Review

Can sequence determine function?

John A Gerlt* and Patricia C Babbitt†

Addresses: *Departments of Biochemistry and Chemistry, University of Illinois, Urbana, IL 61801, USA. †Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, University of California, San Francisco, CA 94143, USA. E-mail: j-gerlt@uiuc.edu; babbitt@cgl.ucsf.edu

Published: 8 November 2000

Genome ***Biology*** 2000, **1(5)**:reviews0005.1–0005.10

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2000/1/5/reviews/0005

Abstract

The functional annotation of proteins identified in genome sequencing projects is based on similarities to homologs in the databases. As a result of the possible strategies for divergent evolution, homologous enzymes frequently do not catalyze the same reaction, and we conclude that assignment of function from sequence information alone should be viewed with some skepticism.

The functional assignment of proteins identified in genome sequencing projects is a major problem in post-genomic biology: approximately 40% of the open reading frames (ORFs) initially identified in the genomes of even the most intensively studied eubacteria, notably *Escherichia coli* [1] and *Bacillus subtilis* [2], as well as in other microbial species and the more complex eukaryotes, have unknown functions, because their sequences are judged to be unrelated to those of functionally characterized proteins. Although the initial deficits are being resolved slowly in prokaryotes, the functions of an even larger percentage of ORFs in eukaryotes remain unpredictable [3]. Until functions are assigned to the unknown ORFs, an organism's capabilities cannot be completely described: the unknown ORFs may encode genes in essential metabolism that remain to be discovered; other unknown ORFs that are species-specific probably determine the unique characteristics of that organism. The promise of post-genomic biology will be elusive until functional assignment to the unknown ORFs is complete. At that point, integration of the functions encoded by complete genomes can be performed.

In some cases, functional assignment of unknown eubacterial ORFs is assisted by their physical proximity to ORFs of known function, because these genes are found either in operons or clustered in proximal transcriptional units that encode proteins that participate in related metabolic functions. In other cases, 'phylogenomic' analyses [4] may allow functional assignment of unknown genes: the intermediates in metabolic pathways are likely to be more conserved than the enzymes that catalyze the reactions, thereby focusing the search for 'missing' enzymes. In other cases, single-domain proteins in some organisms may be present as components of multidomain proteins in other organisms [5], thereby providing insights into the functions of unknown domains in the multidomain proteins. In eukaryotes, ORFs encoding proteins of linked function are not clustered, so in these genomes unequivocal functional assignment will probably require a number of diverse but complementary experimental approaches, including analysis using DNA microarrays, two-hybrid and other approaches for detecting protein-protein interactions, exhaustive screening of substrate libraries for catalytic activity, and gene knockouts.

In this brief review, we draw attention to an important, but we believe insufficiently appreciated, problem of functional assignment: homologous proteins need not catalyze the same chemical reaction. Homologs result from divergent evolution: a progenitor gene is duplicated and its copy assumes a new function in response to selective pressure. Because the outcomes of divergent evolution can be quite different, functional assignment on the basis of homology alone should be performed cautiously, although our experience indicates that genome annotation usually assigns the function of the 'closest' homolog to an unknown ORF. Unfortunately, most annotators and users of the protein databases appear to be unaware

Box 1

A glossary for divergent evolution

Analogs: proteins that catalyze the same reaction but are not structurally related.

Homologs: proteins derived from a common ancestor. By definition, these are structurally related.

Orthologs: homologs in different species that catalyze the same reaction.

Paralogs: homologs in the same species that diverged after speciation and do not catalyze the same reaction.

Family: a group of orthologous enzymes that catalyze the same reaction; often, they share more than 30% pairwise sequence identity, but in some cases structural information may be required to detect their homology.

Specificity diverse superfamily: homologous enzymes that often have less than 30% pairwise sequence identity and catalyze the same reaction with different substrate specificities.

Mechanistically diverse superfamily: homologous enzymes, generally having less than 50% pairwise sequence identity, which catalyze different overall reactions with common mechanistic attributes. Such superfamilies also have conserved active-site elements that perform these common mechanistic functions in all members of the superfamily.

Metabolically linked suprafamily: homologous enzymes that catalyze mechanistically distinct reactions in the same metabolic pathways and have conserved active-site residues that perform different functions in different members of the suprafamily.

Metabolically distinct suprafamily: homologous enzymes that catalyze mechanistically distinct reactions in different metabolic pathways and have conserved active-site residues that perform different functions in different members of the suprafamily.

of this problem and will regard such functional assignments as established fact. As new genomes are sequenced and homologs of functionally annotated proteins are identified, problems in functional assignment will escalate, perhaps to the point where annotation is meaningless [6].

The origins of homologous enzymes

The available information about sequence, structure, and function for homologous enzymes can be correlated with three distinct strategies that have led to divergent evolution of enzyme function. Each is initiated by the duplication of a progenitor gene; the strategies then diverge in significant detail. The terminology we use in this review is defined in Box 1.

Substrate specificity is dominant

In this case, the duplicated gene evolves to provide the substrate for the enzyme encoded by the progenitor gene [7,8]. The corollary is that (anabolic) pathways evolve backward: when the precursor metabolite is depleted, a new enzyme is required to synthesize it from available molecules. The original and evolved enzyme will share the ability to bind the same molecule (as substrate and product) but the mechanisms of the transformations need not be related. Such homologous proteins are members of metabolically linked suprafamilies. Examples include successive steps in both the tryptophan [9] and histidine biosynthetic pathways [10].

Chemical mechanism is dominant

Here, the progenitor gene is selected because its encoded protein provides the structural strategy for stabilizing intermediates/transition states required for the desired new transformation [11,12]. The simplest situation is that the progenitor and new enzyme catalyze the same reaction but with differing specificities, as is the case for the members of the specificity diverse serine protease superfamily [13]. Alternatively, the progenitor and new enzyme may catalyze different reactions that utilize the same type of intermediate, such as the enolic intermediates in reactions catalyzed by members of the mechanistically diverse enolase superfamily [14]. Irrespective of the situation, the progenitor may be 'promiscuous' and may catalyze low levels of the new reaction [15,16]. Divergent evolution will enhance the promiscuous reaction, probably at the expense of the original reaction. Such homologous proteins are members of mechanistically diverse superfamilies.

Active-site structure is dominant

In this case, the progenitor gene is selected because the functional groups present in the active site can catalyze a different reaction, perhaps with no or limited change in the identities of the amino-acid residues directly involved in catalysis [17]. The reactions catalyzed by the original and evolved enzyme differ in both mechanism and substrate specificity. Such homologous proteins are members of metabolically distinct suprafamilies. An example of this paradigm is apparently provided by the homologous orotidine 5-phosphate decarboxylase and D-*arabino*-hex-3-ulose 6-phosphate synthase (J.A.G. and P.C.B., unpublished observations) [18].

