Leveraging Enzyme Structure−Function Relationships for Functional Inference and Experimental Design: The Structure−Function Linkage Database†

Scott C.-H. Pegg,‡ Shoshana D. Brown,‡ Sunil Ojha,‡ Jennifer Seffernick,‖ Elaine C. Meng,§ John H. Morris,§ Patricia J. Chang,‡ Conrad C. Huang,§ Thomas E. Ferrin,‡,§ and Patricia C. Babbitt*,‡,§

Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, University of California, San Francisco, 1700 Fourth Street, San Francisco, California 94143-2250, and Department of Biochemistry, Molecular Biology, and Biophysics, Biological Process Technology Institute, and Center for Microbial and Plant Genomics, University of Minnesota, St. Paul, Minnesota 55108

Received October 14, 2005; Revised Manuscript Received December 8, 2005

ABSTRACT: The study of mechanistically diverse enzyme superfamilies—collections of enzymes that perform different overall reactions but share both a common fold and a distinct mechanistic step performed by key conserved residues—helps elucidate the structure−function relationships of enzymes. We have developed a resource, the structure−function linkage database (SFLD), to analyze these structure−function relationships. Unique to the SFLD is its hierarchical classification scheme based on linking the specific partial reactions (or other chemical capabilities) that are conserved at the superfamily, subgroup, and family levels with the conserved structural elements that mediate them. We present the results of analyses using the SFLD in correcting misannotations, guiding protein engineering experiments, and elucidating the function of recently solved enzyme structures from the structural genomics initiative. The SFLD is freely accessible at http://sfld.rbvi.ucsf.edu.

An important goal of the study of enzymes is a knowledge of their sequence, structure, and function relationships that is deductive, allowing us to rapidly determine without physical experimentation, the substrate, chemical mechanism, and product of a given sequence or structure (*1*). Such knowledge should also be predictive, enabling us to design enzymes that will perform desired reactions on substrates of our choosing. For genome projects, our current abilities to deduce function rely primarily on how similar a new protein sequence is to that of a characterized enzyme. These methods provide what is essentially a binary classification. (The protein either performs the same reaction as the characterized enzyme or it does not.) Depending upon the similarity thresholds used, this approach can produce many erroneous annotations. The annotation of enzymes by breaking down functional representations to correspond to structural similarities at different levels (e.g., superfamily, subgroup, and family levels) enhances the precision with which we can predict functional characteristics and aids in functional inference. In this article we present and demonstrate the application of a resource for performing such multilevel annotations, the structure−function linkage database (SFLD)[1].

Our strategy begins with the study of how enzymes have evolved in nature, seeking an understanding of how proteins of identical fold are used in many contexts to deliver a wide variety of functions. Horowitz (*2*, *3*) proposed that as an enzyme evolved, it maintained the ability to bind a particular substrate although the structural regions of the protein involved in delivering chemistry changed. Although there are some instances that appear to fit this early model (*4*), there are now many more cases observed in which it appears that in the evolution of new functions, the conserved aspect of the enzyme structure−function relationship is not the ability to bind a specific substrate but rather the ability to perform a key mechanistic step in the chemical reaction (*5*, *6*). Recent surveys of enzyme structure and function (*7*−*9*) as well as related theoretical analyses of pathway evolution (*10*−*12*) support this paradigm of chemistry-constrained enzyme evolution. The result of millions of years of this type of evolution has produced mechanistically diverse enzyme superfamilies (*7*), each consisting of homologous enzymes that perform a wide variety of overall chemical reactions on an equally wide variety of substrates but maintain a key mechanistic step mediated by conserved active-site features.

The enolase superfamily (*13*, *14*) provides a good example of how the analysis of structure−function relationships in the context of mechanistically diverse superfamilies can lead to valuable biological insights. The 1000+ proteins of this superfamily perform many distinct and, often, quite different

† This work was performed with support from NIH grant GM60595 (to PCB), NSF grant 0234768 (to PCB), and NIH resource grant P41-RR01081 (to TEF).

* To whom correspondence should be addressed. Tel: (415) 476-3784. Fax: (415) 514-4260. E-mail: babbitt@cgl.ucsf.edu.

‡ Department of Biopharmaceutical Sciences, University of California, San Francisco.

§ Department of Pharmaceutical Chemistry, University of California, San Francisco.

‖ University of Minnesota.

[1] Abbreviations: SFLD, Structure−Function Linkage Database; HMM, hidden Markov model; CSA, Catalytic Site Atlas; AEE, L-Ala-D/L-Glu epimerase; MLEII, chloromuconate lactonizing enzyme, also called muconate lactonizing enzyme II; OSBS, *o*-succinylbenzoate synthase; E. C., Enzyme Commission.

10.1021/bi052101l CCC: $33.50 © 2006 American Chemical Society
Published on Web 02/04/2006

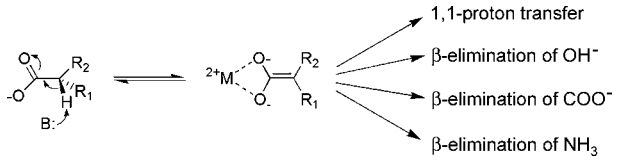

FIGURE 1: Partial reaction conserved in all members of the enolase superfamily, the enolization of a carboxylic acid via the abstraction of a proton alpha to a carboxylate leading to metal-assisted stabilization of an enolic intermediate. Following this initial conserved step, reactions within the superfamily can go down a variety of paths, requiring, in some cases, additional partial reactions as needed to complete the reactions shown. The structure−function relationships of those subsequent steps are distinguished in the SFLD at the subgroup and family levels.

overall chemical reactions. Thirteen of these distinct catalytic reactions have been experimentally characterized, and additional reactions can be predicted from phylogenetic and other bioinformatic and experimental analyses. Each of these reactions includes a common mechanistic step, the abstraction of a proton alpha to a carboxylate, leading to stabilization of a similar type of enolate anion intermediate. This reaction is shown in Figure 1. In all of the different superfamily members, this partial reaction is mediated by active-site residues conserved throughout the superfamily. Knowledge of how these enzymes deliver function, that is, how specific aspects of their chemical mechanisms are delivered via conserved residues, has allowed us and our collaborators to correct mistakes in the annotation of members of the superfamily, to predict the functions of newly sequenced members, and to design protein engineering experiments in which two members of the superfamily were modified to perform the very different overall reaction of a third.

As the rate of information generation from genomic sequencing and high-throughput structure determination projects increases, access to such information in the superfamily context becomes a pressing issue (*15*). What is required is a computational framework that provides more than information storage and the tools to access it. That is, we require a level of organization and curation that allows users to leverage the data for functional inference and biological insight. Several well-known databases address different aspects of this need, storing the raw sequence, structure, and function data in a manner that allows users to put it into a useful biological context. For example, the Superfamily (*16*) and PFAM (*17*) databases group related sequences together and provide hidden Markov models (HMMs) (*18*) that can be used in the assignment of homology to new protein sequences. Databases such as SCOP (*19*, *20*) and CATH (*21*) sort proteins into a structural hierarchy, enabling users to examine distant evolutionary relationships that are recognizable only from similarities in their 3D structures. The KEGG (*22*) and MetaCyc (*23*, *24*) databases provide a catalog of metabolic pathways and reactions, including many types of information about their component enzymes and their small molecule substrates and products. The more comprehensive database BRENDA (*25*) provides both structural and functional information, including the overall reactions of enzymes as well as their pathways and structures, which is organized using the overall reaction-based Enzyme Commission (E. C.) (*26*) numbering hierarchy. Yet none of these, including BRENDA, provides information about how these properties are linked, particularly, how a given protein sequence or structure performs its molecular function. For example, using the KEGG or BRENDA databases, one can easily find nucleotide and amino acid sequences, crystal structures, and representations of the overall reaction for the enzyme subtilisin, but no information is provided regarding the mechanism by which subtilisin performs proteolytic cleavage using the catalytic triad Ser-His-Asp (*27*). This is not surprising, however, as these databases were not specifically designed to contain information about how proteins deliver function at the molecular level but rather as tools that provide biological insights at higher levels of analysis.

