MINIREVIEW

⌘ *Author's Choice*

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 287, NO. 1, pp. 21–28, January 2, 2012
Published in the U.S.A.

Divergence and Convergence in Enzyme Evolution*[S]

Published, JBC Papers in Press, November 8, 2011, DOI 10.1074/jbc.R111.241976

Michael Y. Galperin and Eugene V. Koonin[1]

From the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894

Comparative analysis of the sequences of enzymes encoded in a variety of prokaryotic and eukaryotic genomes reveals convergence and divergence at several levels. Functional convergence can be inferred when structurally distinct and hence non-homologous enzymes show the ability to catalyze the same biochemical reaction. In contrast, as a result of functional diversification, many structurally similar enzyme molecules act on substantially distinct substrates and catalyze diverse biochemical reactions. Here, we present updates on the ATP-grasp, alkaline phosphatase, cupin, HD hydrolase, and N-terminal nucleophile (Ntn) hydrolase enzyme superfamilies and discuss the patterns of sequence and structural conservation and diversity within these superfamilies. Typically, enzymes within a superfamily possess common sequence motifs and key active site residues, as well as (predicted) reaction mechanisms. These observations suggest that the strained conformation (the entatic state) of the active site, which is responsible for the substrate binding and formation of the transition complex, tends to be conserved within enzyme superfamilies. The subsequent fate of the transition complex is not necessarily conserved and depends on the details of the structures of the enzyme and the substrate. This variability of reaction outcomes limits the ability of sequence analysis to predict the exact enzymatic activities of newly sequenced gene products. Nevertheless, sequence-based (super)family assignments and generic functional predictions, even if imprecise, provide valuable leads for experimental studies and remain the best approach to the functional annotation of uncharacterized proteins from new genomes.

The availability of complete genome sequences of numerous bacteria, archaea, and eukaryotes has fundamentally transformed modern biology. With complete genomes, it is a realistic goal to catalog all proteins that are responsible for every essential cellular function, *i.e.* to create a "genomic parts list." Comparative genomics revealed a surprising flexibility of the key metabolic pathways, including numerous biochemical reactions catalyzed by highly diverged or previously uncharacterized enzyme forms (1). Due to the combination of computational and experimental approaches, many such "missing" enzymes have been identified and characterized in some detail, uncovering many cases in which consecutive steps of the pathway are catalyzed by enzymes with different evolutionary histories (reviewed in Refs. 1–5). In addition, it has been shown that many key biochemical steps can be catalyzed by two or more diverse, often unrelated enzyme forms (6, 7), a phenomenon known as non-orthologous gene displacement. This patchy complex distribution of enzymes reflects a long evolutionary history of the enzymes with numerous events of gene duplication, followed by diversification, gene loss, and non-orthologous gene displacement, often via horizontal gene transfer (1, 3). Here, we consider the two key processes in enzyme evolution, namely sequence divergence, which leads to functional diversification within the same protein superfamily, and functional convergence, which results in members of distinct superfamilies being recruited to catalyze the same metabolic reaction. We also briefly discuss how sequence comparison can assist the experimental research in enzymology.

Functional Diversification of Protein Superfamilies

Historically, proteins were unified in families based on sequence similarity (8). Protein families were combined into superfamilies based on similar catalytic activities, sequence motifs, and other conserved features (9, 10). The rapid growth of protein structural data, brought about in part by the structural genomics initiatives, has put identification of protein superfamilies on a firm(er) basis. The current classifications of protein structural (super)families, implemented in the popular SCOP, CATH, and Dali databases, are generally compatible with each other despite the differences between the underlying methodologies (11–13). Furthermore, these superfamilies often correspond to sequence-based domain families (or clans) in the Pfam database (14) and contain conserved sequence motifs that are represented in such databases as InterPro (15). Therefore, proteins within the same superfamily can be confidently inferred to have evolved from a common ancestor, even though they might have dramatically different enzymatic activities or no (known) activity at all (16–18). Most studies on enzyme evolution consider evolution only within families of closely related enzymes, which typically involves changes in the enzymatic specificity without any major changes in protein structure. Here, we focus instead on the evolution of functional diversity within large protein superfamilies that are unified by common sequence motifs and structural cores. Table 1 lists catalytic activities and three-dimensional structures, where known, for members of five representative protein superfamilies that we discuss in this minireview and that, to the best of our knowledge, have not been recently reviewed from an evolutionary standpoint. These superfamilies span a wide range of sequence and structure conservation and provide multiple examples of divergence and convergence in the evolution of enzymes. We use these examples as leads for a general discussion of evolutionary trends in enzymes (see Refs. 10, 17, and 19–25 for in-depth reviews of several other enzyme superfamilies).

* This work was supported, in whole or in part, by the National Institutes of Health Intramural Research Program at the National Library of Medicine. This is the third article in the Thematic Minireview Series on Enzyme Evolution in the Post-genomic Era.

⌘ *Author's Choice*—Final version full access.

[S] This article contains supplemental Table S1.

[1] To whom correspondence should be addressed. E-mail: koonin@ncbi.nlm.nih.gov.

TABLE 1

Common features of proteins from several structural superfamilies

Root mean square deviation (r.m.s.d.) values of Cα traces were taken from the Dali and Molecular Modeling Databases (12, 82). An expanded version of this table that includes EC numbers, references, and hyperlinks to related databases is available in supplemental Table S1 as well as on the NCBI ftp site (ftp.ncbi.nih.gov/pub/galperin/EnzymeSuperfamilies.html). aa, amino acids; GPI, glycosylphosphatidylinositol; fGly, formylglycine.

