Loss of quaternary structure is associated with rapid sequence divergence in the OSBS family

Denis Odokonyero[a], Ayano Sakai[b], Yury Patskovsky[c], Vladimir N. Malashkevich[c], Alexander A. Fedorov[c], Jeffrey B. Bonanno[c], Elena V. Fedorov[c], Rafael Toro[c], Rakhi Agarwal[d], Chenxi Wang[a], Nicole D. S. Ozerova[a], Wen Shan Yew[e], J. Michael Sauder[f], Subramanyam Swaminathan[d], Stephen K. Burley[g,h,i,j,k], Steven C. Almo[c,l], and Margaret E. Glasner[a,1]

[a]Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX 77843-2128; [b]Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801; Departments of [c]Biochemistry and [l]Physiology and Biophysics, Albert Einstein College of Medicine, Bronx, NY 10461; [d]Biosciences Department, Brookhaven National Laboratory, Upton, NY 11973; [e]Department of Biochemistry, National University of Singapore, Singapore 117597; [f]Lilly Biotechnology Center, San Diego, CA 92121; [g]BioMaPS Institute for Quantitative Biology, [h]Research Collaboratory for Structural Bioinformatics Protein Data Bank, [i]Center for Integrative Proteomics Research, [j]Rutgers Cancer Institute of New Jersey, and [k]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854-8076

Edited by Douglas L. Theobald, Brandeis University, Waltham, MA, and accepted by the Editorial Board May 2, 2014 (received for review October 3, 2013)

The rate of protein evolution is determined by a combination of selective pressure on protein function and biophysical constraints on protein folding and structure. Determining the relative contributions of these properties is an unsolved problem in molecular evolution with broad implications for protein engineering and function prediction. As a case study, we examined the structural divergence of the rapidly evolving *o*-succinylbenzoate synthase (OSBS) family, which catalyzes a step in menaquinone synthesis in diverse microorganisms and plants. On average, the OSBS family is much more divergent than other protein families from the same set of species, with the most divergent family members sharing <15% sequence identity. Comparing 11 representative structures revealed that loss of quaternary structure and large deletions or insertions are associated with the family's rapid evolution. Neither of these properties has been investigated in previous studies to identify factors that affect the rate of protein evolution. Intriguingly, one subfamily retained a multimeric quaternary structure and has small insertions and deletions compared with related enzymes that catalyze diverse reactions. Many proteins in this subfamily catalyze both OSBS and *N*-succinylamino acid racemization (NSAR). Retention of ancestral structural characteristics in the NSAR/OSBS subfamily suggests that the rate of protein evolution is not proportional to the capacity to evolve new protein functions. Instead, structural features that are conserved among proteins with diverse functions might contribute to the evolution of new functions.

enolase superfamily | protein structure | protein structure-function relationships

Investigating the causes and effects of protein sequence divergence is the key to identifying properties that enable proteins to evolve new functions. Previous studies found that constraints imposed by biophysical properties such as protein folding and stability, translational accuracy, and interactions with other proteins make a large contribution to the rate of protein evolution (1–11). However, the relative contributions of biophysical properties versus functional constraints is an open question (2). Given that the rate of protein evolution varies over several orders of magnitude, the evolutionary rate of each protein is probably determined by a unique blend of biophysical and functional constraints (12–15). Thus, the evolutionary simulations and statistical analyses of large protein datasets that comprise the primary focus of this field need to be supplemented with case studies.

Here, we present the extraordinarily diverse *o*-succinylbenzoate synthase (OSBS) family as such a case study. The OSBS family belongs to the enolase superfamily, a group of evolutionarily related protein families that have a common fold but catalyze diverse reactions (16). The rate of sequence divergence in the OSBS family is much faster than other families in the enolase superfamily. For example, the average pairwise amino acid sequence identity of OSBSs from 66 species is 26%, and the most divergent family members share <15% identity. Enzymes from the same set of species that belong to the enolase family, for which the superfamily is named, average 56% amino acid sequence identity (17). These numbers are inversely proportional to the evolutionary rate because the proteins are from the same set of species and thus have diverged for the same amount of time. However, these numbers underestimate the difference in evolutionary rate between these families, because sequence identity does not account for the occurrence of multiple mutations per site.

All enzymes in the enolase superfamily consist of a C-terminal $(\beta/\alpha)_8$-barrel linked to a capping domain composed of the N terminus and the last section of the C terminus (Fig. 1*A*). The conserved catalytic residues are in the barrel domain, and two loops from the capping domain form the rest of the active site. The only residues conserved in the whole OSBS family are short motifs surrounding the catalytic residues (17). These motifs are

Significance

The rate at which proteins accumulate amino acid substitutions during evolution depends on the likelihood that mutations will disrupt structure or affect function. Many mutations affect the ability of proteins to fold correctly, and previous studies showed that the burden imposed by misfolded proteins in cells heavily influences evolutionary rates of proteins. However, these studies could not examine the influence of function on evolutionary rates. The work described here examines the relationship between structural and functional divergence in a rapidly evolving protein family. This analysis revealed that family members that evolved a new function retained more ancestral sequence and structural characteristics, suggesting that the rate of protein evolution is not proportional to the capacity to evolve new functions.

Author contributions: M.E.G. designed research; D.O., A.S., Y.P., V.N.M., A.A.F., J.B.B., E.V.F., R.T., R.A., C.W., N.D.S.O., J.M.S., S.S., and M.E.G. performed research; A.S., W.S.Y., J.M.S., S.S., S.K.B., and S.C.A. contributed new reagents/analytic tools; D.O., A.S., Y.P., V.N.M., A.A.F., J.B.B., E.V.F., R.T., R.A., J.M.S., S.S., S.K.B., S.C.A., and M.E.G. analyzed data; and D.O., A.S., J.B.B., W.S.Y., and M.E.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. D.L.T. is a guest editor invited by the Editorial Board.