Each of these strategies for divergent evolution of enzyme function can therefore be seen to make the functional assignment of homologs unreliable: the available sequence/function information indicates that incorrect annotations are possible even when the sequence divergence is low (more than 80%

sequence identity [19]) but more frequently when the sequence divergence is large (less than 25% sequence identity). Another source of difficulty is that orthologs can share surprisingly low levels of sequence homology that are often indistinguishable from those found in paralogs. Because the level of sequence divergence at which functional divergence occurs can be highly variable [15], even within homologous families of a single superfamily, the level of sequence similarity required for reliable prediction of function from sequence cannot be specified with confidence.

In the remainder of this review, we describe examples of why sequence homology alone is an insufficient criterion for making functional assignments. We will focus on groups of enzymes that share elements of mechanism but catalyze different reactions (mechanistically diverse families and superfamilies) and those that catalyze mechanistically unrelated reactions (metabolically linked and metabolically distinct suprafamilies).

A mechanistically diverse 'family': different oxidation reactions without significant sequence divergence

Although functional diversity is almost always associated with significant divergence in sequence so that the definition of 'superfamily' or 'suprafamily' (less than 50% sequence identity) clearly applies, Somerville and coworkers [19] have described a 'family' of closely related enzymes found in plants that catalyze different oxidative reactions and appear to be responsible for the diversity in seed storage fatty acids. These enzymes use oleate as their common substrate but catalyze desaturation, hydroxylation and epoxidation reactions (Figure 1). They are members of a larger functionally diverse superfamily that includes enzymes such as alkane hydroxylase, xylene monooxygenase, carotene ketolase, and sterol methyloxidase, all of which are nonheme-iron enzymes that utilize a diiron center for catalysis and have three conserved histidine residues that probably bind the metal ions.

An oleate hydroxylase from *Lesquerella fendleri* shares 81% sequence identity with an oleate desaturase from *Arabidopsis thaliana* but 71% identity with an oleate hydroxylase from *Ricinus communis*. These relationships suggest that divergence of function is facile and may have occurred many times in the speciation of higher plants. Noting that seven residues were conserved in oleate desaturases from multiple species but diverged in hydroxylases, Somerville and coworkers [19] constructed libraries of mutants in which these conserved residues were replaced. Interestingly, as few as four substitutions were able to convert an oleate 12-desaturase into a hydroxylase; in the opposite direction, six substitutions converted a hydroxylase into a desaturase. Bearing in mind that the substrates for the hydroxylases and desaturases are the same, we conclude that different reaction mechanisms and products can result from subtle changes in active-site geometry that may reposition the substrate relative to the diiron center or alter the coordination geometry or hydrogen-bonding network in the active site. Irrespective of the precise structural explanation, these findings illustrate that mechanistic diversity does not require a large significant divergence in sequence, and underscore that high levels of sequence identity do not 'guarantee' the same enzymatic function.

Mechanistically diverse superfamilies: different reactions forming thioester enolate anions as intermediates

The β-oxidation pathway for the catabolism of fatty acids is encoded by most genomes. A key step in the pathway is the hydration of enoyl CoA esters, which is catalyzed by enoyl

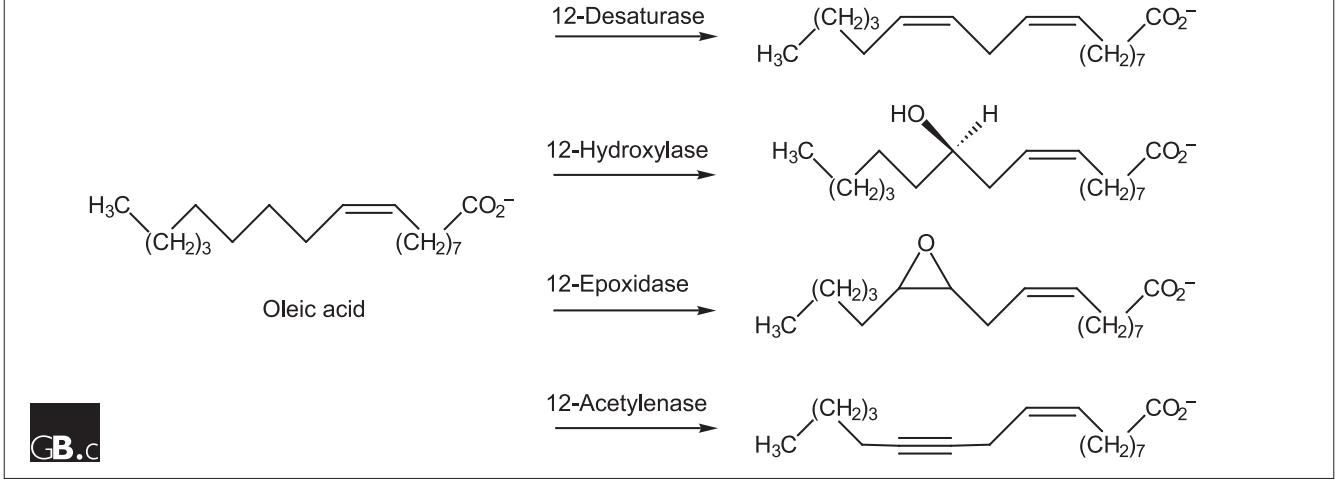

Figure 1
Different oxidation reactions from proteins of similar sequence: reactions catalyzed by members of the oleic-acid-oxidizing 'family'.

CoA hydratase (also called crotonase). The reaction catalyzed by crotonase has been subjected to mechanistic scrutiny [20-22]; several high-resolution structures are available for the rat mitochondrial enzyme [23,24]. As a result of these studies, the mechanism is known to involve two glutamate residues that function as acid/base catalysts: in the direction of hydration of enoyl CoAs, Glu144 activates a water molecule for nucleophilic attack and Glu164 delivers a proton to the α carbon. Although never observed directly, the mechanism probably involves the transient formation of a thioester enolase anion intermediate that is stabilized by enhanced hydrogen-bonding interactions with an 'oxyanion hole' [25]. As judged by the structures of complexes with substrate analogs/products, the 'oxyanion hole' is formed by two peptidic NH groups that bind the thioester carbonyl group.

Database searches with the sequence of the structurally characterized crotonase yield many 'hits', as judged by conserved consensus sequences for both 'halves' of the oxyanion hole [26,27]. The reactions catalyzed by some of these hits have been biochemically characterized, leading to the discovery that crotonase is a member of a mechanistically diverse superfamily: the reactions can be 'explained' by a mechanism that involves the transient formation and stabilization of a thioester enolate anion intermediate.

