Prediction of function in protein superfamilies
Shelley D Copley

Address: Department of Molecular, Cellular & Developmental Biology and Cooperative Institute for Research in Environmental Sciences, University of Colorado at Boulder, Campus Box 216, Boulder, CO 80309, USA
Email: shelley@cires.colorado.edu

F1000 Biology Reports 2009, 1:91 (doi:10.3410/B1-91)
This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/3.0/legalcode), which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes provided the original work is properly cited. You may not use this work for commercial purposes.
The electronic version of this article is the complete one and can be found at: http://F1000.com/Reports/Biology/content/1/91

Abstract
Assignment of function for enzymes encoded in sequenced genomes is a challenging task. Predictions of enzyme function can be made using clues from superfamily assignment, structure, genome context, phylogenetic conservation, and virtual screening to identify potential ligands. Ultimately, confident assignment of function requires experimental verification as well as an understanding of the physiological role of an enzyme in the context of the metabolic network.

Introduction and context
Genome sequences are now available for over 900 microbes and 6 multicellular organisms, providing genetic blueprints for organisms that differ enormously in morphology, physiology, and habitat. Unfortunately, our ability to interpret these blueprints is hampered by the lack of assigned function for one-third or more of the proteins in every organism. This commentary will focus specifically on the assignment of enzyme function. Automated assignment of enzyme function is notoriously difficult, as many enzymes with very low sequence identity catalyze the same reaction, and even enzymes that share 98% identity can have different substrate specificities [1].

Efforts to define the roles of enzymes of unknown function often begin with assignment to a superfamily based upon sequence analysis. Enzymes in a superfamily share a common ancestor. In some cases, the ancestral catalytic activity has been retained and divergence has resulted in different substrate specificities. In others, divergence has generated enzymes that catalyze mechanistically distinct reactions, although structural and mechanistic features of the ancestor are conserved.

Superfamily assignment provides clues to enzyme function by indicating the overall fold of the protein, the location of the active site, and the range of known functions found in superfamily members. Further clues can be provided by conserved sequence motifs. Superfamily members generally share conserved motifs that are important for structure or function or both. Families within a superfamily often have additional motifs and/or patterns of distinct residues within motifs that are involved in substrate specificity or family-specific catalytic functions (Figure 1) [2-6]. Our ability to capitalize on such clues is growing as structural and functional studies expand our knowledge of specific superfamilies. The enolase [3], amidohydrolase [7], and haloalkanoic acid dehalogenase [8] superfamilies are the most thoroughly characterized at this point. However, even in these superfamilies, numerous enzymes fall into families for which there is no known function. Furthermore, some superfamilies do not have easily recognized signatures indicating family membership. The hotdog fold superfamily exhibits little or no conservation of catalytic residues and poorly defined substrate-binding pockets, hindering efforts to use sequence and structural information for the prediction of function [9].

Information about potential functions derived from superfamily affiliation can be exploited along with clues from genome context, phylogenetic conservation, and an understanding of microbial physiology to assign enzyme function [10]. A few of many examples of the use of such information include the identification of...
function for o-succinylbenzoate synthase from *Amycolaptosis* sp. [11], 2,6-dichlorohydroquinone dioxygenase from *Sphingobium chlorophenolicum* [12], N-formimino-L-glutamate deiminase from *Pseudomonas aeruginosa* [13], and D-galacturonate isomerase from *Bacillus halodurans* [14]. However, in many cases, these clues are not enough. For example, protein Cg10062 from *Corynebacterium glutamicum* belongs to the tautomerase superfamily. The protein has six active site residues that are conserved in the superfamily and catalyzes three reactions typical of the superfamily at low rates, but its physiological role still cannot be identified [15].

Additional clues to enzyme function can be obtained by screening libraries of potential substrates for activity (e.g., [16-18]). An example is the identification of function of *Bacillus subtilis* BC0371 [19], which belongs to the muconate-lactonizing enzyme subgroup of the enolase superfamily. This enzyme clusters with the L-Ala-D/L-Glu epimerase family, but three residues typical of that family are missing, suggesting that BC0371 has a different function. The enzyme was incubated with a library of L,L-dipeptides, and epimerization was detected by incorporation of deuterium from the solvent into the substrate. Subsequent kinetic analysis using molecules that were substrates showed that values for $k_{cat}/K_M$ were suspiciously low – at best $10^4$ M$^{-1}$ s$^{-1}$. Since N-acyl amino acid racemases are also found in the muconate-lactonizing enzyme subgroup, a second screening was carried out with a library of N-succinyl l-amino acids. N-succinyl-l-Arg was found to be the best substrate; $k_{cat}/K_M$ was $1.4 \times 10^5$ M$^{-1}$ s$^{-1}$, which is well in the range of values seen for physiologically relevant reactions.