Data deposition: The atomic coordinates and structure factors have been deposited in the Protein Data Bank, www.pdb.org (PDB ID codes 2OKT, 2OLA, 3H70, 2OZT, 3H7V, 2PGE, 3CAW, 1WUE, and 1WUF).

[1]To whom correspondence should be addressed. E-mail: margy.glasner@tamu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1318703111/-/DCSupplemental.

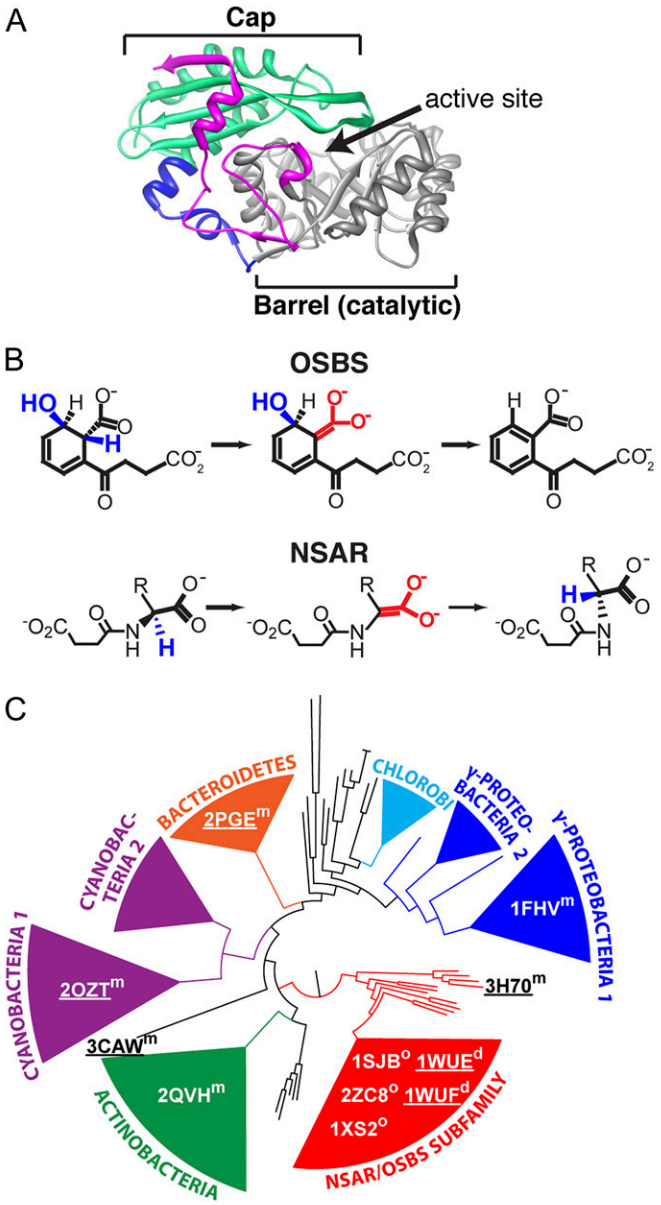

Fig. 1. (*A*) Canonical structure of enolase superfamily proteins. The catalytic barrel domain is gray, the N-terminal part of the capping domain is green, the C-terminal part of the capping domain is magenta, and the linker between the domains is blue. *Amycolatopsis* NSAR/OSBS [Protein Data Bank (PDB) ID code 1SJB] is shown. (*B*) The *o*-succinylbenzoate synthase (OSBS) and *N*-succinylamino acid racemase (NSAR) reactions. Structural similarities of the intermediates are red; blue atoms are lost or rearranged in the reactions. R, hydrophobic amino acid side chain. (*C*) Distribution of crystal structures in the OSBS family (53). Subfamilies (shown as wedges) were defined by grouping sequences whose pairwise amino acid sequence identity is >40%. Underlined PDB ID codes are structures that were determined in this study. Table 1 lists the species that encode each protein. Quaternary structure is indicated with a superscript letter (d, dimer; m, monomer; o, octamer).

also conserved in members of the enolase superfamily that catalyze other reactions, so they are not sufficient to determine specificity for OSBS activity.

We have ruled out several factors that could have contributed to the high sequence divergence of the OSBS family. First, high sequence divergence is not a general property of the enolase superfamily, as mentioned above. Second, the family's sequence diversity is not due to convergent evolution, because it has a monophyletic origin (17). Third, the OSBS family did not evolve earlier than related families, as demonstrated by comparing OSBS enzymes to paralogs from the same species (17).

Finally, sequence divergence is not due to functional divergence: most proteins in the family catalyze a conserved step in menaquinone (Vitamin K) synthesis (Fig. 1*B*). The exceptions are promiscuous enzymes that catalyze both the OSBS reaction and a second reaction, *N*-succinylamino acid racemization (NSAR), which is part of a pathway that converts D-amino acids to L-amino acids (18, 19). The NSAR/OSBS enzymes originated in a single branch of the OSBS family, the NSAR/OSBS subfamily, which also includes proteins that only have OSBS activity (Fig. 1*C*) (17, 18). Proteins within the NSAR/OSBS subfamily share >40% sequence identity, so the high sequence divergence in the OSBS family was not required to evolve the new activity.

In this work, we compared structural and sequence divergence in the OSBS family by determining the structures of a representative set of OSBS enzymes (Fig. 1*C*). The most significant difference among these enzymes is their quaternary structure. All OSBS enzymes, except those in the NSAR/OSBS subfamily, are monomers. Monomeric OSBS enzymes accumulated insertions, deletions, and other mutations that caused them to diverge from each other and the rest of the enolase superfamily. In contrast, proteins in the NSAR/OSBS subfamily are multimeric, like nearly all other members of the enolase superfamily. Because of structural constraints imposed by their quaternary structure, the sequences and structures of proteins in the NSAR/OSBS subfamily are much more similar to other members of the enolase superfamily than they are to other OSBS enzymes. Thus, structural divergence is associated with the high evolutionary rate of the OSBS family, whereas functional divergence in the NSAR/OSBS subfamily is associated with retention of ancestral structural characteristics.

