MINIREVIEW

⌘ *Author's Choice*

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 289, NO. 44, pp. 30221–30228, October 31, 2014

New Insights about Enzyme Evolution from Large Scale Studies of Sequence and Structure Relationships*

Published, JBC Papers in Press, September 10, 2014, DOI 10.1074/jbc.R114.569350

Shoshana D. Brown[‡] **and Patricia C. Babbitt**[‡§¶1]

From the Departments of [‡]*Bioengineering and Therapeutic Sciences and* [§]*Pharmaceutical Chemistry, School of Pharmacy, and the* [¶]*California Institute for Quantitative Biosciences, University of California, San Francisco, California 94158-2330*

Understanding how enzymes have evolved offers clues about their structure-function relationships and mechanisms. Here, we describe evolution of functionally diverse enzyme superfamilies, each representing a large set of sequences that evolved from a common ancestor and that retain conserved features of their structures and active sites. Using several examples, we describe the different structural strategies nature has used to evolve new reaction and substrate specificities in each unique superfamily. The results provide insight about enzyme evolution that is not easily obtained from studies of one or only a few enzymes.

Although we have long assumed that there are many more protein functions in living organisms than fold types (1), *how* a modest number of structural scaffolds (2) have been remodeled by nature to produce the proteins required by living organisms is not well understood. This minireview focuses on functionally diverse enzyme superfamilies, groups of proteins that offer special insight about how nature has solved this challenge.

Functionally (or mechanistically) diverse superfamilies are evolutionarily related sets of enzymes that may be quite diverse in sequence, structure, and overall reaction, but share a conserved constellation of active site residues used for a common partial reaction or chemical capability (3–5). Knowing the fundamental chemical capability and associated substrate substructure(s) that typify each such superfamily constrains the search space for predicting the molecular function of superfamily members of unknown function (unknowns). Comparison among all of the sequences and/or structures in a superfamily can then be used to deduce how evolution has varied these features to produce new enzyme functions from the ancestral structural scaffold. These analyses are valuable for gaining functional clues for the enormous number of sequenced genes that do not have experimental information.

A better understanding of natural enzyme evolution in these types of superfamilies has many other applications as well. For example, understanding how nature has engineered new reactions using the conserved structural features typifying each superfamily could be used to help guide enzyme design in the laboratory (6). Further, assignment of sequences associated with unusual chemical reactions to a superfamily with mechanistically well characterized members may provide clues useful for determining the mechanism of such "outlier" reactions.

Functionally diverse superfamilies represent a significant proportion of the enzyme universe, making up more than one-third of all structurally characterized enzyme superfamilies (7). Because these superfamilies may represent many thousands of sequences and sometimes dozens of different reactions, an inventory of their properties typically requires computational analysis. Many different types of large scale computational studies, focusing on one or multiple superfamilies, have been carried out. See Refs. 8–10 for a few examples. Recently, some of these studies have used network-based approaches (2, 11–13).

Reflecting this relatively new approach, sequence similarity networks are used in some figures in this review (see Figs. 1 and 4) to enable exploration of structure-function relationships in enzyme superfamilies from a large scale perspective. In these networks, nodes represent one or more proteins, and edges between them represent a measure of sequence or structural similarity. Although not a substitute for phylogenetic trees, similarity networks provide several advantages over trees and multiple alignments for developing new hypotheses about the evolution of functional features in superfamilies. They are quick to construct, do not require an accurate multiple sequence alignment, and can summarize in one network relationships among thousands of sequences. The networks can also be visualized and interactively manipulated and explored using such software packages as Cytoscape (14). Although they are not based on an explicit evolutionary model, initial validation studies show that similarity networks correlate well with results from phylogenetic trees (15).

We illustrate here some major themes emerging from large scale studies of functionally diverse enzyme superfamilies that impact our understanding of the evolution of enzyme function. First, studies of a number of these enzyme superfamilies suggest that experimental knowledge of their functions is sparse and that we know very little about the functions of a large proportion of enzymes in each. This lack of knowledge limits our understanding of the evolution of new reactions in significant ways. Second, the patterns of structural variation associated with the evolution of diverse functions in these superfamilies are many and varied and include, for example, structural reorganization of domains, addition of inserts, and even major modifications in active site architecture. Many of these patterns are difficult to deduce from small scale comparisons. Third, deducing how differences in reaction and substrate specificity have evolved within a functionally diverse superfamily can be complicated by issues that are challenging to address. Functional promiscuity (2) and evolutionary invention of the same reaction more than once from intermediate ancestors in a superfamily phylogeny (16–18) provide relevant examples.

* This work was supported, in whole or in part, by National Institutes of Health Grant R01 GM60595 (to P. C. B.). This is the fourth article in the Thematic Minireview series "Enzyme Evolution."

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

[1] To whom correspondence should be addressed: 1700 4th St., UCSF MC2550, University of California, San Francisco CA 94158-2330. Tel.: 415-476-3784; Fax: 415-476-6022; E-mail: babbitt@cgl.ucsf.edu.

