OPEN ACCESS

BIO-complexity

Research Article

Enzyme Families–Shared Evolutionary History or Shared Design? A Study of the GABA-Aminotransferase Family

Mariclair A. Reeves, Ann K. Gauger,* and Douglas D. Axe*

Biologic Institute, Redmond, WA, USA

Abstract

The functional diversity of enzyme families is thought to have been caused by repeated *recruitment* events—gene duplications followed by conversions to new functions. However, mathematical models show this can only work if beneficial new functions are achievable by just one or two base changes in the duplicate genes. Having found no convincing demonstration that this is feasible, we previously chose a highly similar pair of *E. coli* enzymes from the GABA-aminotransferase-like (GAT) family, 2-amino-3-ketobutyrate CoA ligase ($KbiI_2$) and 8-amino-7-oxononanoate synthase ($BioF_2$), and attempted to convert the first to perform the function of the second by site-directed mutagenesis. In the end we were unable to achieve functional conversion by that rational approach. Here we take a complementary approach based on random mutagenesis. Focusing first on single mutations, we prepared mutated libraries of nine genes from the GAT family and tested for $BioF_2$ function in vivo. None of the singly mutated genes had this function. Focusing next on double mutations, we prepared and tested 70% of the 6.5 million possible mutation *pairs* for $KbiI_2$ and for BIKB, an enzyme described as having both $KbiI_2$ and $BioF_2$ activities in vitro. Again, no $BioF_2$ activity was detected in vivo. Based on these results, we conclude that conversion to $BioF_2$ function would require at least two changes in the starting gene and probably more, since most double mutations do not work for two promising starting genes. The most favorable recruitment scenario would therefore require three genetic changes after the duplication event: two to achieve low-level $BioF_2$ activity and one to boost that activity by overexpression. But even this best case would require about 10^{15} years in a natural population, making it unrealistic. Considering this along with the whole body of evidence on enzyme conversions, we think structural similarities among enzymes with distinct functions are better interpreted as supporting shared design principles than shared evolutionary histories.

Cite as: Reeves MA, Gauger AK, Axe DD (2014) Enzyme families—Shared evolutionary history or shared design? A study of the GABA-aminotransferase family. BIO-Complexity 2014 (4):1–16. **doi:**10.5048/BIO-C.2014.4.

Editor: Matti Leisola

Received: July 30, 2014; **Accepted:** October 23, 2014; **Published:** December 1, 2014

Notes: A *Critique* of this paper, when available, will be assigned **doi:**10.5048/BIO-C.2014.4.c.

*email: agauger@biologicinstitute.org, daxe@biologicinstitute.org

INTRODUCTION

It is commonly assumed that because proteins share sequence and structural similarity and can be grouped hierarchically into families and superfamilies based on that similarity, their evolutionary history is plain [1–6]. For many years, protein engineers reasoned from this that it should be relatively easy to shift enzymes to new functions, but experience has forced them to rethink this. In the words of two prominent protein engineers: "many attempts at interchanging activities in mechanistically diverse superfamilies have since been attempted, but few successes have been realized." [7] The results of these projects present two challenges for the Darwinian explanation of protein origins. First, when functional conversions are found to be possible in the laboratory, they almost always require many mutations [8–12], which would make them unfeasible as natural occurrences. Second, the conversions achieved in the laboratory are invariably very weak [7,10–15]. For example, studying a natural bacterial aspartate decarboxylase, Wilson and Kornberg [13] measured a K_m of 80 μM for L-aspartate and a maximal reaction rate (V_{max}) of 5.3×10^3 moles of aspartate decarboxylated per minute per mole of active-site PLP (pyridoxal-5′-phosphate), corresponding to a k_{cat} value of $88s^{-1}$.

The three amino acid substitutions that Graber et al. [10] describe as converting aspartate aminotransferase into an aspartate decarboxylase require seven nucleotide substitutions to achieve an activity that is some 100,000-fold lower, based on k_{cat}/K_m. Very weak activities like this must be amplified in order for them to have any selective effect [7,14–16], which means they could easily confer a selective *disadvantage* in a natural

setting when the metabolic cost of overexpression is taken into account [17,18].

While new enzyme chemistry has proven very difficult to obtain, shifting the substrate or reaction preference of an enzyme toward a minor activity it already possesses is much easier. Many studies have demonstrated, for example, that one or two mutations can make certain enzymes favor a previously minor substrate or reaction [6,19,20], and the relative ease of these transitions has led to speculation that they may have been important for the origins of enzyme diversity. The favored idea is that ancient enzymes were *promiscuous*, catalyzing side reactions in addition to their beneficial primary reaction [1,6,15,20]. With changing circumstances, some of these side reactions could have proven useful, which might have led to their protein catalysts being recruited to specialize in those functions. Duplication of the encoding gene is thought to have provided the genetic material for this process of evolutionary specialization to proceed without loss of the original function.

This idea of functional divergence from promiscuous ancestral enzymes, though possibly of some value, leaves the most fundamental aspects of the origins question unchanged. Whether ancient enzymes resembled modern ones closely or only loosely, the modern ones are both the things to be explained and the things we can observe. The importance of grounding evolutionary explanations in actual study of modern enzymes therefore remains as important now as ever. Yet the promiscuity hypothesis leaves the actual origin of genuinely new enzyme functions unexplained. In the end, then, functional conversion by mutation is still the only evolutionary explanation for those first appearances.

Because of this, we focus here on the feasibility of the classical recruitment scenario where enzymes adopt an entirely new function. The process begins when a gene duplication event provides a spare gene encoding a structurally sound protein that need not perform any function. It is expected that the cost of carrying this superfluous gene, including possible effects on the expression of neighboring genes, would be at least a mild selective disadvantage to the host cell, but the idea is that on rare occasions mutations may cause the spare gene to provide a useful new function before purifying selection takes its course. If the new benefit outweighs the cost, then the net effect would be a selective advantage.

This scenario has some difficulties, however. Recognizing that conversions to new enzyme functions seem to require multiple mutations and that evolutionary feasibility drops precipitously with each required change, several investigators have calculated how the expected waiting time for these events depends on the number of required mutations [21–25]. Using estimated rates of mutation and gene duplication, along with average sizes and generation times for the population of interest, these calculations provide an upper limit on the number of nucleotide changes that a recruitment scenario can assume. This limit is surprisingly low even for huge bacterial populations. Under the most realistic assumptions, where the duplicate gene reduces bacterial fitness measurably prior to functional conversion [17], it appears that the new function must be produced with no more than *two* specific mutations [21,25].

It is important to ask whether new enzyme functions can really evolve within this tight constraint. To frame the question more precisely, we previously gave a precise description of what would qualify as a genuinely new evolved function [26]. The two conditions we described were: 1) that the pre-evolved enzyme should have no detectable activity with respect to the evolved function, and 2) that this evolved function should not be capable of representation by a generalized chemical reaction (i.e., one using R groups) that also describes the pre-evolved function. Certainly, no explanation of enzyme diversity can be considered complete unless it handles newness of this kind.

With that in mind, the key question to be addressed here is this:

> Are enzymes readily converted to new functions by just one or two mutations in their encoding genes?

The importance of considering conversion by just one mutation is that the overexpression needed for a weak conversion to be beneficial would itself typically require upstream genetic modifications. For example, to the best of our knowledge, the only demonstrations of conversion to a genuinely new enzyme function by single nucleotide substitutions are the conversions of two members of the MLE[1] subgroup of the enolase superfamily to a third function within this subgroup, that of *o*-succinylbenzote synthase (OSBS) [27].[2] But with activities measuring only 0.06% and 0.0004% of wild-type OSBS activity [27], the converted enzymes must be substantially overproduced in order for them to substitute for the wild-type enzyme in vivo. Since every additional mutation needed for this overexpression increases evolutionary waiting times dramatically, it is unclear whether recruitment can explain the origin even of OSBS function. The idea of recruitment being a general explanation of functional diversity within enzyme families therefore needs critical evaluation.

Notice that we deliberately pose the above question in the present tense. We are asking whether the enzymes now present in life are as amenable to functional conversion as ancient enzymes must have been in order for evolution to work. If the answer to this is *No*, then the classical recruitment scenario does not work today. That realization would be of considerable importance, in that it would call for careful consideration of how processes that do not work today somehow did work long ago. In other words, it would reinforce the importance of grounding evolutionary explanations in observable science.

We previously chose a test case for experimental examination of this crucial question. Our study [26] focused on the most structurally similar pair of enzymes from the GABA-aminotransferase-like[3] (GAT) family in *E. coli* that have distinct catalytic functions. One of these was 8-amino-7-oxononanoate synthase, which we designated $BioF_2$ (the subscript indicating the functional dimeric form), and the other was

[1] Muconate lactonizing enzyme.

[2] Näsvall et al. [28] have reported spontaneous mutations that enable the HisA of *Salmonella enterica* to substitute for TrpF, which the wild-type HisA cannot do. Although this is an example of a genuine change of substrate, it is not a genuine conversion of function, according to our definition, because the TrpF and HisA functions are represented by the same generalized chemical reaction.

[3] GABA is an abbreviation for gamma-aminobutyric acid.