Results

Comparison of Activities in the OSBS Family. This study analyzes a representative set of enzymes from the OSBS family whose pairwise amino acid sequence identity is <20%. Previously, we determined that these divergent enzymes belong to the OSBS family based on phylogeny and/or the presence of their genes in menaquinone synthesis gene clusters (17). We verified their activities by enzymatic assays (Table 1 and Table S1). All family members had similar catalytic efficiencies for the OSBS reaction ($k_{cat}/K_M = 10^5$ to 10^6 $M^{-1}\cdot s^{-1}$).

NSAR activity was only detected in the NSAR/OSBS subfamily. The two previously uncharacterized members of the NSAR/OSBS subfamily also catalyze the NSAR reaction, like other members of this subfamily (20). *Enterococcus faecalis* NSAR/OSBS is encoded in a menaquinone synthesis operon, indicating that OSBS is its biological function (17). A pathway that requires NSAR activity has not been identified in this species, so whether NSAR is also a biological activity is unknown. *Listeria innocua* has the menaquinone synthesis pathway, indicating that the species requires OSBS activity. However, the NSAR/OSBS is not encoded in the menaquinone operon, raising the possibility that both NSAR and OSBS are biological functions, as observed in the NSAR/OSBS from *Geobacillus kaustophilus* (17, 19).

Quaternary Structure of OSBS Enzymes. Crystal structures of OSBS family members from 12 species have been determined, including 6 reported in this work (Table 1). All of them have the canonical enolase superfamily fold, but their quaternary structures are not conserved, as determined from crystal packing and size exclusion chromatography (Fig. S1). Like most other members of the enolase superfamily, the five enzymes from the NSAR/OSBS subfamily are multimers (21–40). The three previously characterized NSAR/OSBS subfamily enzymes are

Table 1. **Enzymatic activity and quaternary structure of enzymes in the OSBS family**

		OSBS	NSAR*		
Species	Subfamily	k_{cat}/K_M, $M^{-1}\cdot s^{-1}$	k_{cat}/K_M, $M^{-1}\cdot s^{-1}$	ID code(s)	Quaternary structure
Escherichia coli	γ-Proteobacteria 1	2.0 x 10^{6}[†]	n.a.[‡]	1FHV (21)	Monomer
Desulfotalea psychrophila	Bacteroidetes	1.1 x 10^{6}	n.a.	2PGE	Monomer
Thermosynechococcus elongatus	Cyanobacteria 1	1.0 x 10^{6}	n.a.	3H7V, 2OZT	Monomer
Bdellovibrio bacteriovorus	Not assigned	3.1 x 10^{5}	n.a.	3CAW	Monomer
Thermobifida fusca	Actinobacteria	6.7 x 10^{5}[§]	n.a.	2QVH, 2OPJ (52)	Monomer
Staphylococcus aureus	Not assigned	1.1 x 10^{6}	n.a.	3H70, 2OKT, 2OLA	Monomer
Amycolatopsis sp. T-1–60	NSAR/OSBS	2.5 x 10^{5}[¶]	2.0 x 10^{5}	1SJB (22)	Octamer
Deinococcus radiodurans	NSAR/OSBS	3.1 x 10^{5}	3.7 x 10^{5}	1XS2 (24)	Octamer
Thermus thermophilus	NSAR/OSBS	6.5 x 10^{5}	7.5 x 10^{4}	2ZC8 (23)	Octamer
Enterococcus faecalis	NSAR/OSBS	1.6 x 10^{6}	1.4 x 10^{5}	1WUE	Dimer
Listeria innocua	NSAR/OSBS	2.9 x 10^{6}	2.6 x 10^{3}	1WUF	Dimer[‖]

Table S1 lists all kinetic parameters.

[*] *N*-Succinyl-L-phenylglycine was the substrate.

[†] OSBS activity was measured in ref. 53.

[‡] n.a., not active. NSAR activity was measured using 10 μM enzyme and 20 mM *N*-succinyl-L-phenylglycine.

[§] OSBS activity was measured in ref. 52.

[¶] OSBS and NSAR activity were measured in refs. 18 and 20.

[‖] On size exclusion chromatography, it primarily elutes as a dimer, although it also has a significant monomer peak (Fig. S1).

octamers, and the NSAR/OSBS subfamily enzymes from *E. faecalis* and *L. innocua* are dimers (22–24). In contrast, the OSBSs from other subfamilies are all monomers.

Structural Comparison of OSBS Monomers. We compared OSBS family structures to 51 other members of the enolase superfamily using TM-align (Fig. 2) (41). The TM score was used because it considers both RMSD between aligned residues and coverage (fraction of residues in the proteins that were aligned) (42). The TM score is 1 for identical structures, >0.5 for structures that have the same fold, and <0.2 for unrelated structures. As expected from sequence divergence between OSBS subfamilies, structural divergence between OSBS subfamilies is much higher than the divergence within the NSAR/OSBS subfamily (columns 1 and 2 versus column 3 in Fig. 2*A*).

Given that a new function evolved in the NSAR/OSBS subfamily, one might expect that structures from this subfamily would have diverged from the rest of the enolase superfamily as much or more than other OSBS enzymes. The opposite is true: proteins in the NSAR/OSBS subfamily are more similar to proteins from other families in the enolase superfamily than to other members of the OSBS family (columns 4–6 in Fig. 2*A*). The other OSBS subfamilies have diverged both from each other and from the rest of the enolase superfamily.