A newer resource, the Catalytic Site Atlas (CSA) (*28*), provides a useful first step toward exploring enzyme sequence−structure−function relationships by storing the identities of catalytic residues for enzymes with solved crystal structures. One of the strengths of the CSA is the use of a strict definition of the term catalytic, which includes only those residues thought to be directly involved in the reaction catalyzed by an enzyme. The CSA has recently been demonstrated as a useful tool for enzyme function prediction (*29*). However, the CSA does not yet provide information regarding the reaction mechanism of enzymes. Helping to fill this gap, the recently developed databases EzCatDB (*30*) and MACiE (*31*) catalog a broad range of enzyme reactions and include information regarding the specific amino acids involved. The MaCiE database also describes each overall reaction in terms of its constituent steps to aid in the representation of enzyme mechanism. Still, none of these databases provide an organizational framework that directly represents the connections between enzymes with similar sequence−structure−function relationships.

We describe here a database we have designed to provide information and analysis of the links between structure and function information of enzymes, the structure−function linkage database (SFLD). Specifically, the SFLD captures structure−function relationships in mechanistically diverse enzyme superfamilies. It is unique in identifying not only the overall reactions of specific enzymes but also the partial reactions (or other chemical capabilities) that represent conserved functional capabilities shared among highly divergent members of a superfamily. These are the functional attributes that can be specifically correlated with conserved structural elements (residues or networks of residues) in superfamilies that follow the chemistry-constrained model of enzyme evolution. This database is meant to be a resource of information that can be leveraged to aid researchers in the analysis and engineering of protein function. In particular, correlation between structure and function at the superfamily level can be used for rules-based predictions of some functional capabilities of any new sequence or structure identified as a member of a superfamily in the database. Finer granularity assignment of a specific overall reaction, including substrate specificity, can also be achieved for sequences or structures that can be confidently assigned to an individual family within a superfamily. We demonstrate here the use of the SFLD for correcting misannotations, guiding protein engineering experiments, and elucidating the function of recently solved enzyme structures. We also describe the organization, architecture, and searching abilities of the SFLD.

MATERIALS AND METHODS

An initial report describing key aspects of the schema and methods relevant to the issue of structure−function linkage in mechanistically diverse enzyme superfamilies has been published (*32*). For convenience, we provide a summary of that information at the end of the Results section.

In the SFLD, three levels of functional granularity are defined. A superfamily is defined as a set of evolutionarily related proteins that perform in common a fundamental partial reaction or share a mechanistic attribute using conserved structural elements. The subgroup designation distinguishes structural variations in how the aspects of shared chemistry are delivered in subsets of the superfamily proteins. A family is defined as a set of proteins that perform the same overall reaction using an identical mechanism. Thus, families represent orthologous proteins.

RESULTS

The SFLD not only contains specific structure−function relationships for a set of homologous enzymes but also organizes them in a context focused on the shared chemistry performed by conserved active-site residues. This grouping of enzymes into functional fold superfamilies provides a more relevant theoretical representation of structure−function relationships in mechanistically diverse enzyme superfamilies than has previously been available. The resulting framework provides different levels of granularity at which functions can be assigned and thereby is less prone to the overprediction of functional characteristics than are representations based only on overall reactions. Thus, if only the active-site residues associated with a partial reaction are conserved in a member of a superfamily, then that protein is annotated only with the superfamily common partial reaction. Only if other family-specific structural characteristics are additionally conserved is a protein annotated with the family-specific overall reaction. In the sections that follow, examples of how the information and tools provided in the SFLD can be leveraged for functional insight are illustrated. Following this, we provide a short summary of the organization, contents, and analysis tools in the SFLD.

Correcting Misannotations

The need for rapid, high-throughput annotation of whole genomes has resulted in increasing reliance upon electronic annotation schemes that assign functional similarity according to sequence similarity. Thus, a new protein sequence is assigned the function of the annotated protein(s) closest to it in sequence (typically using a similarity threshold). For enzymes belonging to a widely divergent superfamily in which the members may perform many different reactions, this approach is often problematic because the protein (or protein family) showing the closest sequence similarity may not perform the same overall reaction. Such situations may be more common than initially expected, as recent experiments have demonstrated that even single amino acid changes in members of diverse enzyme superfamilies are enough to alter substrate specificity, mechanism, or both (*33*). In these cases, the specific correlations between conserved active-site residues and the aspects of function they mediate can sometimes be used to identify annotations that have been incorrectly assigned by high-throughput methods.

For example, the sequence annotated in Genbank as muconate cycloisomerase from the organism *Oceanobacillus iheyensis* (gi 23100420) shows as its closest characterized homologues the MLEI family of enzymes (members of the enolase superfamily) in both BLAST and PFAM searches. This sequence displays those active-site residues generally conserved across the superfamily to perform the conserved partial reaction of the abstraction of a proton alpha to a carboxylate. Yet, it lacks a key glutamic acid residue found in all experimentally characterized members and orthologs that are thought to perform the specific overall MLE reaction (*34*). Aligning the sequence to other families within the superfamily, armed with the knowledge of which residues are conserved at the superfamily level and which ones are conserved only at the family level, we believe this sequence is more likely to be a dipeptide epimerase.[2]

As sequencing projects expand to genomes that represent more diverse biological niches, we are likely to see a greater number of sequences that are incorrectly assigned the function of a distantly related but characterized homologue via the use of automated high-throughput annotation methods. Such errors and their propagation into secondary databases pose a significant problem to biologists at many levels, from the study of individual enzymes to the larger analyses of systems biology. Some studies have estimated the rates of misannotation in public databases to be as high as 15−30% (*35*, *36*). As shown in the example above, knowledge of how families and superfamilies of related enzymes carry out conserved mechanistic steps such as that provided by the SFLD can be leveraged to provide methods for correcting such misannotations. To gain a more accurate estimate of levels of misannotation for mechanistically diverse enzyme superfamilies in public databases, a systematic study of misannotation using the superfamilies in the SFLD is currently underway in our laboratory.

Protein Engineering

The organization of enzymes into the functional fold hierarchy used by the SFLD has helped guide protein engineering experiments designed to investigate potential pathways for functional evolution. By grouping evolutionarily related but mechanistically diverse enzymes into superfamilies according to their conserved mechanistic steps, we can more easily recognize the structural determinants associated with substrate specificity and distinguish them from residues conserved across the entire superfamily. With a knowledge of how the common mechanistic step is delivered by the structure, we can also recognize specific regions of the active site that may be altered to create the additional functional steps required to perform a new chemical reaction.

[2] We note that the closest BLAST hit reported using gi23100420 as a query against the NCBI "nr" database is gi16078363. This sequence, nominally annotated in the Genbank header as "similar to chloromuconate cycloisomerase," is more likely to be a dipeptide epimerase. In the alignment section of the BLAST output, other gi numbers grouped by the "nr" database as identical to gi16078363 are listed, including the solved structure of the experimentally characterized dipeptide epimerase (gi18158850) as well as identical sequences annotated as "chloromuconate cycloisomerase homologue ykfB" (gi7428454) and "YkfB" (gi2632019 and gi2633652). Automated annotation methods typically do not parse the alignment section of the BLAST output, and are unlikely to be able to resolve the conflicting annotations listed.

