Representing Enzyme Function in Mechanistically Diverse Enzyme Superfamilies

Scott C.-H. Pegg[1], **Shoshana Brown**[1], **Patricia C. Babbitt**[1,2]

[1]Dept. of Biopharmaceutical Sciences and [2]Dept. of Pharmaceutical Chemistry
University of California, San Francisco, 94143, USA

E-Mail: spegg@mako.ucsf.edu

Received: 24th February 2006 / Published: 31st August 2007

Abstract

Computational representation of enzyme function should include the structural elements of enzymes which deliver catalytic ability. This is especially important in mechanistically diverse enzyme superfamilies, whose members catalyze different overall reactions. In such superfamilies, evolutionarily conserved elements of structure can be correlated with only conserved aspects of function. The representation of enzyme function in the Structure-Function Linkage Database, in particular the specific structure-function relationships, at multiple levels of evolutionary conservation, aids in the annotation of enzyme function and in designing enzyme engineering experiments.

Introduction

Computational representations of enzyme function, especially the specific ways in which enzyme structure delivers catalytic function, aids our ability to predict the function of newly sequenced enzymes [1, 2] and in efforts to engineer new functions into existing enzymes. [3] Any such computational representation should have at three main properties. First, it should be rapidly searchable. Second, there should be valid similarity metrics defined between any two reactions, allowing users to identify reactions (or substrates or products) that are similar to other reactions (substrates, products). This ability is especially important in enzyme engineering, where a user desires as a starting structural scaffold an enzyme with a functionality similar to the function being engineered. Third, the specific contributions to function by structural elements of the enzyme (e. g. active site residues) should be represented. This allows users to search for specific mechanistic abilities in potential engineering scaffolds and aids in the annotation of newly sequenced or structurally characterized enzymes.

Several representations of enzyme function are currently available, but they fail to make the explicit connection between enzyme structure and function, especially with regard to how conserved structural elements deliver catalytic abilities. The Enzyme Classification (E.C.) system [4], developed before the wide availability and diversity of crystal structures or enzyme sequences, classifies enzyme function according to the overall reaction catalyzed by an enzyme. While the E.C. representation, a series of four hierarchical numbers, allows for rapid computation and simple similarity functions, it doesn't include the contributions of the enzyme structure. Reactions in the E.C. system are considered independent of enzyme structure, leading to cases where enzymes with very different structure-function relationships are classified as similar and vice versa [5]. Recently developed databases of enzyme reactions such as EzCatDB [6] and MACiE [7] have cataloged a large number of enzyme reactions and, where available, the individual mechanistic steps they're comprised of, including the specific amino acids involved in the reactions. These resources, however, do not provide a representation that has similarity metrics defined upon it, nor do they represent some of the more subtle ways in which enzyme structure can contribute to function (e. g. stabilization of a charged intermediate via backbone dipoles). A computational framework for representing enzyme function in a platform independent manner using XML has recently been proposed [8]. While the motivation behind CMLReact is reasonable, it's unclear how well such a scheme, which remains largely undeveloped, will be able to provide similarity metrics and capture the contributions of enzyme structure. As an extensible scheme, however, there remains the potential for other computational representations of function to be absorbed into the CMLReact format.

We focus here on the computational representation of enzyme function within mechanistically diverse enzyme superfamilies [9]. These superfamilies are sets of homologous enzymes which, while often sharing very little sequence similarity to each other, and often catalyzing different overall reactions with a variety of substrates and products, share the same fold and conserve a specific partial reaction (of some other aspect of mechanism) enabled by a conserved set of residues. Study of these superfamilies, especially their conserved structure-function relationships, provides insights into enzyme evolution and significantly aids enzyme engineering efforts.

Methods

We have created an online resource for the study of mechanistically diverse enzyme superfamilies, the Structure-Function Linkage Database (SFLD) [10, 11]. This database structures enzymes into a three level hierarchy using both structural and functional criteria. At the top (superfamily) level are enzymes that share a common partial reaction step, mediated by conserved elements of structure, while at the bottom (family) level are enzymes that catalyze identical overall reactions, via identical mechanisms, using the same conserved aspects of enzyme structure. The middle (subgroup) level contains sets of enzymes where particular structure-function relationships are shared, and are specific to each superfamily. An example of this hierarchy is shown in Figure 1. The SFLD is a rich resource, containing curated alignments, mechanisms, structures, and sequences of widely divergent enzymes that share conserved structure-function relationships. Most fields are also annotated with evidence codes similar to the Gene Ontology (GO) [12] evidence codes and links to relevant literature references. Due to the time and effort involved in the curation of mechanistically diverse enzyme superfamilies, the SFLD remains a deep resource, containing a wealth of structure-function information about particular superfamilies, as opposed to a broad resource that covers all of enzyme space, although more superfamilies are in the process of being added. The SFLD is freely accessible at http://sfld.rbvi.ucsf.edu.

