Research Article

The Evolutionary Accessibility of New Enzyme Functions: A Case Study from the Biotin Pathway

Ann K. Gauger and Douglas D. Axe*

Biologic Institute, Redmond, WA, USA

Abstract

Enzymes group naturally into families according to similarity of sequence, structure, and underlying mechanism. Enzymes belonging to the same family are considered to be *homologs*—the products of evolutionary divergence, whereby the first family member provided a starting point for conversions to new but related functions. In fact, despite their similarities, these families can include remarkable functional diversity. Here we focus not on minor functional variations within families, but rather on *innovations*—transitions to genuinely new catalytic functions. Prior experimental attempts to reproduce such transitions have typically found that many mutational changes are needed to achieve even weak functional conversion, which raises the question of their evolutionary feasibility. To further investigate this, we examined the members of a large enzyme superfamily, the PLP-dependent transferases, to find a pair with distinct reaction chemistries and high structural similarity. We then set out to convert one of these enzymes, 2-amino-3-ketobutyrate CoA ligase (Kbl_2), to perform the metabolic function of the other, 8-amino-7-oxononanoate synthase ($BioF_2$). After identifying and testing 29 amino-acid changes, we found three groups of active-site positions and one single position where Kbl_2 side chains are incompatible with $BioF_2$ function. Converting these side chains in Kbl_2 makes the residues in the active-site cavity identical to those of $BioF_2$, but nonetheless fails to produce detectable $BioF_2$-like function in *vivo*. We infer from the mutants examined that successful functional conversion would in this case require seven or more nucleotide substitutions. But evolutionary innovations requiring that many changes would be extraordinarily rare, becoming probable only on timescales much longer than the age of life on earth. Considering that Kbl_2 and $BioF_2$ are judged to be close homologs by the usual similarity measures, this result and others like it challenge the conventional practice of inferring from similarity alone that transitions to new functions occurred by Darwinian evolution.

Cite as: 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

Editor: Michael Behe

Received: December 17, 2010; **Accepted:** March 21, 2011; **Published:** April 11, 2011

Notes: A *Critique* of this paper, when available, will be assigned **doi:**10.5048/BIO-C.2011.1.c.

* Email: daxe@biologicinstitute.org

INTRODUCTION

Enzymes are proteins or protein complexes that carry out the chemical transformations necessary for life. Because their functional properties follow directly from their genetically encoded amino-acid sequences, enzymes link genotypes to phenotypes in a relatively simple way. This simplicity provides a valuable opportunity to examine the problem of biological innovation—the origin of completely new functions—in ways that cannot normally be achieved with high-level functions that depend on many genes [1].

Functional innovations throughout the history of enzymes may be divided into two categories based on the degree to which they depend upon *structural* innovation. The first category, which we call large-scale innovation, includes all cases where the new function is provided by a fundamentally new structure—a new protein *fold*. Innovation on this large scale seems to have occurred well over a thousand times, judging by the number of distinct folds known to exist[1]. The second category, small-scale innovation, includes all cases where new function is provided by relatively small structural adjustments to an existing protein fold. We likewise infer from the many examples of different enzyme functions being accomplished by similar structures that these small-scale innovations have occurred many times in the history of life.

However, whether the standard neo-Darwinian model adequately explains enzymatic innovation on either scale remains an open question, as does its adequacy for explaining innovation generally [2–4]. One of us (DDA) has recently described the difficulties that the standard model encounters in attempting to explain large-scale enzymatic innovation, concluding that the model is inadequate [5]. Its adequacy with respect to the small-scale problem is therefore a matter of further interest, to which we turn here.

[1] See http://scop.mrc-lmb.cam.ac.uk/scop/count.html

A high degree of structural similarity between two proteins is taken as strong evidence for their homology, meaning their evolutionary relatedness. If the genes encoding the proteins were separated by a speciation event, then the proteins (and their genes) are known as *orthologs*. Alternatively, they may have been separated by a gene duplication event, in which case they are known as *paralogs*. Orthologs typically continue to serve the same functional role, whereas paralogs are usually found to have different roles, as otherwise their functional redundancy would tend to favor elimination of the duplicate. Paralogous divergence is therefore thought to be the main way that small-scale enzymatic innovations are produced [6–10].

There are two views regarding the sequence of events by which paralogous divergence occurs. The first is that it happens only after a duplication event has provided a spare gene. Because of their functional redundancy, these spare genes are able to accumulate mutations with no selective cost. The cost of redundancy itself usually leads to the elimination of the duplicates, but occasionally mutations may endow them with a new adaptive function—a small-scale innovation. The second view, known as the *promiscuity* hypothesis, holds that functional diversity may be present before duplication occurs, in the form of a bi-functional enzyme [11,12]. Duplication, by this view, simply provides a way for genes to specialize by becoming optimized separately for the two pre-existing functions.

Either way, the underlying assumption is that the structural requirements for small-scale innovation are not prohibitively stringent. More specifically, it is assumed that when species encounter circumstances that present a new need for some enzymatic task to be performed, that need has a reasonably good chance of being met, either by exploiting existing promiscuous functions or by generating new functions. It is further assumed that these beneficial functions, which may initially be performed quite poorly, readily evolve to become the highly efficient functions we associate with natural enzymes.

These ideas have motivated a great many experimental projects aimed at harnessing the supposed power of mutation and selection. The results, however, have generally fallen well short of what might have been expected. Gerlt and Babbitt, for example, gave this sobering assessment of attempts to interconvert enzyme functions:

> Interchanging reactions catalyzed by members of mechanistically diverse superfamilies might be envisioned as "easy" exercises in (re)design: if Nature did it, why can't we? [...] Anecdotally, many attempts at interchanging activities in mechanistically diverse superfamilies have since been attempted, but few successes have been realized [13].

Functional conversion encounters problems both with the catalytic efficiencies achieved and with the number of mutations required to achieve them. Aspartate aminotransferase, for example, has been converted by seven base changes into an aspartate decarboxylase that is some 100,000-fold less efficient (based on k_{cat}/K_m) than a natural aspartate decarboxylase [14,15][2]. In another study, eleven base changes were used to convert a dehalogenase into a crotonase, with only 0.005% of wild-type crotonase activity achieved [16]. Somewhat better conversion, reaching 0.25% of wild-type activity, was accomplished between two hydroxysteroid dehydrogenases (HSDs), but even more extensive change was required. Having identified six residues in the binding pocket as the most likely determinants of specificity, the authors of that study reported that "Even when all the predicted mutations necessary to convert 3α-HSD to 20α-HSD were introduced, the resultant mutant T24Y/F129L/T226Y/W227C/N306F/Y310M had no 20α-HSD activity" [17]. Weak conversion was eventually achieved by transferring entire loops (consisting of 20, 32, and 63 amino-acid residues) from the target protein to the source protein [17].

Results like these raise a question that tends to be overlooked in the papers describing them. Namely, how many changes can the Darwinian mechanism feasibly combine in order to reach a new function? According to a recent analysis of the time required for complex adaptations[3] to appear by duplication and divergence, the answer is no higher than six base changes, with two probably being more realistic [18]. On that basis we consider the above conversions to be evolutionarily implausible simply because of the number of changes they required, whether or not the reported activities would have selective value in the wild.