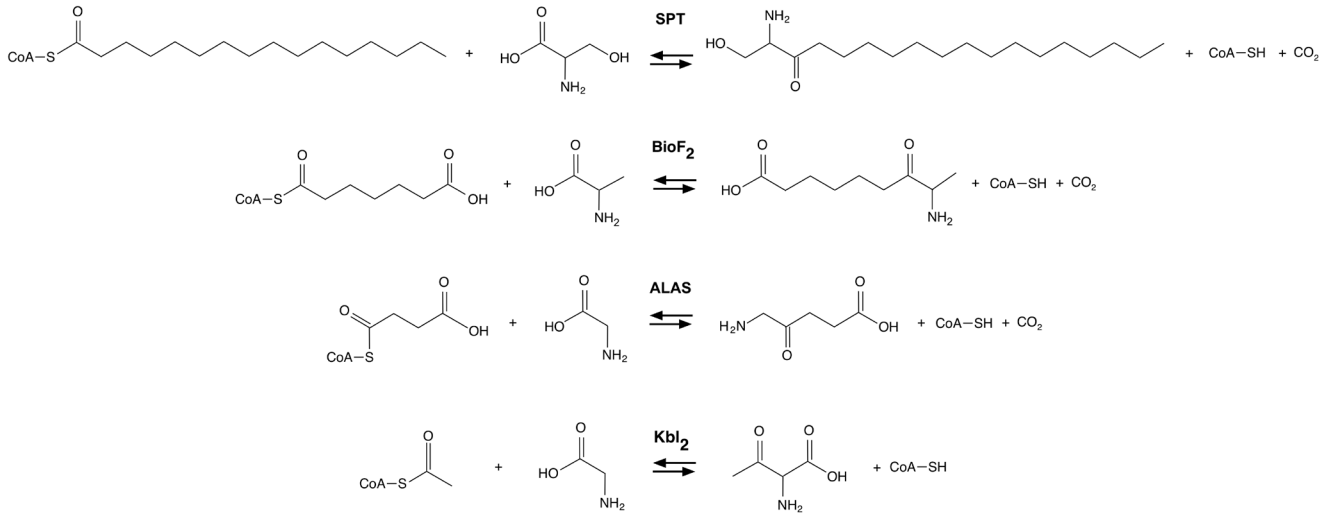

Figure 1: The α-oxoamine synthase enzyme reactions. SPT designates serine palmitoyltransferase, and ALAS designates 5-aminolevulinate synthase. In all cases the second reactant shown is an amino acid (serine for SPT, alanine for $BioF_2$, and glycine for ALAS and Kbl_2), and in all cases except Kbl_2, the carboxylic-acid group of the amino acid is liberated as CO_2. **doi:**10.5048/BIO-C.2014.4.f1

2-amino-3-ketobutyrate CoA ligase, which we designated Kbl_2. $BioF_2$ carries out the first dedicated step in biotin biosynthesis [29,30], whereas Kbl_2 is a non-essential enzyme involved in threonine metabolism [31]. The chemical conversions catalyzed by the two enzymes are similar, both being classified as α-oxoamine synthase reactions (Figure 1), but they are genuinely distinct in that they cannot be represented by a single generalized reaction.

The Kbl and BioF monomers are 34% identical in sequence over an alignment that nearly spans their entire lengths (381 aligned positions; BioF length = 384; Kbl length = 398). As would be expected from full-length sequence identity at that level, they have similar overall fold structures (Figure 2A, B, and C). In fact, the structural similarity is in this case unexpectedly striking, as seen by the matching identity and placement of side chains within the two active sites (Figure 2D). Figure 3 places the structural similarity of this enzyme pair within the broader context of structural similarities among members of the whole PLP-transferase superfamily to which the GAT family belongs. Nodes in this graph are labeled with PDB accession codes with edges connecting nearest structural neighbors, sized according to the structural distance metric we defined previously, δ_S [26]. At 0.44, the structural distance between BioF and Kbl is seen to be within the range of distances separating functionally identical enzymes from different species, and much smaller than many of the nearest-neighbor distances between functionally distinct members of this superfamily.

Having on this basis identified Kbl_2 and $BioF_2$ as a suitable pair for study, we set out in our prior study to estimate how many mutations would be needed to convert Kbl_2 to perform the function of $BioF_2$. Based on sequence and structural alignments of these two proteins, and drawing on alignments of other bacterial proteins with the same functions, we tried to identify a small subset of amino acid positions where changing the Kbl residue to match its BioF counterpart would have the best chance of causing functional conversion. In the end, we were unable to achieve conversion, even after simultaneously changing nearly all side chains in the Kbl_2 binding pocket to match those of $BioF_2$ [26]. Based on this result, we argued that the answer to the above question does indeed seem to be *No*.

Here, we extend that prior work by examining two conceivable ways in which this *No* might be incorrect. First, the complexity of the connection between enzyme sequences and functions makes it possible that the rational approach we used may have missed a conversion that can actually occur within the two-mutation limit. And second, it is possible that some

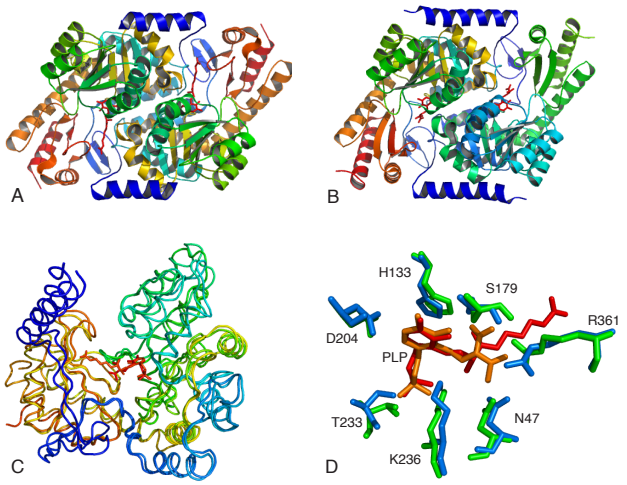

Figure 2: Structural similarity of Kbl_2 and $BioF_2$. Dimeric enzymes A) $BioF_2$ (1DJ9) [32] and B) Kbl_2 (1FC4) [33] viewed along axes of symmetry with external aldimine complexes (PLP covalently linked to enzyme product) in red. Active sites are at the monomer interfaces. C) Aligned backbones of BioF and Kbl monomers. D) Identical side chains in the $BioF_2$ (blue) and Kbl_2 (green) active sites, labeled according to BioF positions. The external aldimine of $BioF_2$ is red (orange for Kbl_2). Reproduced from [26]. **doi:**10.5048/BIO-C.2014.4.f2

member of the GAT family other than Kbl_2, though perhaps less similar to $BioF_2$, may nonetheless happen to be better suited to this conversion. That these two possibilities are conceivable does not imply that they are likely. Nevertheless, we are proceeding with their examination for the sake of thoroughness. To the extent that they can be ruled out, we will have strengthened our initial conclusion.

The present study unfolds in three parts. In the first part, we complete our previous rational approach by examining individually those mutations that were previously examined only in groups. This should clarify the picture of the effects of single mutations that look as though they ought to be among the most important. In the second and third parts, we move from the rational approach to a random approach where large libraries of mutant genes are produced and tested. The second part focuses on single mutations and the third focuses on double mutations. Because random mutations enable the possibilities to be tested in an unbiased way, they should give a true picture of whether functional conversion is readily achievable in just one or two mutations.

RESULTS

Part 1: Completing the examination of rational single mutations

Correction of a previously reported single knock-out mutation. Among the amino acid positions we examined in the prior study were fifteen that show high conservation among bacterial BioF proteins but that differ from the amino acids in the *E. coli* Kbl protein [26]. After examining the effects of replacing these residues in BioF with their Kbl counterparts (referred to as BioF→Kbl substitutions), we reported that one such change, H152N, caused complete loss of $BioF_2$ function, resulting in biotin auxotrophy (i.e., Bio^- phenotype). In the process of preparing for the present study, we discovered that the gene encoding this H152N variant of BioF had an additional mutation that we mistakenly overlooked. This mutation replaces serine at position 265 (adjacent to the PLP cofactor, as seen in Figure 4) with glycine. The complete loss of $BioF_2$ function we previously attributed to the single H152N substitution should therefore be attributed to the double substitution H152N + S265G. To find out whether either of these two substitutions might explain the inactivation alone, we constructed new plasmids carrying each mutation singly. After testing, we found that both of these plasmids confer the Bio^+ phenotype, enabling cells without a chromosomal *bioF* gene to grow normally on minimal medium without biotin. So, neither H152N nor S265G eliminates $BioF_2$ function singlehandedly.

Having previously found H152 to be completely conserved across a sampling of fifty microbial BioF sequences from the Concise Microbial Protein Database (CMPD)[4], we examined the conservation of S265. Although less strongly conserved than H152, this serine is conserved in ninety-nine of the hundred gammaproteobacterial sequences that show greatest similarity to the *E. coli* sequence, the one exception being an asparagine

[4] https://www.ncbi.nlm.nih.gov/genomes/prokhits.cgi

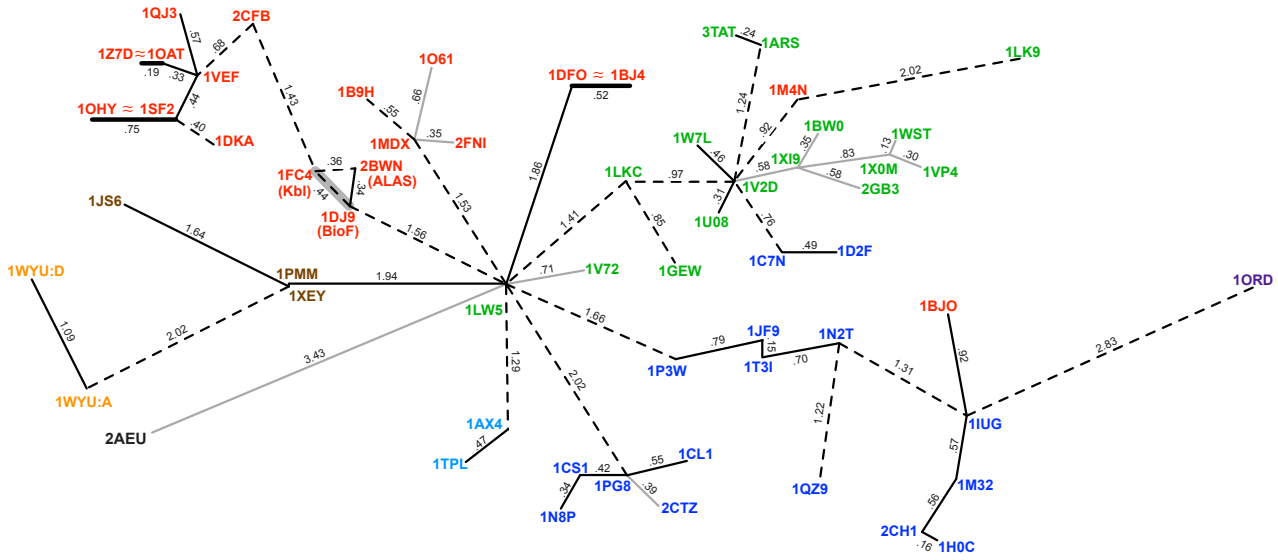

Figure 3: Near-neighbor structural distance graph (δ_S) showing representative members of the PLP-dependent transferase superfamily. Nodes show PDB entry names colored according to SCOP family assignments (http://scop.berkeley.edu/sunid=53417): green = aspartate aminotransferase-like; blue = cystathionine synthase-like; brown = pyridoxal-dependent decarboxylase; red = GABA aminotransferase-like; cyan = beta-eliminating lyases; gold = glycine dehydrogenase subunits; purple = ornithine decarboxylase major domain; black = SelA-like. Edge lengths and connectivity are based on structural data as described previously [26], with dashed edges connecting enzymes having different chemistries, grey edges radiating from nodes with poorly characterized functions, and bold black edges connecting enzymes that perform identical functions in different species. The grey-highlighted edge in the upper left connects the BioF structure with the Kbl structure. Other aspects of geometry (e.g., layout and distances between unjoined nodes) are arbitrary. Adapted from Figure 2 of reference 26. **doi:**10.5048/BIO-C.2014.4.f3