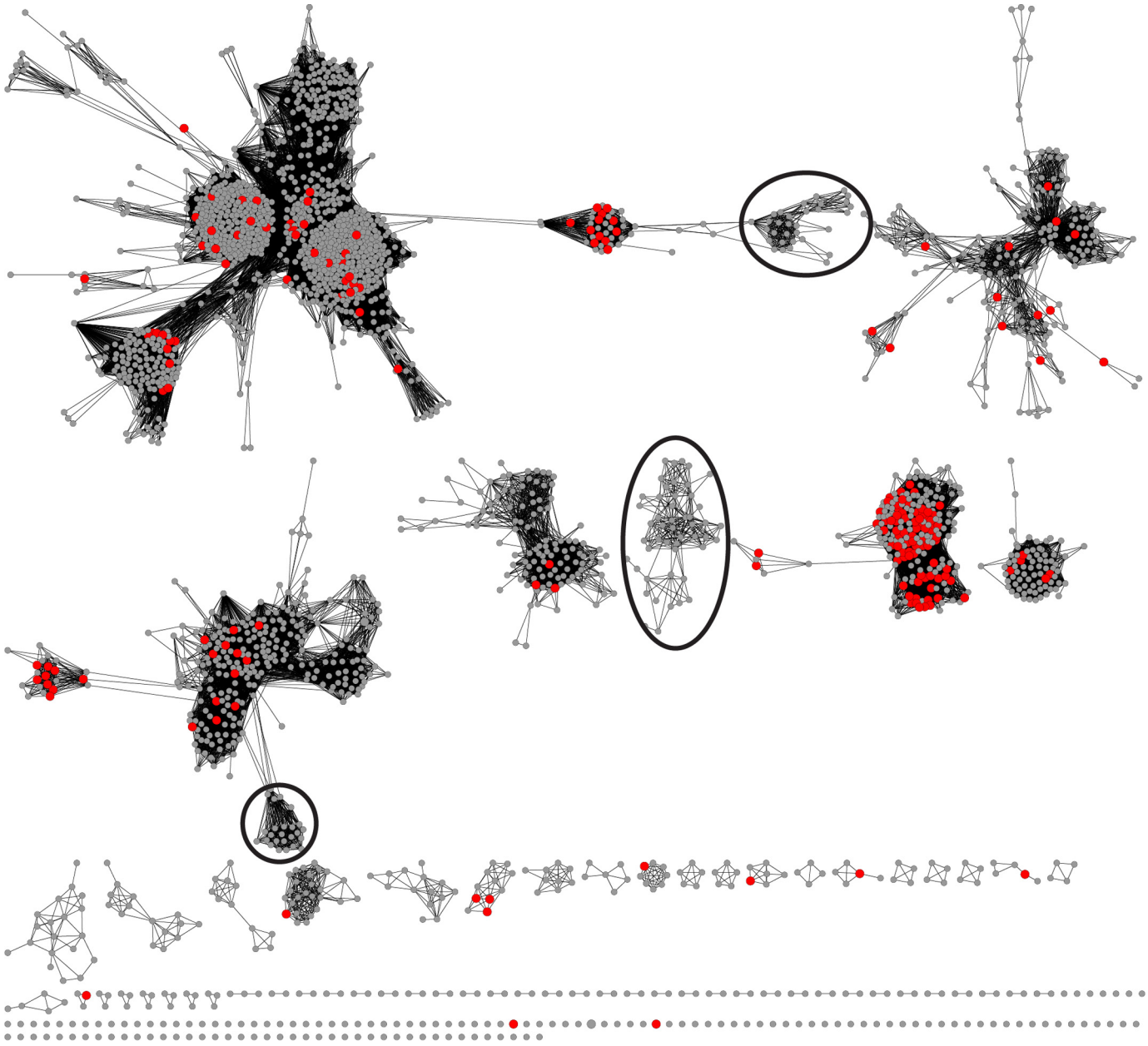

FIGURE 1. **Representative sequence similarity network for the isoprenoid synthase I superfamily that is available from the Structure-Function Linkage Database (SFLD) (62).** Each node (*circle*) represents a group of 1–732 sequences, where each sequence in a node is at least 50% identical to a seed sequence that defines that node (computed using the CD-HIT program (63)). The 2,499 nodes in this network represent over 16,000 sequences. Each edge (*line*) between two nodes indicates that the sequences represented by the connected nodes have a BLAST similarity score with an average −log(E-value) of 30 or more significance. At this −log (E-value) cutoff, alignments have an average length of 273 amino acids, and an average percent identity of 31%. Nodes are laid out in Cytoscape using the yFiles organic layout. A node is colored *red* if at least one constituent sequence represented by that node has a functional annotation in the Swiss-Prot database. A node is colored *gray* if no sequence in that representative node has a functional annotation in Swiss-Prot. Several clusters of nodes where no corresponding sequence has a functional annotation in Swiss-Prot are indicated with *black ovals*.

We Know Very Little about Structure-Function Relationships in Large Enzyme Superfamilies

When examining the available functional information for a superfamily, one of the most striking observations is how much we do not know. This is due in part to the rapid increase in sequence information that continues to accrue at a rapid rate. As a result, the members of many superfamilies now contain many thousands of sequences for which no functional information is available. Even in well studied superfamilies, there are large swaths of protein space where reliable predictions of even general functional features may be difficult. For example, the sequence similarity network in Fig. 1 shows members of the "isoprenoid synthase I" superfamily (19, 20) mapped with functional annotations from the Swiss-Prot database (21). Swiss-Prot annotations are reviewed by curators, preferably based on experimental information, and have been found to be highly accurate when compared with annotations from other major protein databases (22). Although many of the sequences in Fig. 1 have a functional annotation in Swiss-Prot (*red nodes*), there are many other nodes that do not, including sequences in the