In addition to crotonase, structurally characterized members of the superfamily include (Figure 2): 4-chlorobenzoyl CoA dehalogenase, which catalyzes a nucleophilic aromatic substitution [28]; $\Delta^{3,5}$, $\Delta^{2,4}$-dienoyl CoA isomerase, which catalyzes a 1,5-proton transfer reaction [29]; and methylmalonyl CoA decarboxylase, which catalyzes a decarboxylation reaction [30]. In the case of the dehalogenase, homologs of Glu144 and Glu164 in crotonase are missing, but the functional groups involved have been identified and subjected to mechanistic scrutiny assisted by site-directed mutagenesis. In the case of the isomerase and decarboxylase, the sequences of both contain a homolog for either Glu144 or Glu164 but not both; the mechanisms of the reactions are not yet certain.

Biochemically characterized members of the superfamily for which no structure is yet available include (Figure 2): 1,4-dihydroxynaphthoyl CoA synthase, which catalyzes formation of a carbon-carbon bond in a Dieckman reaction [31]; 2-ketocyclohexyl CoA hydrolases, which catalyzes cleavage of a carbon-carbon bond in a retro Dieckman reaction [32]; 3,2-*trans*-enoyl CoA isomerase, which catalyzes a 1,3-proton transfer reaction [33]; feruloyl CoA hydratase/lyase, which catalyzes successive hydration and carbon-carbon bond cleavage reactions [34]; and carnitinyl CoA epimerase, which inverts the configuration of an unactivated carbon by an unknown mechanism [35]. The hydratase/lyase and epimerase contain homologs of both Glu144 and Glu164 in crotonase; as a result, the mechanisms of both may involve hydration/dehydration partial reactions

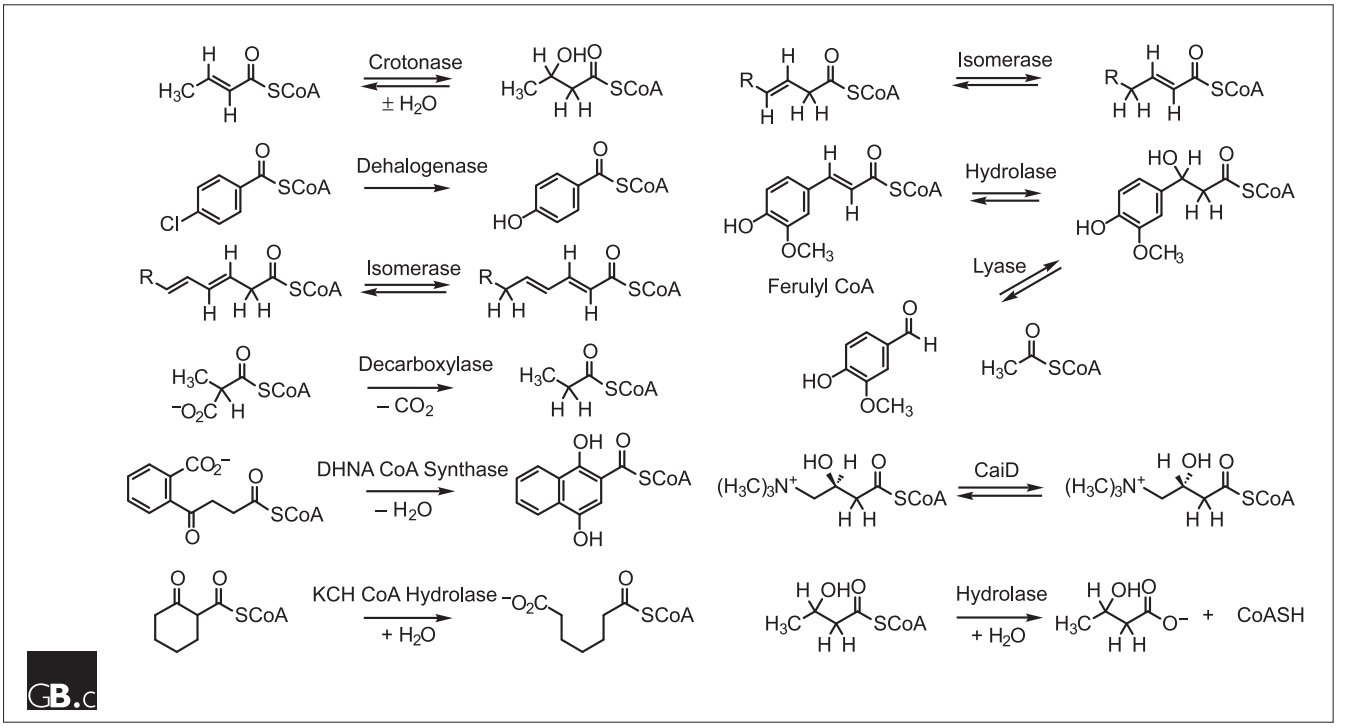

Figure 2
Different reactions catalyzed by members of a mechanistically diverse superfamily: reactions catalyzed by members of the crotonase superfamily.

involving a thioester enolase anion intermediate. Mechanisms involving carbon-carbon bond cleavage reactions via retroaldol mechanisms are also possible, however. Irrespective of the precise reaction coordinates that describe the reactions catalyzed by these enzymes, all probably involve the formation and transient stabilization of a thioester enolate anion intermediate.

The database searches also yield members of the crotonase superfamily that catalyze reactions that probably do not involve formation of a thioester enolate anion but, instead, an anionic tetrahedral intermediate in thioester or peptide bond hydrolysis that would be stabilized by the conserved oxyanion hole (Figure 2): 3-OH isobutyryl CoA hydrolase [36] and the ClpP protease [37,38]. The hydrolase occurs in the catabolic pathway for valine; the preceding enzyme in the pathway, also a member of the crotonase superfamily, catalyzes the hydration of α-methacryl CoA using homologs of Glu144 and Glu164 in rat mitochondrial crotonase. Although structure/function relationships have not yet been characterized in detail for the protease, the structure allows the suggestion that the functional groups are configured in a catalytic triad (Ser-His-Glu) structurally analogous to those found in other examples of analogous serine proteases [38]. That the hydrolase and protease are members of the crotonase superfamily provides compelling evidence that the underlying chemical strategy supported by the α/β fold is stabilization of oxyanion intermediates.

We therefore conclude that structure/function relationships in the crotonase superfamily cannot be extrapolated simply to newly discovered members from the sequence and available structures. Database annotations indicate, however, that newly discovered members of this superfamily are 'probable enoyl CoA hydratases/isomerases'. Table 1 compares the number of members of the crotonase superfamily encoded by several microbial genomes, as assessed by the presence of consensus sequences for the oxyanion hole, with the number of proteins that could catalyze an enoyl CoA hydratase reaction, as assessed by the presence of homologs of both Glu144 and Glu164 in rat mitochondrial crotonase. Despite the differences, the databases invariably include the members of the superfamily that do not contain homologs of Glu144 and Glu164 in rat mitochondrial crotonase as 'likely' or 'probable' 'enoyl CoA hydratases/isomerases'. We believe that, in many cases, these annotations are incorrect or misleading. Although we recognize this problem, given our expertise in structure/function relationships for several members of this superfamily, most users of the sequence databases will not. The likely consequences are that annotations of genome sequences completed in the future will use these incorrect annotations and continue to misrepresent the functions of orthologous ORFs that are identified, and that novel metabolic pathways involving members of the crotonase superfamily will be ignored or assigned incorrect functions. Both types of incorrect assignment will undermine the considerable potential of genome sequencing projects in defining new metabolic capabilities.