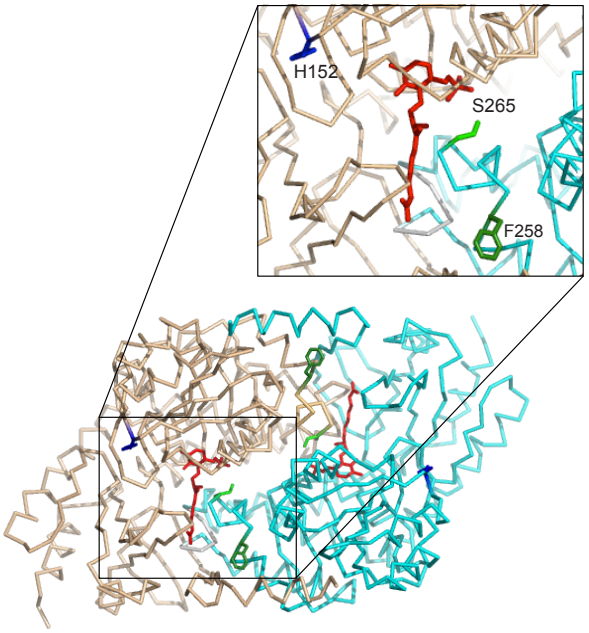

Figure 4: Proximity of H152, S265, and F258 to the BioF$_2$ active site. The PLP external aldimine is shown red, and the F258, S265, H152 side-chains are shown dark green, light green, and blue, respectively. The view is a slice through the dimer to expose the active site, with the two backbone chains colored teal and tan. **doi:**10.5048/BIO-C.2014.4.f4

(see Supplement S1[5]). Similarly, this position is semi-conserved (serine or threonine) among the alpha- and betaproteobacteria we examined (see Supplement S2[6]). This pattern is consistent with our finding that positions 152 and 265 of BioF are both sensitive to substitution, though neither of the native amino acids at these positions is absolutely required for function.

Phenotypic analysis of rational single mutations. This correction raises the possibility that of the 250 residues by which Kbl differs from BioF, there may be none that are absolutely incompatible with BioF$_2$ function. That is, if substitutions in BioF that increase its resemblance to Kbl must be combined in order to eliminate the BioF$_2$ function, then there may be no "wrong" amino acids in Kbl with respect to that function. Converting Kbl$_2$ to work like BioF$_2$ may be more a matter of making several changes that are helpful to that end than correcting any side chains that preclude it. Our previous findings hinted at this possibility. We found three sets of substitutions that eliminate BioF$_2$ function. However, when we tested the constituent substitutions in one of these sets (designated group 3) we found that none of them eliminates function on its own [26].

To see whether the other sets (group 1 and group 2) behave the same way, we constructed low-copy plasmids with mutant *bioF* genes encoding each of the single amino-acid substitutions. As shown in Table 1, no single BioF→Kbl substitution from group 1 is inactivating. However, in group 2 the mutation F258R does completely eliminate BioF$_2$ function. Phe 258 is located near the enzyme surface, just underneath the amino-terminal arm of the opposite monomer (Figure 4). The Phe side chain is solvent-exposed, which makes radical destabilization upon mutation to Arg seem unlikely. F258R may instead interfere with the amino terminal arm's positioning, perhaps affecting dimerization or conformational change when the active site is occupied. None of the other single substitutions in group 2 eliminate function, despite their closer proximity to the substrate binding pocket.

Position 258 shows strong conservation of phenylalanine among the hundred gammaproteobacterial BioF sequences (87%), with a few threonine residues, one alanine residue, and nine tyrosine residues at that position (see Supplement S1[5]). However, when the pool of sequences is expanded to include members of firmicutes and betaproteobacteria, only 49% have phenylalanine at the corresponding position. In fact, eleven sequences from the more distantly related bacteria (<43% sequence identity) have arginine at this position (see Supplement S2[6]). The genes encoding these Arg-containing versions of BioF are located near other genes annotated as biotin biosynthetic enzymes in their respective genomes, so their designation as *bioF* is almost certainly correct. Evidently, then, the disruption caused by F258R in *E. coli* BioF is a matter of context rather than any indispensible role for phenylalanine at this position.

In the end, then, all single BioF→Kbl substitutions we have tested to date have failed to eliminate BioF$_2$ function except for F258R, which appears to exert its effect in a context-dependent way. It therefore seems quite possible that none of the amino acids in Kbl are wholly incompatible with BioF function in all sequence contexts.

Part 2: Examination of random single mutations

Choosing candidates for recruitment to BioF$_2$ function. To test whether a GAT-family enzyme other than Kbl$_2$ might be a better candidate for conversion to the function of BioF$_2$ than

Table 1: Effects of single mutations within previously described mutation groups

Group 1		Group 2		Group 3*	
Mutation	Bio Phenotype	Mutation	Bio Phenotype	Mutation	Bio Phenotype
G75S	+	F258R	-	W344Y	+
S76V	+	A259S	+	A347G	+
G77R	+	H261P	+	I348F	+
H78F	+	L262Y	+	R349F	+
V79I	+	I263L	+	P350Y	+
S80C	+	Y264F	+	T352V	+
		T266N	+		

* As reported previously [26].

[5] Supplement S1: gamma_proteo_alignment.txt (**doi:**10.5048/BIO-C.2014.4.s1)
[6] Supplement S2: CMPD_alignment.txt (**doi:**10.5048/BIO-C.2014.4.s2)

Table 2: Comparison of bacterial enzymes from the GAT family

Enzyme	Present in *E. coli*?	Gene	PDB (species)	EC #	BioF sequence identity	δ_S	Included in mutagenesis screen?
$BioF_2$ (8-amino-7-oxononanoate synthase)	Yes	*bioF*	1DJ9 (*E. coli*)	2.3.1.47	100%	0	n.a.
Kbl_2 (2-amino-3-ketobutyrate CoA ligase)	Yes	*kbl*	1FC4 (*E. coli*)	2.3.1.29	34%	0.44	Yes
Glutamate-1-semialdehyde aminotransferase	Yes	*hemL*	2CFB (*S. elongatus*)	5.4.3.8	19%	1.5	Yes
Adenosylmethionine-8-amino-7-oxononanoate aminotransferase	Yes	*bioA*	1QJ3 (*E. coli*)	2.6.1.62	19%	1.9	Yes
GABA-aminotransferase*	Yes	*gabT*	1SF2 (*E. coli*)	2.6.1.19	18%	1.9	Yes
Putrescine-inducible 4-aminobutyrate aminotransferase*	Yes	*puuE*	n.d.	2.6.1.19	18%	n.d.	No
Acetylornithine aminotransferase†	Yes	*argD*	1VEF (*T. thermophilus*)	2.6.1.11 2.6.1.17	20%	1.5	Yes
Succinylornithine transaminase†	Yes	*astC*	n.d.	2.6.1.81 2.6.1.11	20%	n.d.	Yes
Putrescine aminotransferase	Yes	*ygjG*	n.d.	2.6.1.29	18%	n.d.	Yes
Serine hydroxymethyltransferase	Yes	*glyA*	1DFO (*E. coli*)	2.1.2.1	15%	2.0	No
Phosphoserine aminotransferase	Yes	*serC*	1BJO (*E. coli*)	2.6.1.52	16%	3.0	No
ALAS (5-aminolevulinate synthase)	No	*hemA*	2BWN (*R. capsulatus*)	2.3.1.37	30%	0.36	Yes
SPT (Serine palmitoyltransferase)	No	*SPT1*	2JGT (*S. paucimobilis*)	2.3.1.50	29%	0.70	No
BIKB§	No	*bikb*	n.d. (*T. thermophilus*)	n.d.	29%	n.d.	Yes

* These two enzymes are identical with respect to function and highly similar in sequence (54.2% sequence identity).
† These two multifunctional enzymes have one function in common but differ in secondary functions. Their sequences are 58.4% identical.
§ Previously unnamed; we assigned the name BIKB by combining BioF and Kbl. We likewise refer to the gene as *bikb* (the ORF having been designated TTHA1582 [36]).

Kbl_2, we considered the various possibilities.[7] Table 2 lists all members of the GAT family from *E. coli*, along with their sequence and structural similarity (when known) to BioF [32], and the variety of reactions they catalyze, as indicated by their EC numbers. The more promising candidates from other bacterial species are also included. Enzymes with known structures were compared to BioF using the structural distance metric δ_S, and the calculated distances were used to construct a nearest-neighbor tree (Figure 5A).

Of all these enzymes, $BioF_2$ groups most closely with Kbl_2, 5-aminolevulinate synthase (ALAS), and serine palmitoyltransferase (SPT). These four enzymes have close structural and functional similarities, belonging to the α-oxoamine synthase family [32]. As shown in Figure 1, they all catalyze a Claisen condensation between an amino acid and acyl-CoA, and except for Kbl_2 this condensation reaction is coupled with a decarboxylation reaction [30–33, 37–39].