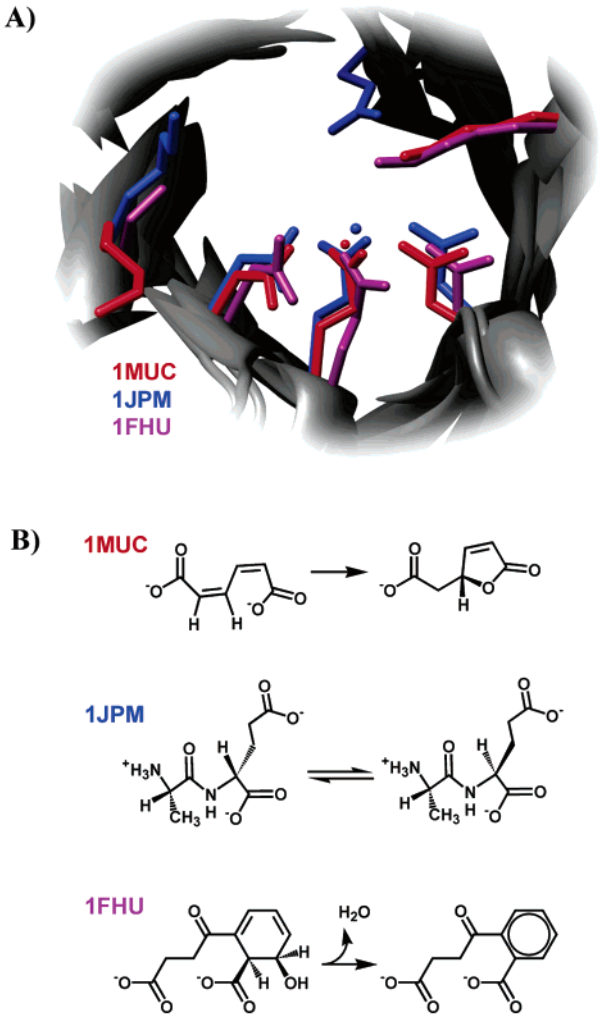

FIGURE 2: (A) Superimposed active sites of crystal structures of 1MUC (muconate lactonizing enzyme), 1JPM (L-Ala-D/L-Glu epimerase), and 1FHU (*o*-succinylbenzoate synthase). The highly conserved active-site residues involved in catalysis of the proton abstraction step common to all three enzymes are represented in color (1MUC in red, 1JPM in blue, and 1FHU in purple). (B) Different overall reactions catalyzed by these enzymes.

Proof of concept for this idea was recently obtained from a protein engineering experiment in which two members of the enolase superfamily were re-engineered to perform a quite different overall reaction of a third member (*33*). Both of the reactions performed by the template enzymes, L-Ala-D/L-Glu epimerase (AEE) and chloromuconate lactonizing enzyme (MLEII), and the target reaction, *o*-succinylbenzoate synthesis, represent different overall reactions within the enolase superfamily but share the common mechanistic step of the abstraction of a proton alpha to a carboxylate. Figure 2A shows the structural conservation of the residues in the active sites of these enzymes that perform this step. Aside from this common ability, however, all three of these enzymes perform quite different overall reactions (Figure 2B). By comparing the structures of the AEE and OSBS enzymes and considering the known structure−function relationships both have in common, that is, shared by the superfamily, contrasted with those unique to each enzyme, a single active-site mutant of AEE (D297G) was designed and found to perform the reaction catalyzed by OSBS. This choice was guided by the knowledge that substrates in the enolase superfamily must bind in an orientation conducive to the stabilization of the enolate anion intermediate by the conserved constellation of residues shown in Figure 2A. A similar result was obtained via directed-evolution experiments in which a single active-site mutation in MLEII was found to confer the OSBS function. Although neither the wild-type AEE nor MLEII performs the OSBS reaction, enzyme efficiencies for the OSBS reaction of the single-site mutants are sufficient to complement an OSBS knockout strain and show measurable levels of OSBS activity (K_{cat}/K_m (M^{-1} s^{-1}) of 12.5 and 2×10^3 for the AEE and MLEII mutants, respectively). We suggest here that choosing a starting template for protein engineering that already "knows" how to perform a fundamental aspect of a new chemical reaction of interest may increase the success of protein design strategies in general. Analysis of the structure−function relationships among members of the enolase and other mechanistically diverse superfamilies thus can provide guidance for such protein engineering experiments and aid in our efforts to gain a broader understanding of the structural determinants of enzyme specificity. The SFLD stores these relationships in a functional fold superfamily hierarchy that provides users with the tools to leverage this information for their own protein engineering work.

Elucidation of Function: Structural Genomics Targets

The notion that the elucidation of a protein's 3D structure can aid in the determination of the protein's function is one of the principal ideas underlying the vision for the structural genomics initiative (*37*) projects. For many proteins, however, the determination of the structure does not immediately lead to inference of function. Of the 1605 publicly available protein structures determined from the structural genomics initiative (as of June 21, 2005), 590 are annotated as having unknown functions. Using the HMMs and curated alignments of the SFLD, we examined all 1605 structures and compared our functional predictions for these proteins to their annotations in the Protein Data Bank (pdb) (*38*). We were also able to predict some functional properties for the structures whose functions are annotated as unknown. Table 1 shows all of the structural genomics initiative targets that matched at least one hidden Markov model in the SFLD and the level(s) of granularity (superfamily, subgroup, or family) at which the target's function can be described. For each target that matched a family HMM (families in the SFLD represent isofunctional groups of enzymes), Table 1 also gives the fraction of conserved functional residues that align between the target and the curated family alignment in the SFLD.

Target 1HZD. The functional predictions made using the HMMs in the SFLD agree with a majority of the annotations provided by the pdb, with some notable differences. Target 1HZD, although annotated as an RNA-binding homologue of enoyl-CoA hydratase by the pdb, matches the methylglutaconyl-CoA hydratase family in the crotonase (also called the enoyl-CoA hydratase) superfamily of the SFLD. Although the literature suggests that the methylglutaconyl-CoA hydratase function is the most biologically relevant (loss of this function via mutation in the human gene causes the disorder 3-methylglutaconic aciduria type I (*39*)), this particular protein has been experimentally determined to perform at least three functions: RNA binding, hydration of enoyl-CoA, and hydration of methylglutaconyl-CoA (*40*,