**Major recent advances**

Virtual screening has been widely used to identify potential enzyme inhibitors for drug discovery efforts by docking a set of ligands into an active site and predicting binding energies based upon van der Waals and electrostatic interactions and solvation effects. This approach has been adapted in recent years to predict substrates for enzymes of unknown function. The true substrates for enzymes are usually found among the high-scoring hits, though often not at the top of the list [20,21]. The correct substrates for 11 enolase superfamily members were found in the top 1% of 19,000 ligands [18]. The substrates for *Pseudomonas putida* mandelate racemase ranked 77 and 140 for docking to a structure of the enzyme-inhibitor complex. Pinpointing the correct substrate is difficult because docking algorithms predict binding affinity, but not propensity for turnover, which requires correct positioning of the substrate with respect to catalytic groups. Furthermore, approximations are required for the scoring function. Finally, it can be difficult to account for conformational changes in the protein that must occur for ligand binding. The docking algorithm can be adapted to allow some flexibility in the protein, but this is not...
successful for large conformational changes. Thus, the primary value of virtual screening is in providing clues to structural characteristics of the substrate and thereby limiting the number of potential substrates that must be screened experimentally.

The results of virtual screening are often improved by use of a database containing high-energy intermediates, rather than ground-state substrates, because enzymes generally bind more tightly to transition states or high-energy intermediates than to substrates. To explore this approach, Hermann et al. [22] generated 21,000 high-energy forms of 3770 potential substrates for amidohydrolase superfamily enzymes. The docking procedure evaluated up to 1 million poses for each molecule. High-energy intermediates corresponding to the known substrates for seven enzymes were found in the top 100 molecules. For five of the seven enzymes, screening using the ground-state metabolite dataset was considerably less successful. A subsequent study used this approach to predict the function of Thermatoga maritima Tm0936, which belongs to the amidohydrolase superfamily [23]. High-energy intermediates based on 4207 potential substrates for amidohydrolase superfamily members were docked into the active site. Nine of the top ten hits were derivatives of adenine, strongly suggesting that the true substrate is some sort of adenine derivative. Only four adenine derivatives were tested experimentally. Two of these (5-methylthioadenosine and 5-adenosylhomocysteine, which were ranked 5 and 6, respectively) were efficiently deaminated, with values of $k_{\text{cat}}/K_M$ of greater than $10^5/M\cdot s^{-1}$.

A homology model can be used for virtual screening, although the docking algorithm should be modified to allow subtle rearrangements of side chains in the active site that may not be correctly predicted by the homology model. Song et al. [19] generated a homology model for the Bacillus cereus protein BC0371, which had been discovered by experimental screening of potential substrates to be an N-succinyl-l-Arg racemase (discussed in the 'Introduction and context' section). When a set of 420 L, D-dipeptides and N-succinyl L-amino acids was docked, N-succinyl-L-Arg ranked 147 when the homology model was used but was the best hit when a flexible-receptor docking protocol was used. Thus, the flexible-receptor protocol allows the virtual screening approach to be extended to enzymes for which a structure is not available. A recent implementation of this strategy led to the prediction and experimental verification that T. maritima TM0006 is a dipeptide isomerase with an unusual substrate specificity, preferring dipeptides containing Phe, Tyr, or His in the C-terminal position, in contrast to the structurally characterized l-Ala-d/L-Glu epimerase from B. subtilis used as the template for the TM0006 homology model [24].

**Future directions**

Identifying a substrate that is turned over efficiently in vitro is an important step toward understanding the function of an enzyme. Taking this understanding to the next level requires fitting an observed activity into the metabolism of the organism. This is straightforward when an enzyme serves a function in a core catabolic or anabolic pathway, particularly when deletion of the gene encoding the enzyme causes a noticeable phenotype. However, many enzymes for which functions are not known will not play such obvious roles. Some may be involved in unusual metabolic pathways that have not yet been recognized. Others may serve more subtle functions whose discovery requires substantial imagination and understanding of chemistry and metabolism.