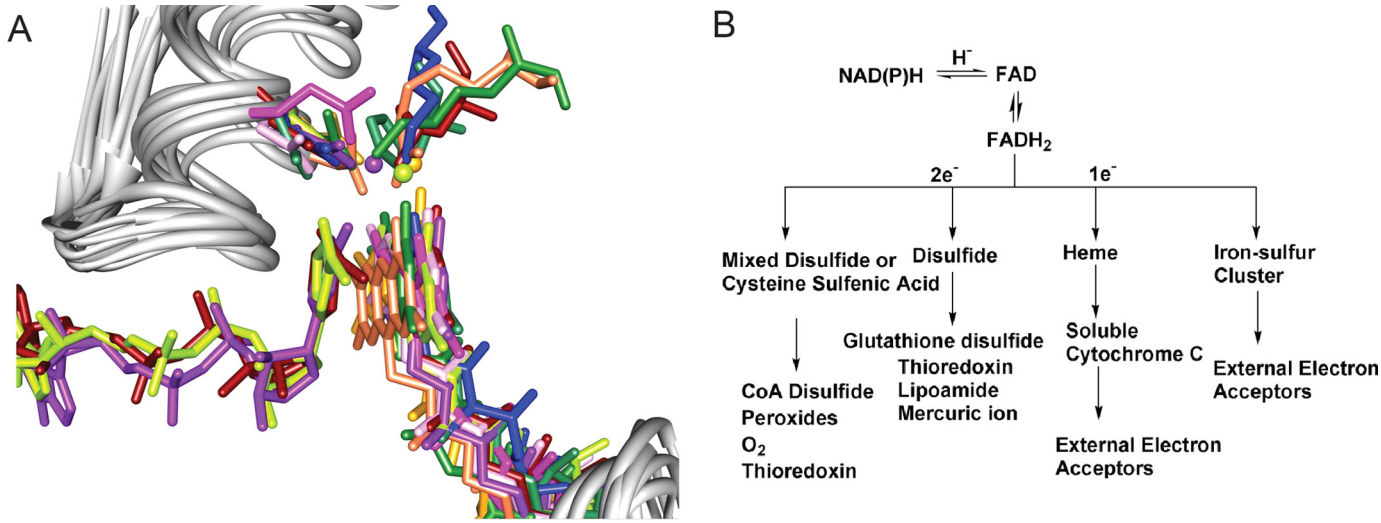

FIGURE 2. **Transfer of electrons from members of the tDBDF superfamily to acceptors.** *A*, superimposed active sites showing 10 members of the tDBDF superfamily. The cofactors and conserved side chains important for stabilizing the isoalloxazine and nicotinamide ring complex are shown in color with each color representing a different reaction family structure. Water residues involved in stabilizing the complex are shown as *balls*. *B*, superfamily members can transfer electrons to acceptors one or two at a time. Intermediate acceptors can be small molecules or proteins, which in turn transfer electrons to a variety of small molecule acceptors or external protein partners. Figure and legend adapted from Ref. 29.

large clusters highlighted in the figure. With so many of these sequences representing unknowns, deducing their contributions to our understanding of the evolution of function in this superfamily will remain challenging, perhaps for years to come. This trend is general, with 20% of protein domains in the Pfam database (23) annotated as "domains of unknown function". A recent study of model bacterial organisms suggests that many of these domains of unknown function are essential proteins (24).

In other superfamilies, even those of broad importance to many organisms, the proportion of unknowns can be even higher. In the cytosolic glutathione transferases, the large majority of sequences are unknowns, with only a minority of superfamily members confirmed to catalyze glutathione transferase-like reactions (16). Knowledge of the physiological reaction(s) catalyzed by these important and heavily studied enzymes is even more sparse as most enzymes experimentally confirmed to catalyze glutathione transferase activity have relied on assays using synthetic compounds as substrates. Likewise, only a relatively few of the enzymes of the "radical *S*-adenosyl methionine (SAM)" superfamily have been structurally characterized (25, 26). As many of the 50,000 nonredundant sequences of this superfamily are highly divergent from each other, understanding their evolution from the currently available structural information is still substantially limited. Further complicating this task, the few structures that are available exhibit such major variations that inference of general features of even the folds of many unknowns is difficult.

Another confounding issue for understanding the evolution of enzyme superfamilies is that the proteins of known function are often not evenly distributed across the sequence space of a superfamily. This is in part due to bias in targets chosen for experimental characterization, with characterization favored for proteins from model organisms or from organisms such as pathogens where the need for functional knowledge may be especially compelling. In the cytosolic glutathione transferases, for example, the human enzymes are far better studied than those from other species. Moreover, high throughput experiments, in which a very small number of studies currently account for a significant proportion of electronically available annotations, show biases in the types of functional information they provide as well as in the types of proteins that are targeted (27).

Many Types of Structural Variation Accompany Evolution of New Functions

How common functional features are retained during evolution while also allowing for the sequence and structural variation required to produce new reactions remains a major question for understanding enzyme evolution in functionally diverse superfamilies. Variations in oligomeric state and protein-protein interactions among sets of homologous proteins are of course well known. Large scale studies of enzyme superfamilies have more recently begun to reveal in greater detail broad patterns by which the divergence in function of each unique superfamily may be accompanied by significant structural modifications. Here, three different superfamilies illustrate strategies nature has used to maintain the fundamental chemistry "hardwired" into a structural and active-site architecture while enabling the evolution of many different reactions.

The cofactor-dependent "two dinucleotide-binding domains flavoproteins" (tDBDF)[2] superfamily is composed of many different reaction families that include several types of monooxygenases, reductases, and dehydrogenases. Comparison of their sequences and structures illustrates how variations in protein-protein interactions can enable a diverse set of overall reactions while the specific organization of the cofactors within the active site is stringently constrained by an active site architecture required for binding the dinucleotide cofactors (28–30). This

[2] The abbreviations used are: tDBDF, two dinucleotide-binding domains flavoproteins; N6P, nucleophilic attack, 6-bladed β-propeller; OSBS, *o*-succinylbenzoate synthase.

ensures that all enzymes of the superfamily share a unidirectional electron flow from the *re*-side to the *si*-side of the isoalloxazine ring of the FAD cofactor so that electron acceptors unique to each member family access the FAD cofactor from the *si*-side of the isoalloxazine ring (Fig. 2*A*). Diversity in the functions of the different reaction families has evolved in part by pairing the delivery of electrons out of the tDBDF member active sites with varied electron acceptors presented via protein-small molecule or protein-protein interactions (Fig. 2*B*) (29). Many of the penultimate or ultimate acceptor proteins come from different fold classes, resulting in a number of solutions for the evolution of these important oxidation/reduction systems.

The "vicinal oxygen chelate fold" (VOC) superfamily represents a quite different structural paradigm (31). All of these enzymes share a common $\beta\alpha\beta\beta\beta$ fold module that provides the environment for metal coordination promoting direct electrophilic participation of the metal ion in catalysis (with the notable exception of the non-enzymatic bleomycin- and mitomycin-binding proteins (32)). This common module is combined and permuted in at least six distinct ways in the different enzymes in the superfamily (33), each associated with different types of reactions, including nucleophilic opening of epoxide, oxidative cleavage of a C–C bond, isomerization, and epimerization.

Another type of structural variation in enzyme superfamilies involves addition of inserts of varying size and functional roles within a conserved core domain. These can play a functional role in enabling diversity in overall reactions while maintaining the chemical capability common to all members of the superfamily. In the "haloalkanoic acid dehalogenase" (HAD) superfamily, an aspartate nucleophile, which forms a covalent intermediate with the substrate, is well conserved (34). Other catalytic residues are found in somewhat different configurations depending on the function of the enzyme, but are also relatively well conserved. Although the core Rossmann fold provides much of the fundamental chemistry that typifies the superfamily, many substrate-binding residues are contributed by variable capping domains, with the active sites of these enzymes situated between the core and capping domains. The cap domains may be inserted at two different points within the conserved core structure and can come from different fold classes (34–36). Although the capping domains clearly play a role in function, functional type does not cleanly correlate with cap type. Thus, although co-evolution between core and cap domains offers hints about how this large superfamily has evolved structural diversity, mapping of these variations to functional properties remains a difficult challenge.

