, *Trends Biochem Sci* 25, 261-5.
  92. Raushel, F. M., and Holden, H. M. (2000) Phosphotriesterase: an enzyme in search of its natural substrate, *Adv Enzymol Relat Areas Mol Biol* 74, 51-93.
  93. Dumas, D. P., Caldwell, S. R., Wild, J. R., and Raushel, F. M. (1989) Purification and properties of the phosphotriesterase from *Pseudomonas diminuta*, *J Biol Chem* 264, 19659-65.
  94. Fersht, A. (1999) *Structure and Mechanism in Protein Science*, W. H. Freeman and Co., New York.
  95. Scanlan, T. S., and Reid, R. C. (1995) Evolution in action, *Chem Biol* 2, 71-5.

96. Anandarajah, K., Kiefer, P. M., Jr., Donohoe, B. S., and Copley, S. D. (2000) Recruitment of a double bond isomerase to serve as a reductive dehalogenase during biodegradation of pentachlorophenol, *Biochemistry* 39, 5303-11.
97. Habash, M. B., Beaudette, L. A., Cassidy, M. B., Leung, K. T., Hoang, T. A., Vogel, H. J., Trevors, J. T., and Lee, H. (2002) Characterization of tetrachlorohydroquinone reductive dehalogenase from *Sphingomonas* sp. UG30, *Biochem Biophys Res Commun* 299, 634-40.
98. James, L. C., and Tawfik, D. S. (2001) Catalytic and binding poly-reactivities shared by two unrelated proteins: The potential role of promiscuity in enzyme evolution, *Protein Sci* 10, 2600-7.
99. Yano, T., Oue, S., and Kagamiyama, H. (1998) Directed evolution of an aspartate aminotransferase with new substrate specificities, *Proc Natl Acad Sci U S A* 95, 5511-5.
100. Zhang, J. H., Dawes, G., and Stemmer, W. P. (1997) Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening, *Proc Natl Acad Sci U S A* 94, 4504-9.
101. James, L. C., and Tawfik, D. S. (2003) Conformational diversity and protein evolution--a 60-year-old hypothesis revisited, *Trends Biochem Sci* 28, 361-8.
102. Vick, J. E., Schmidt, D. M. Z., and Gerlt, J. A. (2005) Evolutionary potential of ( $\alpha/\beta$ )<sub>8</sub>-barrels: in vitro enhancement of a "new" reaction in the enolase superfamily, *Biochemistry* 44, 11722-11729.
103. Tamburini, E., and Mastromei, G. (2000) Do bacterial cryptic genes really exist?, *Res Microbiol* 151, 179-82.
104. Innes, D., Beacham, I. R., Beven, C. A., Douglas, M., Laird, M. W., Joly, J. C., and Burns, D. M. (2001) The cryptic *ushA* gene (*ushA(c)*) in natural isolates of *Salmonella enterica* (serotype Typhimurium) has been inactivated by a single missense mutation, *Microbiology* 147, 1887-96.
105. Shimamoto, T., Xu, X. J., Okazaki, N., Kawakami, H., and Tsuchiya, T. (2001) A cryptic melibiose transporter gene possessing a frameshift from *Citrobacter freundii*, *J Biochem (Tokyo)* 129, 607-13.
106. Plumbridge, J., and Vimr, E. (1999) Convergent pathways for utilization of the amino sugars N-acetylglucosamine, N-acetylmannosamine, and N-acetylneuraminic acid by *Escherichia coli*, *J Bacteriol* 181, 47-54.
107. Serres, M. H., and Riley, M. (2000) MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products, *Microb Comp Genomics* 5, 205-22.
108. Parker, L. L., and Hall, B. G. (1988) A fourth *Escherichia coli* gene system with the potential to evolve beta-glucoside utilization, *Genetics* 119, 485-90.
109. Parker, L. L., and Hall, B. G. (1990) Mechanisms of activation of the cryptic *cel* operon of *Escherichia coli* K12, *Genetics* 124, 473-82.
110. Keyhani, N. O., and Roseman, S. (1997) Wild-type *Escherichia coli* grows on the chitin disaccharide, N,N'-diacetylchitobiose, by expressing the *cel* operon, *Proc Natl Acad Sci U S A* 94, 14367-71.
111. True, H. L., and Lindquist, S. L. (2000) A yeast prion provides a mechanism for genetic variation and phenotypic diversity, *Nature* 407, 477-83.

112. Rutherford, S. L., and Lindquist, S. (1998) Hsp90 as a capacitor for morphological evolution, *Nature* 396, 336-42.
113. Queitsch, C., Sangster, T. A., and Lindquist, S. (2002) Hsp90 as a capacitor of phenotypic variation, *Nature* 417, 618-24. Epub 2002 May 12.
114. Koch, A. L. (1972) Enzyme evolution. I. The importance of untranslatable intermediates, *Genetics* 72, 297-316.
115. Harrison, P. M., and Gerstein, M. (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution, *J Mol Biol* 318, 1155-74.
116. Rogalla, P., Kazmierczak, B., Flohr, A. M., Hauke, S., and Bullerdiek, J. (2000) Back to the roots of a new exon--the molecular archaeology of a SP100 splice variant, *Genomics* 63, 117-22.
117. Arakawa, H., and Buerstedde, J. M. (2004) Immunoglobulin gene conversion: insights from bursal B cells and the DT40 cell line, *Dev Dyn* 229, 458-64.
118. Trabesinger-Ruef, N., Jermann, T., Zankel, T., Durrant, B., Frank, G., and Benner, S. A. (1996) Pseudogenes in ribonuclease evolution: a source of new biomacromolecular function?, *FEBS Lett* 382, 319-22.
119. Freilich, S., Spriggs, R. V., George, R. A., Al-Lazikani, B., Swindells, M., and Thornton, J. M. (2005) The complement of enzymatic sets in different species, *J Mol Biol* 349, 745-63.
120. Buchanan, C. L., Connaris, H., Danson, M. J., Reeve, C. D., and Hough, D. W. (1999) An extremely thermostable aldolase from *Sulfolobus solfataricus* with specificity for non-phosphorylated substrates, *Biochem J* 343 Pt 3, 563-70.
121. Wu, N., Mo, Y., Gao, J., and Pai, E. F. (2000) Electrostatic stress in catalysis: structure and mechanism of the enzyme orotidine monophosphate decarboxylase, *Proc Natl Acad Sci U S A* 97, 2017-22.