The similarity of these enzymes is further confirmed by sequence alignment of α-oxoamine synthase family enzymes from the Conserved Domain Database (CDD),[8] chosen for wide phylogenic representation. The alignment reveals that they share broadly similar amino acid sequences, interspersed with regions of dissimilar sequence (Figure 6). Interestingly, of the three groups of amino acid positions we identified as candidates for change in our previous work, group 1 and group 2 fall within those dissimilar regions, these groups being in loops in or near the active site. In contrast, group 3 is part of the more conserved carboxy-terminus.

Of the ten non-$BioF_2$ *E. coli* enzymes listed in Table 2, we excluded two from testing on the basis of their relatively

[7] Classification systems for the PLP-dependent transferases are numerous and differ in the details, though the overall picture is consistent. We have chosen to use the classification system defined by the curated structural database SCOP [34] for the enzymes whose structures have been determined, but have also included those listed as paralogs of GabT by Ecocyc [35] whose structures have not yet been determined.

[8] http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

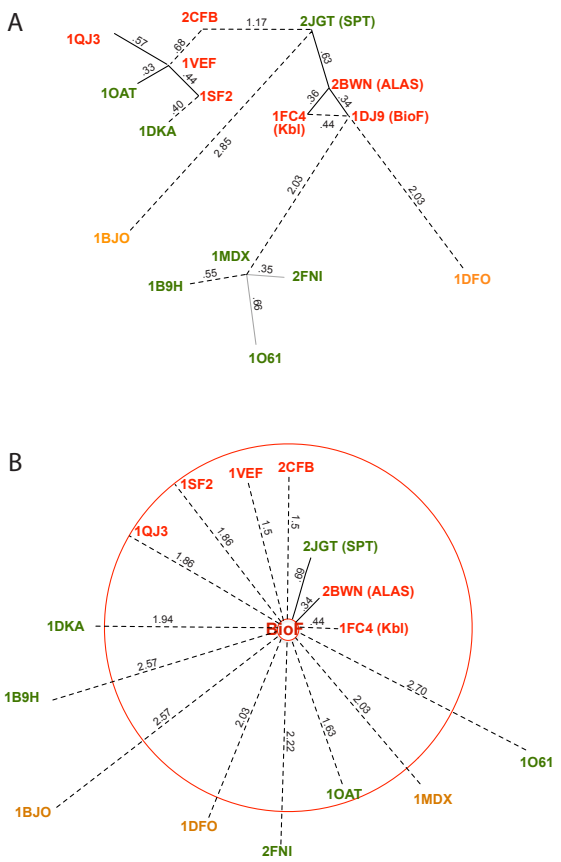

Figure 5: Structural distances (δ_S) of enzymes in the GAT enzyme family. Enzymes are indicated by PDB accession codes, with one structure chosen to represent each distinct enzyme activity. Enzymes included in this study are shown in red, with yellow indicating *E. coli* enzymes from Table 2 that we considered to be too structurally distant, and green indicating enzymes from other species (not included in Table 2). 1VEF and 2CFB are taken to be indicative of the structures of the corresponding *E. coli* enzymes, although they are from other bacterial species. A) Nearest neighbor structural distance graph. Edges between nodes are proportional to the structural distance between that pair. The arrangement is otherwise arbitrary. Where neighbors share the same general reaction, the edges between them are solid. Where their chemistry differs the edges are dashed lines. B) Radial lines proportional to each enzyme's structural distance from BioF (1DJ9) were drawn with BioF at the center. The circle drawn at 1.86 units shows our chosen distance cut-off for this study. **doi:**10.5048/BIO-C.2014.4.f5

low similarity to $BioF_2$ (see Figure 5B), these being serine hydroxymethyl transferase (PDB:1DFO) and phosphoserine aminotransferase (PDB:1BJO). Another (putrescine-inducible 4-aminobutyrate aminotransferase) was excluded because it is substantially similar to GABA-aminotransferase, the two having identical functions and 54% sequence identity. To the remaining seven candidates we added two from other species. ALAS from *R. capsulatus* (PDB:2BWN) was added because of its structural similarity to BioF (Figure 5) [40] and because its chemistry is particularly close to that of $BioF_2$ (Figure 1) [38]. BIKB from *T. thermophilus* was added to our list of candidates because it has been described as having both Kbl_2 and $BioF_2$ activities in vitro [36]. As such, it is the most promising example of a potential promiscuous enzyme with both of these functions. Neither of these enzymes is as similar in sequence to BioF as Kbl is (Figure 7), but structural and functional considerations are equally important.

The seven candidate genes from *E. coli* were obtained from the ASKA plasmid library [41], which contains all *E. coli* open reading frames inserted into the inducible expression vector pCA24N [41]. Genes encoding the two enzymes not from *E. coli*, the *hemA* gene encoding ALAS from *R. capsulatus* and the *bikb* gene from *T. thermophilus*, were cloned into the same vector in our laboratory as described in Methods. After confirming all nine genes by sequencing, we used the plasmids to transform KeioΔ*bioF* [42], an *E. coli* strain that cannot grow without biotin (Bio^- phenotype) because it lacks the *bioF* gene. After two weeks of incubation on minimal medium without biotin, none of the overexpressed genes were found to confer the Bio^+ phenotype. This confirms the prior finding of Patrick et al. [15] that no *E. coli* gene can substitute for *bioF* simply by overexpression, and it extends that result to the two non-*E. coli* genes *hemA* and *bikb*. The fact that cell growth requires only a tiny amount of biotin [43] combined with the fact that these genes were substantially overexpressed makes this is a highly sensitive test. We are therefore confident that none of these tested enzymes has any significant tendency to catalyze the $BioF_2$ reaction in vivo, despite the fact that one of them (BIKB) was described as having that activity in vitro [36][9].

Searching for random single mutations that rescue. To see whether any of the nine candidate genes can be made to confer the Bio^+ phenotype with a single mutation, each was taken through a single round of random mutagenesis using error-prone PCR with conditions optimized to give about one mutation per gene. After inserting the resulting gene libraries into clean pCA24N vector, each library was transferred by electroporation into KeioΔ*bioF* and plated to minimal medium without biotin (see Methods). Trays were incubated for two weeks at 37° C to test for rescue, with no colonies appearing (see Table 3).

By sequencing the gene inserts in ten to twenty pre-selection transformants from each of the nine library transformations, we determined the average number of nucleotide substitutions per kilobase (kb) for each library. Assuming uniform distributions, those mutation rates were used to estimate the number of genes produced with exactly one base change (see Methods). As shown in Table 3, in all cases the library diversity exceeds the total number of possible single mutations (about 3,600) by at least a factor of ten. Oversampling to this extent assures that it is unlikely that any specific single mutation would have gone untested in these experiments. Consequently, the absence of rescue shows that none of the candidate genes can be made to confer the Bio^+ phenotype with one point mutation.

To verify that specific single mutations can be recovered from libraries of mutant genes prepared by our method, we made two variants of the *bioF* plasmid where translation is terminated

[9] Because Kubota et al. [36] assayed both activities by detecting release of their common byproduct CoA–SH (see Figure 1) instead of detecting the products of biological significance, it is possible that their conclusion that both activities are present in vitro may be in error.

prematurely by TAA stop codons. Both of these TAA codons were constructed by altering a single DNA base, one in the codon for K236, and in the other in the codon for Q257. After verifying that KeioΔ*bioF* cells carrying either of these mutant plasmids are unable to grow on minimal medium without added biotin, we took the two genes through one round of random mutagenesis using the same protocol as before. This time about one in a thousand transformants formed colonies on biotin-free medium for each of the two libraries. The recovery of these revertants confirms that our libraries have good coverage of single-base substitutions.