Member enzymes (Protein Data Bank code, where available)	Common traits of superfamily members	Refs.
ATP-grasp superfamily Glutathione synthetase (1gsh, 2hgs), D-ala-D-Ala ligase (1iov), D-Ala-D-lactate ligase (1e4e), biotin carboxylase (1dv1), carbamoyl-phosphate synthase (1jdb), pyruvate-phosphate dikinase (1dik), phosphoribosylamine-glycine ligase PurD (1gso), phosphoribosylglycinamide formyltransferase PurT (1eyz), N^5-carboxyaminoimidazole ribonucleotide synthase PurK (1b6s), 5-formaminoimidazole-4-carboxamide ribonucleotide synthase PurP (2r7k), tubulin-tyrosine ligase, tubulin glycylase, tubulin polyglutamylase, ribosomal protein S6-glutamate ligase RimK, succinate-CoA ligase (1jkj), ATP-citrate synthase (3mwd), malate-CoA ligase, synapsin (1aux), α-aminoadipate-LysW ligase LysX (1uc9), glutathionylspermidine synthetase GspS (2io9), D-aspartate ligase Asl_{fm}, carnosine synthase, γ-F420–2:α-L-glutamate ligase CofF, tetrahydromethanopterin:α-L-glutamate ligase MptN, alanine-anticapsin ligase BacD/YwfE, L-amino acid ligase, *N*-acetylaspartylglutamate synthase, β-citrylglutamate synthase, nikkomycin biosynthesis carboxylase SanS, inositol-1,3,4-trisphosphate 5/6-kinase (1z2n), mycosporine glycine synthetase Ava_3856	Conserved structural core (≤4.3 Å Cα r.m.s.d. on ≥230 aa); common ATP-binding residues, which include two conserved Lys/Arg residues that bind α- and β-phosphates of ATP, Glx/Asp residue that interacts with adenine amino group and N6 atom, hydrophobic residues that bind adenine ring, and three Glx/Asx residues that coordinate Mg^{2+} ions; common catalytic mechanism that includes formation of phosphoacyl intermediate	27, 29, 63–65
AlkP superfamily Alkaline phosphatase (1alk), phosphoglycerate mutase (1o98, 2zkt), phosphopentomutase (3ot9), acid phosphatase (2d1g), nucleotide pyrophosphatase/phosphodiesterase (2gso), arylsulfatase (1auk), *N*-acetylgalactosamine 4-sulfatase (1fsu), steryl-sulfatase (1p49), phosphonoacetate hydrolase (1ei6), phosphoglycerol transferase MdoB, phosphonate monoester hydrolase/phosphodiesterase (2vqr), GPI phosphoethanolamine transferase PIG-N/Mcd4, LPS:phosphoethanolamine transferase EptB, polyglycerol-phosphate lipoteichoic acid synthase LtaS (2w8d), pilin phospho-form transferase PptA, inorganic pyrophosphatase	Conserved structural core (≤3.6 Å Cα r.m.s.d. on ≥220 aa); conserved metal (Zn^{2+}, Mn^{2+}, or Mg^{2+})-binding His and Asp residues; common catalytic mechanism that includes phosphorylation (sulfatation) of active site Ser/Thr/fGly residue	31–35, 38–40
Cupin superfamily Oxalate oxidase (1fi2), oxalate decarboxylase (1uw8), gentisate 1,2-dioxygenase (2d40), homogentisate 1,2-dioxygenase (1ey2), 3-hydroxyanthranilate 3,4-dioxygenase (1yfu), cysteine dioxygenase (3eln), quercetin 2,3-dioxygenase (1juh), acetylacetone dioxygenase Dke1 (3bal), 1,2-dihydroxy-3-keto-5-methylthiopentene (acireductone) dioxygenase (1vr3), 1-hydroxy-2-naphthoate dioxygenase, phosphomannose isomerase (1pmi), glucose-6-phosphate isomerase (1qy4), D-lyxose isomerase, 5-keto-4-deoxyuronate isomerase KduI (1xru), dTDP-4-dehydrorhamnose 3,5-epimerase RmlC (1dzr), dTDP-4-keto-6-deoxyglucose 5-epimerase EvaD (1oi6), dTDP-6-deoxy-3,4-ketohexulose isomerase FdtA (2pa7), ectoine synthase, ureidoglycolate hydrolase (1yqc), hydroxypropylphosphonic acid epoxidase (2bnm), dimethylsulfoniopropionate lyase DddL, phaseolin (2phl), canavalin (2cau), pirin (1j1l), auxin-binding protein (1lrh), ethanolamine utilization protein EutQ (2pyt), polyketide cyclase RemF (3ht1), bacilysin biosynthesis protein BacB (3h7j), cuproprotein CucA (2xla), vitamin K-dependent γ-carboxylase	Conserved structural core (<4.6 Å Cα r.m.s.d. on >99 aa); partly conserved metal (Mn^{2+}, Fe^{2+}, Cu^{2+}, Ni^{2+}, or Zn^{2+})-binding His residues that often form G X_5H XH $X_{3,4}$E X_6G and GDX_4PXGX_2HX_3N motifs; common catalytic mechanism that includes binding of dioxygen to metal atom and substrate with formation of peroxidic intermediate	41–44
HD domain phosphohydrolase superfamily 3′,5′-cAMP/cGMP phosphodiesterase (2hd1), (p)ppGpp hydrolase (1vj7), cyclic di-GMP phosphodiesterase (3tm8), exopolyphosphatase (1u6z), dNMP 5′-nucleotidase YfbR (2par), dNTP triphosphohydrolase (2dqb), dGTPase (3bg2), cyanamide hydratase, 7,8-dihydro-D-neopterin 2′,3′-cyclic phosphate phosphodiesterase MJ0837, 2′,3′-cAMP/cGMP hydrolase, 3′-5′ exoribonuclease YhaM, uridylyl-removing enzyme GlnD, *myo*-inositol oxygenase MioX (2huo)	Conserved structural core (<3.6 Å Cα r.m.s.d. on >105 aa); conserved metal (Mn^{2+}, Mg^{2+}, Co^{2+}, or Fe^{2+})-binding His and Asp residues organized into HX_{20-50}HDX_{60-140}D motif	45–50
Ntn hydrolase superfamily Penicillin acylase (1pnl), glutamine 5-phosphoribosyl-1-pyrophosphate amidotransferase (1ecc), proteasome subunit (1pma), glucosamine-6-phosphate synthase (1xff), protease Hs1V (1m4y), aspartylglucosaminidase (1apy), γ-glutamyltranspeptidase (2dg5), asparagine synthetase B (1ct9), β-lactam synthetase (1jgt), glutamate synthase (1ea0), L-asparaginase (2gez), threonine aspartase (2a8i), acyl-CoA:isopenicillin *N*-acyltransferase (2x1c), bile salt hydrolase (2hez), *N*-acylhomoserine lactone acylase PvdQ (2wyb), acid ceramidase, IMP cyclohydrolase PurO (2ntk)	Common structural core (<4.1 Å Cα r.m.s.d. on >96 aa) decorated with variety of structural elements; sequence conservation limited to N-terminal β-hairpin that contains catalytic Ser, Cys, or Thr residue	51–55