Profound variations in the active sites of members of functionally diverse superfamilies may also allow for diversity in the overall reactions catalyzed across a superfamily. The members of the "nucleophilic attack, 6-bladed β-propeller" (N6P) superfamily share a general catalytic strategy involving nucleophilic attack on an sp^2-hybridized electrophilic atom (37). One major sequence similarity subgroup of the superfamily includes only a few characterized proteins that all catalyze esterase, lactonase, and/or phosphotriesterase reactions; a second subgroup is predicted to catalyze arylesterase-like reactions, typified by only one well characterized organophosphatase reaction, human paraoxonase (38). In the experimentally and structurally characterized proteins of these two subgroups, four conserved active site residues serve as ligands to a divalent metal ion required for catalysis in these hydrolytic reactions (39–42) (Fig. 3*A*). These conserved metal ligands can be identified from sequence comparisons so that the structural and active site similarities predicted for the sequences in these two subgroups suggest that most of the unknowns likely catalyze similar types of hydrolytic reactions.

A third sequence-similar subgroup of ~600 sequences, the "strictosidine synthase-like" proteins, was named for the function of the only experimentally characterized sequences in the subgroup. These enzymes catalyze the metal-independent condensation of tryptamine and secologanin to form strictosidine (43) (Fig. 3*B*). Because the sequences of this subgroup are more similar to characterized strictosidine synthases than to the sequences of the other two subgroups, they have been annotated in public databases as strictosidine synthases or strictosidine synthase-like proteins. It was not until these sequences were examined as part of a large scale analysis of the entire superfamily (37) that it became clear that the experimentally characterized strictosidine synthases were outliers even in the so-called strictosidine synthase-like subgroup. Unlike the seven proteins experimentally confirmed to catalyze the strictosidine synthase reaction, the huge majority of the other sequences in this subgroup appear to conserve four metal-binding ligands, and are thus more likely to catalyze hydrolytic reactions rather than the condensation reaction catalyzed by strictosidine synthase (Fig. 3*B*). Indeed, a strictosidine synthase-like protein conserving only three of the four typical metal-binding ligands and identified from phylogenetic analysis to be among the most similar to the experimentally characterized strictosidine synthases was shown to have hydrolytic activity, but no detectable strictosidine synthase activity (37).

The evolutionary trajectory resulting in both the contemporary hydrolytic enzymes and their metal-independent strictosidine synthase homologs remains a mystery. Although both catalytic types can be assigned to the same superfamily based on sequence and structural similarities, the substantive differences in their active sites offer stunning evidence for how little we understand about how new enzymatic reactions evolve. The most parsimonious explanation for the results of this study suggests that strictosidine synthase may have evolved from a metal-dependent ancestor catalyzing hydrolytic chemistry. However, a later comparison of proteins of the larger Pfam clan to which the N6P superfamily belongs suggests that most of those enzymes *lack* the four metal-binding residues that might be expected of a metal-dependent common ancestor (44), raising questions about this simple hypothesis and suggesting a more complicated path for the evolution of these enzymes.

Challenges for Understanding the Evolution of Varied Functions in Functionally Diverse Superfamilies

Because the different reaction families of a functionally diverse enzyme superfamily all "look alike" with respect to superfamily common active site features, they are difficult to annotate and easy to misannotate (22). Understanding how their different functions evolve while conserving a fundamental