Table 1: Structures Solved by the Structural Genomics Initiative that Match Hidden Markov Models of the SFLD

pdb	pdb annotation	superfamily	subgroup	family	CFR[a]
1j6o	Tatd-related deoxyribonuclease	amidohydrolase	uncharacterized-147		
1j6p	metal-dependent hydrolase of cytosinedemaniase chlorohydrolase family	amidohydrolase	uncharacterized-66		
1kcx	collapsin response mediator protein 1	amidohydrolase	collapsin response mediator	D-hydantoinase[b]	1/6
1o12	*N*-acetylglucosamine-6-phosphate deacetylase	amidohydrolase	*N*-acetylglucosamine-6-phosphate	*N*-acetylglucosamine-6-phosphate deacetylase	5/5
1xwy	Tatd deoxyribonuclease	amidohydrolase	TatD_MttC		
1yix	Tatd homolog, hydrolase	amidohydrolase	uncharacterized-147		
1ymy	*N*-acetylglucosamine-6-phosphate deacetylase	amidohydrolase	*N*-acetylglucosamine-6-phosphate	*N*-acetylglucosamine-6-phosphate deacetylase	5/5
1hzd	RNA-binding homologue of enoyl-CoA hydratase	crotonase		methylglutaconyl-CoA hydratase	7/7
1rjn	MenB−naphthoate synthase	crotonase		1,4-dihydroxy-2-naphthoyl-CoA synthase	4/4
1uiy	enoyl-CoA hydratase	crotonase		enoyl-CoA hydratase[b]	3/4
1rvk	hypothetical protein, unknown function	enolase	mandelate racemase		
1tzz	unknown member of enolase superfamily	enolase	mandelate racemase		
1wue	unknown member of enolase superfamily	enolase	muconate cycloisomerase	*o*-succinylbenzoate synthase	5/5
1wuf	member of enolase superfamily, unknown function	enolase	muconate cycloisomerase	*o*-succinylbenzoate synthase	5/5
1yey	L-fuconate dehydratase	enolase	mandelate racemase	L-fuconate dehydratase	6/6
1k1e	deoxy-D-mannose -octulosonate 8-phosphate phosphatase	HAD	phosphatase-like2	deoxy-D-mannose-octulosonate 8-phosphate phosphatase	6/6
1l7p	phosphoserine phosphatase	HAD	phosphatase-like2	phosphoserine phosphatase	6/6
1pw5	putative Nagd protein	HAD	phosphatase-like4		
1te2	putative phosphatase	HAD	phosphatase-like1		
1vjr	4-nitrophenylphosphatase	HAD	phosphatase-like4		
1wvi	putative phosphatase	HAD	phosphatase-like4		
1xvi	putative mannosyl-3-phosphoglycerate phosphatase	HAD	phosphatase-like3	mannosyl-3-phosphoglycerate phosphatase	4/4
1ydf	hydrolase, haloacid dehalogenase-like family	HAD	phosphatase-like4		
1ys9	hypothetical protein, unknown function	HAD	phosphatase-like4		
1k4n	unknown function	VOC	YecM-like		
1zsw	metallo protein from glyoxalase family, unknown function	VOC	2,6-dichlorohydroquinone dioxygenase		

[a] CFR: the fraction of conserved active-site residues present when aligned to curated family alignments of the SFLD. [b] Although these protein sequences match a family HMM in the SFLD, the fact that they are missing at least one functionally important residue suggests that they do not perform the designated family reaction.

41). This target represents a good example of the difficulties in assigning a function to multifunctional enzymes. The SFLD adds additional knowledge to the annotation of 1HZD in the pdb, that is, the assignment of an additional specific enzymatic function.

Targets 1RVK and 1TZZ. Targets 1RVK and 1TZZ are annotated in the pdb as having unknown function, but they both match very well the mandelate racemase (MR) subgroup of the enolase superfamily in the SFLD. From the coarsest granularity prediction at the superfamily level, we know that both proteins perform proton abstraction on a carbon α to a carboxylate in the unknown substrate. Further, assignment to the MR subgroup specifies proton abstraction machinery using a His-Asp dyad for a substrate of R stereochemistry and proton abstraction machinery usually using a Lys-X-X (usually Lys) for a substrate of S stereochemistry. Thus, even though we cannot assign the specific substrates and overall reactions for either 1RVK or 1TZZ, the SFLD provides additional information not found in the pdb, including a context in which to interpret the existence of conserved active-site residues. Using the conserved structure−function relationships we associate with the MR subgroup, we have, along with our collaborators, made further predictions of overall function that are currently being tested experimentally by our collaborators.

Targets 1WUE and 1WUF. Targets 1WUE and 1WUF, both annotated in the pdb as having unknown function, match the *o*-succinylbenzoate synthase (OSBS) family of the SFLD. When aligned with the curated family alignment in the SFLD, which indicates the location, type, and function of the conserved residues required to perform the OSBS reaction, all five (out of five) residues are conserved in both targets. An examination of the genomic context of these targets adds confidence to these functional predictions. The gene encoding 1WUE (in *Enterococcus faecalis*) lies within an operon encoding five other genes of the menaquinone synthesis pathway in which the OSBS reaction is the third (out of seven) step (*42*). Whereas the gene encoding 1WUF (*Listeria innocua*) does not appear to be localized within an operon, the other six genes of the menaquinone synthesis pathway are present in the organism. We note that a BLASTP (*43*) search using each of these sequences shows that the corresponding gi numbers at the NCBI GenBank database (*44*) for each of these structures annotates these sequences as similar to an OSBS (gi 16804558, 1WUE) or as an OSBS homologue (gi 25514206, 1WUF). The predictions that both of these enzymes perform the OSBS reaction has recently been confirmed experimentally (Yew, W. S., and Gerlt, J. A., personal communication).

Target 1ZSW. Another target of unknown function 1ZSW can be classified as belonging to the 2,6-dichlorohydroquinone dioxygenase subgroup of the vicinal oxygen chelate (VOC) fold superfamily on the basis of its match to this subgroup HMM from the SFLD. The pdb describes 1ZSW as a glyoxalase family protein of unknown function. In the SFLD, the VOC superfamily is represented by seven subgroups, together representing three groups of broadly different functionality: dioxygenases, antibiotic resistance proteins such as the fosfomycin resistance protein, and glyoxalases. We used the analysis tools of the SFLD to align 1ZSW to the curated alignment of the 2,6-dichlorohydroquinone dioxygenase subgroup and found it to contain all residues previously inferred to be important to the only characterized member of this subgroup, 2,6-dichlorohydroquinone dioxygenase (*45*). Using the Dali server (*46*), comparison of the 1ZSW structure to other structures in the pdb shows that its closest structural match is 1MPY, a previously characterized catechol 2,3-dioxygenase (*47*).

Some details of the relationship between 2,6-dichlorohydroquinone dioxygenase and the canonical catechol 2,3-dioxygenases were explored previously by Xu et al., who pointed out salient differences between the two. First, as described by Armstrong (*48*) and shown in Figure 3A, although the canonical extradiol dioxygenases can chelate Fe^{2+} by virtue of the ortho arrangement of their alcoholic groups, 2,6-dichlorohydroquinone dioxygenase cannot because its −OH groups are para to each other on the aromatic ring of the substrate. The question of how this para substrate interacts with the metal in the active site has remained one of the most interesting and unresolved questions regarding the mechanism of this enzyme. Second, the family of sequences represented by 2,6-dichlorohydroquinone dioxygenase shows altered patterns of subdomain organization relative to the canonical extradiol dioxygenases. Such variations in subdomain organization in the 3D structures of VOC superfamily proteins is common (*49*) and represents a special architectural feature of this superfamily that contributes to the use of this scaffold for the evolution of new functions (*48*).

In addition to providing a higher granularity annotation for this protein than is currently available, assignment of 1ZSW to the 2,6-dichlorohydroquinone dioxygenase subgroup provides new clues for answering questions about structure−function relationships for the entire subgroup, particularly in the context of its interesting subdomain organization. Indeed, the analysis of 1ZSW provided here supports the hypothesis previously advanced by Xu et al. (*44*) that the 2,6-dichlorohydroquinone dioxygenase subgroup represents a new variant in the VOC superfamily. We can now use 1ZSW to elaborate this hypothesis further.