ATP-grasp—The original description of the ATP-grasp superfamily featured five enzymes with very similar three-domain structures (each featuring an α+β-sandwich) and several other enzymes assigned to that superfamily based solely on conserved sequence motifs (26, 27). Since then, crystal structures of some of these enzymes have been solved, confirming sequence-based predictions and expanding the ATP-grasp superfamily to include, among others, a variety of peptide synthetases (amino acid ligases) (28); tubulin glycylase and tubulin polyglutamylase, which regulate ciliary motility; and the synthetases of carnosine and *N*-acetylaspartylglutamate, dipeptides that are abundant in muscle and brain tissues, respectively (Table 1). In addition, the ATP-grasp fold and the conserved mode of ATP binding, albeit without an apparent enzymatic activity, have been identified in synapsin I, a regulator of neurotransmitter release.

The diverse members of the ATP-grasp superfamily share the conserved structural fold, typically retain a similar arrangement of the active site residues, and appear to have a common reaction mechanism that includes interaction of ATP with a carboxyl group of one substrate, followed by formation of a phosphoacyl intermediate and nucleophilic attack by an amino group of the second substrate (27). Until recently, the only deviations from this pattern were succinyl-CoA synthetase and pyruvate-phosphate dikinase, which form phosphohistidine intermediates that are attacked by a thiol or carbonyl group, respectively. However, two newly described ATP-grasp enzymes act on substrates that contain hydroxyl groups instead of carboxyl groups. Inositol-1,3,4-trisphosphate 5/6-kinase catalyzes phosphorylation of a hydroxyl group at position 5 or 6 of the inositol ring with a likely involvement of a phosphohistidine intermediate (29), whereas mycosporine glycine synthetase has a 4-deoxygadusol substrate that contains a hydroxyl group attached to an aromatic ring (30). These examples reveal substantial plasticity of the ATP-grasp fold that allows its members to evolve a variety of specific activities while preserving the key features of the superfamily.

Alkaline Phosphatase—Enzymes of the AlkP (alkaline phosphatase) superfamily share a core domain that consists of an eight-strand β-sheet surrounded by α-helices (31, 32). The recent expansion of this superfamily included both new enzymes and identification of new enzymatic activities in the known members of the superfamily. The family prototype, *Escherichia coli* AlkP, has been shown to possess phosphodiesterase, phosphonate monoesterase, and even phosphite-dependent hydrogenase activities, in addition to its well known phosphatase and sulfatase activities (33). Conversely, the enzyme originally characterized as a phosphonate monoesterase has been shown to have also phosphatase, sulfatase, sulfonate monoesterase, and phosphodiesterase activities (34). This catalytic promiscuity appears to be a characteristic feature of the AlkP superfamily enzymes (35). However, this is not a property of the entire superfamily, as a recently described member appears to be a highly specific inorganic pyrophosphatase (36). Other highly specific members of the AlkP superfamily are phosphotransferases that transfer phosphoglycerol, phosphoethanolamine, and phosphocholine moieties of the respective phospholipids to such acceptors as bacterial lipopolysaccharide or eukaryotic glycosylphosphatidylinositol (37). A particularly important example is the lipoteichoic acid synthase (phosphoglycerol transferase) LtaS, an essential enzyme in Gram-positive bacteria and a potential drug target (38, 39).

Despite the variety of their catalytic activities, AlkP superfamily members share a conserved structural fold (decorated with a variety of additional structural elements), similarly organized active sites, and the general catalytic mechanism that includes phosphorylation (or sulfation) of the active site residue, which can be Ser, Thr, or formylglycine (formed post-translationally from Cys or Ser). In phosphopentomutase, the phosphorylated Thr residue appears to be present in the ground state, leading to the suggestion that the substrate enters this enzyme at a different point in the catalytic cycle than in AlkP (40). A phosphorylated Thr residue has also been reported in the active site of LtaS (38).

Cupins—The cupin superfamily, together with the 2-ketoglutarate- and iron-dependent dioxygenase superfamily, belongs to the double-stranded β-helix fold, and members of both superfamilies have been occasionally referred to as cupins (41, 42). However, even cupins *sensu stricto* are extremely diverse, ranging from metal-binding proteins with dioxygenase, hydroxylase, and other activities to sugar isomerases (epimerases), some of which are metal-dependent and some not, to catalytically inactive seed storage and sugar-binding proteins. A recent analysis of the evolution of this fold suggested an early divergence of metal-dependent and metal-independent cupins with subsequent re-emergence of metal binding in various lineages (43). For metal-dependent cupins, the proposed reaction mechanisms typically include sequential binding of the substrate and dioxygen to the catalytic divalent metal cation (44).

HD Domain Phosphohydrolases—Members of the HD domain superfamily were originally described as (putative) metal-dependent phosphatases and phosphodiesterases (45). However, this superfamily also included a Zn^{2+}-dependent cyanamide hydratase (urea hydro-lyase), which suggested that it might possess additional catalytic activities (45). In the past several years, the HD domain has been identified in several phosphohydrolases, including the widespread HD-GYP domain phosphodiesterase that specifically hydrolyzes bacterial second messenger cyclic di-GMP (46–48). In addition, structural comparisons unexpectedly identified this domain in the iron-dependent enzyme *myo*-inositol oxygenase (49). Structures of more than a dozen HD domain-containing enzymes have been solved by structural genomics projects. However, few of these enzymes have been biochemically characterized, so the full range of catalytic activities evolved in this superfamily remains unknown. A plausible catalytic mechanism has been proposed for the 5′-nucleotidase (50) and might prove applicable to the whole superfamily.