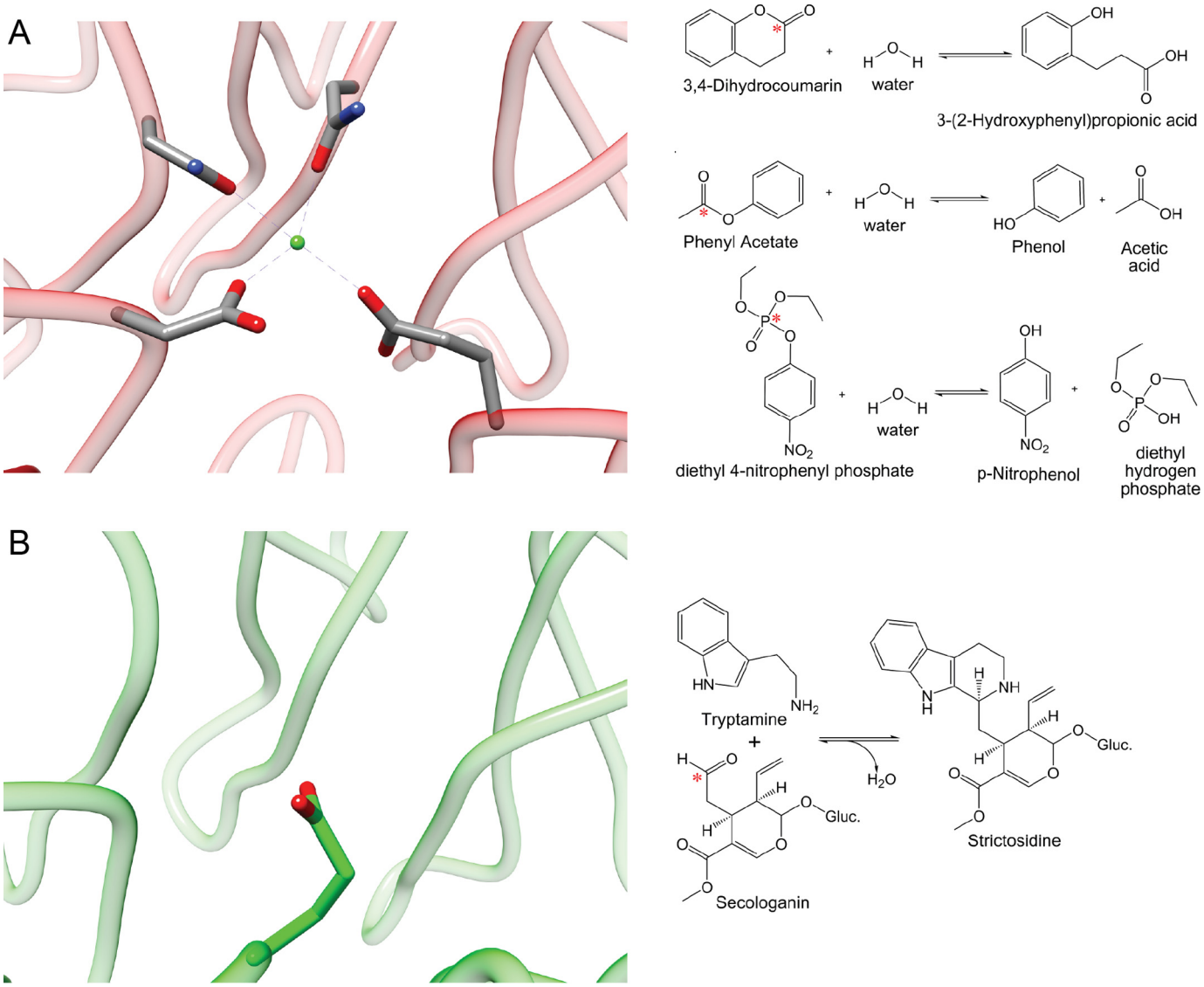

FIGURE 3. **A related catalytic strategy unites the strictosidine synthase enzymes with the rest of the N6P SF.** *A*, *left panel*, active site of diisopropylfluorophosphatase (Protein Data Bank (PDB) id: 2gvv). The four metal-binding ligands conserved in the majority of superfamily members are shown coordinated to a divalent metal ion. *Right panel*, examples of reactions catalyzed by characterized metal-dependent proteins. B, *left panel*, active site of strictosidine synthase (PDB id: 2fpb). *Right panel*, the metal-independent strictosidine synthase reaction. (Figure and legend adapted from Ref. 37 with permission.) In both *right panels*, a *red asterisk* indicates the electrophilic atom that is attacked in the reactions catalyzed by characterized members of the superfamily.

chemical strategy is yet more difficult, in part because each superfamily is unique with respect to the linked sequence, structural, and functional features its members share and the ways in which those sequences and structures have been modified by evolution for new functions.

Two other important themes complicate our understanding and at the same time offer new clues about the ways that new reactions may evolve. The first is functional promiscuity, which offers insight about the capabilities of the same active site to support different reactions. The second comes from observations that the same reaction can evolve independently from different intermediate ancestors in a superfamily tree.

Promiscuity and "Moonlighting" Enzymes

Catalytic promiscuity, the ability of an enzyme to catalyze different types of reactions using the same active site, has been observed in many enzymes. The seminal study by O'Brien and Herschlag (45) described this phenomenon in multiple systems and tied it to the evolution of new activities via enzyme duplication, a concept further elaborated by others, for example, in Ref. 46. Although some promiscuous enzymes evolved to catalyze the same reaction with different substrates, such as cytochrome P450s, others catalyze reactions that appear to be quite different from each other. The *o*-succinylbenzoate synthase (OSBS) enzyme from *Amycolatopsis* sp., a member of the "enolase" superfamily, was originally characterized as an *N*-acylamino acid racemase (47). This annotation was propagated to other related sequences, and only later was it determined that the biologically relevant function of the original enzyme was actually OSBS (48). Other enzymes from this superfamily have now been characterized that catalyze both the OSBS and the *N*-succinyl amino acid racemase reactions, and the evolution of both reactions continues to be a topic of investigation (49–51).

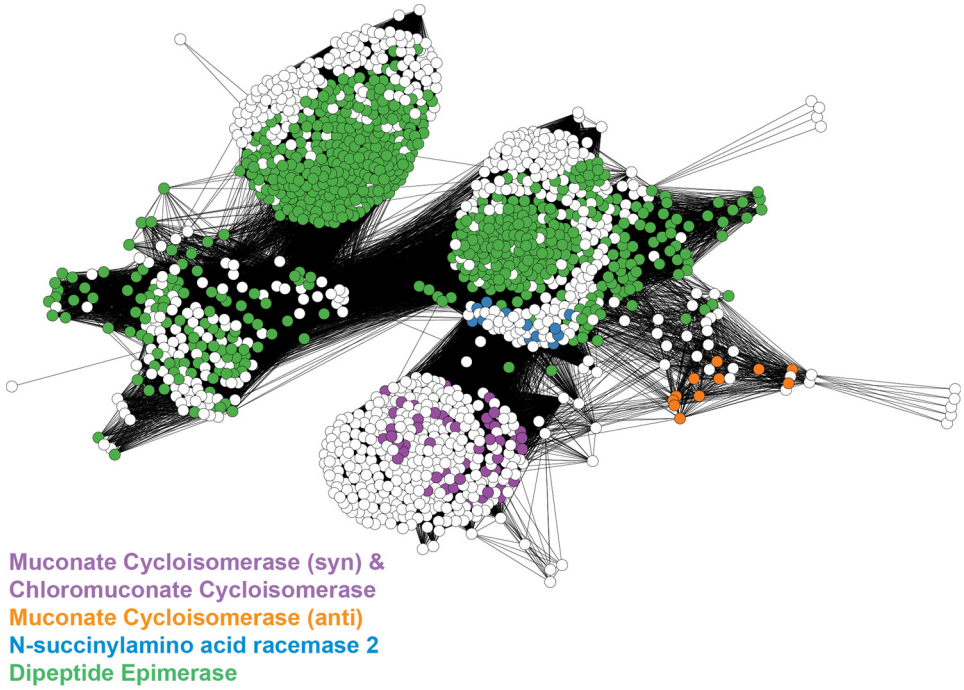

FIGURE 4. **Full sequence similarity network for a subset of the enolase superfamily, including the two muconate cycloisomerase families and their closest neighbors.** Each node (*circle*) represents a single sequence, and edges (*lines*) represent BLAST connections with a −log (E-value) of 47 or more significant. Nodes are laid out using the yFiles organic layout and colored by SFLD family assignment.