Computational representation of enzyme function in the SFLD is accomplished primarily through the SMILES/SMARTS [13] representation of small molecules and reactions. Overall reactions are stored as well as their constituent partial reactions. These reactions can be searched rapidly using SMARTS queries, allowing users to search for substructures in substrates and/or products. Figure 2 shows some examples of this type of query. Individual residues involved in delivering function, as well as their specific participation, where known, are stored for every structure, and across all sets of proteins at each level of the SFLD hierarchy. This allows users to quickly align a sequence to a curated alignment and determine from the annotated residue positions if the query sequence is likely to have a similar structure-function relationship.

The value of these capabilities is illustrated by our experiments in annotating structures solved by the Structural Genomics Initiatives (SGI) [14]. We scanned 1,605 structures solved by the SGI using hidden Markov Models (HMMs) [15] built on the curated sequence alignments in the SFLD, and compared their Protein Data Bank (PDB) [16]] annotations to our own predictions of function. Our predictions were made according to the level(s) of the SFLD hierarchy for which a HMM matched an SGI sequence and the fraction of annotated conserved active site residues that were matched in the alignment of the sequence to the curated multiple alignment upon which the HMM was built. In some cases, we were able to make very specific predictions of enzyme function which have been validated experimentally by our collaborators. (Gerlt, JA, unpublished)

Results

Table 1 shows the SGI structures which matched at least one HMM in the SFLD. Targets for which our predictions of function agree with the current PDB annotations have a white background. In green are cases where we were able to increase the knowledge about the target protein, adding some information about the reaction the enzyme is likely to perform. In the case of 1WUE and 1WUF, targets annotated as being of unknown function, we accurately predicted their ability to catalyze the synthesis of o-succinylbenzoate, a function that was subsequently confirmed experimentally [11]. Our analysis was also able to identify target 1UIY as having been misannotated (orange background in Table 1). This target, while aligning well to the enoyl-CoA hydratase family of the SFLD, is missing a critical glutamic acid residue required for catalysis [17].

A key aspect of the organization of enzyme structure-function relationships within the SFLD is that it allows annotation at multiple levels of granularity. In some cases we can make predictions of overall function with some certainty (such as with 1WUF), but in others we can only state that the enzyme performs a partial reaction conserved throughout the subgroup or superfamily (such as with 1RVK). A more comprehensive discussion of our annotation of several of the SGI targets listed in Table 1 has recently been published [11].

Conclusion

The hierarchy of conserved structure-function relationships within an enzyme superfamily helps us not only avoid the overprediction of enzyme function, but also to make guided decisions when performing enzyme engineering. Our representation of enzyme function in the SFLD allows users to rapidly search for similar substrates and products, and through the annotation of functional residues at each level of the SFLD hierarchy to obtain information about how particular aspects of enzyme structure deliver catalytic function. This information can then be used to identify appropriate starting scaffolds [3, 18].

Our current representation of function is somewhat incomplete, however. While rapidly searchable and with adequately defined similarity metrics based upon small molecule chemical similarity, it lacks a formal representation of some aspects of enzyme participation. For example, the terpene synthase superfamily displays a variety of methods of stabilizing the positive charge on the carbocation intermediates of its reactions, including dipole-charge interactions from sidechains and backbone carbonyls, and cation-pi interactions with aromatic sidechains [19]. These aspects are currently stored as text descriptions in a table of conserved residues, a representation that is not amenable to the sort of similarity queries we'd like to make. Ultimately, we desire a representation of enzyme function in which we can quickly answer such queries as, “what are the enzymes that use a backbone carbonyl to stabilize a positive charge?” and “what are the partial reactions in which an lysine acts as a Schiff base?” While such queries can be answered through string matching of the text descriptions of conserved residue function, the results are inconsistent due to the freeform nature of the text field-different curators will describe identical functions in different ways. A more structured representation of the contributions of a particular aspect of an enzyme structure to a given catalytic step is required to accurately answer the sorts of questions posed above.