Although the problem of too many changes is common in studies of functional conversion, there are a few examples where genuinely new chemistry appears to be achievable within the limit of two changes. Cited examples can be misleading, however. For example, atrazine chlorohydrolase is sometimes described as a recently evolved enzyme with a completely new function. It degrades atrazine, an unnatural chemical used as a crop herbicide, but the inferred ancestral enzyme (melamine deaminase) does likewise, albeit more slowly [19]. It appears, then, that this is an example of a preexisting (promiscuous) activity being refined rather than a genuinely new activity appearing. The most clear case of new catalytic activity involves an enzyme function (*o*-succinylbenzoate synthase, or OSBS) that can be achieved from two different starting points (different natural enzymes) by single base changes [20]. But the converted activities are very weak, amounting to only 0.0004% or 0.06% of wild-type activity [20]. So again it is unclear whether the converted functions would provide enough benefit to be of evolutionary significance.

Although it is possible to compensate for poor functional conversion to OSBS by over-expressing the converted genes, this reduces the evolutionary plausibility in two respects. First, the over-expression itself would require particular genetic modifications, making the total complexity of the adaptation greater than the single change to the enzyme. Second, because over-expression involves a significant metabolic cost [21,22], adaptive evolution may eliminate over-expressed genes more readily than it tinkers with them [23]. This presents a catch-

[2] Reference 15 reports 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 88 s^{-1}.

[3] The term complex adaptation refers to adaptations requiring multiple base changes, with the incomplete stages being non-adaptive.

22 situation for the fate of duplicate genes. If they are strongly expressed they are vulnerable to rapid elimination, but if they are weakly expressed the new function would need to appear with high proficiency in order to have a selective effect.

The promiscuity hypothesis seems to offer a way out of this by positing that small-scale innovations can originate as secondary functions in enzymes that are already highly beneficial because of their primary functions. The primary function guarantees that the gene is preserved and expressed, potentially making it a good platform for secondary functions to 'hitchhike' their way to selective success. The obvious difficulty, though, is that efficient performance of the primary function seems to require that hitchhiking be minimized. Indeed, an important study by Patrick *et al.* [24] shows this promiscuous hitchhiking to be a limited exception rather than a rule. They used 104 auxotrophic *E. coli* strains, each with a single-gene knockout, and a plasmid library in which all *E. coli* genes are individually over-expressed to find out how many of the missing gene functions can be filled in by other genes. Functional rescue was found to be possible for 21 of the knockouts, with fifteen of these cases appearing to involve metabolic workarounds of various kinds and only six appearing to involve catalytic promiscuity [24]. This shows that promiscuous activities do exist in modern enzymes, but it also indicates that they are rare. Furthermore, considering the high expression levels of the rescuing genes and the poor growth of the rescued strains,[4] it is again unclear whether the activities demonstrated are of evolutionary significance. An attempt at evolutionary optimization of one of these activities fell ten-million-fold short of wild-type proficiency [25], suggesting that they may actually be evolutionary dead ends.

On the whole, then, it is far from clear whether the structural changes we can expect from random mutations can accomplish the many small-scale enzymatic innovations attributed to them. Here, we explore this question by asking how many mutations are needed to achieve a genuine functional conversion in a case where the necessary structural change is known to be small relative to the changes commonly attributed to paralogous divergence. We focus not on minor functional adjustments, like shifts in substrate profiles, but rather on true innovations—the jumps to new chemistry that must have happened but which seem to defy gradualistic explanation. The relative difficulty of these innovations has already been acknowledged [26]:

> Some functions, however, simply cannot be reached through a series of small uphill steps and instead require longer jumps that include mutations that would be neutral or even deleterious when made individually. Examples of functions that might require multiple simultaneous mutations include the appearance of a new catalytic activity....

Our aim is to get a better understanding of how difficult these jumps really are.

We begin by analyzing the structural similarities among pairs of enzymes in a large 'superfamily' of presumed homologs, the pyridoxal-5′-phosphate (PLP) dependent transferases. With over fifty structurally characterized enzymes that share a common fold but catalyze distinct reactions, these proteins provide an exceptionally rich picture of the structural basis for functional diversity among enzymes [27]. After identifying a pair with very close structural similarity but no functional overlap, we used a direct experimental approach to test the importance of various amino-acid side chains as determinants of the respective functions. The results enabled us to estimate how many nucleotide substitutions are likely to be required to achieve a functional conversion.

APPROACH

Figure 1 illustrates for a hypothetical family of enzymes how functional divergence depends on structural divergence. The outer circle encloses a region in protein structure space corresponding to a protein fold that can support a variety of enzymatic functions (represented by colors). In this example, the gene encoding the earliest function (bronze) gave rise to three new functions by separate small-scale innovations (dashed arrows). For the bronze function to have been retained in the process, the innovations would have appeared by paralogous divergence, each starting with a duplication of the bronze gene that enabled one copy to lose that function while the other kept it. The new paralogs underwent rapid refinement, depicted as migration toward the centers of their respective colored regions. In some cases, functional overlap would allow new functions to appear as promiscuous secondary functions, as illustrated by the overlapping violet shades in Figure 1.

In terms of this picture, the question we aim to address for the PLP-dependent transferase superfamily is whether neighboring functions are generally accessible to evolutionary explo-

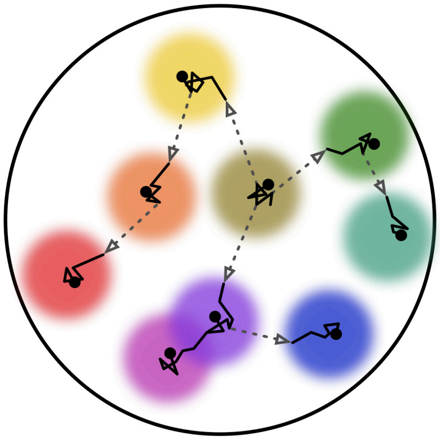


Figure 1: A structure-space mapping of the evolutionary history of a hypothetical enzyme family. Colors represent functions, as described in the text. Solid lines represent the structural effects of mutations, with black dots representing extant structures and dashed arrows representing jumps to new functions (i.e., transitions that pass through non-functional structural intermediates). The number of mutational steps required for these conversions is the subject of this study. In the case of overlapping functions (shades of violet) functional transition can occur without passing through non-functional intermediates. doi:10.5048/BIO-C.2011.1.f1

[4] In cases of promiscuous rescue, colony formation required up to fifteen days, the mean being eight days (see supplement to reference 24). Normal growth would produce colonies in about a day.

ration. We have no way to delineate the boundaries of the various functions in structure space, as represented in Figure 1, but we do have a large collection of extant enzyme structures for comparison. A reasonable assumption, consistent with methods used for reconstructing evolutionary histories, is that enzyme pairs with high structural similarity should be most amenable to functional conversion. Whether or not a particular conversion ever occurred as a paralogous innovation (or the direction in which it occurred if it did) is not the point of interest here. Rather, the point is to identify the kind of functional innovation that ought to be among the most feasible within this superfamily and then to assess how feasible this innovation is.

The metric we use to quantify pairwise structural similarity, which we call *structural distance* (δ_s), is based on the scoring function used by the SSM structure comparison algorithm [28]. That algorithm uses secondary structure matching to produce rough structural alignments, which are then scored with a quality-of-fit function, Q, defined as:

$$Q = N_{\text{align}}^2 \Big/ \left\{ \left[1 + (\text{RMSD}/3)^2 \right] N_1 N_2 \right\}, \qquad (1)$$

where RMSD is the root-mean-square deviation (Å) of aligned alpha carbons (numbering N_{align}), and N_1 and N_2 are the lengths of the two proteins. Structural alignments are optimized by maximizing Q. The maximal value, $Q_{\max}$, has been found to be a good metric for quantifying structural similarity on a scale ranging from zero, indicating no common secondary structure, to one for identical backbone structures [28].