As described above, because the organization of their subdomains differs, it is not straightforward to superpose 1ZSW with canonical extradiol dioxygenase structures. However, using the structure superposition algorithm MultiProt (*50*), which can perform superpositions independent of residue sequence order, we were able to obtain a good overall alignment of 1ZSW with the extradiol dioxygenase structure 1MPY using default parameters. This structural alignment results from the permutation of the N-terminal subdomain of 1ZSW relative to 1MPY. As shown in Figure 3B, residues 2−68 of 1ZSW are permuted such that they

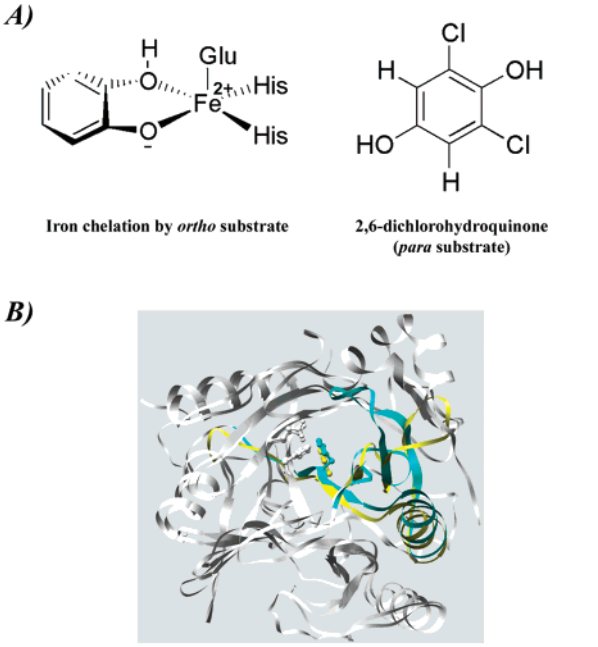

FIGURE 3: Comparison of 2,6-dichlorohydroquinone dioxygenase with canonical extradiol dioxygenases. (A) Left: general form of chelation interations in canonical extradiol dioxygenases between ligand and metal. Right: 2,6-dichlorohydroquinone, the substrate for 2,6-dichlorohydroquinone dioxygenase, showing the para relationship of the −OH groups. (B) Superposition of 1ZSW with 1MPY generated via permutation using MultiProt. Structures are shown in white except for the region showing the permuted subdomains. Cyan: residues 2−68 of 1ZSW; yellow: residues 142−207 of 1MPY. Side chains are shown for reference and described in C. (C) Close-up of the active-site region shown in B. Labeled residues are 1MPY, metal ligands H153 (yellow), H214, and E265; 1ZSW, corresponding residues predicted to be metal ligands, H222, and E270 (*44*). Residue H9 (cyan) is conserved in the 2,6-dichlorohydroquinone dioxygenase subgroup. For ease in viewing, some of the C-terminal residues are hidden only in part C.

align with residues 142−207 of 1MPY (and region 153−212 of 1ZSW is permuted to align with region 4−62 of 1MPY), with all other segments of the alignment following sequence-dependent ordering.

This analysis allows the identification of candidates for at least two of the metal binding ligands. As illustrated in the close-up of the active-site region shown in Figure 3C, H222 and E270 of 1ZSW align well with their counterparts in 1MPY, H214 and E265, respectively, suggesting that they may play similar roles. Interestingly, H9 in 1ZSW, a histidine highly conserved in the N-terminal subdomain of all members of the 2,6-dichlorohydroquinone dioxygenase sub-

group, aligns well with the third metal-binding ligand in canonical extradiol dioxygenases (H153 in Figure 3C). In the absence of a structure of 2,6-dichlorohydroquinone dioxygenase with a biologically relevant ligand bound to it, it is difficult to infer from the structurally permuted alignment the possible role of this histidine side chain in 1ZSW or in any other members of this subgroup. We expect, however, that a more detailed examination of 1ZSW may lead to useful hypotheses that may help in answering these questions.

Targets 1YS9 and 1K4N. Two of the structural genomics initiative targets with unknown function, 1YS9 and 1K4N, did not match a specific family HMM of the SFLD but matched superfamily and subgroup HMMs in the haloacid dehalogenase (HAD) (*51, 52*) and vicinal oxygen chelate (VOC) (*48*) superfamilies, respectively. As a result, we are unable to make a prediction of the overall reactions catalyzed by either of these enzymes. But we can predict with some confidence that the reactions will share the common mechanistic steps conserved across their respective subgroup and superfamily, that is, form covalent enzyme−substrate intermediates via a conserved active-site aspartic acid, thus facilitating cleavage of the relevant bonds in the different subgroups of the HAD superfamily and the direct electrophilic participation of a metal ion in catalysis for members of the VOC superfamily.

Target 1UIY. Target 1UIY, although aligning well with the enoyl-CoA hydratase family in the SFLD and annotated in the pdb as an enoyl-CoA hydratase, is missing a critical glutamic acid residue required for catalysis (*53*). For this target, use of the explicit structure−function information stored in the SFLD allows us to recognize a potential misannotation that could result from the simple use of overall sequence and structural similarity for the annotation.

Target 1KCX. Target 1KCX matched the D-hydantoinase family HMM of the SFLD; however, in an alignment with this family, 1KCX lacks five out of the six residues critical in performing the D-hydantoinase reaction. It has recently been determined experimentally that 1KCX functions as a collapsin response mediator protein, a nonenzymatic function (*54*). This example highlights the difficulty of using overall sequence or structure similarity alone in the inference of enzyme function, especially when the protein in question may be a descendent from an enzyme ancestor but has lost its catalytic abilities. The SFLD provides added information regarding the positions and natures of key functional residues that, as with 1KCX and 1UIY, helps us avoid annotating a protein with a function it cannot perform.

Determination of enzyme function from structure remains a difficult task. However, in cases where an enzyme has diverged within a functional fold superfamily, the SFLD can aid in predicting at least the capabilities of the enzyme at the superfamily level, the coarsest level of granularity in the SFLD hierarchy. Even when substantial divergence has led to big differences in the overall substrate, product, and reactions among related enzymes, when enough information is available, we can often recognize key structural features that are responsible for delivering the specific mechanistic aspect of chemistry common to the superfamily. Thus, despite the relatively small slice of the enzyme universe currently represented in the SFLD, we have been able to place several functionally uncharacterized proteins from the structural genomics initiative into functional fold superfamilies. For some of these unknowns, we can provide more precise annotation as well, specifying additional functional properties common to a specific subgroup or even the overall function by assignment to a specific family.

Organization of the Structure−Function Linkage Database. The SFLD is organized around a functional fold definition of a mechanistically diverse enzyme superfamily. All enzymes within a superfamily share both a common fold and a distinct mechanistic step performed by key conserved residues (*6*). The SFLD represents the first attempt to consolidate information about multiple mechanistically diverse enzyme superfamilies and include the specific structure−function relationships conserved throughout each superfamily.

Formally, the hierarchy of the SFLD consists of four levels: superfamily, subgroup, family, and enzyme functional domain. Each superfamily contains homologous enzymes that share a common fold and a mechanistic step performed by residues conserved throughout the superfamily. Subgroups contain enzymes that share additional reaction steps or use additional unique sets of residues or residue positions to perform the step conserved by the superfamily. For example, in the terpene cyclase superfamily, all enzymes utilize a conserved mechanistic step resulting in the formation of a carbocation intermediate (*55*). Within the sesquiterpene subgroup, the same substrate (farnesyl diphosphate) is used by each member of the subgroup. (Subgroups within the terpene cyclase superfamily are defined by the number of the isoprenyl units in their substrates.) Alternatively, in subgroups of the amidohydrolase superfamily (*56*), somewhat varied constellations of metal binding ligands can be distinguished and associated with specific groups of functionally distinct enzymes. Enzymes predicted to perform exactly the same overall function, utilizing an identical mechanism, are grouped at the family level (e.g., all aristolochene synthases). Enzyme functional domains are defined in the SFLD as the functional domains of individual enzymatic proteins. For the enolase superfamily, for example, the enzyme functional domain includes both the N-terminal and C-terminal domains because both are required for enzymatic function. This is in contrast to the representation of this superfamily in SCOP (*17*) or CATH (*19*), both of which represent the N-terminal and C-terminal domains separately, consistent with the design of those structure-based hierarchies. Figure 4 shows an example of the three major levels of the SFLD hierarchy with which an individual enzyme functional domain can be associated.