To determine which parts of the structure have diverged the most, we analyzed the barrel and capping domains separately. The structural divergence of the barrel domain is similar to the full-length protein (Fig. 2*B*). However, the capping domain is much more divergent, both within the OSBS family and among other members of the enolase superfamily (Fig. 2*C*). Restricting the analysis to structures bound to substrate or product analogs produced the same result, indicating that divergence of the capping domain is not an artifact from comparing apo- and ligand-bound structures.

Subdividing the capping domain into smaller regions showed that the linker between the capping and barrel domains and the C-terminal portion of the capping domain have extremely low TM scores (~0.3 within the OSBS family compared with ~0.45 for the whole enolase superfamily; Table S2 and Fig. 1*A*). The conformation of the linker in the NSAR/OSBS subfamily is similar to that of other enolase superfamily members (Fig. 3). In other OSBS subfamilies, deletions in the linker resulted in loss of a short helix and an extended conformation. The length of the C-terminal section of the capping domain is especially variable, and extensions at the C terminus lie in a variety of directions relative to the rest of the protein (Fig. S2).

Insertions and Deletions in the OSBS Family. Insertions and deletions (indels) are mainly responsible for structural divergence of the capping domain in the OSBS family (Fig. 4). The average

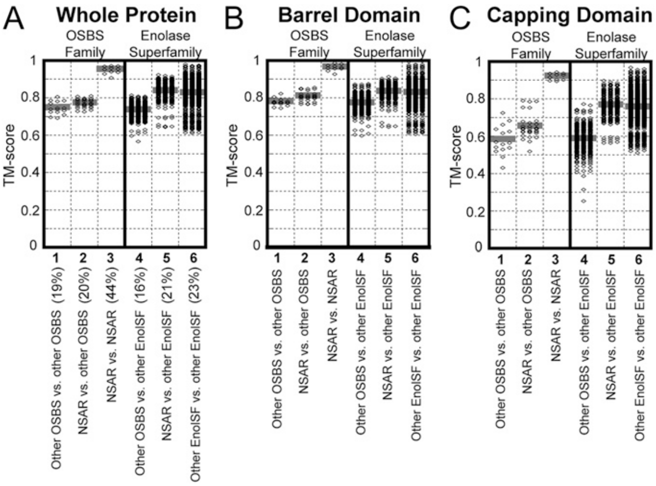

Fig. 2. Structural divergence of the OSBS family comparing (*A*) full-length proteins, (*B*) barrel domains, or (*C*) capping domains. Each point represents the TM score of a pair of proteins. The gray bars are the average TM score of each set. The average percentage sequence identity (number of identical residues divided by the length of the structural alignment) are shown in *A* to illustrate that the sequence divergence in the OSBS family is similar to the divergence in the whole superfamily. Because calculated percentage identity varies by several percent depending on how the sequences are aligned, the difference between 16% and 23% might not be significant (60). The proteins compared in each column are as follows: (1) OSBS family structures, excluding the NSAR/OSBS subfamily; (2) NSAR/OSBS subfamily structures compared with OSBSs from other subfamilies; (3) NSAR/OSBS subfamily structures, excluding other OSBS subfamilies; (4) OSBS family structures, excluding the NSAR/OSBS subfamily, compared with proteins from other families in the enolase superfamily; (5) NSAR/OSBS subfamily structures compared with proteins from other families in the enolase superfamily; (6) structures from other families in the enolase superfamily, excluding the OSBS family.

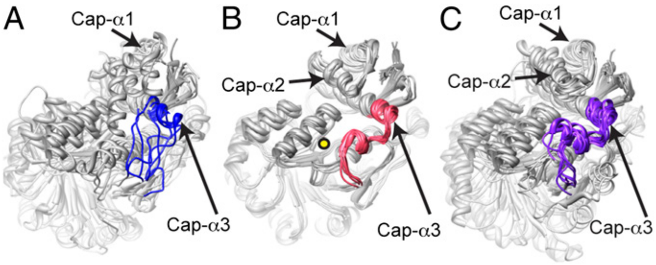

Fig. 3. Structural divergence of the linker between the barrel and capping domains. (*A*) OSBS proteins, excluding the NSAR/OSBS subfamily (blue; PDB ID codes 1FHV, 2OKT, 2OZT, 2PGE, 2QVH, and 3CAW). (*B*) The NSAR/OSBS subfamily (pink; PDB ID codes 1SJB, 1WUE, 1WUF, 1XS2, and 2ZC8). (*C*) Other members of the enolase superfamily (purple; PDB ID codes 1EBG, 1EC8, 1KKR, 2PMQ, 2QJN, 1TKK, 2MNR, and 3DGB). The entrance to the active site is behind the structures and is marked with a yellow circle in *B*.

number and length of indels in the enolase superfamily are 7.5 and 4.0, respectively (Table S3). Similarly, the monomeric OSBSs have 8.8 indels that are 4.8 residues long, on average. Although the average number of indels in the NSAR/OSBS subfamily is similar (8.6), the average length (1.4) is much shorter. Within OSBS subfamilies, the positions of most indels are conserved, although the length can vary.

Most indels in the capping domain are distant from the active site. The second α-helix of the N-terminal capping domain is missing or truncated in four OSBSs (*Escherichia coli*, *Thermosynechococcus elongatus*, *Bdellovibrio bacteriovorus*, and *Thermobifida fusca*), but it is present in most other members of the enolase superfamily (Figs. 3 and 4). Both the first and second α-helices are deleted in *T. fusca* OSBS. In *Desulfotalea psychrophila* OSBS, another helix is inserted after the second α-helix. This insertion is uncommon in the Bacteroidetes subfamily, and enzymes without the insertion lack both the first and second α-helices of the capping domain. Deletions also occur in the linker between the capping and barrel domains in many OSBSs, which is at the end of the third α-helix. Strikingly, the positions of indels in the capping domain helices and linker in monomeric OSBSs are at the interface between subunits in NSAR/OSBS enzymes and other multimeric enolase superfamily members (Fig. 5).