From this, we define the structural distance between two structures as:

$$\delta_s \equiv \frac{1}{Q_{\max}} - 1 \, . \qquad (2)$$

From the above properties of $Q_{\max}$, it follows that δ_s is a unitless quantity ranging in value from zero for identical structures to substantially greater than one for structures without significant similarity. The singularity at $Q_{\max} = 0$ is of no practical significance, since it is only encountered in the case of structures lacking any common secondary structure (as would be the case if an all-helix structure were compared to an all-sheet structure).

The SCOP structural classification [29][5] provided us a nonredundant set of structures from the PLP-dependent transferases superfamily. Taking one structure from each named protein within this superfamily gave 57 coordinate files covering the whole spectrum of functions known to use this fold. After calculating δ_s for all possible pairs within this set[6], we used the neighbor-joining method [30] to construct a graph with nodes representing specific entries in the Protein Data Bank (i.e., structural coordinate files) and edges representing their structural distance (Figure 2). The graph's near-neighbor distances range from about 0.1 to 3.4 (mean = 0.89, median = 0.62) and furthest-neighbor distances (not shown) range from about four to nine. Figure 3 gives a visual sense of the structural differences corresponding to δ_s values ranging from 0.16 to 6.4.

[5] This work is based on SCOP release 1.73 (http://scop.mrc-lmb.cam.ac.uk/scop-1.73/data/scop.b.d.jc.b.html).

[6] Using $Q_{\max}$ scores from the online tool at http://www.ebi.ac.uk/msd-srv/ssm/.

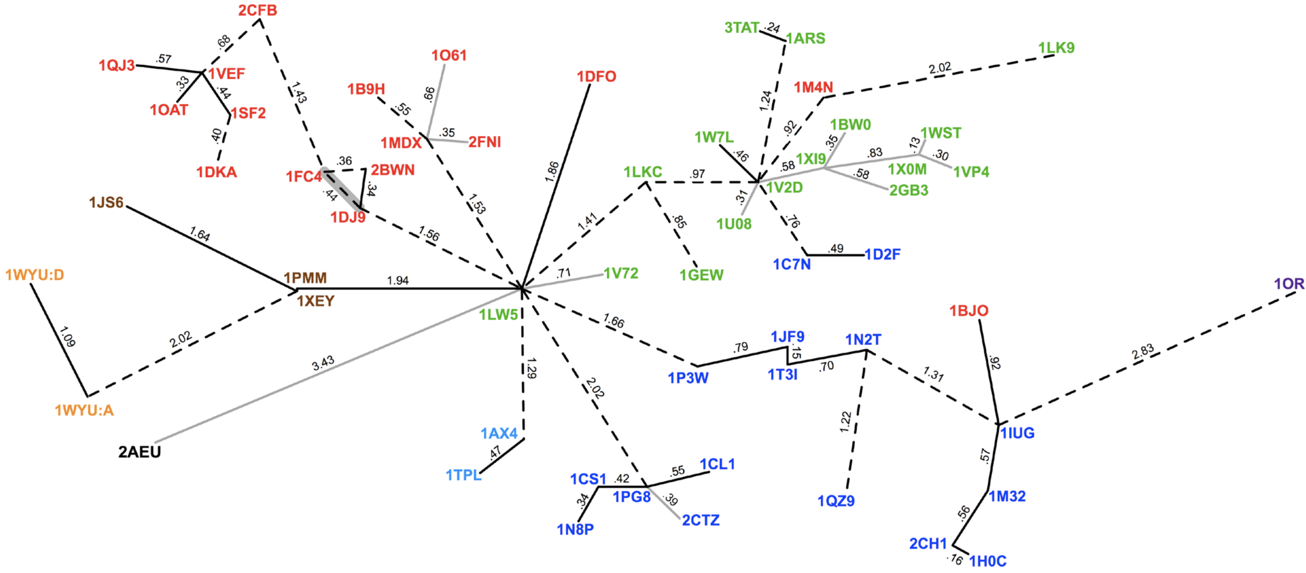


Figure 2: Near-neighbor structural distance graph for the SCOP PLP-dependent transferases superfamily. Nodes show PDB entry names colored according to SCOP family assignments (http://scop.mrc-lmb.cam.ac.uk/scop-1.73/data/scop.b.d.jc.b.html): green = aspartate aminotransferase-like, blue = cystathionine synthase-like, brown = pyridoxal-dependent decarboxylase, red = GABA aminotransferase-like, cyan = beta-eliminating lyases, gold = glycine dehydrogenase subunits, purple = ornithine decarboxylase major domain, and black = SelA-like. Edge lengths and connectivity are based on structural data as described, with dashed edges connecting enzymes having different chemistries, grey edges radiating from nodes with poorly characterized functions, and the back-shaded edge showing the functional transition examined here (as described in the text). Other aspects of geometry (e.g., layout and distances between unjoined nodes) are arbitrary. **doi:**10.5048/BIO-C.2011.1.f2

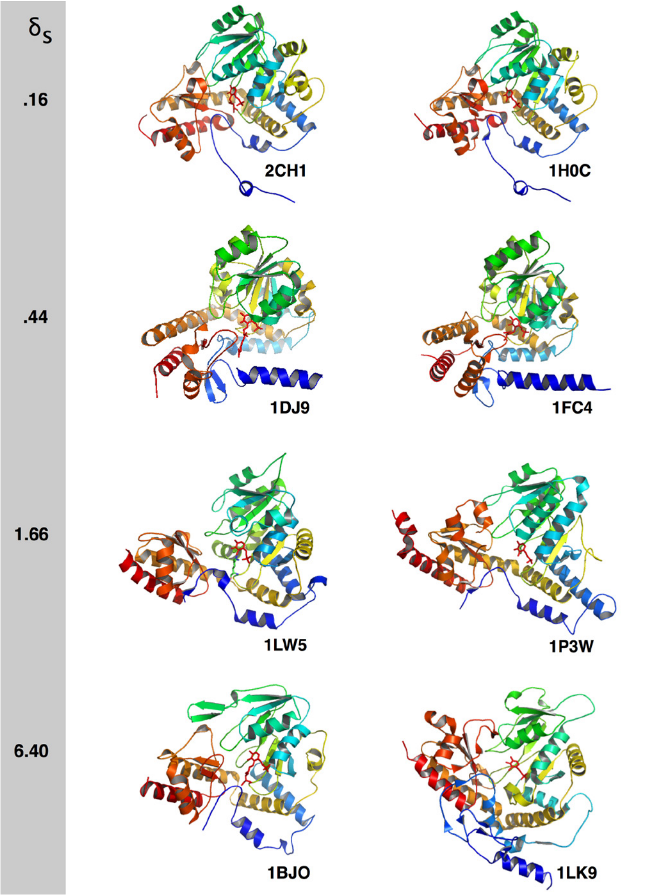


Figure 3: Visual comparison of selected monomer pairs ranging in structural distance. PLP is shown red. Rainbow coloring runs from blue at N termini to red at C termini. **doi:**10.5048/BIO-C.2011.1.f3

As would be expected, very close structural neighbors have substantially overlapping functions. Figure 4 describes the functional overlap, if any, for all neighbor pairs having δ_s values below half the mean value. Isozymes like 1XEY and 1PMM overlap completely, whereas other pairs overlap by sharing some but not all reactions (e.g., 1ARS and 3TAT) or by sharing a reaction mechanism but differing in substrates (e.g., 1OAT and 1VEF). As structural distance increases, examples with no overlap (i.e., no shared general reaction) appear. The first example of this in Figure 4 is 2BWN and 1FC4, with $\delta_s = 0.36$. Edges connecting functionally distinct pairs like this are indicated by dashed lines in Figure 2. Clearly, any proposed cause of this superfamily's diversity needs to be able to produce innovations that cross these functional divisions.