Mechanistically diverse superfamilies: different reactions forming metal-coordinated enediolate anions as intermediates

Members of mechanistically diverse superfamilies are common in the ORFs encoded by microbial genomes. For example, the microbial sequence databases contain many members of the enolase superfamily [14], with virtually all sharing three conserved carboxylate ligands for an essential divalent metal ion. The conservation of these elements is persuasive evidence that the underlying catalytic strategy employed by all members of the enolase superfamily is electrostatic stabilization of a geminal enediolate anion derived from a carboxylate substrate by abstraction of a proton from the α carbon. The members of the superfamily can be divided into three groups: those most homologous to enolase, those most homologous to mandelate racemase (MR), and those most homologous to muconate lactonizing enzyme (MLE). High-resolution structures are available for each of these, and reveal that, in enolase Lys345 is the general base; in MR, Lys166 in a Lys164-X-Lys166 motif is the (S)-specific base and His297 hydrogen-bonded to Asp270 is the (R)-specific base; and in MLE, Lys169 in a Lys 167-X-Lys 169 motif and Lys 273 are positioned on opposite faces of the active site, although which of these is the expected general base has not been unequivocally assigned. Alignments of sequence and structural superpositions reveal that Lys345 (enolase), Asp270 (MR), and Lys273 (MLE) are homologous residues, and that the Lys164-X-Lys166 motif (MR) and the Lys167-X-Lys169 motif are also homologous.

Table 1

Members of the crotonase superfamily encoded by microbial genomes, and the likely enoyl CoA hydratases

Organism	Number of members of superfamily	Number of likely enoyl CoA hydratases
Aeropyrum pernix	3	3
Archaeoglobus fulgidus	10	6
Bacillus subtilis	7	3
Deinococcus radiodurans	6	4
Escherichia coli	7	3
Haemophilus influenzae	1	0
Mycobacterium tuberculosis	24	4
Pseudomonas aeruginosa	17	9
Synechocystis sp.	1	0
Vibrio cholerae	3	2

Database searches disclose that the *E. coli* genome encodes three members of the enolase superfamily that are most homologous to MR (Figure 3). MR catalyzes a 1,1-proton transfer reaction using Lys166 and His297 of the His297-Asp270 dyad as general/acid base catalysts that abstract protons from and deliver protons to the α carbon of the carboxylate anion substrate/product; the mechanism of the reaction is well understood in terms of the structure of the active site. Biochemical studies have established, however, that the three homologs of MR encoded by the *E. coli* genome catalyze the dehydration of acid sugars: (D)-glucarate/(L)-idarate dehydratase (GlucD) [39,40]; (D)-galactonate dehydratase (GalD) [41,42]; and (L)-rhamnonate dehydratase (RhamD; B.K. Hubbard, J. Delli and J.A.G., unpublished observations). We note that the incomplete *Streptomyces coelicolor* genome-sequencing project has discovered at least seven homologs of MR, some annotated as 'racemases' and others as 'dehydratases'. *A priori*, we and, we assume, others cannot assign the correct function to these ORFs without biochemical evidence that must involve screening these ORFs for both racemization and dehydration reactions on a library of acid sugars.

The *E. coli* genome also encodes two members of the enolase superfamily that are most homologous to MLE (Figure 3). MLE catalyzes a reversible cycloisomerization reaction involving intramolecular addition/elimination to a conjugated enoic acid. Biochemical studies have established that one of these homologs is *o*-succinylbenzoate synthase (OSBS), which catalyzes an exergonic dehydration reaction in the biosynthesis of menaquinone [43]. The other homolog is a dipeptide epimerase that is specific for L-Ala-(D/L)-Glu (AE Epim), although other dipeptides are epimerized (D.Z. Schmidt, B.K. Hubbard and J.A.G., unpublished observations); we hypothesize that the epimerase is involved in either recycling or remodeling of cell-wall-derived peptides. We have noted that characterization of the gene encoding OSBS in *E. coli* has not allowed the functional assignment of orthologs in other microbial genomes [15]. For example, the genome of *B. subtilis* encodes a pathway for menaquinone biosynthesis, but the gene encoding OSBS was variously annotated as 'similar to muconate cycloisomerase' and 'similar to N-acylamino acid racemase'. The sequences of orthologous OSBSs are highly