N-terminal Nucleophile Hydrolases—The N-terminal nucleophile (Ntn)[2] hydrolase superfamily unifies diverse amidohydrolases that share a four-layered $\alpha\beta\beta\alpha$-structure and a common catalytic mechanism but do not have recognizable sequence similarity (51). Members of this superfamily are typically synthesized as catalytically inactive precursors that undergo autocatalytic processing to generate active enzymes. Their common reaction mechanism includes deprotonation of the hydroxyl or thiol group of the side chain of the N-terminal residue (Ser, Thr, or Cys) of the enzyme molecule by the free amino group of the same residue (52, 53). This stage is followed by nucleophilic attack on the carbonyl carbon of the amide bond of the substrate, formation of an acyl-enzyme intermediate coupled with the release of an amino group-containing part of the substrate, and subsequent hydrolysis of the acyl-enzyme, leading to the release of the carboxyl group-containing portion of the substrate (reviewed in Refs. 53 and 54).

The Ntn hydrolase-like fold is also present in the archaeal IMP cyclohydrolase PurO, which catalyzes the final step of purine biosynthesis (55). This enzyme retains all the structural

[2] The abbreviations used are: Ntn, N-terminal nucleophile; NISE, non-homologous isofunctional enzyme; AMP-PCP, adenosine 5′-(β,γ-methylene)diphosphonate.

features of the Ntn hydrolase superfamily but is not proteolytically processed, lacks a nucleophilic residue at the N terminus, and does not function as an amidohydrolase. Accordingly, the SCOP database assigns it to a separate superfamily (11). This enzyme is found only in a small set of methano- and haloarchaea and represents an unusual variant of extreme divergence within the common structural core.

Two other enzymes, DmpA (L-aminopeptidase D-Ala-esterase/amidase) and ornithine acetyltransferase, share with the Ntn hydrolase superfamily the $\alpha\beta\beta\alpha$-structure and catalyze a similar amidohydrolase reaction; the former enzyme also undergoes proteolytic activation. However, these proteins display a substantially different directionality and connectivity of the structural elements, indicating that their similarity to Ntn hydrolases results from convergent rather than divergent evolution (56).

Did Evolution Favor Conservation of Entatic State?

Although the abovementioned enzyme superfamilies have been defined based primarily on the structural similarity of their members, most of these members share additional properties beyond the structural fold. Such conserved features include the overall organization of the active sites, conservation of certain (although not all) active site residues, and (where known) common reaction intermediates (Table 1). To discuss the interplay of common and unique features among enzymes, it is instrumental to consider the concept of the entatic state that was originally proposed by Vallee and Williams in 1968 (57) and developed in greater detail in subsequent reviews (58, 59). The term "entatic," meaning a stretched (or otherwise stressed) state, was used to describe "a catalytically poised state intrinsic to the active site." This concept implied "the possibility that enzymes might be poised for catalytic action in the absence of substrate, *i.e.* are in an entatic state" (57). The authors acknowledged the difficulties in the interpretation of the potential indications of the entatic state, such as an exceptionally high reactivity or anomalous pK_a values of particular amino acid side chains, for most (non-metallo)enzymes (57) and concentrated on demonstrating the existence of the entatic state for catalytically active metal atoms (58, 59). As a result, entatic state is often viewed as a specific property of metalloenzymes, despite well documented instances of steric strain and perturbed pK_a values in a variety of enzyme active sites (60–62).

The concept of entatic state helps to define the characteristic features of an enzyme superfamily and explain their evolutionary conservation. Each member of the superfamily has its own range of substrates that need to be tightly bound, attacked, brought to the transition complex stage, and finally converted into the products. Although some amino acid residues, apparently those responsible for the unique specificity of the enzyme, vary from one enzyme family to another, certain residues are conserved within the superfamily as a whole. Although conservation of certain residues, *e.g.* glycines in the various cupin domains, appears to be related to the unique folding patterns of the respective proteins, the most conserved residues in the ATP-grasp superfamily are responsible for binding ATP; in the AlkP, cupin, and HD domain superfamilies for binding active site metal ions; and in the Ntn hydrolase superfamily for providing the N-terminal nucleophile and the oxyanion hole (Fig. 1 and Table 1). In these five superfamilies, sequence conservation apparently extends to the residues that directly participate in the initial attack on the substrate and stabilization of the transition complex. The proper positioning of these residues is provided by a variety of conserved structural elements. In ATP-grasp enzymes, for example, these include a helix-turn-helix structure connecting the first two domains (63, 64); a conserved flexible loop with a sharp turn (designated the T-loop by Thoden *et al.* (65)), which follows the ATP/α-phosphate-binding Lys/Arg residue (Lys-136 in Fig. 1*A*); a *cis*-peptide bond in the backbone just upstream of that Lys/Arg residue; and other rare structural features.[3] As a result, members of these superfamilies typically share the initial stages of the catalytic process. On the other hand, the breakdown of the transition complex in different enzymes (or, as discussed by Jencks (66), in the same enzyme under different conditions) can follow a number of different paths, yielding, for example within the AlkP superfamily, substrate hydrolysis, isomerization, or phosphate group transfer (see also Refs. 10 and 19–21).

One could argue that the emergence of each distinct entatic state conformation was a major evolutionary event, opening the door to the utilization of new classes of substrates or to the catalysis of new classes of reactions. During the subsequent evolution, major changes in protein structure were restricted by the likelihood of the formation of toxic (or inactive) misfolded molecules (67). Thus, only those sequence changes would prove viable that preserved the structural fold and accordingly the mechanism of formation of the same entatic state. These constraints led to the formation of series of structurally and catalytically (albeit not necessarily functionally) related protein molecules, which later evolved into the current superfamilies. A somewhat similar conclusion was reached by Warshel and Florián (68), who singled out pre-oriented dipoles as the source of the catalytic power of enzymes and argued that evolutionary optimization of enzymes increased their "preorganization effect", *i.e.* the ability of enzymes to "minimize the reorganization energy associated with the formation of the charged transition state."