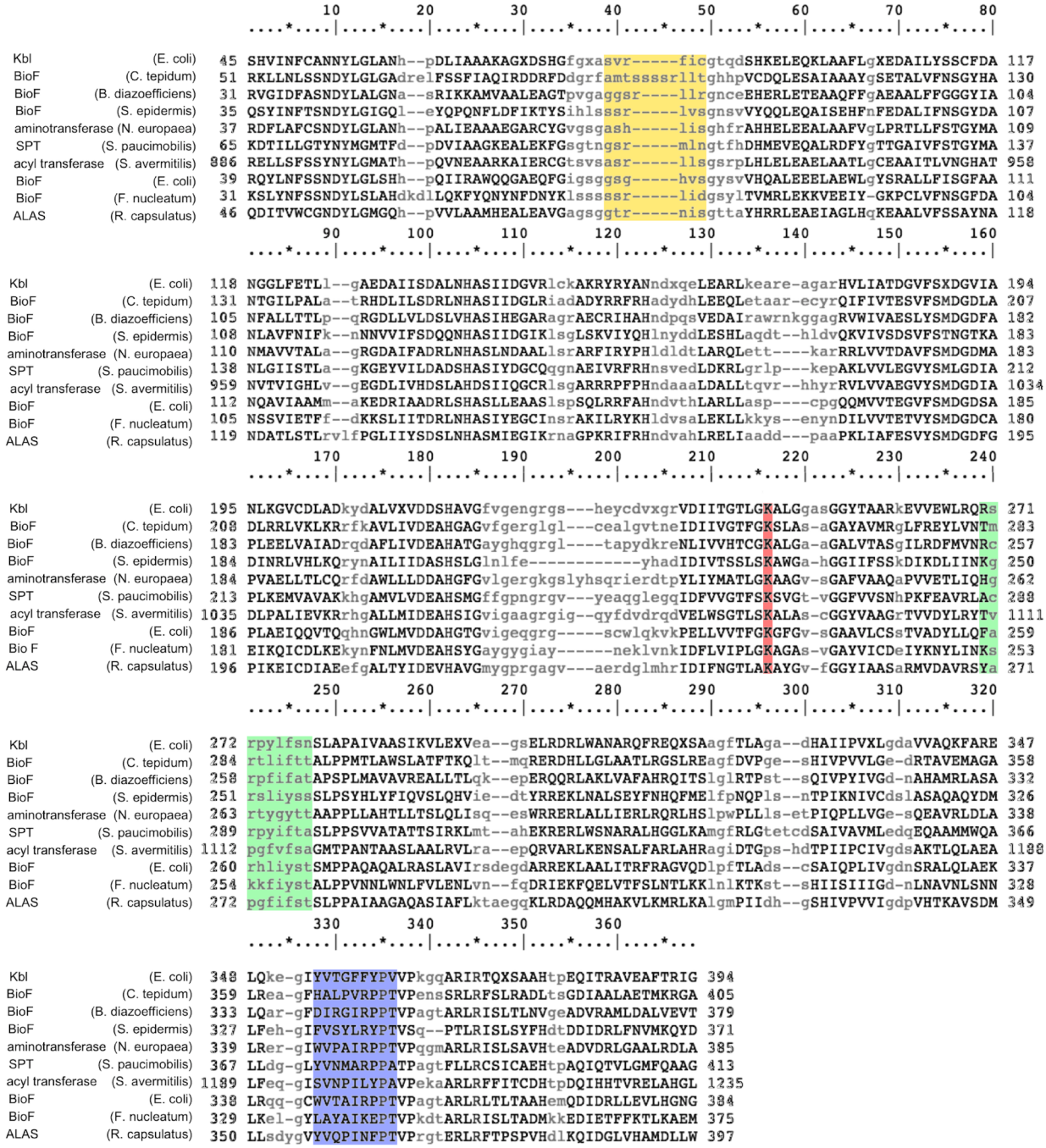

Figure 6: Sequence alignment of α-oxamine enzymes from the CDD Kbl-like family. Sequences were chosen to represent the widest possible diversity among bacteria. The enzyme activities of Kbl, BioF, SPT and ALAS have been experimentally verified [30,31,38,39] while the activities of the other encoded enzymes are annotated either as BioF or as 'undetermined oxoamine transferase.' Greyed lower case indicates regions of high variability. Red highlight shows the PLP-binding lysine. Yellow, green, and blue highlights show groups 1, 2, and 3, respectively, described in Part 1 of Results. **doi:**10.5048/BIO-C.2014.4.f6

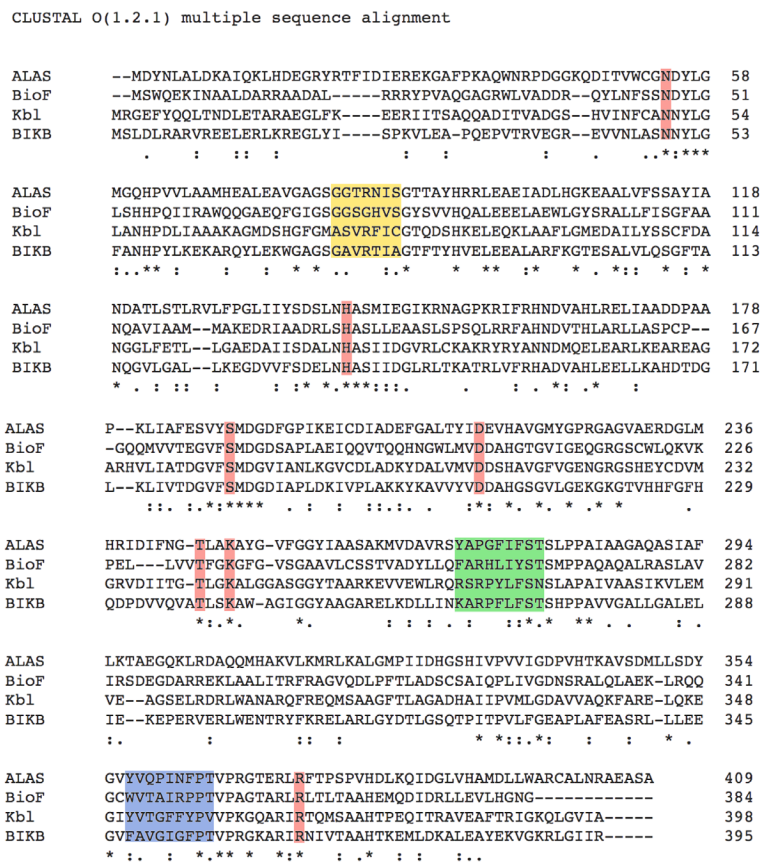

Figure 7: Sequence alignment of α-oxamine enzymes examined in this study: ALAS, BioF, Kbl, and BIKB. Asterisks indicate complete conservation with double dots and single dots indicating lower degrees of conservation. Yellow, green, and blue highlights show groups 1, 2, and 3, respectively, described in Part 1 of Results. Active site residues from Figure 2D are highlighted red. **doi:**10.5048/BIO-C.2014.4.f7

Part 3: Examination of random double mutations

As discussed in the introduction, it is conceivable that the rational approach we used previously [26] to identify the most promising amino-acid substitutions in Kbl_2 might have missed substitutions that actually work. From the above result, we can now be confident that we did not miss any amino-acid substitutions that can be achieved with a single DNA base change, because all of these should have been present in our library of mutant *kbl* genes. Moreover, it is now clear that the difficulty of conversion to $BioF_2$ function is not at all peculiar to Kbl_2, as this conversion is now seen to require more than a single mutation from eight other seemingly plausible starting points.

To see whether two mutations might succeed where one did not, we chose two of our nine candidate genes for more extensive mutational analysis (knowing that the amount of work would make it impractical to examine all nine). Among *E. coli* genes, Kbl_2 is still the favored candidate for conversion to $BioF_2$ function for all the reasons that we first identified it as such. In addition to Kbl_2, we decided to examine double mutations in BIKB because its putative promiscuous catalysis of both the Kbl_2 and the $BioF_2$ reactions in vitro [36] suggests that it might be within two mutations of in vivo conversion.

Having chosen these two, we took the *kbl* and *bikb* genes through extensive additional random mutagenesis, this time under conditions optimized to produce two mutations per gene. By performing five separate error-prone PCR reactions on *kbl*, the products of which were used in thirteen electroporations of KeioΔ*bioF*, we were able to screen a library of forty million transformants. Sequence analysis of 101 genes from the pre-selection library revealed that about 20% (20 of 101) had two base substitutions, which means our library contains about 7.9 million doubly mutated genes. This corresponds to about 71% coverage of the 6.4 million possible double mutants, taking repeats into account. None of these transformants were able to grow on minimal medium without biotin.

Similarly, for BIKB we screened a library of 48 million randomly mutated plasmids for their ability to replace the missing $BioF_2$ activity in KeioΔ*bioF*. Sequence analysis performed on 27 clones from the naïve library showed an average base substitution rate of about 1.1 per gene and an insertion/deletion rate of about 0.18 per gene. Using these rates, we estimate the throughput of double mutants without insertions or deletions to be 16%, which means that 7.7 million double mutants were examined, corresponding to 70% coverage of the possibilities

(see Methods). No conversion to the Bio$^+$ phenotype was obtained after 2 weeks of incubation at 37°C, this despite BIKB's putative functional promiscuity in vitro (see footnote 9 for a possible explanation).

DISCUSSION

The greatest challenge facing evolutionary accounts of enzyme origins is explaining how enzymes with new fold structures first appeared. Having made the case that this challenge is insurmountable in Darwinian terms [44] we turned our attention several years ago to the more modest challenge of explaining how enzymes that existed long ago might have been coaxed into putting their structures to new uses. Certainly there is no shortage of modern enzymes that use similar structures to perform different functions, and at first glance this may seem to fit the evolutionary account of enzyme origins.

However, because the point of studying protein origins is to explain how the many different functions arose, a successful explanation of enzyme diversity will have to focus more on the differences than on the similarities. The fact that subtle structural differences among the members of enzyme families cause profound functional differences might suggest that these functional differences are easily achieved, but the accompanying sequence differences, which are substantial, could equally support the opposite conclusion. That is, of the many amino acid differences (often hundreds) that distinguish any two enzymes with different functions, if more than a tiny fraction of these are important for making those functions different, then it may be effectively impossible for undirected mutations to stumble upon the right combinations for functional conversions.

Furthermore, careful inspection of the full biological context in which recruitment would have to occur turns up several other similarly weighty concerns [45]. For example, gene expression carries a measurable metabolic cost in bacterial populations, and natural selection has a proven tendency to curtail this cost by favoring mutations that halt expression of useless genes [17,18]. Combined with the fact that reports of successful functional conversions typically depend upon overexpression to compensate for very weak activities, this raises serious questions about the relevance of these success stories to actual evolutionary processes. If overexpression is needed for a weak new function to have a beneficial effect, then the rarity of semi-stable gene duplicates that might initiate recruitment is compounded by the rarity of mutations that cause overexpression, over and above the rarity of mutations needed for functional conversion. Moreover, in order for maladaptive intermediates to be avoided, these mutations must occur in the right order, with conversion preceding overexpression.

The present study has added to our previous examination of these problems in several respects. We have shown, based on sequence alignment of α-oxoamine synthases (a subset of the GAT family), that our previous use of rational design did indeed target regions of Kbl_2 that are likely to be functionally

Table 3: Library statistics for random single mutations

Gene	Library size	Kilobases sequenced*	Base changes found*	Indels found*	Estimated throughput†	*bioF* rescue?
argD	2.1×10^5	7.8	11	1	5.6×10^4	No
astC	1.7×10^5	8.3	8	0	6.2×10^4	No
bioA	3.7×10^5	12.3	8	1	1.2×10^5	No
gabT	1.6×10^5	16.1	20	0	5.4×10^4	No
hemL	1.7×10^5	9.3	12	0	5.6×10^4	No
kbl	2.2×10^5	13.9	12	2	6.8×10^4	No
ygjG	2.1×10^5	10.4	9	0	7.7×10^4	No
hemA	1.8×10^5	17.6	10	0	6.2×10^4	No
bikb	1.6×10^5	17.4	13	1	5.2×10^4	No
*bioF*Stop236	1.7×10^5	6.9	8	0	5.9×10^4	Yes
*bioF*Stop257	2.6×10^5	6.7	7	3	5.4×10^4	Yes

* In total, based upon partial sequencing with either forward or reverse read (approximately 900 bases per read; orf lengths being about 1200 bases) of from ten to twenty gene inserts from the library. Indel and base change frequencies were calculated from these results on a per-kilobase basis.