Discussion

Our study highlights two structural features that affected the rate of protein evolution in the OSBS family: quaternary structure and indels. Most studies to identify factors that affect the rate of protein evolution have not considered these features because they used large, multifamily datasets that lack experimental information about quaternary structure or accurate alignments to determine positions of indels. As a result, case studies on model systems like the OSBS family provide critical insights into factors affecting the structural and functional evolution of proteins. The large insertions and deletions in the capping domains of several OSBSs could have accelerated the evolutionary rate of amino acid substitutions to compensate for structural perturbations. Indeed, the evolutionary rates of inserted residues and residues flanking deletions are higher than expected based on their solvent accessibility in several monomeric OSBS subfamilies (Fig. S3*A*). This result agrees with previous studies that found higher mutation rates near the sites of indels (43, 44).

Previous studies also did not consider the role of homomeric quaternary structure in determining evolutionary rates. However, several studies have shown that protein–protein interactions decrease evolutionary rates, partly by decreasing the fraction of surface-exposed residues (5–9, 45). Likewise, interactions with large capping domains in the haloalkanoate dehalogenase superfamily constrain the structural divergence of their Rossman-fold core domain (46). Our observation that the OSBS family, which is primarily composed of monomers, evolved at a faster rate than related, homomultimeric families is consistent with these studies. These studies would also suggest that buried residues at the interface between subunits in NSAR/OSBS enzymes would have slower evolutionary rates than homologous, solvent-exposed residues in monomeric OSBS enzymes. Our results offer some support for this idea, but only a small number of sites fit these criteria (Fig. S3*B*).

Other studies have calculated frequencies of deletions in protein superfamilies and assessed their effect on functional divergence. Reeves et al. (47) reported that the average indel length of proteins with <20% identity is 6.6, which declines to 3.5 for proteins having 20–40% sequence identity. The OSBS family is in the middle of this range, with an average indel length of 4.3 among proteins that have ~20% sequence identity. Previous studies also noted a steep decline in both structural and functional similarity below ~30% sequence identity (47, 48). However,

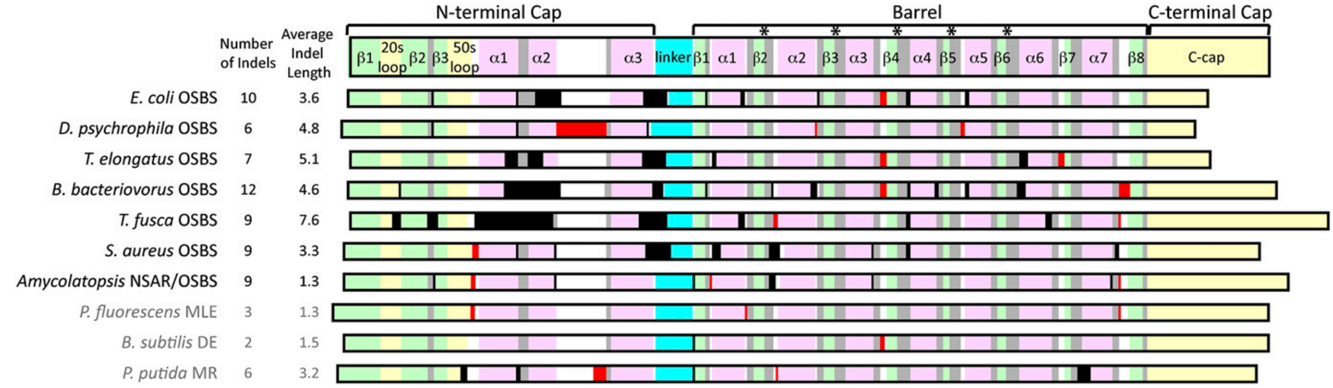

Fig. 4. Minimum number of insertions and deletions (indels) in OSBS enzymes. Seven OSBS family members (black) are compared with three other members of the enolase superfamily (gray): muconate lactonizing enzyme (MLE) from *Pseudomonas fluorescens* (PDB ID code 3DGB), dipeptide epimerase (DE) from *Bacillus subtilis* (PDB ID code 1TKK), and mandelate racemase (MR) from *Pseudomonas putida* (PDB ID code 2MNR). The first schematic shows the typical secondary structure of enolase superfamily proteins, with green representing β-sheets, pink representing α-helices, and yellow, cyan, and gray representing loops and linkers. Deletions are black, and insertions are red. White regions are gaps that align with insertions in other sequences. The length of each colored segment is proportional to the number of amino acids. Asterisks indicate the positions of the conserved catalytic residues. The total number of indels listed excludes length heterogeneity at the N and C termini.

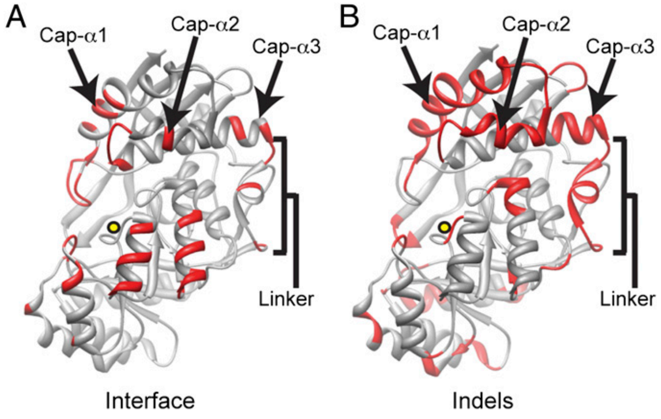

Fig. 5. Many deletions are located at lost subunit interfaces. (*A*) Residues that are at subunit interfaces in the octameric *Amycolatopsis* NSAR/OSBS (1SJB) are red. (*B*) Positions at which insertions or deletions occur in monomeric OSBS enzymes from other subfamilies are red. The active site is marked with a yellow circle.