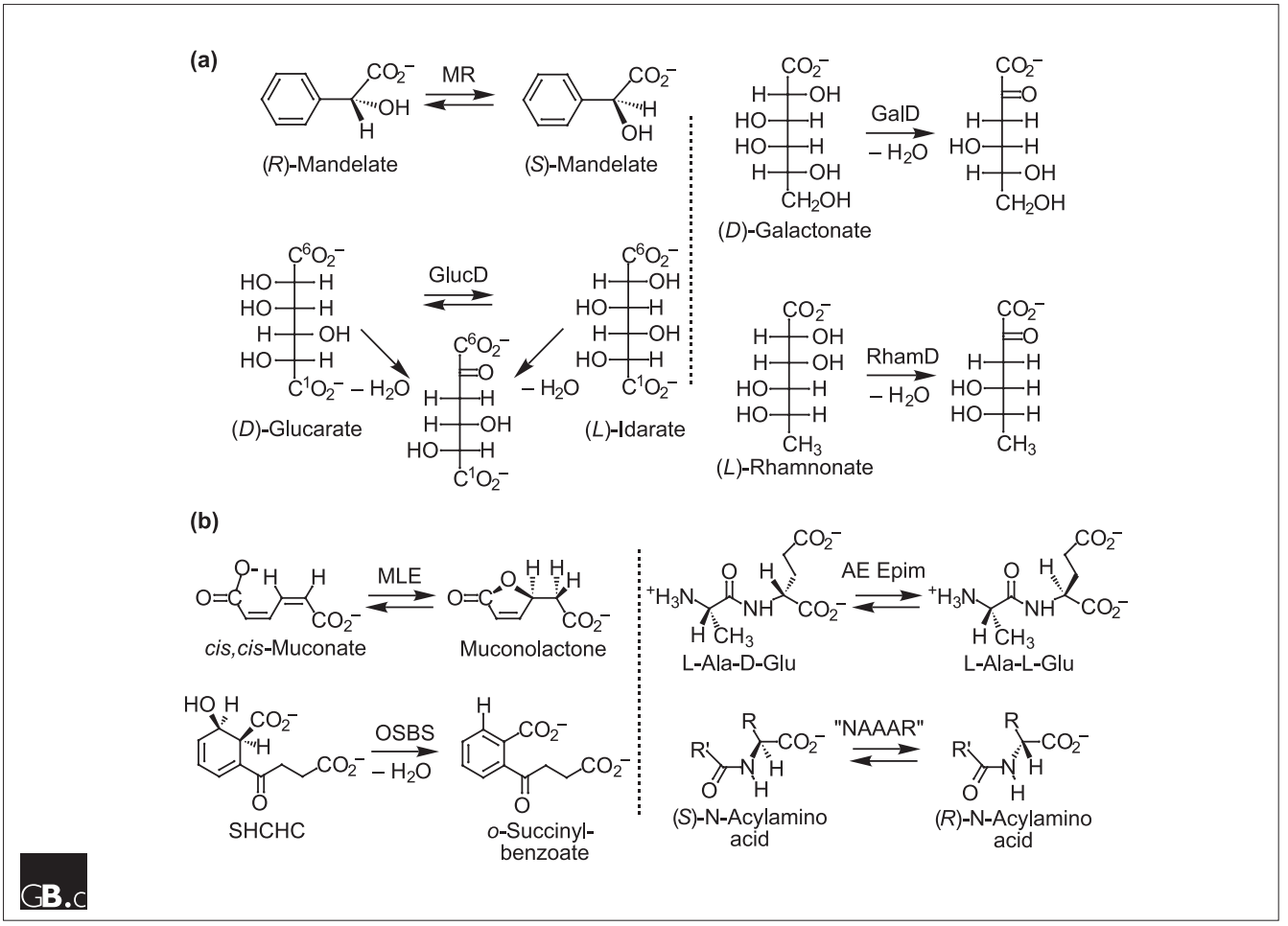

Figure 3
Reactions catalyzed by homologs of **(a)** MR and **(b)** MLE in the enolase superfamily. See text for details.

diverged but can be identified unequivocally when their proximal genome context includes genes encoding other enzymes in the menaquinone biosynthetic pathway [44]. The orthologous OSBSs retain invariant residues only for the three metal ion ligands and the Lys167-X-Lys169 motif and Lys273 in MLE. The reason for this extreme divergence may be related to the extreme exergonicity of this OSBS reaction. Interestingly, the OSBS from a species of *Amycolaptosis* was initially characterized as an N-acylamino acid racemase, an enzyme catalyzing a reaction needed to achieve a commercial enzymatic conversion of racemic N-acylamino acids to chiral amino acids [45]. Kinetic characterization of this protein revealed that it is promiscuous, catalyzing both the racemase and OSBS reactions, although the latter reaction is approximately 1,000-fold more efficient, as assessed by the values of k_{cat}/K_m [15]. Unfortunately, the less efficient and, presumably, physiologically irrelevant racemase reaction continues to be used in the annotation of newly sequenced OSBSs.

A metabolically linked suprafamily: the evolution of biosynthetic pathways for tryptophan and histidine

We expect that functional annotations of newly discovered members of mechanistically diverse superfamilies will continue to be problematic. Mechanism-oriented biologists will still be able to derive some value from such annotations, however, if they are aware of the existence of such superfamilies. An essential element of mechanism is retained as a consequence of the chemical mechanism-dominant strategy that nature used in their evolution; presumably, the same catalytic strategy will be employed by the misannotated gene products. A more serious problem of functional ambiguity is associated with annotation of members of suprafamilies in which conserved elements of active-site architecture are used in different ways to catalyze different reactions.

Two structurally documented examples of divergent evolution supporting Horowitz's hypothesis [7,8] that metabolic pathways evolve backward are available. Interestingly, these examples are functionally intertwined, further illustrating difficulties in assigning function from sequence. In both the tryptophan and histidine biosynthetic pathways (Figure 4), the immediate precursors of the cyclization reactions that form the pyrrole and imidazole rings, respectively, are obtained by Amadori rearrangements in which substituted 1-amino ribose 5-phosphates are isomerized to 1-amino ketose 5-phosphates (TrpF and HisA, respectively). In the tryptophan pathway, the Amadori product undergoes a poorly understood decarboxylation/dehydration reaction to form a carbon-carbon bond, yielding the pyrrole ring (TrpC); in the histidine pathway, the Amadori product with ammonia derived from a glutamine cosubstrate undergoes amination/transaldimination to form the imidazole ring (HisF). Thus, while the Amadori rearrangements employ the same mechanisms and logically could be catalyzed by homologous enzymes (members of a specificity diverse superfamily), the cyclization reactions are mechanistically distinct.

The sequences of the enzymes catalyzing the Amadori rearrangements, HisA and TrpF in the histidine and tryptophan biosynthetic pathways, show little, if any, homology, however. In contrast, the sequences of TrpF and TrpC show significant levels of sequence identity: 22% in the case of the individual domains in the bifunctional and structurally characterized protein encoded by the *E. coli* genome; both domains of this protein have the $(\beta/\alpha)_8$ fold [46]. Independent evidence that these enzymes are related by divergent evolution was provided by Fersht and coworkers [9], who retained the barrel of TrpC but randomized the sequences of two of the loops that connect the β strands with the following α helices. Selection by complementation of a mutant strain of *E. coli* and repeated rounds of DNA shuffling yielded an '*in vitro* evolved' TrpF that exceeded the wild-type TrpF in the measured value of k_{cat}/K_m.

Presumably an analogous directed evolution strategy could mimic nature by converting HisF to HisA. The structurally characterized enzymes from *Thermotoga maritima* share 25% sequence identity; each has the $(\beta/\alpha)_8$ fold [10]. Interestingly, these structures confirm the presence of a two-fold repeat first predicted from sequence analysis, in which the first and second halves of the barrel are symmetrically disposed [47,48]. Presumably, the ancestral gene encoded a four-stranded half-barrel, which was duplicated and fused in tandem to produce each gene that encodes HisA and HisF.

Remarkably, despite undetectable levels of sequence identity, TrpF and HisA that catalyze the Amadori rearrangements can be interconverted by directed evolution: starting with TrpF, a single point mutation at the end of the fifth β strand was found to confer the ability to catalyze the HisA reaction while retaining the TrpF reaction [49]. This experiment, which relates the functions of enzymes in distinct pathways, implies that HisA/F and TrpF/C could have been derived from a common ancestor and provides persuasive evidence for the suggestion that the design of the $(\beta/\alpha)_8$ fold, including its possible origin from half-barrels, was exploited by nature in the divergent evolution of new enzymatic functions.

These examples confirm the importance of Horowitz's hypothesis [8,9] that substrate specificity can dominate divergent evolution to produce members of metabolically linked suprafamilies. Although the levels of sequence identity that relate TrpF with TrpC and HisA with HisF, approximately 25%, are likely to be sufficient for correct functional assignment of orthologs in newly sequenced genomes, the established importance of this strategy for divergent evolution increases the possibility that members of metabolically linked suprafamilies occur in other metabolic pathways, thereby confounding the functional annotations of the genes that encode these proteins.