Practical Aspects of Superfamily Assignment

The conservation of catalytic elements within enzyme superfamilies makes sequence analysis an extremely useful tool in enzymology: assignment of a poorly characterized enzyme to a specific superfamily immediately predicts the structural fold, active site residues, a range of its potential catalytic activities, and even the likely catalytic mechanism. This could be particularly valuable for enzymes with complex substrates, for which direct assays are complicated and cumbersome. Thus, measuring the activity of tubulin-modifying enzymes, which are involved in tumor progression and have a vital role in neuronal organization, is certainly not an easy task. The assignment of tubulin-tyrosine ligase to the ATP-grasp superfamily (27) led to prediction of its active site residues and suggested a plausible catalytic mechanism for this enzyme (69). Likewise, assignment of the glycosylphosphatidylinositol phosphoethanolamine transferase PIG-N (Mcd4) to the AlkP superfamily was instrumental for the studies of this and related enzymes (37).

[3] M. Y. Galperin, unpublished data.

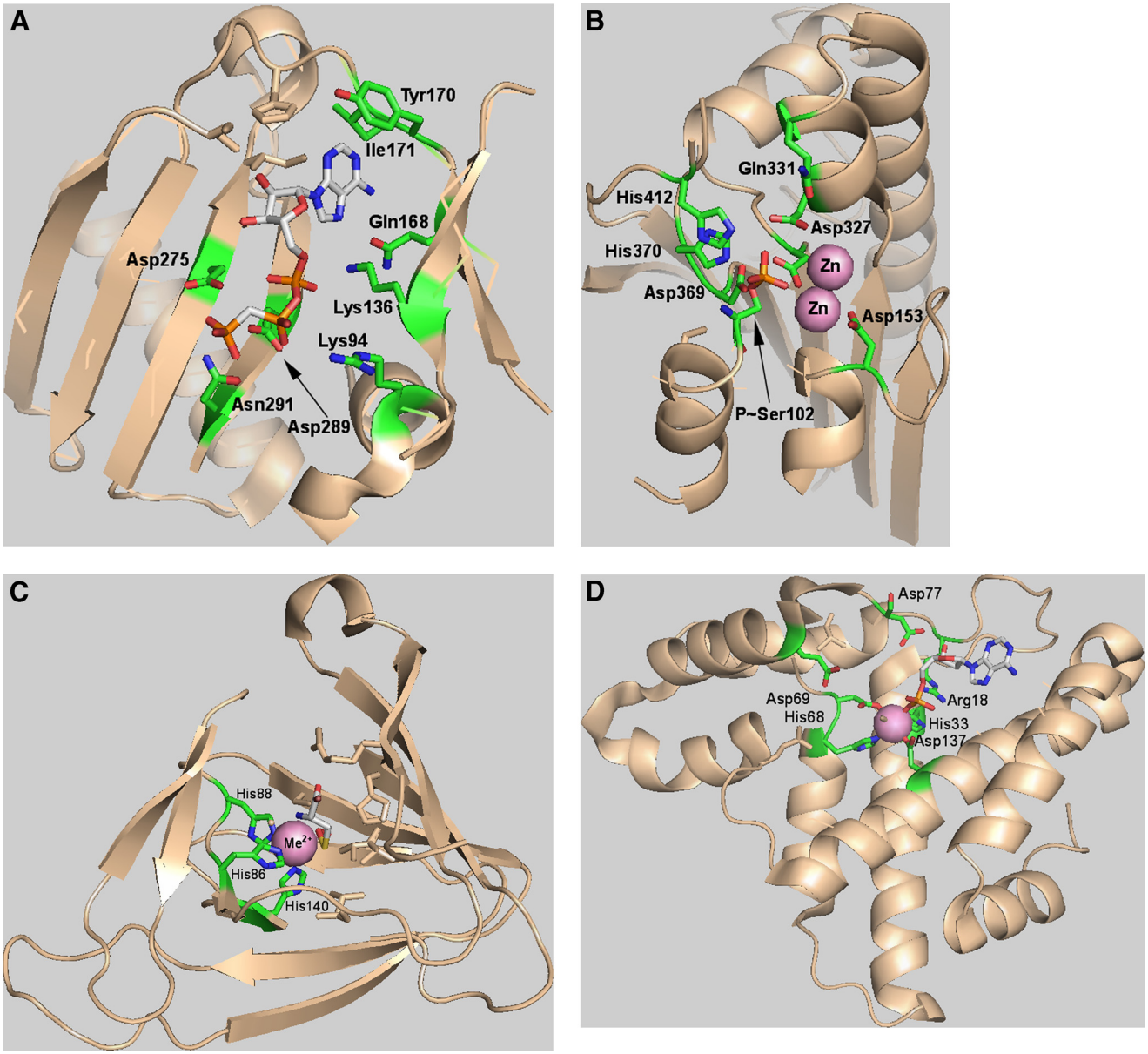

FIGURE 1. **Conservation of structural core and active site residues in ATP-grasp (*A*), AlkP (*B*), cupin (*C*), and HD phosphohydrolase (*D*) superfamily enzymes.** Conserved structural elements, identified through VAST alignments (83), are shown in *tan*, active site residues are shown as *sticks*, the most conserved residues are shown in *bright colors* (with carbon atoms shown in *green*), and catalytic metal atoms are shown as *pink spheres*. *A*, inositol-1,3,4-trisphosphate 5/6-kinase (Protein Data Bank code 1z2p (29)) with bound ATP analog AMP-PCP. Carbon atoms are in *silver*. *B*, AlkP(H331Q) mutant with a phosphoserine intermediate (code 1hjk (84)). *C*, cysteine dioxygenase with a persulfenate intermediate (code 3eln (85)). *D*, 5′-deoxyribonucleotidase with bound dAMP (code 2pau (50)).

Superfamily assignments proved most valuable when used for the analysis of metabolic pathways where the nature of the "missing" enzyme could be used to look for suitable candidates among uncharacterized genes (3, 4). Reconstruction of purine biosynthesis in archaea, which included characterization of the PurP and PurO gene products (55, 70), provides an impressive example of the power of the integrative approach that combines sequence analysis with biochemical assays and structural studies (reviewed in Ref. 71).