Contents of the Structure−Function Linkage Database. The SFLD currently contains six superfamilies: amidohydrolase, crotonase, enolase, haloacid dehalogenase, terpene cyclase, and vicinal oxygen chelate. These comprise roughly 6000 enzymes, 302 crystal structures, and over 140 different overall reactions. Although this is likely a small proportion of the mechanistically diverse enzyme superfamilies that exist, these are well studied, with clear structure−function relationships determined by experiment. They include some of the most divergent superfamilies described to date, each of which displays the remarkable ability to perform a wide variety of overall chemical reactions while utilizing a common scaffold and conserved mechanistic step. As such, the SFLD serves as a valuable repository of information for many of the best characterized of these special types of

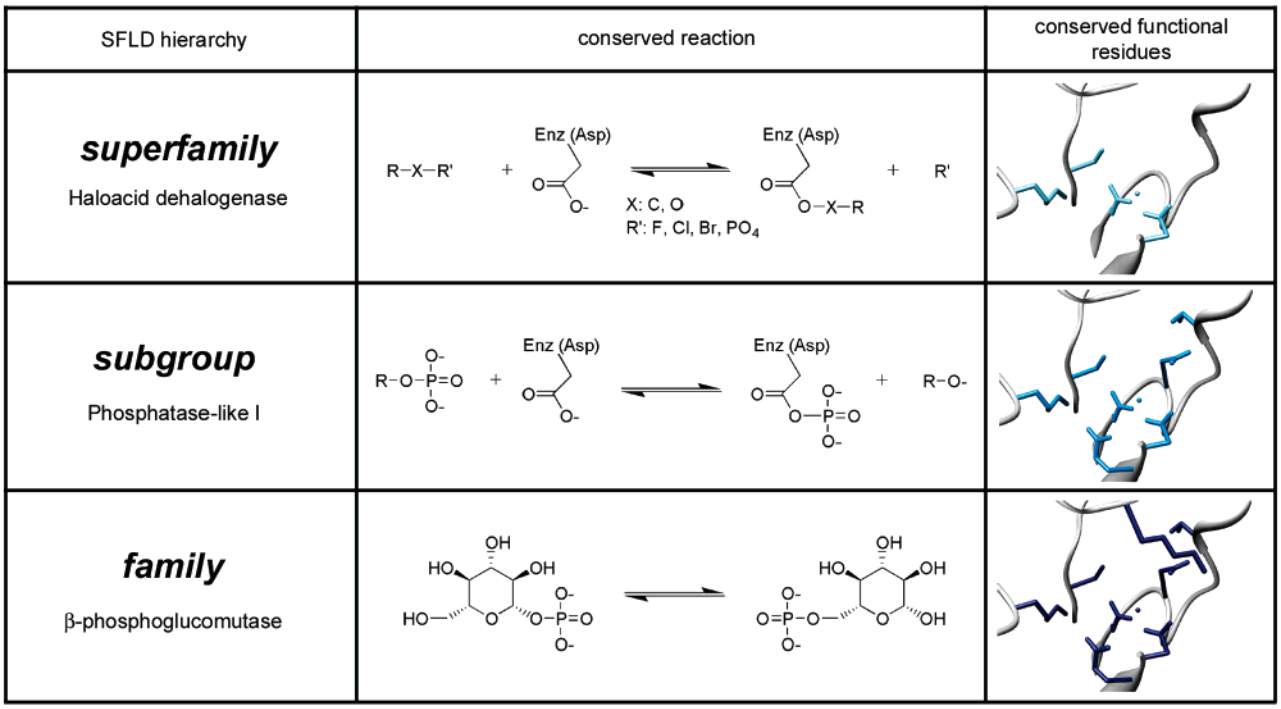

FIGURE 4: Example of the SFLD hierarchy. This example shows the β-phosphoglucomutase family, which belongs to the phosphatase-like I subgroup, which in turn belongs to the haloacid dehalogenase superfamily. The middle column shows the conserved reaction across all members of the hierarchical level (row), and the rightmost column shows the active-site residues conserved at each level.

superfamilies. As we have described here, this information can be leveraged to gain valuable biological insights that would be difficult to obtain in the absence of such a resource.

Where available, X-ray crystal structures are stored in the SFLD along with information regarding the specific residues involved in performing enzyme chemistry and their individual roles. For many enzymes that have not been structurally characterized experimentally, links to modeled structures from ModBase (*57*) are provided. The structures can be downloaded directly from the database in pdb format. Via a single hyperlink, they can also be visualized with their conserved active-site residues highlighted using the Chimera program (*58*). Chimera can also be used to simultaneously view these structures and multiple sequence alignments in which functionally important residues are highlighted, allowing a flexible method of visualizing the links between protein structure and function.[3]

The decisions made by curators of the SFLD regarding such issues as the superfamily assignment of a new enzyme sequence, the partial reaction(s) performed by a given enzyme, or the specific role an individual active-site residue plays in a given reaction are not all made at equal levels of certainty. Some enzymes are well characterized experimentally, whereas others have not been experimentally characterized at all, so that the evidence for a functional assignment rests only on sequence similarity to a characterized enzyme. The SFLD uses a set of evidence codes similar to those used by the Gene Ontology Consortium (*59*) to provide users with information about how annotation decisions were made and the levels of uncertainty involved. In addition, nearly all tables of the SFLD contain metadata fields consisting of free text written by database curators, intended to augment the evidence codes. For each superfamily in the SFLD, an experimental collaborator, knowledgeable about specific enzymes in the superfamily, has provided expert aid in curation.

Searching the Structure−Function Linkage Database. Users can query the database using a variety of methods, all via a web-based interface, at http://sfld.rbvi.ucsf.edu. Users can enter a protein sequence, which is then matched to pregenerated HMMs representing the superfamilies, subgroups, and families in the SFLD. The resulting matches (with their expectation values) are displayed, along with a hyperlink for each match that leads to a dynamically generated alignment of the query sequence to the multiple sequence alignment used to construct the HMM. As described above, this has proved useful in functional characterization of newly discovered proteins. The alignment highlights the conserved superfamily/family residues that participate in enzymatic function, and links are provided to literature references for the experiments by which the structure−function relationship was determined.

Reflecting the organizational concepts describing enzyme superfamilies following the chemistry-constrained evolution model, the SFLD can also be searched by specifying a reaction or partial reaction, or by specifying the structure (or substructure) of a substrate or product. Searches are performed using the flexible SMILES/SMARTS patterns (*60*). This feature is most important for revealing the reaction capabilities of a superfamily, as well as identifying potential protein and ligand structural templates to be used in the engineering of new functions. Reactions in the SFLD can also be searched via keyword, gi number, or Enzyme Commission (E. C.) number, providing one has been assigned.

[3] Chimera must be installed on a user's computer to access these functionalities. It can be obtained for use on many different platforms from http://www.cgl.ucsf.edu/chimera.