OSBS activity has been conserved despite divergence of the tertiary and quaternary structure. Counterintuitively, functional divergence occurred in the subfamily that retained the most structural similarity to functionally diverse members of the enolase superfamily.

Analyzing the structures of 11 OSBS enzymes revealed that absence of quaternary structure is associated with the structural and sequence diversification of the OSBS family. The fact that nearly all other proteins in the enolase superfamily are multimers suggests that the common ancestor of the OSBS family was also a multimer. If so, loss of quaternary structure permitted the extreme structural and sequence divergence seen in the OSBS family. This scenario is supported by rooting the phylogenetic tree using closely related families as the outgroup (17, 49). The root falls between the NSAR/OSBS subfamily and the other OSBS subfamilies, suggesting that quaternary structure was lost once (Fig. 1*C*). The alternative scenario is that the root falls within the OSBS family. If so, an ancestral, monomeric OSBS would have given rise to homomultimeric descendants that subsequently experienced functional divergence to give rise to proteins with OSBS, NSAR, dipeptide epimerase, muconate cycloisomerase, and other activities. We cannot exclude this possibility because of challenges associated with rooting the phylogeny of paralogous proteins. However, it is less parsimonious than invoking loss of quaternary structure as the driving force for divergence of the OSBS family.

Is loss of quaternary structure sufficient to explain the extreme sequence divergence of monomeric OSBSs? The uncatalyzed rate of the OSBS reaction is 1,000 times faster than the uncatalyzed rate of mandelate racemization, a reaction catalyzed by a related family (50, 51). Authors of these studies suggested that the relatively high uncatalyzed rate of the OSBS reaction might be associated with greater tolerance of mutations and thus a higher evolutionary rate (50). To date, our data do not support this idea, although experiments have been limited to a small number of active-site residues. Mutations of active-site residues have similar effects in *E. coli* OSBS, *T. fusca* OSBS, and *P. putida* mandelate racemase (MR), reducing k_{cat}/K_M by ∼10- to 500-fold (52–55).

Instead, our data suggest a model in which the active sites of OSBS enzymes diverged to compensate for (or were permitted to diverge by) mutations that affected the structure outside the active site, such as deletions at former subunit interfaces. Given the large structural changes associated with loss of quaternary structure and indels, the divergence of OSBS enzymes is probably irreversible. Indeed, mutagenesis to swap amino acids at homologous positions in *E. coli* and *T. fusca* OSBS enzymes was deleterious (52). This is similar to the observed mutational epistasis in other proteins, such as the glucocorticoid receptors, although structure, not specificity, has diverged among monomeric OSBSs (56).

The only members of the OSBS family that have NSAR activity are in the NSAR/OSBS subfamily, which includes both promiscuous NSAR/OSBS enzymes and enzymes that catalyze only the OSBS reaction. Remarkably, enzymes in the NSAR/OSBS subfamily are more similar to members of the enolase superfamily that have diverse functions than they are to proteins in other OSBS subfamilies. This raises the possibility that retention of ancestral sequence and structural features contributed to the evolution of NSAR activity.

This idea contrasts with the concept of designability as proposed by England and Shakhnovich (57). They define designability as the number of sequences that are capable of folding into a specific topology below a certain energy threshold. Bloom et al. (58) related designability to high rates of protein evolution, which are enhanced due to structural features such as higher densities of interresidue contacts. This concept might be useful when considering designing protein structures, but it may not be applicable to designing new protein functions. Instead, our results show that family members that evolved a new function retained more ancestral sequence and structural characteristics, suggesting that the rate of protein evolution is not proportional to the capacity to evolve new functions.

Materials and Methods

Biochemical Methods. Genes for OSBS family enzymes from *Staphylococcus aureus* (menC), *T. elongatus* (Tlr1174), *D. psychrophila* (DP0251), *B. bacteriovorus* (Bd0547), *E. faecalis* (EF0450), and *L. innocua* (lin2664) were cloned into N- or C-terminal His-tag vectors for protein expression and purification. Detailed methods for protein production, structure determination, size exclusion chromatography, and catalytic activity assays are in *SI Materials and Methods* (Table S4).

Mapping Insertions and Deletions. To accurately map insertions and deletions, 62 proteins from the enolase superfamily were aligned using University of California San Francisco Chimera (59). The alignment was manually refined based on visual inspection of the structural alignment. Positions of insertions and deletions were determined by comparing each protein to the consensus of the structural alignment. Ideally, the consensus would represent the ancestral structure of the enolase superfamily. However, the lengths of some regions are heterogenous throughout the enolase superfamily, making it difficult to determine the ancestral state. The C-terminal section of the capping domain has additional indels, but they were not enumerated because this region is difficult to align (Fig. S2). To determine the number of indels in each sequence, indels that were separated by more than five residues were considered a single indel, to account for inaccuracies in the alignment. Also, indels longer than one residue could represent multiple insertion and deletion events. Consequently, Fig. 4 and Table S3 report the minimum number of indels.