Superfamily assignments could also be useful for the functional annotation of newly sequenced genes that do not show clear sequence similarity to any well characterized enzymes. In such cases, searching new gene products against superfamily-specific sequence profiles (available, for example, in the NCBI Conserved Domain Database (72)) provides hints that can be used for generic functional prediction and as guidance for subsequent experiments, *e.g.* by predicting catalytic residues that are targets of choice for site-specific mutagenesis. For example, identification of the cupin domain in the sequence of the vitamin K-dependent γ-glutamyl carboxylase (residues 524–625 of VKGC_HUMAN (15)) could open new avenues for studying this important but still enigmatic enzyme.

Convergent Evolution: Similar Active Sites in Analogous Enzymes

Diversification of enzyme families can result in functional overlap when members of two or more distinct families end up catalyzing the same biochemical reaction. In some cases, such

enzyme isoforms are distantly related, and the low sequence similarity conceivably stems from rapid divergence of homologous protein sequences that accompanies adaptation to different environmental conditions. Examples of such enzyme pairs include the thermostable and mesophilic forms of β-glucosidase and adenylate kinase, which have retained very similar structures but share only a limited number of conserved residues (7).

There are cases, however, in which distinct enzyme forms catalyzing the same reaction share no detectable sequence similarity or even belong to two or more distinct structural superfamilies or folds (6, 7). The best known examples include superoxide dismutase, for which four distinct structural forms have been described, and cellulase, which is found in at least five structurally distinct forms. For such analogous (as opposed to homologous) enzymes, adoption of different structural folds indicates independent evolutionary origins; we have recently proposed a more precise designation for these enzymes, non-homologous isofunctional enzymes (NISEs) (7).

As in the textbook example of trypsin and subtilisin, diverse enzymes that act on related substrates might still share similarities in the organization of their active sites. Such similarities have been noted, for example, in the similar configurations of the ATP-binding residues in the ATP-grasp enzyme D-Ala-D-Ala ligase and enzymes that adopt two other folds, cAMP-dependent protein kinase and ribonucleotide reductase (73). A subsequent comparison of the adenine-binding sites revealed a common structural framework with similar polar and hydrophobic interactions in representatives of eight different folds (74). A similar pattern of structural convergence of evolutionarily unrelated enzymes has been revealed in the organization of pyridoxal phosphate-interacting residues of pyridoxal phosphate-dependent enzymes representing five distinct folds (75). Similar examples of functional convergence can be seen in NISEs, which, by definition, act on the same substrates. A recent comparison of the enzyme-substrate complexes of the phosphorylated chemotaxis protein CheY with two structurally distinct phosphatases, CheZ and CheX, revealed a very similar organization of the catalytic residues involved in the dephosphorylation of phospho-CheY (76).

An interesting evolutionary feature is the often skewed phylogenetic distribution of distinct isoforms of the same enzyme. For example, the archaeal shikimate kinase (77) has not yet been detected outside of the archaeal domain, whereas the other form of shikimate kinase is found in bacteria and eukaryotes. Similarly, the recently described cupin form of glucose-6-phosphate isomerase is found only in certain bacteria and archaea, whereas the other form of this enzyme, a member of the sugar isomerase family, is widespread. For the cases in which a particular enzyme is confined to a certain taxonomic group, recent evolutionary emergence from an enzyme of a different specificity seems to be the easiest explanation (7). Archaeal shikimate kinase, for example, is a member of the GHMP kinase superfamily and could have evolved from homoserine kinases or similar enzymes (77).

Recruitment for Non-enzymatic Functions: Moonlighting Enzymes

As discussed above, certain members of the ATP-grasp and cupin superfamilies lack (known) enzymatic activities and act solely as ATP-binding (synapsin), auxin-binding, or seed storage proteins. Many enzymes acquire such additional (typically non-enzymatic) functions even without the loss of their catalytic activity. This phenomenon, referred to as "moonlighting" (78, 79), is usually observed when the respective genes are expressed in atypical environments (tissue, cell organelle, or secreted). First observed in eye lens crystallins, where lactate dehydrogenase, enolase, argininosuccinate lyase, aldehyde dehydrogenase, and a variety of other proteins play predominantly or exclusively structural roles (80), moonlighting has since been demonstrated for a variety of glycolytic, TCA cycle, and other metabolic enzymes. It is now clear that moonlighting represents an important source of protein diversity in multicellular eukaryotes and is relevant for certain human diseases (79, 81).

Conclusions

The availability of complete genome sequences of diverse bacteria, archaea, and eukaryotes illuminated the unexpected diversity of protein sequences encoded in those genomes. Many genomes turned out to lack genes for well known enzymes involved in key steps of certain metabolic pathways. Identification of the alternative enzymes that catalyzed those steps was made possible only through a detailed computational analysis of the respective genomes, followed by experimental study of the plausible candidates (3, 4, 71). Alternative enzyme forms often appear to be recruited from distinct superfamilies, so NISE is a common evolutionary phenomenon (6, 7). Notably, some of the NISEs share not only the reactions they catalyze but also the configurations of the catalytic residues, although in these cases, the catalytic centers are embedded in distinct unrelated folds. Complementary to the cases of functional and even structural convergence and despite extensive diversification, superfamilies of enzymes show remarkable evolutionary conservation. It appears that the common denominator behind this conservation is the persistence of amino acid residues that are required to maintain the strained (entatic) state involved in the formation of transition complexes. To summarize, evolutionary approaches are critically important for the analysis of metabolic pathways, especially in poorly studied organisms. These approaches also provide valuable clues to the catalytic properties of even relatively well characterized enzymes.

Acknowledgments—We thank Drs. Armen Mulkidjanian, Alexey Murzin, Andrei Osterman, and R. J. P. Williams for helpful discussions.