† The number of indel-free genes carrying exactly one nucleotide substitution in the library, estimated as described in Methods.

significant. Furthermore we have now shown that the lack of a simple evolutionary transition to $BioF_2$ function is not at all unique to our initial choice of Kbl_2 as the starting point. Single mutations cannot convert any of eight other members of the GAT family to that function, despite the fact that all of these enzymes are regarded as close evolutionary relatives.

Finally, we have demonstrated that converting either Kbl_2 or BIKB to perform the function $BioF_2$ with two DNA base substitutions is at least mildly unlikely, in that neither conversion was found after examining over two thirds of the possibilities. Of course, many possibilities remain unexamined. Although it is certainly possible for a working combination to be among those unchecked possibilities, we think it is more informative at this point to ask whether the available evidence as a whole really supports the idea that evolutionary recruitment is the cause of functional diversity in enzyme families.

To that end, we return to our initial question, to which we give the same answer as before with greater evidential weight: *No*, enzymes are not readily converted to new functions by just one or two mutations in their encoding genes. All things considered, we should still expect that Kbl_2 and BIKB are among the best starting points for evolutionary conversion to the function of $BioF_2$, and therefore that successful conversion from any good starting point would probably require no fewer than three changes to the coding region of the starting gene. Nevertheless, if we take two changes to be an open possibility, we should expect that any new activity achieved with so few changes will have to be amplified by overexpression. This adds a third change to the list of requirements, namely a change to the upstream region of the starting gene that elevates expression. Preceding these three changes is the initial duplication event, the change that sets the stage for recruitment.

Equation 10 of the population genetics analysis by Axe [25] tells us how long it would take for all of these necessary changes to come together to produce a successful evolutionary adaptation by the classical recruitment mechanism. Using the parameter values listed in Table 4, we estimate that this would happen about once in 10^{15} years. This timescale is multiplied a million-fold if three changes to the coding region are required, as seems probable in light of our inability to find conversion with two changes. Either way, the classical recruitment scenario is clearly problematic as an explanation of the origin of the $BioF_2$ function.

Table 4: Parameter values used for calculation of evolutionary timescale

Parameter	Value	Reference
Total population size	10^{20}	46
Effective population size	10^{9}	47
Specific base mutation rate	10^{-9} per site per cell	48*
Gene duplication rate	3×10^{-6} per cell	49†
Adaptive selection coefficient	+0.01	
Maladaptive selection coefficient	-0.04	49†

* Based on the expected total mutation rate for bacteria with a genome size of 10^6 base pairs, as found by Drake et al. [48].

† Based on data for *pyrD* in *Salmonella enterica* as reported by Reams et al. [49].

As tempting as it is to think this problem can be solved by adjusting the numbers that go into the calculation, the reality is that every adjustment that helps in one respect hurts in another. For example, an elevated mutation rate helps the situation by increasing the number of cells carrying precursors to the converted gene, but it simultaneously hurts by decreasing overall fitness through mutation load. The finding of Drake and coworkers that bacterial mutation rates tend to hover at about one mutation per 300 cells [48] is consistent with there being a selective disadvantage for sustained rates much higher than this. Indeed, even under conditions where elevated mutation rates enhance adaptation, there is a tendency for lower rates to be restored in the long run [50]. In *E. coli*, with a genome size of about 5×10^6, a mutation rate of 10^{-7} per nucleotide site per cell would cause about half of newly divided cells to carry a new mutation, which would certainly entail a fitness cost. Small pockets of a global bacterial population can easily sustain rates that high for short periods, but these subpopulations are continually replaced by the wild-type, which is of higher fitness when averaged over the conditions experienced by the entire population. To put this in perspective, *if it were* possible for this higher mutation rate to be sustained throughout a global population, then we calculate the waiting time for conversion to be about a billion years. Since that is definitely a long term, it only underscores the need to use a realistic long-term global average mutation rate rather than an exceptional one.

The possibility that other evolutionary scenarios, such as divergence from promiscuous ancestral enzymes, could perform better cannot be ruled out by this work. It is also conceivable that the function of $BioF_2$ is exceptionally hard to achieve[10] and that successful conversion would therefore require a particularly favorable starting point, perhaps with much higher sequence identity to BioF. But it is one thing to ask what is needed for conversions to work in the lab and another to ask whether we should expect similar conversions to have happened naturally. *The problem for evolutionary explanations is that the very special circumstances needed to achieve even weak conversions in the lab translate into highly unrealistic evolutionary scenarios.* If the promiscuity hypothesis seems to avoid this critique, it is not because the circumstances it assumes are any less special but only because they are much more hypothetical. In the end, then, when all laboratory experience with enzyme conversion is considered collectively in this light, it seems quite clear both that the classical recruitment explanation of enzyme diversity is

[10] More probably, the function of OSBS (described in the Introduction) is exceptionally easy to achieve. As far as we know, this is the only enzyme activity shown to be acquired by single nucleotide substitutions in genes that encode enzymes with genuinely different functions and no prior OSBS function [27].

severely undermined and that there is no credible evolutionary alternative.

Toward a correct interpretation of structural similarity

The claim that enzymes are related by descent is interesting only if those enzymes have significant differences in addition to their similarities. For this reason the claim that gene recruitment explains not only the functional diversity of the GAT family but also that of the greater PLP-transferase superfamily is indeed interesting. However, when we examine enzymes within this superfamily where the differences are as small as they can be without functional overlap, we find functional transitions to be evolutionarily implausible.

One way around this problem would be to restrict claims of recruitment to cases of even tighter similarity, where the functions do overlap. But this leaves all the more interesting transitions unexplained, which means it does little to account for the whole picture of enzyme diversity. The proposal that functional overlap was much more common in the remote past than it is now is at least an acknowledgement of the problem, but again, unless that idea finds much more evidential support that it has so far, it ought to be viewed with skepticism.

Although there is as yet no satisfactory theory of biology to take the place of Darwinism, we believe the time has come for serious pursuit of such a theory. To quote one of our previous papers [45]:

> The insights we gain from the critique of neo-Darwinism can and should inform the construction of a new theory to take its place. That is, in pinpointing the key problems with the old theory we are identifying crucial respects in which its replacement must differ from it. We ourselves have become convinced that intelligent causation is essential as a starting point for any successful theory of biological innovation. If this is so, what is needed now is an elaboration of the general principles by which living things have been designed.

To that end, one of our inferred principles of design is this [45]:

> The substantial reworking of a homologous structure needed to give it a genuinely new function is more suggestive of reapplication of a concept than adjustment of a physical thing.

And another is this [45]:

> The implementation of innovation is nearly the opposite of ordinary physical causation. It is the top-down arrangement of matter in such a way that the resulting bottom-up behavior of that matter serves the intended purpose of the innovator.

Taking these two ideas together, it may be that our prior attempts to convert Kbl_2 to perform the function of $BioF_2$ failed not because we made the wrong alterations but rather because it is misguided even to think of this as an exercise in alteration. Perhaps we should think of this more in the way we think about writing. Sentences that convey different ideas may have similar structures, but when we write a sentence we start with the idea, not the sentence structure. We never take a sentence that conveys some *other* idea and ask which letters can be changed to make it better suited for our present purpose. The fact that different ideas end up being conveyed with sentences of similar structure, then, has nothing to do with recycling of sentences and everything to do with the suitability of certain forms for certain functions.

Might this be the right perspective from which to view Kbl_2 and $BioF_2$? They use similar structures not because they are both adjusted versions of some older enzyme, but instead because the purposes they serve happen to call for similar structures. As we found in this work, it is not that Kbl has amino acid residues that are incompatible with the function of $BioF_2$, but rather that Kbl_2 is comprehensively suited to one function, while $BioF_2$ is comprehensively suited to another. To us this change of perspective has the feel of a turn in the right direction. It does not in itself take us very far, perhaps, but having made the turn, forward progress may become much more likely.

METHODS

Media and solutions

Cultures for plasmid preparation were grown in Terrific Broth (TB) with 20 μg/ml chloramphenicol (TBC20) (Sigma). Standard solid culture was on LB agar (Fluka) for strains without plasmids. All plasmid-bearing strains were maintained on LB with 20 μg/ml chloramphenicol (LBC20). Minimal medium was either Minimal Davis Broth or Davis Minimal Agar from Fluka (abbreviated MDM or MDMC with chloramphenicol) with supplements as follows. Glucose and methionine were added to final concentrations of 0.5% and 0.2%, respectively, and salts to 1× concentration (the 100× stock containing the following in 100 ml water: 10g NH_4Cl; 2g $MgSO_4 \cdot 7H_2O$; 31.43mg $MnCl_2 \cdot 4H_2O$; 100mg $CaCl_2$; 50mg $FeSO_4 \cdot 7H_2O$). For phenotype testing of plasmids, biotin (Fluka) was added just before pouring to a final concentration of 20 ng/ml (MDMCB) or streptavidin (Sigma) to a final concentration of 100 ng/ml (MDMCSA). For screening mutant libraries, isopropyl-β-D-thio-galactopyranoside (IPTG) (Affymetrix) was spread (final concentration 1 mM) on the day of plating onto plates or trays to induce overexpression of the targeted gene. Plasmid preparation, gel purification, and PCR purification kits were from Qiagen.

Strains and plasmids

The strains used in this work (all *E. coli*) are listed in Table 5. Strain AG1, from the National BioResource Project (NIG, Japan): *E. coli* (abbreviated NRBP (NIG, Japan): *E. coli*), was used for cloning *hemA* (the gene encoding ALAS) and *bikb* (the gene encoding BIKB) into plasmid pCA24N. The in-frame single-gene knockout strain KeioΔ*bioF* [42] (obtained from NBRP) was used for phenotype-testing mutant libraries (see Part 2 and Part 3 of Results), while 1D3Δ*bioF* [26] was used for testing single mutations to *bioF* for their effects (see Part 1 of Results).