Other Bioinformatics Methods. Detailed procedures for automated structural alignment and calculation of evolutionary rates are in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We acknowledge the efforts of all New York SGX Research Center for Structural Genomics (NYSGXRC) and New York Structural Genomics Research Consortium (NYSGRC) personnel who contributed to the structure determination and manuscript preparation. We thank Dr. Larry Dangott of the Protein Chemistry Laboratory at Texas A&M University for assistance with size exclusion chromatography. We thank Jacob Jipp and Christine Jones for assistance with subcloning and Dr. Andrew McMillan for insightful comments on the manuscript. This work was supported by Grant A-1758 from the Robert A. Welch Foundation and Grant 000517-0016-2011 from the Norman Hackerman Advanced Research Program (to M.E.G.). The NYSGXRC was supported by National Institutes of Health (NIH) Grant U54 GM074945 (to S.K.B.). The NYSGRC was supported by NIH Grant U54 GM094662 (to S.C.A.). Diffraction data were collected at the National Synchrotron Light Source, Brookhaven National Laboratory [supported by the US Department of Energy (DOE), Office of Science, Office of Basic Energy Sciences, under Contract DE-AC02-98CH10886]: beamlines X29A and X12C (supported by Grant P30-EB-009998 from the National Institute of Biomedical Imaging and Bioengineering to The Center for Synchrotron Biosciences, and the NIH National Center for Research

Resources and the DOE Office of Biological and Environmental Research to the National Synchrotron Light Source Research Resource for Macromolecular Crystallography); beamline X9A (supported by Rockefeller University, Albert Einstein College of Medicine, and Sloan Kettering Institute); and beamline X4A (supported by the New York Structural Biology Center). Diffraction data were also collected at the Advanced Photon Source, Argonne National Laboratory (supported by the DOE, Office of Science, Office of Basic Energy Sciences, under Contract DE-AC02-06CH11357): LRL-CAT 31-ID-D beamline (data collection at sector 31 of the Advanced Photon Source was provided by Eli Lilly, which operates the facility).