The concept of moonlighting enzymes, describing additional structural or regulatory functions performed in addition to their catalytic functions, has also provided novel insight about the evolution of new reactions from existing structures. An important early example describes the conscription of several different enzymes to serve roles as eye lens proteins (52). Enzymes from functionally diverse superfamilies have also been shown to play moonlighting roles. For example, the glycolytic enzyme enolase, another reaction family of the enolase superfamily, plays many moonlighting roles relevant to human health and disease (see Refs. 53 and 54 for some examples). Broader inventories of moonlighting proteins have also recently been collected (see Refs. 55 and 56 and references therein).

As observations of both promiscuous and moonlighting enzymes continue to increase, it seems likely that both phenomena may be far more widespread than previously thought, although many additional studies will be needed to obtain more accurate estimates of their prevalence. The widespread incidence of promiscuous or multifunctional enzymes also suggests that definitions of functional boundaries may be difficult to determine, even with the aid of similarity clustering of large numbers of sequences (see, for example, Refs. 13 and 57). As a corollary, sorting out the evolutionary path by which these functional types emerged may not accurately predict functional boundaries either, especially in functionally diverse superfamilies. Indeed, the idea that each enzyme has a single, specific function may be a more artificial construct than has been recognized. Alternatively, enzyme function space may perhaps be better described as a continuum rather than a set of discrete definitions.

Invention of the Same Reaction within a Superfamily Phylogeny

Convergent evolution, where unrelated enzymes have evolved to catalyze the same overall reaction, has been well studied. (For a recent review, see Ref. 58 and references therein.) Perhaps less widely appreciated is the propensity for proteins with the same, or similar, functions to be invented multiple times from different starting points within a superfamily (13, 16–18, 59–61). Fig. 4 shows a sequence similarity network of a subgroup of enzymes of the enolase superfamily, including the two different families that catalyze the cycloisomerization of muconate. These two families are distinct from each other in the phylogenetic tree of the subgroup to which both belong (49), suggesting that they evolved from different progenitors within the subgroup. Supporting this suggestion, experimental work shows that the stereochemical course of the reaction differs between the two families due to different modes of substrate binding, which result in opposite faces of the enolate anion intermediate being presented to the conserved Lys acid catalyst (17). In the same subgroup, two different families of enzymes that catalyze an *N*-succinyl amino acid racemase reaction have also been identified (18).

Because all enzymes in each functionally diverse superfamily already have the chemical machinery required for a critical part of the reaction, it may not be surprising that enzymes catalyzing the same overall reaction might evolve from different precursors within the superfamily. This has important implications both for understanding enzyme evolution and for predicting the functions for proteins in functionally diverse superfamilies. Knowing that a particular function is found in one region of a phylogenetic tree does not preclude the same function from

occurring in another, quite distant, area of the tree. Further, knowing that two enzymes catalyze the same overall reaction does not necessarily mean that they are close homologs, even if they belong to the same superfamily.

Conclusions and Future Directions

In the breadth of their organismal representation, the range of biochemical reactions they catalyze, and the many ways nature has reused each ancestral scaffold for many different functions, functionally diverse enzyme superfamilies offer a powerful model for understanding how the enzymes required for life have evolved. By determining the conserved structure-function paradigm represented by each of these "privileged scaffolds," we can begin to sort out the structural variations nature has used to evolve a wide array of different chemical reactions and tailor each of them for their specialized biological roles.

Analysis of such superfamilies is daunting, however, as each may contain tens of thousands of sequences, most of which are of unknown reaction or substrate specificities. Thus, the challenges even for managing the data, let alone inferring the functional repertoire each superfamily supports, may appear at first glance to be insurmountable. Offsetting the obvious impossibility of experimentally characterizing the molecular functions of the still rapidly growing volume of newly discovered enzyme sequences, we propose that the context provided by large scale computational characterization of these superfamilies will lead to new types of hypotheses about their structure-function relationships that cannot be accessed by comparison of only a few homologous enzymes. Even in the early forms in which this technology has been used, protein similarity networks have already been shown to provide structure-function mapping on the scale required. Future work by both the biochemical and the computational communities will improve both the robustness and the interpretability of this approach and expand its applications to address pressing issues that range from choosing experimental targets for answering many types of questions to exploiting our understanding of natural evolution to aid in engineering new reactions in the laboratory.