An interesting example is Escherichia coli YghZ, a member of the aldo-ketose reductase superfamily [25]. Over-expression of Ygh allows a strain lacking triose phosphate isomerase (TIM) to grow on lactose, although YghZ does not have TIM activity. A possible function suggested by the superfamily assignment and metabolic context was that YghZ might reduce D-glyceraldehyde 3-phosphate (D-GAP) produced by racemization of D-glyceraldehyde 3-phosphate (d-GAP) (Figure 2). Indeed, YghZ reduces L-GAP with a $k_{\text{cat}}/K_M = 4.2 \times 10^4/M\cdot s^{-1}$. The product, L-glycerol 3-phosphate, can be converted to dihydroxyacetone phosphate by glycerol 3-phosphate dehydrogenase. When TIM is present, YghZ may serve to detoxify L-GAP produced by non-enzymatic racemization of D-GAP.

A particular challenge is understanding the physiological roles of enzymes that have broad substrate specificity. For example, Haemophilus influenza YciA hydrolyzes a wide range of CoA thioesters with high efficiency (values of $k_{\text{cat}}/K_M$ for 14 substrates are greater than $10^5/M\cdot s^{-1}$) [8]. Similarly, E. coli NagD dephosphorylates UMP, CMP, AMP, and ribose 5'-phosphate with values of $k_{\text{cat}}/K_M$ from $2 \times 10^3$ to $3 \times 10^4/M\cdot s^{-1}$. The relatively robust activities of these enzymes with multiple substrates suggest that broad specificity may be part of their function. Understanding the physiological roles of such broad-specificity enzymes will require a more sophisticated understanding of the conditions under which they are expressed and the dynamics and relative concentrations of potential substrates that occur as the metabolic network changes in response to environmental conditions.

Identification of function for enzymes of unknown function has necessarily been an endeavor focused upon individual enzymes, and this will continue to be the case since experimental verification is critical for correct annotation of function. However, genomic
enzymologists should take advantage of clues emerging from new high-throughput proteomic methods. Techniques for immobilizing small molecules on microarrays will allow the identification of proteins in complex mixtures that bind to or transform specific molecules or both [26]. Activity-based protein profiling methods identify proteins that react with affinity probes designed to modify active sites of specific classes of enzymes [27]. Some of the hits obtained in such experiments will undoubtedly be enzymes of unknown function. Such information provides an additional source of clues for identification of function for enzymes for which superfamily analysis, genome context, and phylogenetic analysis have not been sufficient.

Figure 2. Potential role for Escherichia coli YghZ

D-GAP, D-glyceraldehyde 3-phosphate; G3P, glycerol 3-phosphate; L-GAP, L-glyceraldehyde 3-phosphate; TIM, triose phosphate isomerase; non-enzymatic racemization; G3P dehydrogenase; YghZ.
Abbreviations
AMP, adenosine monophosphate; CMP, cytidine monophosphate; d-GAP, d-glyceraldehyde 3-phosphate; GMP, guanosine monophosphate; i-GAP, i-glyceraldehyde 3-phosphate; TIM, triose phosphate isomerase; UMP, uridine monophosphate.

Competing interests
The author declares that she has no competing interests.