Three of the dashed lines in Figure 2 are short enough for inclusion in Figure 4, the paired coordinate files being 2BWN and 1FC4, 1DKA and 1SF2, and 1FC4 and 1DJ9. Because these pairs have comparable structural distances ($\delta_s = 0.40 \pm 0.04$), we chose the one where both proteins come from *E. coli*, the proteins being Kbl (1FC4) and BioF (1DJ9). These proteins form homodimeric enzymes, designated Kbl$_2$ (2-amino-3-ketobutyrate CoA ligase)[7] and BioF$_2$ (8-amino-7-oxononanoate synthase), that show clear structural similarity (Figures 3 and 5) and significant catalytic similarity as well (Figure 6) [42, 44, 48]. Despite these similarities, they contribute to very different metabolic pathways—Kbl$_2$ being involved in threonine metabolism [49], while BioF$_2$ is required for the production of biotin, an essential cofactor in fatty-acid synthesis and other carboxylation reactions [50–52]. Although no functional overlap between the two enzymes is evident in *E. coli*, both functions have been detected *in vitro* in an enzyme isolated from *T. thermophilus* [53]. This affirms the choice of these enzymes as candidates for functional interconversion by showing that a single structure can perform both functions.

RESULTS

Because bacteria require only tiny amounts of biotin for growth (possibly as little as 100 molecules per cell [54]), biotin production can be selected with very high sensitivity. We therefore examined the feasibility of converting Kbl to perform the function of BioF, rather than the reverse. Complete functional conversion can obviously be achieved by complete sequence conversion. But since that would require some 250 amino-acid substitutions (the sequences being 34% identical over 381 aligned positions), it is equally plain that conversion must be achievable with far fewer changes if this sort of task is to be evolutionarily feasible. Our aim is therefore to identify the most important changes among the 250 initial candidates.

We used a three stage process to do this. First we used sequence and structure information to identify a small subset—a short list—of the 250 changes that are apt to be most important in distinguishing the BioF function from the Kbl function. Then we tested changes from that short list, individually or in small groups, to confirm which really are critical with respect to BioF function. This was done in the reverse direction (i.e., *bioF→kbl*) by modifying the *bioF* gene to make its product slightly more Kbl-like, and then testing for biotin auxotrophy (phenotype: Bio$^-$). Finally, based on the results of those tests, we constructed the reciprocal *kbl→bioF* mutants where codons in *kbl* were changed to incorporate the critical BioF residues. These mutant *kbl* genes were then tested for their ability to confer the Bio$^+$ phenotype in a strain lacking a chromosomal *bioF* gene.

Stage 1: Short-listing substitutions

Two approaches were used to determine which of the amino-acid residues that distinguish BioF from Kbl (both as found in *E. coli*) are apt to be the most critical determinants of BioF function. First, by aligning the *E. coli* BioF sequence with BioF sequences from other bacteria, we identified all fully conserved BioF residues that differ from their counterparts in the *E. coli* version of Kbl. The assumption here is that the consensus BioF sequence should carry more information about what is necessary for BioF function than any single sequence does. Second, by aligning the structures of the *E. coli* versions of BioF$_2$ and Kbl$_2$, we determined which BioF active-site residues differ from their structural counterparts in Kbl. Here the reasoning is that side chains forming the substrate-binding and catalytic interface are the most likely to influence catalytic specificity.

[7] CoA is an abbreviation of coenzyme A.

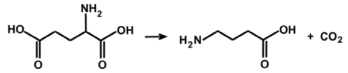
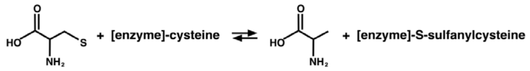
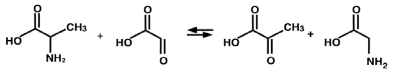
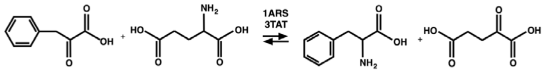
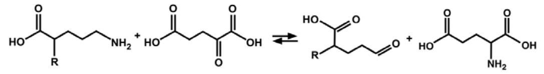
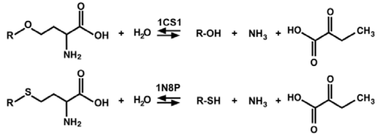
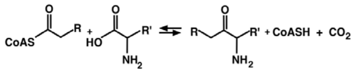
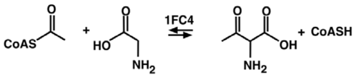
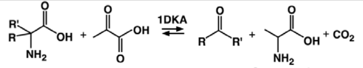
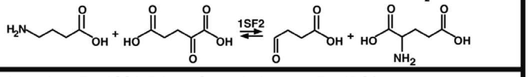
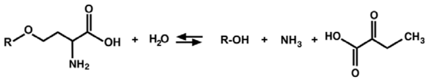
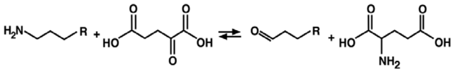
PDB Pair	δ_S	Enzyme Names	Active unit	Comparison of Catalyzed Reactions
1PMM	.02	Glutamate decarboxylase beta	hexamer	Isozymes [31]: 
1XEY		Glutamate decarboxylase beta	hexamer	
1JF9	.15	Selenocysteine lyase with cysteine desulferase activity	dimer	Shared reaction: cysteine desulferase [32, 33]. 
1T3I		Probable cysteine desulferase	dimer	
2CH1	.16	3-hydroxykynurenine transaminase	dimer	Probable shared reaction [34-36]: 
1H0C		Alanine-glyoxylate aminotransferase	dimer	
1ARS	.24	Aspartate aminotransferase	dimer	Shared reaction [37]: 
3TAT		Aromatic aminoacid aminotransferase	dimer	
1OAT	.33	Ornithine aminotransferase	dimer	Shared general reaction [38, 39]: For 1OAT, R = NH$_2$; For 1VEF, R = NHCOCH$_3$ 
1VEF		Acetylornithine aminotransferase	dimer	
1CS1	.34	Cystathionine gamma synthase	tetramer	Similar general reaction [40, 41]: For 1CS1 R = O-succinyl group For 1N8P R = CH$_2$CH(NH$_2$)COOH 
1N8P		Cystathionine gamma lyase	tetramer	
2BWN	.34	5-aminolevulinate synthase	dimer	Shared general reaction [42, 43]: For 2BWN, R = CH$_2$COOH and R' = H. For 1DJ9, R = (CH$_2$)$_4$COOH and R' = CH$_3$. 
1DJ9		8-amino-7-oxononanoate synthase	dimer	
2BWN	.36	5-aminolevulinate synthase	dimer	Different chemistry; same cofactor (for 2BWN see above) [43, 44]: 
1FC4		2-amino-3-ketobutyrate CoA ligase	dimer	
1DKA	.40	Dialkylglycine decarboxylase	tetramer	Different chemistry [45, 46]:  
1SF2		4-aminobutyrate aminotransferase	tetramer	
1CS1	.42	Cystathionine gamma synthase	tetramer	Shared reaction. For 1CS1 R is an O-succinyl group. For 1PG8 R can be an O-acetyl or O-ethyl group [40, 47]: 
1PG8		Methionine alpha-, gamma-lyase	octomer	
1FC4	.44	2-amino-3-ketobutyrate CoA ligase	dimer	Different chemistry; same cofactor. See other pairings above for chemistries and references.
1DJ9		8-amino-7-oxononanoate synthase	dimer	
1VEF	.44	Acetylornithine aminotransferase	dimer	Shared general reaction [39, 46]: For 1VEF, R = CH(NHCOCH$_3$)COOH; For 1SF2, R = COOH 
1SF2		4-aminobutyrate aminotransferase	tetramer	