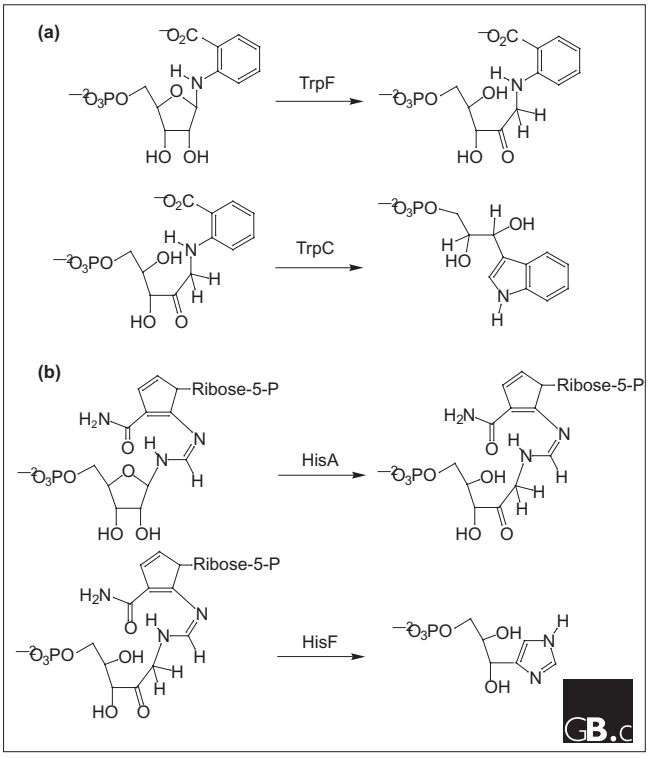

Figure 4
Reactions catalyzed by members of the metabolically linked suprafamilies in the **(a)** tryptophan and **(b)** histidine biosynthetic pathways.

A metabolically distinct suprafamily: the breadth of reactions catalyzed by the $(\beta/\alpha)_8$ fold

Given the confirmation of substrate-specificity-driven divergent evolution of enzymes to create metabolic pathways (generating metabolically linked suprafamilies), the importance of divergent evolution of enzymes that catalyze mechanistically distinct reactions in different metabolic pathways (generating metabolically distinct suprafamilies) must be considered. Although such evolution would confound functional annotations on the basis of sequence alone, it would provide important insights into the processes by which nature selected progenitors for the evolution of 'new' enzymes.

We are aware of one unequivocal example of a metabolically distinct suprafamily and, as expected, annotation of function based on sequence appears problematic. Orotidine 5'-phosphate decarboxylase (OMPDC) catalyzes the last step in the biosynthesis of UMP (Figure 5). This enzyme has attracted considerable recent attention in the enzymological community, because it is the most proficient of any enzyme studied to date [50]. Four independent structural studies of OMPDCs from each of the three kingdoms of life have been published in recent months (*Saccharomyces cerevisiae*, *Methanobacterium thermoautotrophicum*, *B. subtilis* and *E. coli*); each has a dimeric structure in which the active sites are located at the carboxy-terminal ends of $(\beta/\alpha)_8$ domains [51-54]. Remarkably, only five residues are absolutely conserved in the known OMPDCs, and they are located in the active site. Of these, an aspartate is located at the end of the first β strand, a lysine is located at the end of the second β strand, and an Asp-X-Lys-X-X-Asp motif is located at the end of the third β strand; all of these residues directly contact the OMP substrate. Although the mechanism is not yet certain, the functional groups in the Asp-X-Lys-X-X-Asp motif shared between both active sites in the functional dimer probably destabilize the substrate carboxylate group and deliver a proton to carbon-6 in an S_E2 reaction that avoids the formation of an unstable vinyl carbanion intermediate.

Distant homologs of OMPDC that contain some homologs of the residues listed above can be identified in the sequence databases. One of these has been identified as D-*arabino*-hex-3-ulose 6-phosphate synthase (HUPS), the enzyme that 'fixes' formaldehyde in methylotrophic bacteria that can utilize either methylamine or formaldehyde as their sole carbon source (J.A.G. and P.C.B, unpublished observations; Figure 5) [18]. Although OMPDC does not utilize a divalent metal ion [55,56], HUPS is reported to do so [57]. Because the HUPS-catalyzed reaction can be rationalized by a mechanism involving the transient formation and stabilization of an enediolate intermediate, an intermediate that is impossible in the OMPDC-catalyzed reaction, we conclude that OMPDC and HUPS constitute a metabolically distinct suprafamily. While their substrates contain phosphate groups and the active site may contain homologous phosphate-binding sites, the metabolic pathways are distinct and the reactive portions of the substrates are not structurally similar.

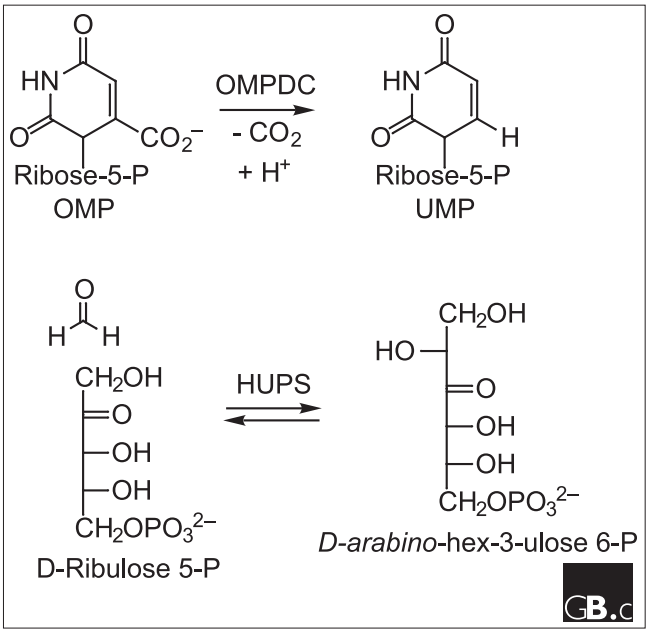

Figure 5
Reactions catalyzed by OMPDC and HUPS in a metabolically distinct suprafamily. See text for details.

The sequence databases contain homologs of OMPDC/HUPS that are annotated as 'probable hexulose phosphate synthases' or hexulose phosphate synthase 'homologs', 'isologs', or 'related proteins'; several of these are under investigation in our laboratories, and we have concluded that they are not hexulose phosphate synthases. Although our studies are too premature to provide more details about these homologs and the reactions they catalyze, we include this example to illustrate that metabolically distinct suprafamilies exist and that correct annotation of the members of these will be difficult.