References
1. Seffernick JL, de Souza ML, Sadowski MJ, Wackett LP. Melamine deaminase and atrazine chlorohydrolase: 98 percent identical but functionally different. J Bacteriol 2001, 183:2405-10.
2. Glaser ME, Gerlt JA, Babbitt PC. Evolution of enzyme superfamilies. Curr Opin Chem Biol 2006, 10:492-7.
3. Gerlt JA, Babbitt PC, Rayment I. Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. Arch Biochem Biophys 2005, 433:59-70.
4. Copley SD, Novak W, Babbitt PC. Divergence of function in the thioredoxin-fold suprafamily: evidence for evolution of peroxiredoxins from thioredoxins. Biochemistry 2004, 43:13981-95.
5. Tremblay LW, Dunaway-Mariano D, Allen KN. Structure and activity analyses of Escherichia coli NAGD provide insights into the evolution of biochemical function in the haloalkane dehalogenase superfamily. Biochemistry 2006, 45:13803-9.
6. Joosten HJ, Han Y, Niu W, Vervoort J, Dunaway-Mariano D, Schaap PJ. Identification of fungal oxaloacetate hydrolyase within the isocitrate lyase/PEP mutase superfamily as a sequence marker-based method. Proteins 2008, 70:157-66.
7. Seibert CM, Raushel FM. Structural and catalytic diversity within the amidohydrolase superfamily. Biochemistry 2005, 44:6383-91.
8. Burroughs AM, Allen KN, Dunaway-Mariano D, Aravind L. Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. J Mol Biol 2006, 361:1003-34.
9. Zhuang Z, Song F, Zhao H, Li L, Cao J, Eisenstein E, Herzberg O, Dunaway-Mariano D. Divergence of function in the hot dog fold enzyme superfamily: the bacterial thioesterase YciA. Biochemistry 2008, 47:2789-96.
10. Rea D, Hoxting R, Rakus JF, Gerlt JA, Fulop V, Bugg TD, Roper DI. Crystal structure and functional assignment of YfaU, a metal ion dependent class II aldolase from Escherichia coli K12. Biochemistry 2008, 47:9955-65.
11. Palmer DR, Garrett JB, Sharma V, Meganathan R, Babbitt PC, Gerlt JA. Unexpected divergence of enzyme function and sequence: “N-acylamino acid racemase” is o-succinylbenzoate synthase. Biochemistry 1999, 38:4252-8.
12. Xu L, Lawson SL, Resing K, Babbitt PC, Copley SD. Evidence that pcpA encodes 2,4-dichlorophenoxyacetic acid dioxygenase, the ring-cleavage enzyme required for pentachlorophenol degradation in Sphingomonas chlorophenolica strain atcc 39723. Biochemistry 1999, 38:7659-69.
13. Marti-Arbona R, Xu C, Steele S, Weeks A, Kuty GF, Seibert CM, Raushel FM. Annotation of enzymes of unknown function: N-formimino-o-glutamate deiminase is a member of the amidohydrolase superfamily. Biochemistry 2006, 45:1997-2005.
14. Nguyen TT, Brown S, Fedorov AA, Fedorov EV, Babbitt PC, Almo SC, Raushel FM. At the periphery of the amidohydrolase superfamily: Bh0493 from Bacillus halodurans catalyzes the isomerization of D-galacturionate to D-tagaturonate. Biochemistry 2008, 47:1194-206.
15. Poelarends GJ, Serrano H, Person MD, Johnson WH Jr, Whitman CP. Characterization of cg10062 from Corynebacterium glutamicum: implications for the evolution of cis-3-chloroacryl acid dehalogenase activity in the tautomerase superfamily. Biochemistry 2008, 47:8139-47.
16. Rakus JF, Fedorov AA, Fedorov EV, Glaser ME, Hubbard BK, Delli JD, Babbitt PC, Almo SC, Gerlt JA. Evolution of enzymatic activities in the enolase superfamily: l-rhamnionate dehydratase. Biochemistry 2008, 47:9944-54.
17. Yew WS, Fedorov AA, Fedorov EV, Wood BM, Almo SC, Gerlt JA. Evolution of enzymatic activities in the enolase superfamily: d-tartrate dehydratase from Brodyribizium japonicum. Biochemistry 2006, 45:14598-608.
18. Yew WS, Fedorov AA, Fedorov EV, Rakus JF, Pierce RW, Almo SC, Gerlt JA. Evolution of enzymatic activities in the enolase superfamily: L-fucose dehydratase from Xanthomonas campestris. Biochemistry 2006, 45:14582-97.
19. Song L, Kalyanaraman C, Fedorov AA, Fedorov EV, Glaser ME, Brown S, Imker Hj, Babbitt PC, Almo SC, Jacobson MP, Gerlt JA. Prediction and assignment of function for a divergent N-succinyl amino acid racemase. Nat Chem Biol 2007, 3:486-91.
20. Kalyanaraman C, Bernacki K, Jacobson MP. Virtual screening against highly charged active sites: identifying substrates of alpha-beta barrel enzymes. Biochemistry 2005, 44:2059-71.
21. Favia AD, Nobeli I, Glaser F, Thornton JM. Molecular docking for substrate identification: the short-chain dehydrogenases/reductases. J Mol Biol 2008, 375:855-74.
22. Hermann JC, Ghanem E, Li Y, Raushel FM, Irwin JJ, Shoichet BK. Predicting substrates by docking high-energy intermediates to enzyme structures. J Am Chem Soc 2006, 128:15882-91.
23. Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov EV, Shoichet BK, Raushel FM. Structure-based activity prediction for an enzyme of unknown function. Nature 2007, 448:775-9.
24. Kalyanaraman C, Imker Hj, Fedorov AA, Fedorov EV, Glaser ME, Babbitt PC, Almo SC, Gerlt JA, Jacobson MP. Discovery of a dipeptide epimerase enzymatic function guided by homology modeling and virtual screening. Structure 2008, 16:1668-77.
25. Desai KK, Miller BG. A metabolic bypass of the triosephosphate isomerase reaction. Biochemistry 2008, 47:7983-5.