REFERENCES

1. Chothia, C. (1992) Proteins: one thousand families for the molecular biologist. *Nature* **357,** 543–544
2. Baier, F., and Tokuriki, N. (2014) Connectivity between catalytic landscapes of the metallo-β-lactamase superfamily. *J. Mol. Biol.* **426,** 2442–2456
3. 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,** 30591–30594
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,** 209–246
5. Glasner, M. E., Gerlt, J. A., and Babbitt, P. C. (2006) Evolution of enzyme superfamilies. *Curr. Opin. Chem. Biol.* **10,** 492–497
6. Schmidt, D. M., 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 (β/α)8-barrels: functional promiscuity produced by single substitutions in the enolase superfamily. *Biochemistry* **42,** 8387–8393
7. Almonacid, D. E., and Babbitt, P. C. (2011) Toward mechanistic classification of enzyme functions. *Curr. Opin. Chem. Biol.* **15,** 435–442
8. Furnham, N., Sillitoe, I., Holliday, G. L., Cuff, A. L., Laskowski, R. A., Orengo, C. A., and Thornton, J. M. (2012) Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies. *PLoS Comput. Biol.* **8,** e1002403
9. Iyer, L. M., Zhang, D., Burroughs, A. M., and Aravind, L. (2013) Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA. *Nucleic Acids Res.* **41,** 7635–7655
10. Nelson, K. J., Knutson, S. T., Soito, L., Klomsiri, C., Poole, L. B., and Fetrow, J. S. (2011) Analysis of the peroxiredoxin family: using active-site structure and sequence information for global classification and residue analysis. *Proteins* **79,** 947–964
11. Atkinson, H. J., and Babbitt, P. C. (2009) An atlas of the thioredoxin fold class reveals the complexity of function-enabling adaptations. *PLoS Comput. Biol.* **5,** e1000541
12. Uberto, R., and Moomaw, E. W. (2013) Protein similarity networks reveal relationships among sequence, structure, and function within the Cupin superfamily. *PLoS One* **8,** e74477
13. Wallrapp, F. H., Pan, J. J., Ramamoorthy, G., Almonacid, D. E., Hillerich, B. S., Seidel, R., Patskovsky, Y., Babbitt, P. C., Almo, S. C., Jacobson, M. P., and Poulter, C. D. (2013) Prediction of function for the polyprenyl transferase subgroup in the isoprenoid synthase superfamily. *Proc. Natl. Acad. Sci. U.S.A.* **110,** E1196–E1202
14. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504
15. Atkinson, H. J., Morris, J. H., Ferrin, T. E., and Babbitt, P. C. (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One* **4,** e4345
16. Mashiyama, S. T., Malabanan, M. M., Akiva, E., Bhosle, R., Branch, M. C., Hillerich, B., Jagessar, K., Kim, J., Patskovsky, Y., Seidel, R. D., Stead, M., Toro, R., Vetting, M. W., Almo, S. C., Armstrong, R. N., and Babbitt, P. C. (2014) Large-scale determination of sequence, structure, and function relationships in cytosolic glutathione transferases across the biosphere. *PLoS Biol.* **12,** e1001843
17. Sakai, A., Fedorov, A. A., Fedorov, E. V., Schnoes, A. M., Glasner, M. E., Brown, S., Rutter, M. E., Bain, K., Chang, S., Gheyi, T., Sauder, J. M., Burley, S. K., Babbitt, P. C., Almo, S. C., and Gerlt, J. A. (2009) Evolution of enzymatic activities in the enolase superfamily: stereochemically distinct mechanisms in two families of *cis*,*cis*-muconate lactonizing enzymes. *Biochemistry* **48,** 1445–1453
18. Song, L., Kalyanaraman, C., Fedorov, A. A., Fedorov, E. V., Glasner, M. E., Brown, S., Imker, H. J., Babbitt, P. C., Almo, S. C., Jacobson, M. P., and Gerlt, J. A. (2007) Prediction and assignment of function for a divergent *N*-succinyl amino acid racemase. *Nat. Chem. Biol.* **3,** 486–491
19. Christianson, D. W. (2006) Structural biology and chemistry of the terpenoid cyclases. *Chem. Rev.* **106,** 3412–3442
20. Oldfield, E., and Lin, F. Y. (2012) Terpene biosynthesis: modularity rules. *Angew Chem. Int. Ed. Engl.* **51,** 1124–1137
21. UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **42,** D191–D198
22. Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* **5,** e1000605
23. Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012) The Pfam protein families database. *Nucleic Acids Res.* **40,** D290–301
24. Goodacre, N. F., Gerloff, D. L., and Uetz, P. (2014) Protein domains of unknown function are essential in bacteria. *MBio* **5,** e00744–00713
25. Goldman, P. J., Grove, T. L., Booker, S. J., and Drennan, C. L. (2013) X-ray analysis of butirosin biosynthetic enzyme BtrN redefines structural motifs for AdoMet radical chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 15949–15954
26. Vey, J. L., and Drennan, C. L. (2011) Structural insights into radical generation by the radical SAM superfamily. *Chem. Rev.* **111,** 2487–2506
27. Schnoes, A. M., Ream, D. C., Thorman, A. W., Babbitt, P. C., and Fried-