Table 5: *E. coli* strains used in this work

Strain	Genotype	Source
1D3	Δ*bioF*, EMG2 derivative, *r*K+, *m*K+	Gauger, Axe [26]
NEB5α	*fhuA2 Δ(argF-lacZ)U169 phoA glnV44 Φ80 Δ (lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17(rK- mK+)*	New England Biolabs
KeioΔ*bioF*	*rrnB ΔlacZ4787 hsdR514 Δ(araBAD)567 Δ(rhaBAD)568 rph-1 ΔbioF*	NBRP*
AG1	*recA1 endA1 gyrA96 thi-1 hsdR17(rK– mK+) supE44 relA1]*	NBRP*
NEB 10-beta	*Δ(ara-leu) 7697 araD139 fhuA ΔlacX74 galK16 galE15 e14- Φ80dlacZΔM15 recA1 relA1 endA1 nupG rpsL (StrR) rph spoT1 Δ(mrr-hsdRMS-mcrBC)*	New England Biolabs

* National BioResource Project of Japan (NBRP-E .coli at NIG).

ASKA plasmids carrying *E. coli* genes *bioF, kbl*, *bioA*, *hemL, gabT, astC, argD*, and *ygjG* and the parent plasmid pCA24N were obtained from the National BioResource Project of Japan (NBRP-*E. coli* at NIG). Their construction is described by Kitagawa et al. [41]. We used plasmid pCA24N for cloning *hemA* and *bikb*, as described below. Plasmid pKBF2 [26] was our starting point for making individual mutations to *bioF* (Part 1 of Results).

Site-directed mutagenesis of *bioF*

Plasmids carrying specific mutations to the *bioF* gene (see Part 1 of Results) were generated by inverse PCR of pKBF2 with primers designed to introduce the desired mutations. Ligated PCR products were used to transform strain NEB 5-alpha to allow DNA methylation without host restriction. Transformed colonies on LBC20 were streaked and their plasmids purified and sequenced to confirm genotype. Plasmids were then used to transform strain 1D3. These transformants were pre-cultured for two days in MDMCB broth to adapt the cells to minimal culture. Phenotypes were tested as described. Briefly, 1 ml of pre-adapted cell culture was washed in ice-cold phosphate-buffered saline (PBS), and spread onto LBC20 plates for cell counts, onto MDMCB plates as a check of efficiency of growth on minimal medium with biotin, and onto MDMCSA plates to assess growth on minimal medium without biotin. Control strains 1D3:pKBF2 (Bio^+) and 1D3:pH152N/S265G (Bio^-) were preconditioned and plated in parallel with every experimental set. All were incubated at 30° C and checked at 42 hours for growth.

Construction of ALAS and BIKB plasmids

To insert the genes encoding ALAS and BIKB into pCA24N, we followed the protocol used to generate the ASKA library [42]. Briefly, genomic DNA from *R. capsulatus*, and *T. thermophilus* (both from ATCC) was amplified using gene-specific primers with short N-terminal tags, as described [41]. After purification, the amplified genes were inserted by blunt end ligation into plasmid pCA24N that had been pre-digested with StuI and gel purified. The N- and C-terminal tags on the primers were designed to create unique SfiI sites at the ligation junctions, but only when insertion is in the proper orientation. All plasmid sequences were verified before experiments were performed.

Preparation of randomly mutagenized gene libraries

For experiments where our target was one random mutation per gene (Part 2 of Results), the desired template was amplified from the appropriate plasmid using Mutazyme II (Stratagene) and flanking primers ASKM-F and ASKM-R (see Figure 8). Conditions for amplification were as specified in the Mutazyme II kit (Stratagene). We determined experimentally that starting with 700 ng of template and amplifying for 22 cycles produced about one base substitution per kilobase of template. Following

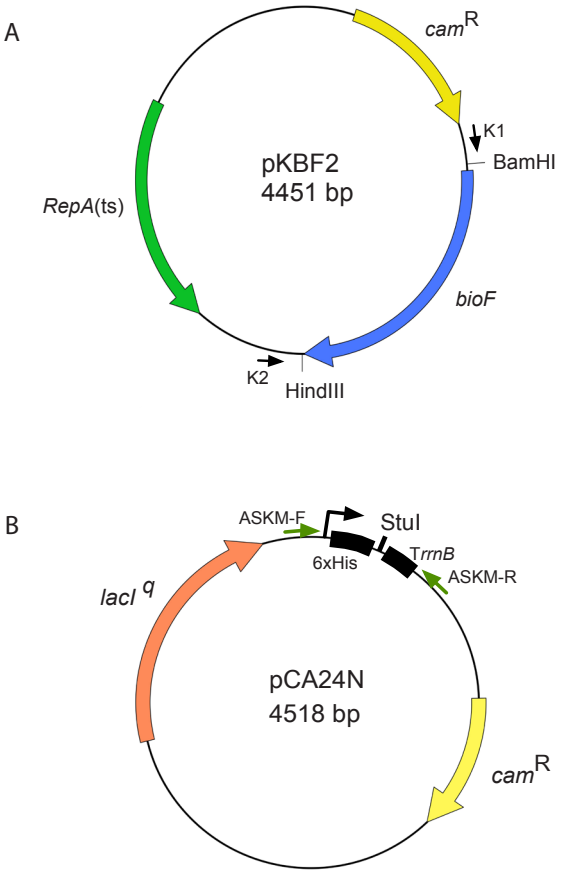

Figure 8: Parental plasmids used in this study. A) The pKBF2 plasmid was designed for low copy phenotype testing of single mutant *bioF* plasmids. B) The pCA24N plasmid [41] was designed to allow easy insertion of genes at the StuI site. When inserted in the proper orientation the flanking sequences combine with the tagged primers to generate distinct SfiI sites at each end of the inserted gene. **doi:**10.5048/BIO-C.2014.4.f8

amplification, mutagenized PCR products were column purified (Qiagen), digested with DpnI (NEB), column purified again, digested with SfiI (NEB), and then column purified one last time. The product was dried completely and resuspended in water.

The vector for cloning the SfiI-digested PCR products was prepared by SfiI digestion of the ASKA plasmid that carries gene *mrdA* (which is large enough to allow clean separation of the vector and insert) followed by gel purification of the linearized vector with a Qiagen kit. This linearized vector has two different SfiI ends that match those of the mutagenized, SfiI-digested PCR products above, guaranteeing insertion in the right direction. Vector and inserts were ligated with Quick Ligase (NEB) and column purified, eluting with water for electroporation.

For experiments where our target was two mutations per gene (Part 3 of Results), the Mutazyme II PCR amplification time was increased to 28 cycles. For the *bikb* library we performed five additional amplification cycles after mutagenesis by adding Taq polymerase (NEB) to the reaction mix. Both *bikb* and *kbl* libraries were then prepared by DpnI and SfiI digestion as described above. Finally, each batch of *kbl* mutant library was gel purified to remove extraneous priming products after the SfiI digest. For *bikb* library batches we used column purification instead to increase the yield. Purified, digested library DNA was ligated into SfiI-digested pCA24N using T4 DNA ligase (NEB) as before.

Library electroporation

Part 2 of our work required fewer than 100,000 transformants per starting gene to cover all single base changes. The ligation product was therefore directly electroporated into 50 µl electrocompetent *E. coli* KeioΔ*bioF* cells, using a BioRad Gene Pulser II with settings of 25 µF, 200 Ω, and 2.5 kV and a 2 mm gap cuvette. Immediately after pulsing, cells were suspended in 975 µl SOC medium and incubated at 37° C, 250 rpm for 90 min, then spread onto a 245×245 mm LBC20 tray (Becton Dickenson) and incubated overnight at 37° C.

To reach our target of forty million transformants for each of the double-mutant libraries prepared in Part 3 of our work, we pooled the products of five rounds of PCR mutagenesis and used the pooled DNA for repeated electroporations (13 for Kbl and 25 for BIKB). In these two cases, the DNA was first used to transform NEB 10-beta electrocompetent cells to increase the yield of initial transformants. Plasmid DNA prepared from these combined transformants was then used to transform electrocompetent *E. coli* KeioΔ*bioF*, spreading onto LBC20 trays as before.

Library screening

Transformed KeioΔ*bioF* cells were washed from the trays and pre-cultured for two days (for single mutations; Part 2) or one day (double mutations; Part 3) in MDMCB broth at 37° C to adapt the cells to minimal medium. One ml of pre-adapted cell culture was washed four times in ice-cold PBS. A small portion of the resulting cell suspension was diluted one million fold and spread on LBC20, MDMCB, and MDMCSA plates for cell counts and positive and negative controls (see below). The remaining undiluted cells (typically numbering in the millions) were spread onto MDMCSA + IPTG agar trays, and incubated for two weeks at 37° C.

Because bacterial cells require only trace quantities of biotin for growth, screening for biotin autotrophy (phenotype Bio^+) requires careful controls. For each test, we used single batches of freshly prepared medium and plated both an appropriate positive control strain (either KeioΔ*bioF*:p*bioF*-ASKA or 1D3:pKBF2) and negative control strain (either KeioΔ*bioF*:p*bioA*-ASKA or 1D3:pH122N S265G) in parallel with experimental strains.

Sequencing and analysis of library samples

For Part 3 of our work (where the focus was random double mutations), individual colonies were transferred from the naïve (i.e., pre-selection) library to fresh plates, and sent as colonies for sequencing of the ASKA insert. For Part 2 (where the focus was random single mutations), sequencing was performed on from ten to twenty samples, either bacterial colonies from the naïve library (as above) or PCR-amplified mutant genes. All sequencing was done by Genewiz, Seattle, WA.

For the double-mutant analysis of *kbl* (Part 3), a total of 101 naïve colonies were sent for sequence analysis. With twenty confirmed double-mutant *kbl* genes among that set, this fraction (20/101) was deemed to be a reasonably accurate measure of the proportion of double mutants within the naïve library. The fractional coverage of the 6.4 million possibilities was calculated from this by assuming a uniform distribution of possibilities within the library.

The number of mutant genes sequenced was substantially lower (in the range of ten to twenty) for all other library experiments (i.e., the nine experiments in Part 2 and the experiment with the BIKB gene in Part 3). In these cases we used the rate of insertion/deletion mutations per kilobase sequenced (fifth column of Table 3) to estimate the fraction of each library that lacked insertions or deletions, and the rate of nucleotide substitutions per kilobase sequenced (fourth column of Table 3) to estimate the fraction of each library with the desired number of mutations, assuming the number of mutations per gene to follow a Poisson distribution.