Figure 4: Functional comparison of enzyme pairs with high structural similarity $(d_s < \bar{d}_s/2)$**.** Because the functions of 1X0M , 1U08, 2FNI, 1XI9, and 2CTZ are poorly characterized, the following structurally similar pairs are omitted: 1X0M and 1WST, 1X0M and 1VP4, 1U08 and 1VP4, 1U08 and 1V2D, 2FNI and 1MDX, 1XI9 and 1BW0, and 2CTZ and 1PG8. doi:10.5048/BIO-C.2011.1.f4

BioF sequences for the first method were found by performing a BLAST search on the Concise Microbial Protein Database[8] (which reduces redundancy by including only one sequence from each genus-level cluster of similar proteins), with the *E. coli* BioF as the query sequence. Specifying a minimum amino-acid identity of 45%, 35 taxonomically distinct BioF sequences were identified. A ClustalW [55] alignment of these shows 53 amino-acid positions with no variation (see Supplement [56]). The *E. coli* Kbl sequence conforms to this BioF consensus at 38 positions, leaving only fifteen candidates short-listed for functional conversion (Figure 7).

For the structural method, we aligned the PLP-bound dimeric structures of the two *E. coli* enzymes such that the functionally central PLP moieties overlap in one of the two symmetry-related active sites. Many active-site residues are seen to be identical in this alignment (Figure 5C). Those that differ, making them candidates for functional conversion, cluster into three groups along the aligned chains (Figures 7, 8). Of the nineteen candidate residues in these groups, five were previously identified by sequence alignment and fourteen are new, bringing the total number of candidate positions identified by the two methods to 29.

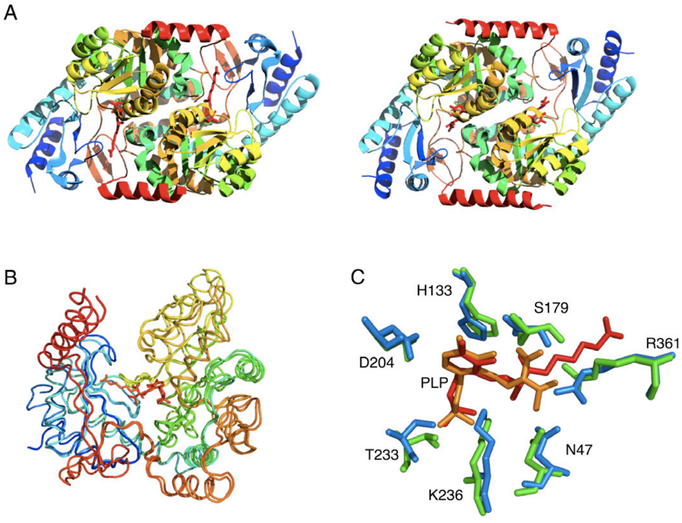


Figure 5: Structural similarity of BioF and Kbl. A) Dimeric enzymes $BioF_2$ (left; 1DJ9 [48]) and Kbl_2 (right; 1FC4 [44]) viewed along axes of symmetry with external aldimine complexes (PLP covalently linked to enzyme product) in red. Active sites are at the monomer interfaces. B) Aligned backbones of BioF and Kbl monomers. C) 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).
doi:10.5048/BIO-C.2011.1.f5

Stage 2: Testing short-listed candidates by BioF→Kbl mutation

For each of the fifteen candidate positions identified by sequence alignment, we constructed a mutant *bioF* gene specifying the Kbl amino acid at the candidate site. Plasmids carrying these mutant genes were introduced into an engineered strain of *E. coli* lacking the chromosomal *bioF* gene (see Methods) in order to test for biotin auxotrophy. Interestingly, of these fifteen single amino-acid substitutions, the only one disruptive enough to produce the Bio^- phenotype is the replacement of histidine 152 with asparagine (see Table 1).

That this substitution is so disruptive is surprising, given that H152 lies on the enzyme surface some distance away from the active site cavity (Figure 9). The role of this histidine is worth considering further, since to our knowledge it has never been identified as functionally significant before. A previous study of the effects of single amino-acid changes on the function of a bacterial ribonuclease [58] found that all inactivating substitutions fell into one of three classes: 1) those that replaced a side chain directly involved in substrate binding or catalysis, 2) those that replaced a buried side chain (<10% solvent exposure), or 3) those that introduced a proline or replaced a glycine. Assuming these rules may hold for other enzymes as well, we see that $BioF_{H152N}$ does not fit into classes 2 or 3, which suggests that H152 may have a direct functional role. If so, its position outside the active-site cavity suggests a binding role rather than a catalytic role.

One of the two substrates in the $BioF_2$ reaction, pimeloyl-CoA, has a long chain-like structure. Although the reported $BioF_2$ structures do not show this molecule, the substrate-bound structure of a very similar PLP-dependent enzyme, 5-aminolevulinate synthase (ALAS)[9], provides an informative comparison. Like $BioF_2$, ALAS uses a CoA derivative (succinyl-CoA) in its reaction. The low structural distance between these

[8] http://www.ncbi.nlm.nih.gov/genomes/prokhits.cgi

[9] Its structure is represented in Figure 2 as 2BWN, the third vertex of the triangle including the structures of $BioF_2$ (1DJ9) and Kbl_2 (1FC4). PDB entry 2BWO [43] describes the substrate-bound structure of ALAS.