Our current attempts at such a representation involve development of extensions to the SMILES representation. This allows us to retain some of the major benefits of SMILES, such as its wide acceptance in third party software, which allows us to implement rapid substructure searching and well developed similarity metrics between the chemical structures represented. Work on this extension of SMILES remains an ongoing research project in our laboratory.

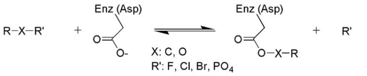
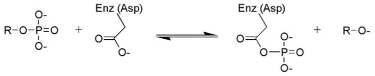
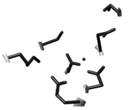
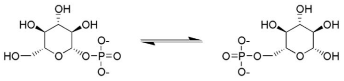
SFLD hierarchy	conserved reaction	conserved functional residues
superfamily Haloacid dehalogenase	 Enz (Asp) R–X–R' + O= O- X: C, O R': F, Cl, Br, PO4 Enz (Asp) O= O–X–R + R'	
subgroup Phosphatase-like I	 O- R–O–P=O O- + Enz (Asp) O= O- Enz (Asp) O= O- O–P=O O- + R–O-	
family β-phosphoglucomutase	 OH HO OH O- HO O O–P=O O- OH HO OH O- O=P–O O- O OH	

Figure 1: An 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 right-most column shows the active site residues conserved at each level.

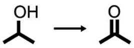
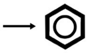
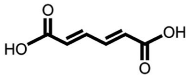
SMARTS Query	Meaning	Schematic
`C([OD1])>>C(=O)`	Reaction contains the conversion of an alcohol to a ketone	
`>>c1ccccc1`	Product of the reaction contains a benzene ring	
`C(=O)([OD1])C=CC=CC(=O)[OD1]`	Either the substrate or the product of the reaction contains mandelate	

Figure 2: Examples of SMARTS queries and their chemical meanings.

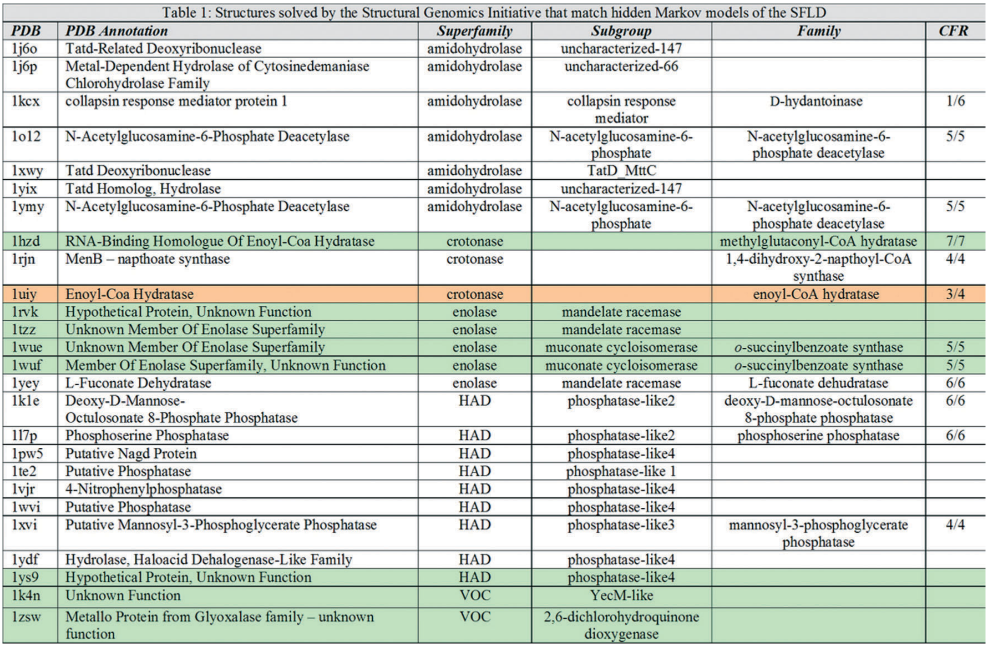


Table 1: Structures solved by the Structural Genomics Initiative that match hidden Markov models of the SFLD