Finally, users can easily navigate through the SFLD's hierarchy representing superfamilies, subgroups, and families as well as browse lists of all enzymes and reactions in the database. The names and descriptions of all enzymes, superfamilies, subgroups, families, and reactions also can be searched using keywords.

DISCUSSION

The cataloging of enzyme reactions is clearly not a novel concept. The most common methodology in current use comes from the Enzyme Nomenclature Commission, which assigns a unique set of numbers to each overall enzymatic reaction based on a curated hierarchy. Although such classifications are useful for analyses at higher levels (e.g., pathway analysis), annotation by E. C. numbers provides at most only rudimentary (and often misleading) information about how enzyme sequence, structure, and function relate. Because the E. C. hierarchy was conceived before sequence and structure information was available on a large scale, it fails to capture structure−function mappings at an appropriate level of granularity. Thus, the E. C. system, because it describes only overall reactions, fails to account for the conserved aspects of chemistry shared among evolutionarily related enzymes. As a result, especially for mechanistically diverse superfamilies, it often classifies proteins within a structurally related superfamily as being functionally dissimilar (*61*).

Databases of related proteins and HMMs are also not a novel idea. Generally, these databases consider only sequence (or structural) similarity information and provide annotation at a coarser level of granularity than that provided by the SFLD. Figure 5 shows the classification of every sequence in the enolase superfamily by both the SFLD and PFAM (*17*). PFAM classifies the vast majority of the sequences into just a few large groups, whereas the SFLD attempts to place those same sequences into many smaller subsets of sequences, providing a more detailed and precise representation of isofunctional groups (families). It annotates those that it cannot place into a family as simply belonging to the subgroup or superfamily. In contrast, PFAM classifications are less hierarchical than SFLD classifications and contain no explicit linkage between protein structure and function.

A full understanding of enzyme structure−function relationships requires a mapping of specific structural features to specific aspects of chemical mechanisms. Grouping enzymes that share conserved structural features that perform a common aspect of a chemical mechanism is a step in this direction. It lets us observe how the overall functions have diverged and which structural elements may be responsible for the less conserved aspects of the reactions, as well as identify those responsible for shared chemical capabilities at the subgroup or superfamily level. The SFLD provides not only structure−function information but is organized around a functional fold superfamily paradigm that uses precisely this grouping. As described here, this organization provides the deductive and predictive capabilities to leverage structure−function information in new ways that can lead to useful biological insights.

The SFLD is still in its infancy and will grow as more superfamilies are curated. Despite representing only a small fraction of the likely number of enzyme superfamilies, the

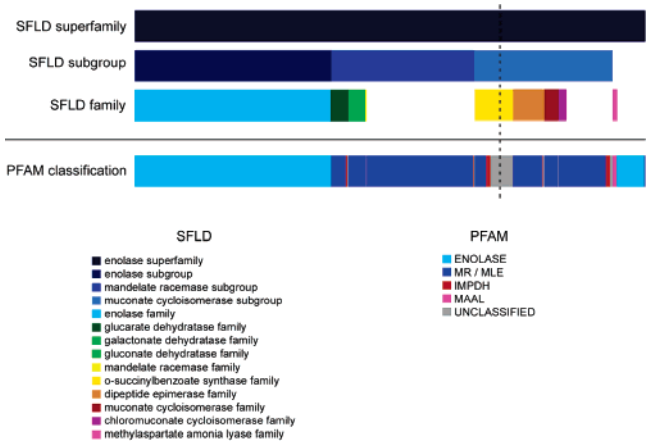

FIGURE 5: Comparison of sequence classifications between the SFLD and PFAM, using all of the sequences in the SFLD superfamily enolase. Each horizontal position represents a single sequence, with the classification of the sequence represented vertically. For example, the dotted line represents a sequence that is classified in the SFLD as belonging to the enolase superfamily, the muconate cycloisomerase subgroup, and the *o*-succinylbenzoate synthase family and is not classified by PFAM. Blank areas mean that a sequence is not classified at the given level in the SFLD. For example, there are many members of the SFLD subgroup mandelate racemase that have no SFLD family classification. The use of such a hierarchical classification scheme helps avoid overprediction of enzyme function. PFAM classifications were obtained by matching sequences from the enolase superfamily of the SFLD to the PFAM hidden Markov models (HMMs), using the gathering cutoff the PFAM curators have established for each HMM. Note that some sequences in the SFLD's enolase superfamily are divergent enough to have not matched any PFAM HMM at a significant score and are thus labeled UNCLASSIFIED. The function represented by the PFAM family IMPDH (IMP dehydrogenase, E. C. 1.1.1.205) does not share the partial reaction conserved throughout the enolase superfamily (abstraction of a proton alpha to a carboxylate), and is therefore a misannotation.

SFLD has already proven useful in studying enzyme structure−function relationships in the special class of mechanistically diverse superfamilies. Several more superfamilies, most notably the *N*-acetyl neuraminate lyase and guanidine kinase superfamilies, are in the process of being added to the SFLD. Additional large superfamilies, particularly those from the $(\beta/\alpha)_8$ barrel and thioredoxin fold classes, are in the process of curation. Along with these data sets, new methods of browsing and searching the data and new methodologies for representing specific structure−function relationships in a computationally useful form are under development. The SFLD is freely accessible at http://sfld.rbvi.ucsf.edu.

ACKNOWLEDGMENT

We thank John Gerlt for expert advice on the enolase and crotonase superfamilies, Richard Armstrong for advice on the VOC superfamily, Frank Raushel for advice on the amidohydrolase superfamily, and Kinkead Reiling for his advice on the terpene cyclase superfamily.

REFERENCES

1. Roberts, R. J. (2004) Identifying protein function—a call for community action, *PLoS Biol. 2*, E42.
2. Horowitz, N. H. (1945) On the evolution of biochemical syntheses, *Proc. Natl. Acad. Sci. U.S.A. 31*, 153−157.
3. Horowitz, N. H. (1965) in *Evolving Genes and Proteins* (Bryson, V., and Vogel, H. J., Eds.) pp 15−, Academic Press, New York.