1. Pál C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. *Nat Rev Genet* 7(5):337–348.
2. Wilke CO, Drummond DA (2010) Signatures of protein biophysics in coding sequence evolution. *Curr Opin Struct Biol* 20(3):385–389.
3. Serohijos AW, Rimas Z, Shakhnovich EI (2012) Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Rep* 2(2):249–256.
4. Lobkovsky AE, Wolf YI, Koonin EV (2010) Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proc Natl Acad Sci USA* 107(7): 2983–2988.
5. Yang JR, Liao BY, Zhuang SM, Zhang J (2012) Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc Natl Acad Sci USA* 109(14):E831–E840.
6. Franzosa EA, Xia Y (2009) Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol* 26(10):2387–2395.
7. Mintseris J, Weng Z (2005) Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci USA* 102(31):10930–10935.
8. Kim PM, Lu LJ, Xia Y, Gerstein MB (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* 314(5807):1938–1941.
9. Eames M, Kortemme T (2007) Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance. *Structure* 15(11):1442–1451.
10. Yang JR, Zhuang SM, Zhang J (2010) Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol Syst Biol* 6:421.
11. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134(2):341–352.
12. Grishin NV, Wolf YI, Koonin EV (2000) From complete genomes to measures of substitution rate variability within and between proteins. *Genome Res* 10(7):991–1000.
13. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA* 102(40):14338–14343.
14. Wall DP, et al. (2005) Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci USA* 102(15):5483–5488.
15. Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ (2009) The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci USA* 106(18):7273–7280.
16. Gerlt JA, Babbitt PC, Rayment I (2005) Divergent evolution in the enolase superfamily: The interplay of mechanism and specificity. *Arch Biochem Biophys* 433(1):59–70.
17. Glasner ME, et al. (2006) Evolution of structure and function in the *o*-succinylbenzoate synthase/*N*-acylamino acid racemase family of the enolase superfamily. *J Mol Biol* 360(1):228–250.
18. Palmer DR, et al. (1999) Unexpected divergence of enzyme function and sequence: "*N*-Acylamino acid racemase" is *o*-succinylbenzoate synthase. *Biochemistry* 38(14): 4252–4258.
19. Sakai A, et al. (2006) Evolution of enzymatic activities in the enolase superfamily: *N*-Succinylamino acid racemase and a new pathway for the irreversible conversion of D- to L-amino acids. *Biochemistry* 45(14):4455–4462.
20. Taylor Ringia EA, et al. (2004) Evolution of enzymatic activity in the enolase superfamily: Functional studies of the promiscuous *o*-succinylbenzoate synthase from *Amycolatopsis*. *Biochemistry* 43(1):224–229.
21. Thompson TB, et al. (2000) 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* 39(35):10662–10676.
22. Thoden JB, et al. (2004) Evolution of enzymatic activity in the enolase superfamily: Structural studies of the promiscuous *o*-succinylbenzoate synthase from *Amycolatopsis*. *Biochemistry* 43(19):5716–5727.
23. Hayashida M, Kim SH, Takeda K, Hisano T, Miki K (2008) Crystal structure of *N*-acylamino acid racemase from *Thermus thermophilus* HB8. *Proteins* 71(1):519–523.
24. Wang WC, et al. (2004) Structural basis for catalytic racemization and substrate specificity of an *N*-acylamino acid racemase homologue from *Deinococcus radiodurans*. *J Mol Biol* 342(1):155–169.
25. Gulick AM, Schmidt DM, Gerlt JA, Rayment I (2001) Evolution of enzymatic activities in the enolase superfamily: Crystal structures of the L-Ala-D/L-Glu epimerases from *Escherichia coli* and *Bacillus subtilis*. *Biochemistry* 40(51):15716–15724.
26. Helin S, Kahn PC, Guha BL, Mallows DG, Goldman A (1995) The refined X-ray structure of muconate lactonizing enzyme from *Pseudomonas putida* PRS2000 at 1.85 A resolution. *J Mol Biol* 254(5):918–941.
27. Kajander T, Lehtiö L, Schlömann M, Goldman A (2003) The structure of *Pseudomonas* P51 Cl-muconate lactonizing enzyme: Co-evolution of structure and dynamics with the dehalogenation function. *Protein Sci* 12(9):1855–1864.
28. Klenchin VA, Schmidt DM, Gerlt JA, Rayment I (2004) Evolution of enzymatic activities in the enolase superfamily: Structure of a substrate-liganded complex of the L-Ala-D/L-Glu epimerase from *Bacillus subtilis*. *Biochemistry* 43(32):10370–10378.
29. Song L, et al. (2007) Prediction and assignment of function for a divergent *N*-succinyl amino acid racemase. *Nat Chem Biol* 3(8):486–491.
30. Kalyanaraman C, et al. (2008) Discovery of a dipeptide epimerase enzymatic function guided by homology modeling and virtual screening. *Structure* 16(11):1668–1677.
31. Gulick AM, Hubbard BK, Gerlt JA, Rayment I (2000) Evolution of enzymatic activities in the enolase superfamily: Crystallographic and mutagenesis studies of the reaction catalyzed by D-glucarate dehydratase from *Escherichia coli*. *Biochemistry* 39(16): 4590–4602.
32. Yew WS, et al. (2006) Evolution of enzymatic activities in the enolase superfamily: L-Fuconate dehydratase from *Xanthomonas campestris*. *Biochemistry* 45(49):14582–14597.
33. Yew WS, et al. (2006) Evolution of enzymatic activities in the enolase superfamily: D-Tartrate dehydratase from *Bradyrhizobium japonicum*. *Biochemistry* 45(49): 14598–14608.
34. Neidhart DJ, et al. (1991) Mechanism of the reaction catalyzed by mandelate racemase. 2. Crystal structure of mandelate racemase at 2.5-A resolution: Identification of the active site and possible catalytic residues. *Biochemistry* 30(38):9264–9273.
35. Yew WS, Fedorov AA, Fedorov EV, Almo SC, Gerlt JA (2007) Evolution of enzymatic activities in the enolase superfamily: L-Talarate/galactarate dehydratase from *Salmonella typhimurium* LT2. *Biochemistry* 46(33):9564–9577.
36. Rakus JF, et al. (2009) Computation-facilitated assignment of the function in the enolase superfamily: A regiochemically distinct galactarate dehydratase from *Oceanobacillus iheyensis*. *Biochemistry* 48(48):11546–11558.
37. Rakus JF, et al. (2008) Evolution of enzymatic activities in the enolase superfamily: L-Rhamnonate dehydratase. *Biochemistry* 47(38):9944–9954.
38. Rakus JF, et al. (2007) Evolution of enzymatic activities in the enolase superfamily: D-Mannonate dehydratase from *Novosphingobium aromaticivorans*. *Biochemistry* 46(45):12896–12908.
39. Levy CW, et al. (2002) Insights into enzyme evolution revealed by the structure of methylaspartate ammonia lyase. *Structure* 10(1):105–113.
40. Wedekind JE, Poyner RR, Reed GH, Rayment I (1994) Chelation of serine 39 to Mg^{2+} latches a gate at the active site of enolase: Structure of the bis(Mg^{2+}) complex of yeast enolase and the intermediate analog phosphonoacetohydroxamate at 2.1-A resolution. *Biochemistry* 33(31):9333–9342.
41. Zhang Y, Skolnick J (2005) TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33(7):2302–2309.
42. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. *Proteins* 57(4):702–710.
43. Zhang Z, Huang J, Wang Z, Wang L, Gao P (2011) Impact of indels on the flanking regions in structural domains. *Mol Biol Evol* 28(1):291–301.
44. Tian D, et al. (2008) Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* 455(7209):105–108.
45. Lin YS, Hsu WL, Hwang JK, Li WH (2007) Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Mol Biol Evol* 24(4):1005–1011.
46. Pandya C, et al. (2013) Consequences of domain insertion on sequence-structure divergence in a superfold. *Proc Natl Acad Sci USA* 110(36):E3381–E3387.
47. Reeves GA, Dallman TJ, Redfern OC, Akpor A, Orengo CA (2006) Structural diversity of domain superfamilies in the CATH database. *J Mol Biol* 360(3):725–741.
48. Sandhya S, et al. (2009) Length variations amongst protein domain superfamilies and consequences on structure and function. *PLoS One* 4(3):e4981.
49. Sakai A, et al. (2009) Evolution of enzymatic activities in the enolase superfamily: Stereochemically distinct mechanisms in two families of *cis,cis*-muconate lactonizing enzymes. *Biochemistry* 48(7):1445–1453.
50. Taylor EA, Palmer DR, Gerlt JA (2001) The lesser "burden borne" by *o*-succinylbenzoate synthase: An "easy" reaction involving a carboxylate carbon acid. *J Am Chem Soc* 123(24):5824–5825.
51. Bearne SL, Wolfenden R (1997) Mandelate racemase in pieces: Effective concentrations of enzyme functional groups in the transition state. *Biochemistry* 36(7): 1646–1656.
52. Odokonyero D, et al. (2013) Divergent evolution of ligand binding in the *o*-succinylbenzoate synthase family. *Biochemistry* 52(42):7512–7521.
53. Zhu WW, et al. (2012) Residues required for activity in *Escherichia coli* *o*-succinylbenzoate synthase (OSBS) are not conserved in all OSBS enzymes. *Biochemistry* 51(31):6171–6181.
54. Bourque JR, Bearne SL (2008) Mutational analysis of the active site flap (20s loop) of mandelate racemase. *Biochemistry* 47(2):566–578.
55. Siddiqi F, et al. (2005) Perturbing the hydrophobic pocket of mandelate racemase to probe phenyl motion during catalysis. *Biochemistry* 44(25):9013–9021.
56. Bridgham JT, Ortlund EA, Thornton JW (2009) An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* 461(7263):515–519.
57. England JL, Shakhnovich EI (2003) Structural determinant of protein designability. *Phys Rev Lett* 90(21):218101.
58. Bloom JD, Drummond DA, Arnold FH, Wilke CO (2006) Structural determinants of the rate of protein evolution in yeast. *Mol Biol Evol* 23(9):1751–1761.
59. Pettersen EF, et al. (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem* 25(13):1605–1612.
60. Raghava GP, Barton GJ (2006) Quantification of the variation in percentage identity for protein sequence alignments. *BMC Bioinformatics* 7:415.