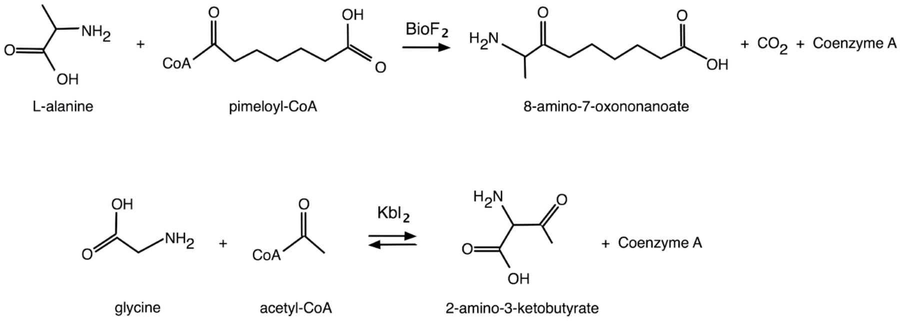


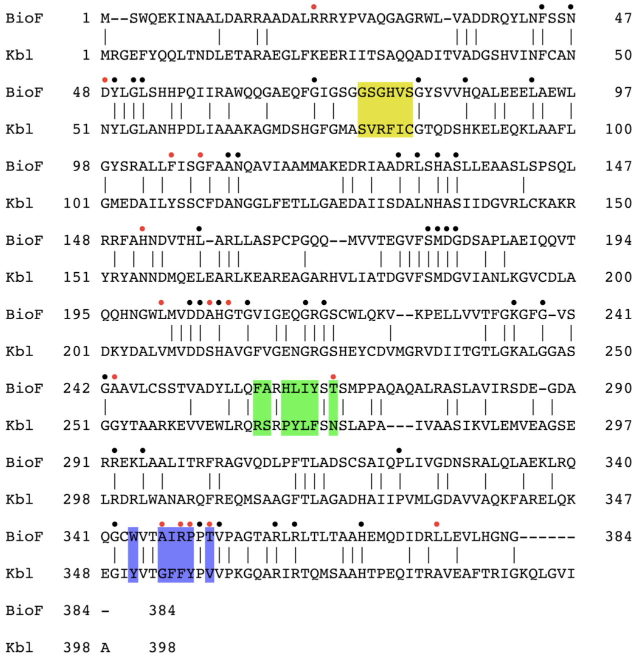
Figure 6: Biological functions of $BioF_2$ and Kbl_2. The Kbl_2 reaction may occur in either direction. **doi:**10.5048/BIO-C.2011.1.f6

enzymes (δ_s = 0.34) and their shared reaction chemistry[10] suggest that CoA should bind to them in very similar ways. In ALAS, CoA binds to an exterior pocket, with its reactive end (carrying the succinyl moiety) entering the active-site cavity through an opening (Figure 9). ALAS also has a histidine residue corresponding to H152 in $BioF_2$, which is seen in Figure 9 not to be in contact with CoA. This suggests that H152 of BioF likewise does not interact directly with CoA in the reaction complex, leaving its functional role unexplained. Two possibilities are that it affects the reaction indirectly by altering the structure of the active site in a decisive way, or that it interacts directly but transiently with pimeloyl-CoA, perhaps by playing a crucial role in getting this relatively large substrate molecule into position for reaction without actually holding it in place during the reaction.

The other fourteen BioF substitutions are significantly less disruptive than $BioF_{H152N}$ (Table 1). This implies that these other residues have less crucial roles than H152, but considering the inherent limitations of phenotype tests [58] they may nonetheless make significant contributions. Unless introducing the equivalent of histidine 152 into Kbl is sufficient in itself to cause functional conversion (to be tested below), a more inclusive way of identifying important residues is needed. A simple solution is to test mutations in small groups. If simultaneous change of several BioF residues to the Kbl amino-acids were to produce the Bio^- phenotype, this would mean that the Kbl sequence is unsuited for BioF function at some position or combination of positions within the group. We refer to the grouped positions in this case as forming a *critical locus*, meaning that important determinants of function reside somewhere within this set of positions.

To form mutation groups, we combined positions that were seen to form natural groups in the structural comparison. This not only limited the extent of simultaneous change to a very modest level (under 2% of BioF positions), but also restricted grouped positions to the active-site cavity, making mechanistic malfunction (rather than structural destabilization) the most likely cause of inactivation. When BioF substitutions were combined in these three natural groups, the resulting mutants each produced the Bio^- phenotype (Table 1), indicating that all three groups contain important determinants of BioF function. In all, then, we have identified four critical loci—three loci consisting of six or seven grouped positions (each), and one locus consisting of position 152 alone.

Pinpointing the cause of functional importance for all the group loci would require considerable further work. However, group 3 provided an opportunity to examine its constituent mutations individually, since four of them had already been examined (see bottom of Figure 7). We therefore constructed mutant *bioF* genes to test the remaining two positions individually and found that they, like the first four, conferred the Bio^+ phenotype (Table 1). This shows that the functional importance of the group 3 critical locus involves multiple positions.



```
BioF    1 M--SWQEKINAALDARRAADALRRRYPVAQGAGRWL-VADDRQYLNFSSN    47
Kbl     1 MRGEFYQQLTNDLETARAEGLFKEERIITSAQQADITVADGSHVINFCAN    50
BioF   48 DYLGLSHHPQIIRAWQQGAEQFGIGSGGSGHVSGYSVVHQALEEELAEWL    97
Kbl    51 NYLGLANHPDLIAAAKAGMDSHGFGMASVRFICGTQDSHKELEQKLAAFL   100
BioF   98 GYSRALLFISGFAANQAVIAAMMAKEDRIAADRLSHASLLEAASLSPSQL   147
Kbl   101 GMEDAILYSSCFDANGGLFETLLGAEDAIISDALNHASIIDGVRLCKAKR   150
BioF  148 RRFAHNDVTHL-ARLLASPCPGQQ--MVVTEGVFSMDGDSAPLAEIQQVT   194
Kbl   151 YRYANNDMQELEARLKEAREAGARHVLIATDGVFSMDGVIANLKGVCDLA   200
BioF  195 QQHNGWLMVDDAHGTGVIGEQGRGSCWLQKV--KPELLVVTFGKGFG-VS   241
Kbl   201 DKYDALVMVDDSHAVGFVGENGRGSHEYCDVMGRVDIITGTLGKALGGAS   250
BioF  242 GAAVLCSSTVADYLLQFARHLIYSTSMPPAQAQALRASLAVIRSDE-GDA   290
Kbl   251 GGYTAARKEVVEWLRQRSRPYLFSNSLAPA---IVAASIKVLEMVEAGSE   297
BioF  291 RREKLAALITRFRAGVQDLPFTLADSCSAIQPLIVGDNSRALQLAEKLRQ   340
Kbl   298 LRDRLWANARQFREQMSAAGFTLAGADHAIIPVMLGDAVVAQKFARELQK   347
BioF  341 QGCWVTAIRPPTVPAGTARLRLTLTAAHEMQDIDRLLEVLHGNG------   384
Kbl   348 EGIYVTGFFYPVVPKGQARIRTQMSAAHTPEQITRAVEAFTRIGKQLGVI
BioF  384 -    384
Kbl   398 A    398
```

Figure 7: Alignment of *E. coli* BioF and Kbl sequences. Vertical lines indicate matches. Dots identify positions where aligned bacterial BioF sequences (see text) show no variation, red ones showing where *E. coli* Kbl residues differ from the invariant BioF residues. Boxes (colored according to Figure 8) show positions identified by structural comparison as described. Aligned with StretcherP [57] (BLOSUM scoring matrix; gap penalty = 12; extend penalty = 2). doi:10.5048/BIO-C.2011.1.f7

Stage 3: Introducing Kbl→BioF mutations at critical loci

We have shown that the normal function of BioF can be greatly impaired by changing certain small subsets of the roughly 250 residues that distinguish it from Kbl. Making BioF slightly more like Kbl in sequence can, in these demonstrated cases, cause it to cease functioning as BioF. If the objective is essentially the reverse—to convert Kbl to the function of BioF by making it slightly more like BioF in sequence—then the most necessary changes will be the reciprocals of those that ruin BioF. That is, the aspects of Kbl that we now know ruin BioF are the first things we should change.