In summary, although an enzyme's membership in mechanistically diverse family or superfamily can provide useful information about the nature of the reaction that is catalyzed, membership in a suprafamily will provide few, if any clues, about the identity of the reaction. Unfortunately, if the annotators of genome sequencing projects are unaware of both the range of strategies nature uses to evolve homologous enzymes and the associated functional implications, annotations of function will provide misleading interpretations of sequence data that will be propagated as additional genomes are annotated. This situation will diminish both the intellectual and practical importance of sequence data that have the potential to provide heretofore impossible-to-achieve insights into the molecular details of cellular function and speciation.

Acknowledgements

We thank the National Institutes of Health for financial support (grants GM-60595 to P.C.B.; GM-40570, GM-52594 and GM-58506 to J.A.G.).

References

1. Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, *et al.*: **The complete genome sequence of *Escherichia coli* K-12.** *Science* 1997, **277:**1453-1474.
2. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, *et al.*: **The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*.** *Nature* 1997, **390:**249-256.
3. **Genome sequence of the nematode *C. elegans*: a platform for investigating biology. The *C. elegans* Sequencing Consortium.** *Science* 1998, **282:**2012-2018.
4. Eisen JA: **Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis.** *Genome Res* 1998, **8:**163-167.
5. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96:**4285-4288.
6. Karp PD: **What we do not know about sequence analysis and sequence databases.** *Bioinformatics* 1998, **14:**753-754.
7. Horowitz NH: **On the evolution of biochemical syntheses.** *Proc Natl Acad Sci USA* 1945, **31:**153-157.
8. Horowitz NH: *Evolving Genes and Proteins*. New York: Academic Press, 1965.
9. Altamirano MM, Blackburn JM, Aguayo C, Fersht AR: **Directed evolution of new catalytic activity using the alpha/beta-barrel scaffold.** *Nature* 2000, **403:**617-622.
10. Lang D, Thoma R, Henn-Sax M, Sterner R, Wilmanns M: **Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion.** *Science* 2000, **289:**1546-1550.
11. Babbitt PC, Gerlt JA: **Understanding enzyme superfamilies. Chemistry as the fundamental determinant in the evolution of new catalytic activities.** *J Biol Chem* 1997, **272:**30591-30594.
12. Gerlt JA, Babbitt PC: **Mechanistically diverse enzyme superfamilies: the importance of chemistry in the evolution of catalysis.** *Curr Opin Chem Biol* 1998, **2:**607-612.
13. Perona JJ, Craik CS: **Evolutionary divergence of substrate specificity within the chymotrypsin-like serine protease fold.** *J Biol Chem* 1997, **272:**29987-29990.
14. Babbitt PC, Hasson MS, Wedekind JE, Palmer DR, Barrett WC, Reed GH, Rayment I, Ringe D, Kenyon GL, Gerlt JA: **The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids.** *Biochemistry* 1996, **35:**16489-16501.
15. 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:**4252-4258.
16. O'Brien PJ, Herschlag D: **Catalytic promiscuity and the evolution of new enzymatic activities.** *Chem Biol* 1999, **6:**R91-R105.
17. Gerlt JA, Babbitt PC: **Divergent evolution of enzyme function: mechanistically diverse superfamilies and functionally distinct suprfamilies.** *Ann Rev Biochem* 2001, **70:** in press.
18. Traut TW, Temple BR: **The chemistry of the reaction determines the invariant amino acids during the evolution and divergence of orotidine 5'-monophosphate decarboxylase.** *J Biol Chem* 2000, **275:**28675-28681.
19. Broun P, Shanklin J, Whittle E, Somerville C: **Catalytic plasticity of fatty acid modification enzymes underlying chemical diversity of plant lipids.** *Science* 1998, **282:**1315-1317.
20. Bahnson BJ, Anderson VE: **Isotope effects on the crotonase reaction.** *Biochemistry* 1989, **28:**4173-4181.
21. D'Ordine RL, Bahnson BJ, Tonge PJ, Anderson VE: **Enoyl-coenzyme A hydratase-catalyzed exchange of the alpha-protons of coenzyme A thiol esters: a model for an enolized intermediate in the enzyme-catalyzed elimination?** *Biochemistry* 1994, **33:**14733-14742.
22. Hofstein HA, Feng Y, Anderson VE, Tonge PJ: **Role of glutamate 144 and glutamate 164 in the catalytic mechanism of enoyl-CoA hydratase.** *Biochemistry* 1999, **38:**9508-9516.
23. Engel CK, Mathieu M, Zeelen JP, Hiltunen JK, Wierenga RK: **Crystal structure of enoyl-coenzyme A (CoA) hydratase at 2.5 angstroms resolution: a spiral fold defines the CoA-binding pocket.** *EMBO J* 1996, **15:**5135-5145.
24. Engel CK, Kiema TR, Hiltunen JK, Wierenga RK: **The crystal structure of enoyl-CoA hydratase complexed with octanoyl-CoA reveals the structural adaptations required for binding of a long chain fatty acid-CoA molecule.** *J Mol Biol* 1998, **275:**859-847.
25. Gerlt JA, Gassman PG: **An explanation for rapid enzyme-catalyzed proton abstraction from carbon acids: the importance of late transition states in concerted mechanisms.** *J Am Chem Soc* 1993, **115:**11552-11569.
26. Xiang H, Luo L, Taylor KL, Dunaway-Mariano D: **Interchange of catalytic activity within the 2-enoyl-coenzyme A hydratase/isomerase superfamily based on a common active site template.** *Biochemistry* 1999, **38:**7638-7652.
27. Haller T, Buckel T, Retey J, Gerlt JA: **Discovering new enzymes and metabolic pathways: conversion of succinate to propionate by *Escherichia coli*.** *Biochemistry* 2000, **39:**4622-4629.
28. Benning MM, Taylor KL, Liu RQ, Yang G, Xiang H, Wesenberg G, Dunaway-Mariano D, Holden HM: **Structure of 4-chlorobenzoyl coenzyme A dehalogenase determined to 1.8 Å resolution: an enzyme catalyst generated via adaptive mutation.** *Biochemistry* 1996, **35:**8103-8109.
29. Modis Y, Filppula SA, Novikov DK, Norledge B, Hiltunen JK, Wierenga RK: **The crystal structure of dienoyl-CoA isomerase at 1.5 Å resolution reveals the importance of aspartate and glutamate sidechains for catalysis.** *Structure* 1998, **6:**957-970.
30. Benning MM, Haller T, Gerlt JA, Holden HM: **New reactions in the crotonase superfamily: structure of methylmalonyl CoA decarboxylase from *Escherichia coli*.** *Biochemistry* 2000, **39:**4630-4639.
31. Sharma V, Suvarna K, Meganathan R, Hudspeth ME: **Menaquinone (vitamin K2) biosynthesis: nucleotide sequence and expression of the menB gene from *Escherichia coli*.** *J Bacteriol* 1992, **174:**5057-5062.
32. Pelletier DA, Harwood CS: **2-Ketocyclohexanecarboxyl coenzyme A hydrolase, the ring cleavage enzyme required for anaerobic benzoate degradation by *Rhodopseudomonas palustris*.** *J Bacteriol* 1998, **180:**2330-2336.