berg, I. (2013) Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS Comput. Biol.* **9,** e1003063
28. Vallon, O. (2000) New sequence motifs in flavoproteins: evidence for common ancestry and tools to predict structure. *Proteins* **38,** 95–114
29. Ojha, S., Meng, E. C., and Babbitt, P. C. (2007) Evolution of function in the "two dinucleotide binding domains" flavoproteins. *PLoS Comput. Biol.* **3,** e121
30. Dym, O., and Eisenberg, D. (2001) Sequence-structure analysis of FAD-containing proteins. *Protein Sci.* **10,** 1712–1728
31. Meng, E. C., and Babbitt, P. C. (2011) Topological variation in the evolution of new reactions in functionally diverse enzyme superfamilies. *Curr. Opin. Struct. Biol.* **21,** 391–397
32. Armstrong, R. N. (2000) Mechanistic diversity in a metalloenzyme superfamily. *Biochemistry* **39,** 13625–13632
33. He, P., and Moran, G. R. (2011) Structural and mechanistic comparisons of the metal-binding members of the vicinal oxygen chelate (VOC) superfamily. *J. Inorg. Biochem.* **105,** 1259–1272
34. Burroughs, A. M., Allen, K. N., Dunaway-Mariano, D., and Aravind, L. (2006) Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. *J. Mol. Biol.* **361,** 1003–1034
35. Allen, K. N., and Dunaway-Mariano, D. (2009) Markers of fitness in a successful enzyme superfamily. *Curr. Opin. Struct. Biol.* **19,** 658–665
36. Pandya, C., Brown, S., Pieper, U., Sali, A., Dunaway-Mariano, D., Babbitt, P. C., Xia, Y., and Allen, K. N. (2013) Consequences of domain insertion on sequence-structure divergence in a superfold. *Proc. Natl. Acad. Sci. U.S.A.* **110,** E3381–3387
37. Hicks, M. A., Barber, A. E., 2nd, Giddings, L. A., Caldwell, J., O'Connor, S. E., and Babbitt, P. C. (2011) The evolution of function in strictosidine synthase-like proteins. *Proteins* **79,** 3082–3098
38. Harel, M., Aharoni, A., Gaidukov, L., Brumshtein, B., Khersonsky, O., Meged, R., Dvir, H., Ravelli, R. B. G., McCarthy, A., Toker, L., Silman, I., Sussman, J. L., and Tawfik, D. S. (2004) Structure and evolution of the serum paraoxonase family of detoxifying and anti-atherosclerotic enzymes. *Nat. Struct. Mol. Biol.* **11,** 412–419
39. Tanaka, Y., Morikawa, K., Ohki, Y., Yao, M., Tsumoto, K., Watanabe, N., Ohta, T., and Tanaka, I. (2007) Structural and mutational analyses of Drp35 from *Staphylococcus aureus*: a possible mechanism for its lactonase activity. *J. Biol. Chem.* **282,** 5770–5780
40. Katsemi, V., Lücke, C., Koepke, J., Löhr, F., Maurer, S., Fritzsch, G., and Rüterjans, H. (2005) Mutational and structural studies of the diisopropyl-fluorophosphatase from *Loligo vulgaris* shed new light on the catalytic mechanism of the enzyme. *Biochemistry* **44,** 9022–9033
41. Blum, M. M., Löhr, F., Richardt, A., Rüterjans, H., and Chen, J. C. (2006) Binding of a designed substrate analogue to diisopropyl fluorophosphatase: implications for the phosphotriesterase mechanism. *J. Am. Chem. Soc.* **128,** 12750–12757
42. Blum, M. M., and Chen, J. C. (2010) Structural characterization of the catalytic calcium-binding site in diisopropyl fluorophosphatase (DFPase): comparison with related β-propeller enzymes. *Chem. Biol. Interact.* **187,** 373–379
43. Kutchan, T. M. (1989) Expression of enzymatically active cloned strictosidine synthase from the higher plant *Rauvolfia serpentina* in *Escherichia coli*. *FEBS Lett.* **257,** 127–130
44. Hicks, M. A., Barber II, A. E., and Babbitt, P. C. (2014) The nucleophilic attack 6-bladed β-propeller (N6P) superfamily. in *Protein Families: Relating Protein Sequence, Structure, and Function* (Orengo, C., and Bateman, A. eds), pp. 127–158, John Wiley & Sons, New York
45. O'Brien, P. J., and Herschlag, D. (1999) Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.* **6,** R91–R105
46. Khersonsky, O., Roodveldt, C., and Tawfik, D. S. (2006) Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr. Opin. Chem. Biol.* **10,** 498–508
47. Tokuyama, S., and Hatano, K. (1995) Cloning, DNA sequencing and heterologous expression of the gene for thermostable *N*-acylamino acid racemase from *Amycolatopsis* sp. TS-1–60 in *Escherichia coli*. *Appl. Microbiol. Biotechnol.* **42,** 884–889
48. 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,** 4252–4258
49. Glasner, M. E., Fayazmanesh, N., Chiang, R. A., Sakai, A., Jacobson, M. P., Gerlt, J. A., and Babbitt, P. C. (2006) Evolution of structure and function in the *o*-succinylbenzoate synthase/*N*-acylamino acid racemase family of the enolase superfamily. *J. Mol. Biol.* **360,** 228–250
50. Odokonyero, D., Ragumani, S., Lopez, M. S., Bonanno, J. B., Ozerova, N. D., Woodard, D. R., Machala, B. W., Swaminathan, S., Burley, S. K., Almo, S. C., and Glasner, M. E. (2013) Divergent evolution of ligand binding in the *o*-succinylbenzoate synthase family. *Biochemistry* **52,** 7512–7521
51. Brizendine, A. M., Odokonyero, D., McMillan, A. W., Zhu, M., Hull, K., Romo, D., and Glasner, M. E. (2014) Promiscuity of Exiguobacterium sp. AT1b o-succinylbenzoate synthase illustrates evolutionary transitions in the OSBS family. *Biochem. Biophys. Res. Commun.* **450,** 679–684
52. Piatigorsky, J., O'Brien, W. E., Norman, B. L., Kalumuck, K., Wistow, G. J., Borras, T., Nickerson, J. M., and Wawrousek, E. F. (1988) Gene sharing by δ-crystallin and argininosuccinate lyase. *Proc. Natl. Acad. Sci. U.S.A.* **85,** 3479–3483
53. Avilán, L., Gualdrón-López, M., Quiñones, W., González-González, L., Hannaert, V., Michels, P. A., and Concepción, J. L. (2011) Enolase: a key player in the metabolism and a probable virulence factor of trypanosomatid parasites-perspectives for its use as a therapeutic target. *Enzyme Res.* **2011,** 932549
54. Butterfield, D. A., and Lange, M. L. (2009) Multifunctional roles of enolase in Alzheimer's disease brain: beyond altered glucose metabolism. *J. Neurochem.* **111,** 915–933
55. Copley, S. D. (2012) Moonlighting is mainstream: paradigm adjustment required. *Bioessays* **34,** 578–588
56. Henderson, B., and Martin, A. (2013) Bacterial moonlighting proteins and bacterial virulence. *Curr. Top. Microbiol. Immunol.* **358,** 155–213
57. Lukk, T., Sakai, A., Kalyanaraman, C., Brown, S. D., Imker, H. J., Song, L., Fedorov, A. A., Fedorov, E. V., Toro, R., Hillerich, B., Seidel, R., Patskovsky, Y., Vetting, M. W., Nair, S. K., Babbitt, P. C., Almo, S. C., Gerlt, J. A., and Jacobson, M. P. (2012) Homology models guide discovery of diverse enzyme specificities among dipeptide epimerases in the enolase superfamily. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 4122–4127
58. Omelchenko, M. V., Galperin, M. Y., Wolf, Y. I., and Koonin, E. V. (2010) Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biol. Direct* **5,** 31
59. Mursula, A. M., van Aalten, D. M., Hiltunen, J. K., and Wierenga, R. K. (2001) The crystal structure of Δ^3-Δ^2-enoyl-CoA isomerase. *J. Mol. Biol.* **309,** 845–853
60. Bruns, C. M., Nowalk, A. J., Arvai, A. S., McTigue, M. A., Vaughan, K. G., Mietzner, T. A., and McRee, D. E. (1997) Structure of *Haemophilus influenzae* Fe^{+3}-binding protein reveals convergent evolution within a superfamily. *Nat. Struct. Biol.* **4,** 919–924
61. Sharkey, T. D., Yeh, S., Wiberley, A. E., Falbel, T. G., Gong, D., and Fernandez, D. E. (2005) Evolution of the isoprene biosynthetic pathway in kudzu. *Plant Physiol.* **137,** 700–712
62. Akiva, E., Brown, S., Almonacid, D. E., Barber, A. E., 2nd, Custer, A. F., Hicks, M. A., Huang, C. C., Lauck, F., Mashiyama, S. T., Meng, E. C., Mischel, D., Morris, J. H., Ojha, S., Schnoes, A. M., Stryke, D., Yunes, J. M., Ferrin, T. E., Holliday, G. L., and Babbitt, P. C. (2014) The structure-function linkage database. *Nucleic Acids Res.* **42,** D521–530
63. Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659