PDB	*PDB Annotation*	*Superfamily*	*Subgroup*	*Family*	*CFR*
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	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 – napthoate synthase	crotonase		1,4-dihydroxy-2-napthoyl-CoA synthase	4/4
1uiy	Enoyl-Coa Hydratase	crotonase		enoyl-CoA hydratase	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 dehudratase	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-like 1		
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		

Table 1: Structures solved by the Structural Genomics Initiative that match hidden Markov models of the SFLD. Targets with a white background have PDB annotations that agree with our annotations using the SFLD. Targets with green backgrounds represent cases in which the SFLD annotations add useful information to the current PDB annotations. Targets with an orange background represent misannotations in the PDB that are corrected by the SFLD annotations. Although targets 1kcx and 1uiy 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. (CFR: Conserved Functional Residue)

Acknowledgements

This work was performed with support from NIH grant GM60595 and NSF grant 0234768 (to PCB).

References

[1] Roberts, R.J. (2004). Identifying protein function–a call for community action. *PLoS Biol* **2**, E42.

[2] Saghatelian, A. & Cravatt, B. (2005). Assignment of protein function in the post-genomic era. *Nat Chem Biol* **1**, 130–142.

[3] Schmidt, D.M., Mundorff, E.C., Dojka, M., Bermudez, E., Ness, J.E., Govindarajan, S., Babbitt, P.C., Minshull, J. & Gerlt, J.A. (2003). Evolutionary potential of (beta/alpha)8-barrels: functional promiscuity produced by single substitutions in the enolase superfamily. *Biochemistry* **42**, 8387–93.

[4] Webb, E.C. (1993). Enzyme nomenclature: a personal retrospective. *Faseb J* **7**, 1192–4.

[5] Babbitt, P.C. (2003). Definitions of enzyme function for the structural genomics era. *Curr Opin Chem Biol* **7**, 230–7.

[6] Nagano, N. (2005). EzCatDB: the Enzyme Catalytic-mechanism Database. *Nucleic Acids Res* **33**, D407–12.

[7] Holliday, G.L., Bartlett, G.J., Almonacid, D.E., O'Boyle N, M., Murray-Rust, P., Thornton, J.M. & Mitchell, J.B. (2005). MACiE: a database of enzyme reaction mechanisms. *Bioinformatics*.

[8] Holliday, G.L., Murray-Rust, P. & Rzepa, H.S. (2006). Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions. *J Chem Inf Model* **46**, 145–57.

[9] Gerlt, J.A. & Babbitt, P.C. (2001). Divergent evolution of enzymatic function: Mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu Rev Biochem* **70**, 209–246.

[10] Pegg, S.C., Brown, S., Ojha, S., Huang, C.C., Ferrin, T.E. & Babbitt, P.C. (2005). Representing structure-function relationships in mechanistically diverse enzyme superfamilies. *Pac Symp Biocomput*, 358–69.

[11] Pegg, S.C., Brown, S., Ojha, S., Seffernick, J., Meng, E.C., Morris, J.H., Chang, P.J., Huang, C.C., Ferrin, T.E., Babbitt, P.C. (2006). Leveraging Enzyme Structure-Function Relationships for Functional Inference and Experimental Design: The Structure-Function Linkage Database. *Biochemistry Epub* ahead of print.

[12] 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. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–9.

[13] Weininger, D.J. (1988). SMILES. 1. Introduction and encoding rules. *Jour. Chem. Inf. Comput. Sci.* **28**, 31–46.

[14] 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. & Wang, L.K. (2002). Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci* **11**, 723–38.

[15] Eddy, S. (1996). Hidden Markov models. *Curr. Op. Struct. Biol.* **6**, 361–365.

[16] Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D. & Zardecki, C. (2002). The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* **58**, 899–907.

[17] Bahnson, B.J., Anderson, V.E. & 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**, 2621–9.

[18] Glasner, M.E., Gerlt, J.A., Babbitt, P.C. (2006). Mechanisms of Protein Evolution and Their Application to Protein Engineering. In Advances in Enzymology and Related Areas of Molecular Biology. Wiley & Sons.

[19] Lesburg, C.A., Zhai, G., Cane, D.E. & Christianson, D.W. (1997). Crystal structure of pentalenene synthase: mechanistic insights on terpenoid cyclization reactions in biology. *Science* **277**, 1820–4.