33. Muller-Newen G, Janssen U, Stoffel W: **Enoyl-CoA hydratase and isomerase form a superfamily with a common active-site glutamate residue.** *Eur J Biochem* 1995, **228:**68-73.
34. Gasson MJ, Kitamura Y, McLauchlan WR, Narbad A, Parr AJ, Parsons ELH, Payne J, Rhodes MJC, Walton NJ: **Metabolism of ferulic acid to vanillin. A bacterial gene of the enoyl-SCoA hydratase/isomerase superfamily encodes an enzyme for the hydration and cleavage of a hydroxycinnamic acid SCoA thioester.** *J Biol Chem* 1998, **273:**4163-4170.
35. Eichler K, Bourgis F, Buchet A, Kleber HP, Mandrand-Berthelot MA: **Molecular characterization of the cai operon necessary for carnitine metabolism in *Escherichia coli*.** *Mol Microbiol* 1994, **13:**775-786.
36. Hawes JW, Jaskiewicz J, Shimomura Y, Huang B, Bunting J, Harper ET, Harris RA: **Primary structure and tissue-specific expression of human beta- hydroxyisobutyryl-coenzyme A hydrolase.** *J Biol Chem* 1996, **271:**26430-26434.
37. Murzin AG: **How far divergent evolution goes in proteins.** *Curr Opin Struct Biol* 1998, **8:**380-387.
38. Wang J, Hartling JA, Flanagan JM: **The structure of ClpP at 2.3 Å resolution suggests a model for ATP- dependent proteolysis.** *Cell* 1997, **91:**447-456.
39. Palmer DR, Gerlt JA: **Evolution of enzymatic activities: multiple pathways for generating and partitioning a common enolic intermediate by glucarate dehydratase from *Pseudomonas putida*.** *J Am Chem Soc* 1996, **118:**10323-10324.
40. Palmer DR, Hubbard BK, Gerlt JA: **Evolution of enzymatic activities in the enolase superfamily: partitioning of reactive intermediates by (D)-glucarate dehydratase from *Pseudomonas putida*.** *Biochemistry* 1998, **37:**14350-14357.
41. Babbitt PC, Mrachko GT, Hasson MS, Huisman GW, Kolter R, Ringe D, Petsko GA, Kenyon GL, Gerlt JA: **A functionally diverse enzyme superfamily that abstracts the alpha protons of carboxylic acids.** *Science* 1995, **267:**1159-1161.
42. Wieczorek SW, Kalivoda KA, Clifton JG, Ringe D, Petsko GA, Gerlt JA: **Evolution of enzymatic activities in the enolase superfamily: identification of a "new" general acid catalyst in the active site of D-galactonate dehydratase from *Escherichia coli*.** *J Am Chem Soc* 1999, **121:**4540-4541.
43. Kwon O, Bhattacharyya DK, Meganathan R: **Menaquinone (vitamin K2) biosynthesis: overexpression, purification, and properties of o-succinylbenzoyl-coenzyme A synthetase from *Escherichia coli*.** *J Bacteriol* 1996, **178:**6778-6781.
44. Thompson TB, Garrett JB, Taylor EA, Meganathan R, Gerlt JA, Rayment I: **Evolution of enzymatic activity in the enolase superfamily: structure of o-succinylbenzoate synthase from *Escherichia coli* in complex with Mg^{2+} and o-succinylbenzoate.** *Biochemistry* 2000, **39:**10662-10676.
45. Tokuyama S, Hatano K: **Purification and properties of thermostable N-acylamino acid racemase from *Amycolatopsis* sp. TS-1-60.** *Appl Microbiol Biotechnol* 1995, **42:**853-859.
46. Wilmanns M, Priestle JP, Niermann T, Jansonius JN: **Three-dimensional structure of the bifunctional enzyme phosphoribosylanthranilate isomerase: indoleglycerolphosphate synthase from *Escherichia coli* refined at 2.0 A resolution.** *J Mol Biol* 1992, **223:**477-507.
47. Fani R, Lio P, Lazcano A: **Molecular evolution of the histidine biosynthetic pathway.** *J Mol Evol* 1995, **41:**760-774.
48. Fani R, Tamburini E, Mori E, Lazcano A, Lio P, Barberio C, Casalone E, Cavalieri D, Perito B, Polsinelli M: **Paralogous histidine biosynthetic genes: evolutionary analysis of the *Saccharomyces cerevisiae* HIS6 and HIS7 genes.** *Gene* 1997, **197:**9-17.
49. Jurgens C, Strom A, Wegener D, Hettwer S, Wilmanns M, Sterner R: **Directed evolution of a (beta alpha)8-barrel enzyme to catalyze related reactions in two different metabolic pathways.** *Proc Natl Acad Sci USA* 2000, **97:**9925-9930.
50. Radzicka A, Wolfenden R: **A proficient enzyme.** *Science* 1995, 267:90-93.
51. Harris P, Navarro Poulsen JC, Jensen KF, Larsen S: **Structural basis for the catalytic mechanism of a proficient enzyme: orotidine 5'-monophosphate decarboxylase.** *Biochemistry* 2000, **39:**4217-4224.
52. Appleby TC, Kinsland C, Begley TP, Ealick SE: **The crystal structure and mechanism of orotidine 5'-monophosphate decarboxylase.** *Proc Natl Acad Sci USA* 2000, **97:**2005-2010.
53. Wu N, Mo Y, Gao J, Pai EF: **Electrostatic stress in catalysis: structure and mechanism of the enzyme orotidine monophosphate decarboxylase.** *Proc Natl Acad Sci USA* 2000, **97:**2017-2022.
54. Miller BG, Hassell AM, Wolfenden R, Milburn MV, Short SA: **Anatomy of a proficient enzyme: the structure of orotidine 5'-monophosphate decarboxylase in the presence and absence of a potential transition state analog.** *Proc Natl Acad Sci USA* 2000, **97:**2011-2016.
55. Miller BG, Smiley JA, Short SA, Wolfenden R: **Activity of yeast orotidine-5'-phosphate decarboxylase in the absence of metals.** *J Biol Chem* 1999, **274:**23841-23843.
56. Cui W, DeWitt JG, Miller SM, Wu W: **No metal cofactor in orotidine 5'-monophosphate decarboxylase.** *Biochem Biophys Res Commun* 1999, **259:**133-135.
57. Kato N, Ohashi H, Tani Y, Ogata K: **3-Hexulosephosphate synthase from *Methylomonas aminofaciens* 77a. Purification, properties and kinetics.** *Biochim Biophys Acta* 1978, **523:**236-244.