REFERENCES

1. Koonin, E. V., and Galperin, M. Y. (2002) *Sequence-Evolution-Function: Computational Approaches in Comparative Genomics*, Kluwer Academic Publishers, Boston
2. Galperin, M. Y., and Koonin, E. V. (1999) *Genetica* **106,** 159–170
3. Osterman, A., and Overbeek, R. (2003) *Curr. Opin. Chem. Biol.* **7,** 238–251

4. Osterman, A. L. (2009) *Nat. Chem. Biol.* **5,** 871–872
5. Hanson, A. D., Pribat, A., Waller, J. C., and de Crécy-Lagard, V. (2010) *Biochem. J.* **425,** 1–11
6. Galperin, M. Y., Walker, D. R., and Koonin, E. V. (1998) *Genome Res.* **8,** 779–790
7. Omelchenko, M. V., Galperin, M. Y., Wolf, Y. I., and Koonin, E. V. (2010) *Biol. Direct* **5,** 31
8. Dayhoff, M. O. (1976) *Fed. Proc.* **35,** 2132–2138
9. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) *J. Mol. Biol.* **247,** 536–540
10. Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001) *J. Mol. Biol.* **307,** 1113–1143
11. Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2008) *Nucleic Acids Res.* **36,** D419–D425
12. Holm, L., and Rosenström, P. (2010) *Nucleic Acids Res.* **38,** W545–W549
13. Cuff, A. L., Sillitoe, I., Lewis, T., Clegg, A. B., Rentzsch, R., Furnham, N., Pellegrini-Calace, M., Jones, D., Thornton, J., and Orengo, C. A. (2011) *Nucleic Acids Res.* **39,** D420–D426
14. Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L., Eddy, S. R., and Bateman, A. (2010) *Nucleic Acids Res.* **38,** D211–D222
15. Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. (2009) *Nucleic Acids Res.* **37,** D211–D215
16. Meng, E. C., and Babbitt, P. C. (2011) *Curr. Opin. Struct. Biol.* **21,** 391–397
17. Gerlt, J. A., Babbitt, P. C., Jacobson, M. P., and Almo, S. C. (2012) *J. Biol. Chem.* **287,** 29–34
18. Copley, S. D. (2012) *J. Biol. Chem.* **287,** 3–10
19. Todd, A. E., Orengo, C. A., and Thornton, J. M. (2002) *Trends Biochem. Sci.* **27,** 419–426
20. Gerlt, J. A., Babbitt, P. C., and Rayment, I. (2005) *Arch. Biochem. Biophys.* **433,** 59–70
21. Glasner, M. E., Gerlt, J. A., and Babbitt, P. C. (2006) *Curr. Opin. Chem. Biol.* **10,** 492–497
22. Scheeff, E. D., and Bourne, P. E. (2005) *PLoS Comput. Biol.* **1,** e49
23. Allen, K. N., and Dunaway-Mariano, D. (2004) *Trends Biochem. Sci.* **29,** 495–503
24. Burroughs, A. M., Allen, K. N., Dunaway-Mariano, D., and Aravind, L. (2006) *J. Mol. Biol.* **361,** 1003–1034
25. Elias, M., and Tawfik, D. S. (2012) *J. Biol. Chem.* **287,** 11–20
26. Murzin, A. G. (1996) *Curr. Opin. Struct. Biol.* **6,** 386–394
27. Galperin, M. Y., and Koonin, E. V. (1997) *Protein Sci.* **6,** 2639–2643
28. Iyer, L. M., Abhiman, S., Maxwell Burroughs, A., and Aravind, L. (2009) *Mol. BioSyst.* **5,** 1636–1660
29. Miller, G. J., Wilson, M. P., Majerus, P. W., and Hurley, J. H. (2005) *Mol. Cell* **18,** 201–212
30. Balskus, E. P., and Walsh, C. T. (2010) *Science* **329,** 1653–1656
31. Galperin, M. Y., Bairoch, A., and Koonin, E. V. (1998) *Protein Sci.* **7,** 1829–1835
32. Galperin, M. Y., and Jedrzejas, M. J. (2001) *Proteins* **45,** 318–324
33. Yang, K., and Metcalf, W. W. (2004) *Proc. Natl. Acad. Sci. U.S.A.* **101,** 7919–7924
34. van Loo, B., Jonas, S., Babtie, A. C., Benjdia, A., Berteau, O., Hyvönen, M., and Hollfelder, F. (2010) *Proc. Natl. Acad. Sci. U.S.A.* **107,** 2740–2745
35. Jonas, S., and Hollfelder, F. (2009) *Pure Appl. Chem.* **81,** 731–742
36. Lee, H. S., Kim, Y. J., Lee, J. H., and Kang, S. G. (2009) *J. Bacteriol.* **191,** 3415–3419
37. Orlean, P., and Menon, A. K. (2007) *J. Lipid Res.* **48,** 993–1011
38. Schirner, K., Marles-Wright, J., Lewis, R. J., and Errington, J. (2009) *EMBO J.* **28,** 830–842
39. Lu, D., Wörmann, M. E., Zhang, X., Schneewind, O., Gründling, A., and Freemont, P. S. (2009) *Proc. Natl. Acad. Sci. U.S.A.* **106,** 1584–1589
40. Panosian, T. D., Nannemann, D. P., Watkins, G. R., Phelan, V. V., McDonald, W. H., Wadzinski, B. E., Bachmann, B. O., and Iverson, T. M. (2011) *J. Biol. Chem.* **286,** 8043–8054
41. Dunwell, J. M., Culham, A., Carter, C. E., Sosa-Aguirre, C. R., and Goodenough, P. W. (2001) *Trends Biochem. Sci.* **26,** 740–746
42. Dunwell, J. M., Purvis, A., and Khuri, S. (2004) *Phytochemistry* **65,** 7–17
43. Iyer, L. M., Abhiman, S., de Souza, R. F., and Aravind, L. (2010) *Nucleic Acids Res.* **38,** 5261–5279
44. McCoy, J. G., Bailey, L. J., Bitto, E., Bingman, C. A., Aceti, D. J., Fox, B. G., and Phillips, G. N., Jr. (2006) *Proc. Natl. Acad. Sci. U.S.A.* **103,** 3084–3089
45. Aravind, L., and Koonin, E. V. (1998) *Trends Biochem. Sci.* **23,** 469–472
46. Galperin, M. Y., Natale, D. A., Aravind, L., and Koonin, E. V. (1999) *J. Mol. Microbiol. Biotechnol.* **1,** 303–305
47. Ryan, R. P., Fouhy, Y., Lucey, J. F., Crossman, L. C., Spiro, S., He, Y. W., Zhang, L. H., Heeb, S., Cámara, M., Williams, P., and Dow, J. M. (2006) *Proc. Natl. Acad. Sci. U.S.A.* **103,** 6712–6717
48. Lovering, A. L., Capeness, M. J., Lambert, C., Hobley, L., and Sockett, R. E. (2011) *MBio* **2,** e00163-e00111
49. Thorsell, A. G., Persson, C., Voevodskaya, N., Busam, R. D., Hammarström, M., Gräslund, S., Gräslund, A., and Hallberg, B. M. (2008) *J. Biol. Chem.* **283,** 15209–15216
50. Zimmerman, M. D., Proudfoot, M., Yakunin, A., and Minor, W. (2008) *J. Mol. Biol.* **378,** 215–226
51. Brannigan, J. A., Dodson, G., Duggleby, H. J., Moody, P. C., Smith, J. L., Tomchick, D. R., and Murzin, A. G. (1995) *Nature* **378,** 416–419
52. Duggleby, H. J., Tolley, S. P., Hill, C. P., Dodson, E. J., Dodson, G., and Moody, P. C. (1995) *Nature* **373,** 264–268
53. Oinonen, C., and Rouvinen, J. (2000) *Protein Sci.* **9,** 2329–2337
54. Pei, J., and Grishin, N. V. (2003) *Protein Sci.* **12,** 1131–1135
55. Kang, Y. N., Tran, A., White, R. H., and Ealick, S. E. (2007) *Biochemistry* **46,** 5050–5062
56. Cheng, H., and Grishin, N. V. (2005) *Protein Sci.* **14,** 1902–1910
57. Vallee, B. L., and Williams, R. J. P. (1968) *Proc. Natl. Acad. Sci. U.S.A.* **59,** 498–505
58. Williams, R. J. P. (1985) *J. Mol. Catal.* **30,** 1–26
59. Williams, R. J. P. (1995) *Eur. J. Biochem.* **234,** 363–381
60. Herzberg, O., and Moult, J. (1991) *Proteins* **11,** 223–229
61. Mulkidjanian, A. Y. (1999) *FEBS Lett.* **463,** 199–204
62. Harris, T. K., and Turner, G. J. (2002) *IUBMB Life* **53,** 85–98
63. Thoden, J. B., Kappock, T. J., Stubbe, J., and Holden, H. M. (1999) *Biochemistry* **38,** 15480–15492
64. Thoden, J. B., Firestine, S., Nixon, A., Benkovic, S. J., and Holden, H. M. (2000) *Biochemistry* **39,** 8791–8802
65. Thoden, J. B., Firestine, S. M., Benkovic, S. J., and Holden, H. M. (2002) *J. Biol. Chem.* **277,** 23898–23908
66. Jencks, W. P. (1987) *Catalysis in Chemistry and Enzymology*, Courier Dover Publications, Mineola, NY
67. Drummond, D. A., and Wilke, C. O. (2008) *Cell* **134,** 341–352
68. Warshel, A., and Florián, J. (1998) *Proc. Natl. Acad. Sci. U.S.A.* **95,** 5950–5955
69. Janke, C., Rogowski, K., Wloga, D., Regnard, C., Kajava, A. V., Strub, J. M., Temurak, N., van Dijk, J., Boucher, D., van Dorsselaer, A., Suryavanshi, S., Gaertig, J., and Eddé, B. (2005) *Science* **308,** 1758–1762
70. Zhang, Y., White, R. H., and Ealick, S. E. (2008) *Biochemistry* **47,** 205–217
71. Zhang, Y., Morar, M., and Ealick, S. E. (2008) *Cell. Mol. Life Sci.* **65,** 3699–3724
72. Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Lu, F., Marchler, G. H., Mullokandov, M., Omelchenko, M. V., Robertson, C. L., Song, J. S., Thanki, N., Yamashita, R. A., Zhang, D., Zhang, N., Zheng, C., and Bryant, S. H. (2011) *Nucleic Acids Res.* **39,** D225–D229
73. Denessiouk, K. A., Lehtonen, J. V., and Johnson, M. S. (1998) *Protein Sci.* **7,** 1768–1771
74. Denessiouk, K. A., and Johnson, M. S. (2000) *Proteins* **38,** 310–326
75. Denessiouk, K. A., Denesyuk, A. I., Lehtonen, J. V., Korpela, T., and Johnson, M. S. (1999) *Proteins* **35,** 250–261