4. Hyde, C. C., and Miles, E. W. (1990) The tryptophan synthase multienzyme complex: exploring structure-function relationships with X-ray crystallography and mutagenesis, *Biotechnology 8*, 27−32.
5. Jensen, R. A. (1976) Enzyme recruitment in evolution of new function, *Annu. Rev. Microbiol. 30*, 409−25.
6. 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.
7. 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.
8. Gerlt, J. A., and Raushel, F. M. (2003) Evolution of function in $(\beta/\alpha)_8$-barrel enzymes, *Curr. Opin. Chem. Biol. 7*, 254−264.
9. 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.
10. 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.
11. Alves, R., Chaleil, R. A., and Sternberg, M. J. (2002) Evolution of enzymes in metabolism: a network perspective, *J. Mol. Biol. 320*, 751−70.
12. 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.
13. 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.
14. 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.
15. Saghatelian, A., and Cravatt, B. (2005) Assignment of protein function in the postgenomic era, *Nat. Chem. Biol. 1*, 130−142.
16. Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure, *J. Mol. Biol. 313*, 903−919.
17. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004) The Pfam protein families database, *Nucleic Acids Res. 32*, D138−141.
18. Eddy, S. R. (1998) Profile hidden Markov models, *Bioinformatics 14*, 755−763.
19. 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.
20. 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*, D226−229.
21. 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*, D247−251.
22. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004) The KEGG resource for deciphering the genome, *Nucleic Acids Res. 32*, D277−280.
23. Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Paley, S. M., and Pellegrini-Toole, A. (2000) The EcoCyc and MetaCyc databases, *Nucleic Acids Res. 28*, 56−69.
24. Krieger, C. J., Zhang, P., Mueller, L. A., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S. Y., and Karp, P. D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes, *Nucleic Acids Res. 32*, D438−442.
25. Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D. (2004) BRENDA, the enzyme database: updates and major new developments, *Nucleic Acids Res. 32*, D431−433.
26. Webb, E. C. (1993) Enzyme nomenclature: a personal retrospective, *FASEB J. 7*, 1192−1194.
27. Dodson, G., and Wlodawer, A. (1998) Catalytic triads and their relatives, *Trends Biochem. Sci. 23*, 347−352.
28. Porter, C. T., Bartlett, G. J., and Thornton, J. M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data, *Nucleic Acids Res. 32*, D129−133.
29. George, R. A., Spriggs, R. V., Bartlett, G. J., Gutteridge, A., Macarthur, M. W., Porter, C. T., Al-Lazikani, B., Thornton, J. M., and Swindells, M. B. (2005) Effective function annotation through catalytic residue conservation, *Proc. Natl. Acad. Sci. U.S.A.*
30. Nagano, N. (2005) EzCatDB: the Enzyme Catalytic-mechanism Database, *Nucleic Acids Res. 33*, D407−412.
31. Holliday, G. L., Bartlett, G. J., Almonacid, D. E., O'Boyle N. M., Murray-Rust, P., Thornton, J. M., and Mitchell, J. B. (2005) MACiE: a database of enzyme reaction mechanisms, *Bioinformatics*.
32. 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. 2005*, 358−369.
33. 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 (beta/alpha)8-barrels: functional promiscuity produced by single substitutions in the enolase superfamily, *Biochemistry 42*, 8387−8393.
34. Brown, S., Gerlt, J. A., Seffernick, J., and Babbitt, P. C., in press. The gold standard set of mechanistically diverse enzyme superfamilies, *GenomeBiology*.
35. Devos, D., and Valencia, A. (2001) Intrinsic errors in genome annotation, *Trends Genet. 17*, 429−431.
36. Brenner, S. E. (1999) Errors in genome annotation, *Trends Genet. 15*, 132−133.
37. Chance, M. R., Bresnick, A. R., Burley, S. K., Jiang, J. S., Lima, C. D., Sali, A., Almo, S. C., Bonanno, J. B., Buglino, J. A., Boulton, S., Chen, H., Eswar, N., He, G., Huang, R., Ilyin, V., McMahan, L., Pieper, U., Ray, S., Vidal, M., and Wang, L. K. (2002) Structural genomics: a pipeline for providing structures for the biologist, *Protein Sci. 11*, 723−738.
38. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The protein data bank, *Nucleic Acids Res. 28*, 235−242.
39. Ijlst, L., Loupatty, F. J., Ruiter, J. P., Duran, M., Lehnert, W., and Wanders, R. J. (2002) 3-Methylglutaconic aciduria type I is caused by mutations in AUH, *Am. J. Hum. Genet. 71*, 1463−1466.
40. Nakagawa, J., Waldner, H., Meyer-Monard, S., Hofsteenge, J., Jeno, P., and Moroni, C. (1995) AUH, a gene encoding an AU-specific RNA binding protein with intrinsic enoyl-CoA hydratase activity, *Proc. Natl. Acad. Sci. U.S.A. 92*, 2051−2055.
41. Kurimoto, K., Fukai, S., Nureki, O., Muto, Y., and Yokoyama, S. (2001) Crystal structure of human AUH protein, a single-stranded RNA binding homolog of enoyl-CoA hydratase, *Structure (Cambridge, MA, U.S.) 9*, 1253−1263.
42. Meganathan, R. (2001) Biosynthesis of menaquinone (vitamin K2) and ubiquinone (coenzyme Q): a perspective on enzymatic mechanisms, *Vitam. Horm. 61*, 173−218.
43. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res. 25*, 3389−3402.
44. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2004) GenBank: update, *Nucleic Acids Res. 32*, D23−26.
45. Xu, L., Resing, K., Lawson, S. L., Babbitt, P. C., and Copley, S. D. (1999) Evidence that pcpA encodes 2,6-dichlorohydroquinone dioxygenase, the ring cleavage enzyme required for pentachlorophenol degradation in *Sphingomonas chlorophenolica* strain ATCC 39723, *Biochemistry 38*, 7659−7669.
46. Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M., and Holm, L. (2001) A fully automatic evolutionary classification of protein folds: Dali domain dictionary version 3, *Nucleic Acids Res. 29*, 55−57.
47. Kita, A., Kita, S., Fujisawa, I., Inaka, K., Ishida, T., Horiike, K., Nozaki, M., and Miki, K. (1999) An archetypical extradiol-

cleaving catecholic dioxygenase: the crystal structure of catechol 2,3-dioxygenase (metapyrocatechase) from *Ppseudomonas putida* mt-2, *Structure Fold Des. 7*, 25−34.
48. Armstrong, R. N. (2000) Mechanistic diversity in a metalloenzyme superfamily, *Biochemistry 39*, 13625−13632.
49. Bergdoll, M., Eltis, L. D., Cameron, A. D., Dumas, P., and Bolin, J. T. (1998) All in the family: structural and evolutionary relationships among three modular proteins with diverse functions and variable assembly, *Protein Sci. 7*, 1661−1670.
50. Shatsky, M., Nussinov, R., and Wolfson, H. J. (2004) A method for simultaneous alignment of multiple protein structures, *Proteins: Struct., Funct., Genet. 56*, 143−156.
51. Koonin, E. V., and Tatusov, R. L. (1994) Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search, *J. Mol. Biol. 244*, 125−132.
52. Allen, K. N., and Dunaway-Mariano, D. (2004) Phosphoryl group transfer: evolution of a catalytic scaffold, *Trends Biochem. Sci. 29*, 495−503.
53. Bahnson, B. J., Anderson, V. E., and Petsko, G. A. (2002) Structural mechanism of enoyl-CoA hydratase: three atoms from a single water are added in either an E1cb stepwise or concerted fashion, *Biochemistry 41*, 2621−2629.
54. Deo, R. C., Schmidt, E. F., Elhabazi, A., Togashi, H., Burley, S. K., and Strittmatter, S. M. (2004) Structural bases for CRMP function in plexin-dependent semaphorin3A signaling, *EMBO J. 23*, 9−22.
55. Segura, M. J., Jackson, B. E., and Matsuda, S. P. (2003) Mutagenesis approaches to deduce structure−function relationships in terpene synthases, *Nat. Prod. Rep. 20*, 304−317.
56. Seibert, C. M., and Raushel, F. M. (2005) Structural and catalytic diversity within the amidohydrolase superfamily, *Biochemistry 44*, 6383−6391.
57. Pieper, U., Eswar, N., Braberg, H., Madhusudhan, M. S., Davis, F. P., Stuart, A. C., Mirkovic, N., Rossi, A., Marti-Renom, M. A., Fiser, A., Webb, B., Greenblatt, D., Huang, C. C., Ferrin, T. E., and Sali, A. (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources, *Nucleic Acids Res. 32*, D217−222.
58. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera-a visualization system for exploratory research and analysis, *J. Comput. Chem. 25*, 1605−1612.
59. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat. Genet. 25*, 25−29.
60. Weininger, D. J. (1988) SMILES.1. Introduction and encoding rules, *J. Chem. Inf. Comput. Sci. 28*, 31−46.
61. Babbitt, P. C. (2003) Definitions of enzyme function for the structural genomics era, *Curr. Opin. Chem. Biol. 7*, 230−237.

BI052101L