Whether that will be enough depends not only on whether we have identified all critical loci, but also on where we draw the line between critical and non-critical. As noted above, mutations leaving BioF functional by the test used here may nonetheless cause some functional impairment. Consequently, if Kbl is made BioF-like only at critical loci, many of these sub-critical effects may add up, resulting in a Bio^- phenotype. The objective, though, is to see whether functional conversion can be achieved with a small number of amino-acid changes. This can be tested with the identified loci (where change appears to be necessary) even if conversion is unsuccessful.

We therefore constructed plasmids encoding mutant versions of Kbl where the four critical loci were made BioF-like, either individually or in combination. Testing these plasmids in the same way that the *bioF* plasmids were tested (see Methods) showed that none confer a Bio^+ phenotype (Table 1). This is true even in the case of $Kbl_{g1,g2,g3,N155H}$, where all side chains within the active-

[10] See the seventh pair compared in Figure 4 (2BWN and 1DJ9).

Table 1: Phenotypes of BioF and Kbl mutants

BioF construct1*	Bio phenotype	Kbl construct2†	Bio phenotype
wild-type	+	wild-type	–
R21K	+	*g*1	–
D48N	+	*g*2	–
F105Y	+	*g*3	–
G108C	+	N155H	–
H152N	–	*g*1, *g*2	–
L201V	+	*g*1, *g*3	–
A206S	+	*g*2, *g*3	–
G208A	+	*g*1, N155H	–
A243G	+	*g*2, N155H	–
T266N	+	*g*3, N155H	–
A347G	+	*g*1, *g*2, *g*3	–
R349F	+	*g*1, *g*2, *g*3, N155H	–
P350Y	+		
T352V	+		
L376A	+		
*g*1	–		
*g*2	–		
*g*3	–		
W344Y	+		
I348F	+		

* Indicated *bioF*→*kbl* mutations were made to the *bioF* gene on plasmid pKBF$_2$, which was then used to transform strain 1D3(Δ*bioF*) for phenotype testing as described in Methods. See Figure 8 legend for substitutions included in groups *g*1, *g*2, and *g*3.

† Indicated *kbl*→*bioF* mutations were made to the *kbl* gene on plasmid pKbl, which was then used to transform strain AG2 (Δ*bioF* Δ*kbl-tdh::kan*^{R}) for phenotype testing as described in Methods. Substitutions in groups *g*1, *g*2, and *g*3 are the reciprocals of those described in the Figure 8 legend, namely (using Kbl position numbering): *g*1 ≡ S78G, V79S, R80G, F81H, I82V, C83S; *g*2 ≡ R267F, S268A, P270H, Y271L, L272I, F273Y, N275T; and *g*3 ≡ Y351W, G354A, F355I, F356R, Y357P, V359T.

site cavity have been changed to resemble BioF (see Figure 8).

Finally, two approaches were taken to see whether some unidentified additional mutation might achieve functional conversion (see Methods for details). First, the gene encoding Kbl$_{g1,g2,g3,N155H}$ was used to generate a gene library containing ~10^6 randomly mutated variants. Among them should be all single-base variants of the parent gene, as well as a significant fraction of the possible two-base variants. Despite this diversity, we were unable to isolate a Bio$^+$ variant from the library. Second, because non-growing cells may enter a stress-induced hypermutable state [60] that might produce a Bio$^+$ variant naturally, we spread ~10^{11} cells carrying the Kbl$_{g1,g2,g3,N155H}$ gene onto minimal agar trays lacking biotin. After incubating 21 days at 25° C, the trays were inspected for Bio$^+$ colonies. None were found, indicating that natural mutations were also unable to produce the Bio$^+$ phenotype.

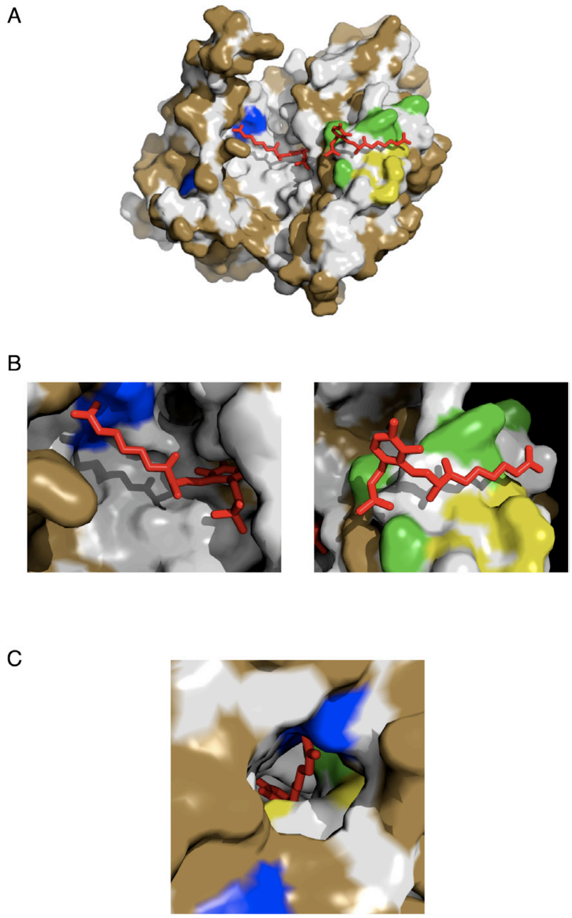


Figure 8: Surface view of BioF (1DJ9) showing candidate residues for functional conversion identified by structural comparison. Side chains (or alpha carbons for Gly residues) of candidate groups are colored in accordance with Figure 7: yellow for the first group (*g*1 ≡ G75S, S76V, G77R, H78F, V79I, S80C); green for the second group (*g*2 ≡ F258R, A259S, H261P, L262Y, I263L, Y264F, T266N); blue for the third group (*g*3 ≡ W344Y, A347G, I348F, R349F, P350Y, T352V). Regions where aligned BioF and Kbl residues are identical, including backbones, are colored light grey. Non-identical side chains not included in any groups are colored brown. A) BioF monomer with active sites indicated by the positions of external aldimine molecules (red). B) Close-up views of the reactant–enzyme interfaces in the monomer structure. C) View through the major opening into the active-site cavity of the BioF$_2$ dimer. The enzyme surfaces forming the cavity are seen to consist entirely of already identical regions (light grey) or regions that can be made identical by converting the side chains in the three groups (yellow, green, and blue). **doi:**10.5048/BIO-C.2011.1.f8

DISCUSSION

Implications for our understanding of enzymes

To explore the implications of these results, we begin by considering in conceptual terms what might account for the failure to achieve functional conversion. Figure 10 illustrates a general model of the structural determinants of enzyme function. The box represents an enzyme, with the network of arrows representing all the side-chain-dependent physical interactions that enable it to convert substrate(s) *S* into product(s) *P* (dots representing the side chains themselves). The right edge of the box represents the physical interface between the enzyme and

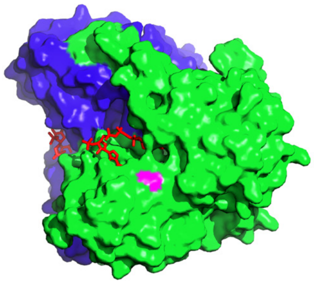


Figure 9: Inferring the site of CoA binding in BioF$_2$ by structural comparison. The dimeric structures of ALAS (left; 2BWO [43]) and BioF$_2$ (right; 1DJE [48]) are shown with corresponding histidine side chains (H161 of ALAS and H152 of BioF$_2$) colored magenta. Succinyl-CoA molecules (red) are shown bound to both active sites of ALAS. BioF$_2$ is believed to change its conformation upon pimeloyl-CoA binding [59]. It is shown here in the open conformation, which would be accessible for entry of pimeloyl-CoA. **doi:**10.5048/BIO-C.2011.1.f9

its ligands, by which we mean not only *S* and *P* but also all the chemical intermediates along the course of the reaction. Within the box, the horizontal dimension represents distance from that ligand interface. Interactions between the active site (shaded red) and its ligands are shown as red arrows protruding from the box, while arrows within the box represent important interactions that occur within the enzyme itself.