76. Pazy, Y., Motaleb, M. A., Guarnieri, M. T., Charon, N. W., Zhao, R., and Silversmith, R. E. (2010) *Proc. Natl. Acad. Sci. U.S.A.* **107,** 1924–1929
77. Daugherty, M., Vonstein, V., Overbeek, R., and Osterman, A. (2001) *J. Bacteriol.* **183,** 292–300
78. Jeffery, C. J. (2009) *Mol. BioSyst.* **5,** 345–350
79. Huberts, D. H., and van der Klei, I. J. (2010) *Biochim. Biophys. Acta* **1803,** 520–525
80. Piatigorsky, J. (2003) *J. Struct. Funct. Genomics* **3,** 131–137
81. Sriram, G., Martinez, J. A., McCabe, E. R., Liao, J. C., and Dipple, K. M. (2005) *Am. J. Hum. Genet.* **76,** 911–924
82. Wang, Y., Addess, K. J., Chen, J., Geer, L. Y., He, J., He, S., Lu, S., Madej, T., Marchler-Bauer, A., Thiessen, P. A., Zhang, N., and Bryant, S. H. (2007) *Nucleic Acids Res.* **35,** D298–D300
83. Gibrat, J. F., Madej, T., and Bryant, S. H. (1996) *Curr. Opin. Struct. Biol.* **6,** 377–385
84. Murphy, J. E., Stec, B., Ma, L., and Kantrowitz, E. R. (1997) *Nat. Struct. Biol.* **4,** 618–622
85. Simmons, C. R., Krishnamoorthy, K., Granett, S. L., Schuller, D. J., Dominy, J. E., Jr., Begley, T. P., Stipanuk, M. H., and Karplus, P. A. (2008) *Biochemistry* **47,** 11390–11392