1. Jensen RA (1976) Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30:409–425.
2. Ohno S (1970) Evolution by Gene Duplication. Springer-Verlag (New York).
3. Ohno S (1973) Ancient linkages and frozen accidents. Nature 244:259–262.
4. Hughes AL (2002) Adaptive evolution after gene duplication. Trends Genet 18:433–434. **doi:** 10.1016/S0168-9525(02)02755-5
5. Chothia C, Gough G, Vogel C, Teichman SA (2003) Evolution of the protein repetoire. Nature 300:1701–1703. **doi:** 10.1126/science.1085371
6. Khersonsky O, Roodveldt C, Tawfik DS (2006) Enzyme promiscuity: Evolutionary and mechanistic aspects. Curr Opin Chem Biol 10:498–508. **doi:** 10.1016/j.cbpa.2006.08.011
7. Gerlt JA, Babbitt PC (2009) Enzyme (re)design: Lessons from natural evolution and computation. Curr Opin Chem Biol 13:10–18. **doi:**10.1016/j.cbpa.2009.01.014
8. Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol 10:866–876. **doi:**10.1038/nrm2805
9. Rothlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J et al. (2008) Kemp elimination catalysts by computational enzyme design. Nature 453:190–195. **doi:**10.1038/nature06879
10. Graber R, Kasper P, Malashkevich VN, Strop P, Gehring H et al. (1999) Conversion of aspartate aminotransferase into an L-aspartate β-decarboxylase by a triple active-site mutation. J Biol Chem 274:31203–31208. **doi:** 10.1074/jbc.274.44.31203
11. Xiang H, Luo L, Taylor KL, Dunaway-Mariano D (1999) Interchange of catalytic activity within the 2-enoyl-coenzyme A hydratase/isomerase superfamily based on a common active site template. Biochemistry 38:7638–7652. **doi:**10.1021/bi9901432
12. Ma H, Penning TM (1999) Conversion of mammalian 3α-hydroxysteroid dehydrogenase to 20α-hydroxysteroid dehydrogenase using loop chimeras: Changing specificity from androgens to progestins. Proc Natl Acad Sci USA. 96:11161–11166. **doi:**10.1073/pnas.96.20.11161
13. Wilson EM, Kornberg HL (1963) Properties of crystalline L-aspartate 4-carboxy-lyase from *Achromobacter sp.* Biochem J 88:578–587.
14. McLoughlin SY, Copley SD (2008) A compromise required by gene sharing enables survival: Implications for evolution of new enzyme activities. Proc Natl Acad Sci U S A 105:13497–13502. **doi:**10.1073/pnas.0804804105
15. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I (2007) Multicopy suppression underpins metabolic evolvability. Mol Biol Evol 24:2716–2722. **doi:**10.1093/molbev/msm204
16. Copley SD (2009) Evolution of efficient pathways for degradation of anthropogenic chemicals. Nat Chem Biol 5:559–566. **doi:**10.1038/nchembio.197
17. Kuo CH, Ochman H (2010) The extinction dynamics of bacterial pseudogenes. PLoS Genet 6:e1001050. **doi:**10.1371/journal.pgen.1001050
18. Gauger AK, Ebnet S, Fahey PF, Seelke R (2010) Reductive evolution can prevent populations from taking simple adaptive paths to high fitness. BIO-Complexity 2010(2):1–9. **doi:**10.50498/BIO-C.2010.2
19. Yoshikuni Y (2006) Designed divergent evolution of enzyme function. Nature 440:1078-1082. **doi:**10.1038/nature04607
20. Aharoni A, Gaidukov L, Khersonsky O, Gould SM, Roodveldt C et al. (2005) The 'evolvability' of promiscuous protein functions. Nat Genet 37:73–76. **doi:**10.1038/ng1482
21. Behe MJ, Snoke DW (2004) Simulating evolution by gene duplication of protein features that require multiple amino acid residues. Protein Sci 13:2651–2664. **doi:**10.1110/ps.04802904
22. Durrett R, Schmidt D (2008) Waiting for two mutations: With applications to regulatory sequence evolution and the limits of Darwinian evolution. Genetics 180:1501–1509. **doi:**10.1534/genetics.107.082610
23. Behe MJ (2009) Waiting longer for two mutations. Genetics 181:819–820; author reply 821–812. **doi:**10.1534/genetics.108.098905
24. Lynch M, Abegg A (2010) The rate of establishment of complex adaptations. Mol Biol Evol 27:1404–1414. **doi:**10.1093/molbev/msq020
25. Axe DD (2010) The limits of complex adaptation: an analysis based on a simple model of structured bacterial populations. BIO-Complexity 2010(4):1–10. **doi:**10.5048/BIO-C.2010.4
26. Gauger AK, Axe DD (2011) The evolutionary accessibility of new enzyme functions: A case study from the biotin pathway. BIO-Complexity 2011(1):1–17. **doi:**10.5048/BIO-C.2011.1
27. Schmidt DM, Mundorff EC, Dojka M, Bermudez E, Ness JE et al. (2003) Evolutionary potential of (beta/alpha)8-barrels: Functional promiscuity produced by single substitutions in the enolase superfamily. Biochemistry 42:8387–8393. **doi:**10.1021/bi034769a
28. Näsvall J, Sun L, Roth JR, Andersson DI (2012) Real-time evolution of new genes by innovation, amplification, and divergence. Science 338:384–387. **doi:**10.1126/science.1226521
29. Cleary PP, Campbell A (1972) Deletion and complementation analysis of biotin gene cluster of Escherichia coli. J Bacteriol 112:830–839.
30. Mann S, Ploux O (2011) Pyridoxal-5'-phosphate-dependent enzymes involved in biotin biosynthesis: Structure, reaction mechanism and inhibition. Biochim Biophys Acta 1814:1459–1466. **doi:** 10.1016/j.bbapap.2010.12.004
31. Marcus JP, Dekker EE (1993) Threonine formation via the coupled activity of 2-amino-3-ketobutyrate coenzyme A lyase and threonine dehydrogenase. J Bacteriol 175:6505–6511.
32. Alexeev D, Alexeeva M, Baxter RL, Campopiano DJ, Webster SP et al. (1998) The crystal structure of 8-amino-7-oxononanoate synthase: A bacterial PLP-dependent, acyl-CoA-condensing enzyme. J Mol Biol 284:401–419. **doi:**10.1006/jmbi.1998.2086
33. Schmidt A, Sivaraman J, Li Y, Larocque R, Barbosa J et al. (2001) Three dimensional structure of 2-amino-3-ketobutyrate CoA ligase from *Escherichia coli* complexed with a PLP-substrate intermediate: inferred reaction mechanism. Biochem 40:5151–5160. **doi:**10.1021/bi002204y
34. Fox NK, Brenner SE, Chandonia JM (2014) SCOPe: Structural Classification of Proteins Extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 42:D304–309. **doi:** 10.1093/nar/gkt1240
35. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S et al. (2013) EcoCyc: Fusing model organism databases with systems biology. Nucleic Acids Res 41:D605–612. **doi:**10.1093/nar/gks1027
36. Kubota T, Shimono J, Kanameda C, Izumi Y (2007) The first thermophilic α-oxoamine synthase family enzyme that has activities of 2-amino-3-ketobutyrate CoA ligase and 7-Keto-8-aminopelargonic acid synthase: Cloning and overexpression of the gene from an extreme thermophile, *Thermus thermophilus*, and characterization of its gene product. Biosci Biotech Biochem 71:3033–3040.
37. Eliot AC, Kirsch JF (2004) Pyridoxal phosphate enzymes: Mechanistic, structural, and evolutionary considerations. Annu Rev Biochem 73:383–415. **doi:**10.1146/annurev.biochem.73.011303.074021
38. Ferreira GC, Gong J (1995) 5-Aminolevulinate synthase and the first step of heme biosynthesis. J Bioenergy Biomembranes 27:151–159.
39. Hanada K (2003) Serine palmitoyltransferase, a key enzyme of sphingolipid metabolism. Biochimica et Biophysica Acta (BBA)—Molecular and cell biology of lipids 1632:16–30. **doi:**10.1016/S1388-1981(03)00059-3
40. Astner I, Schulze JO, Van den Heuvel J, Jahn D, Schubert W-D et al. (2005) Crystal structure of 5-aminolevulinate synthase, the first enzyme of heme biosynthesis, and its link to XLSA in humans. The EMBO Journal 24:3166–3177. **doi:**10.1038/sj.emboj.7600792

41. Kitagawa M, Ara T, Arifuzzaman M, Ioka-Nakamichi T, Inamoto E et al. (2005) Complete set of ORF clones of *Escherichia coli* ASKA library (a complete set of *E. coli* K-12 ORF archive): Unique resources for biological research. DNA Res 12:291–299. **doi:**10.1093/dnares/dsi012
42. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y et al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: The Keio collection. Mol Syst Biol 2:2006–2008. **doi:**10.1038/msb4100050
43. Jarrett JT (2005) Biotin synthase: Enzyme or reactant? Chem Biol 12:409–410. **doi:**10.1016/j.chembiol.2005.04.003
44. Axe DD (2010) The case against a Darwinian origin of protein folds. BIO-Complexity 2010(1):1–12. **doi:**10.5048/BIO-C.2010.1
45. Axe DD, Gauger AK (2013) Explaining metabolic innovation: Neo-Darwinism versus design. In: Marks II RJ, Behe MJ, Dembski WA, Gordon BL, Sanford JC (eds) Biological Information: New Perspectives. World Scientific, pp. 489–507.
46. Milkman R, Stoltzfus A (1988) Molecular evolution of the *Escherichia coli* chromosome. II. Clonal segments. Genetics 120:359–366.
47. Lynch M, Conery JS (2003) The origins of genome complexity. Science 302:1401–1404. **doi:**10.1126/science.1089370
48. Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. Genetics 148:1667–1686.
49. Reams AB, Kofoid E, Savageau M, Roth JR (2010) Duplication frequency in a population of *Salmonella enterica* rapidly approaches steady state with or without recombination. Genetics 184:1077–1094. **doi:**10.1534/genetics.109.111963
50. Wielgoss S, Barrick JE, Tenaillon O, Wiser M, Dittmar WJ et al. (2013) Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. Proc Nat Acad Sci USA 110:222–227. **doi:** 10.1073/pnas.1219574110