It is common to think of enzymes as consisting of an active site that is held in place by a structural scaffold (e.g., [61]). This is represented in Figure 10 by distinguishing the interactions that form the scaffold (dark grey) from those directly involved in the catalytic function (red). By this way of thinking, the fact that two enzymes are classified as sharing a common fold suggests that their scaffolds may be equivalent, with the significant differences residing in their active sites. This in turn leads to the expectation that functional conversions within enzyme families ought to be a simple matter of transplanting the active-site residues from any one onto the scaffold of any other. Since this is, in effect, what was attempted here (without success), our results now add to a larger body of evidence that seems, on the whole, to challenge that expectation. We emphasize again that this assessment of the evidence is not unique to us. John Gerlt and Patricia Babbitt, both prominent contributors to the field, were quoted above as observing that "many attempts at interchanging activities in mechanistically diverse superfamilies have [...] been attempted, but few successes have been realized" [13].

The most obvious explanation for the unexpected difficulty of functional interconversion is that the degree of structural similarity that justifies co-classification within a fold family does not justify the assumption of scaffold equivalence. In other words, many of the relatively small structural differences that are overlooked for the purposes of classification may in fact be important for function. If so, then scaffolds should be thought of not merely as *holding* active sites, but rather as providing them with the precise structural framework that enables them to perform their specific functions. It follows that scaffolds and active sites must be well-matched in order to work together, which explains why simply transplanting active-site residues fails to cause functional conversion, in our study and elsewhere (e.g., [17]).

Implications for enzyme evolution

The finding that functional conversions within fold families are much harder to achieve in the laboratory than was expected raises the question of their evolutionary feasibility. As discussed, the present study was designed with the aim of addressing this question. In particular, because we tested the effects of reciprocal substitutions in the contexts of the source and the target proteins, it is possible to make inferences about the requirements for functional conversion even though conversion was not achieved. For example, apart from knowledge that the substitution H152N inactivates BioF, the fact that the reciprocal change to Kbl (N155H) does not produce a Bio$^+$ phenotype would tell us very little about what is needed to achieve this functional conversion. But with both facts established, we believe it is correct to infer that the shortest route to a Bio$^+$ variant of Kbl will involve multiple changes, one of them being N155H.

Our reasoning here should be laid out in some detail. We begin by noting that everything about BioF$_{H152N}$ is known to be appropriate for producing the BioF$_2$ function *except* the asparagine side chain at position 152, which we know to be decisively inappropriate. By comparison, relatively little about Kbl$_2$ is actually known to be appropriate for producing the BioF$_2$ function. Indeed, it is the similarities and *only* the similarities between these two enzymes that make us think the conversion ought to be feasible. Consequently, logical consistency leads us to attribute failed conversion attempts to insufficient similarity. That is, alterations to Kbl that make it structurally more similar to BioF but still fail to confer the Bio$^+$ phenotype ought to be interpreted as not accomplishing enough change rather than as making the wrong kind of change. To think otherwise is to contradict the fact that similarity is the only basis for believing the conversion to be feasible in the first place.

There is one caveat, though, having to do with the distinction between sequence similarity and structural similarity. Previous work has shown that it is possible to disrupt folding by generating randomly shuffled hybrids of two natural isozymes having 50% sequence identity [62]. Because the parent enzymes in that study have nearly identical active sites that perform the same function, their scaffolds really are equivalent. Nevertheless, all hybrids were found to be nonfunctional, despite the fact that the hybrid sequences are more similar to the parent sequences than the parents are to each other. The explanation for this is that the parents differ in a structurally *coherent* way—each having its own specific way of stabilizing the same fold—whereas the shuffled hybrids differ in a structurally *incoherent*

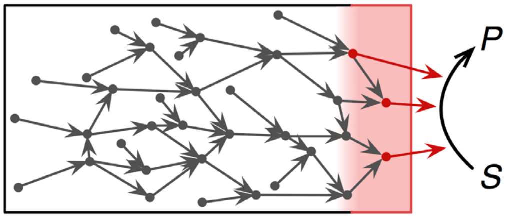


Figure 10: Model of the structural determinants of enzyme function. See text for explanation. **doi:**10.5048/BIO-C.2011.1.f10

way, having borrowed randomly from two different coherent solutions [62]. Because the structural incoherence was extensive in these hybrids, it appears to have caused loss of function by destabilizing the native structure and thereby preventing proper folding. The net result, then, is that an increase in sequence similarity caused a profound decrease in structural similarity (folded versus unfolded).

Based on the above, we think it would be possible to disrupt the folding of Kbl if, for example, half of the amino-acid positions where Kbl differs from BioF were to be chosen at random, and *kbl* →*bioF* changes were to be incorporated at each of the corresponding codons in *kbl* (numbering about 125). The resulting Kbl mutant probably would not fold into a properly formed native-like structure because of the structural incoherence introduced by so many changes. Despite increased *sequence* similarity to BioF, this mutant would suffer from decreased structural coherence and, as a likely consequence, decreased *structural* similarity to BioF.

We think it unlikely, however, that the Kbl variants in this study have been destabilized enough to disrupt folding—even the more highly altered ones like $\mathrm{Kbl}_{\mathrm{g1,g2,g3}}$ and $\mathrm{Kbl}_{\mathrm{g1,g2,g3,N155H}}$. There are two reasons for this. First, apart from substitutions that fall into the potentially highly disruptive classes mentioned previously (random replacements of buried side chains or glycines, or introduction of prolines at random locations [58]), it appears that about 10% or more of the residues in natural proteins need to be changed before the cumulative structural disruption can be expected to cause complete loss of function [62]. The twenty substitutions in $\mathrm{Kbl}_{\mathrm{g1,g2,g3,N155H}}$ alter only 5% of the protein, and because the replacements come from corresponding positions in a protein with a very similar overall structure (BioF), they are not apt to be drastically disruptive. For example, the single proline introduced in this twenty-position mutant (Y357P; see Table 1 legend) is at a turn with very similar backbone geometry to the turn at proline 350 in BioF. Consequently, we would not expect the Y357P substitution in Kbl to have anything like the disruptive effect that a randomly introduced proline might have. Second, if the conventional role distinction between scaffold and active site has any validity, side chains forming the ligand interface (i.e., the "front line" of the active site) must carry little or no responsibility for stabilizing the scaffold. The thinking, in other words, is that these residues are free to be optimized for substrate binding and catalysis precisely because the scaffold residues have been optimized to stabilize the overall folded structure. If this is true, or even approximately true, then the locations of the changes introduced in $\mathrm{Kbl}_{\mathrm{g1,g2,g3,N155H}}$ imply that they are even less apt to cause structural destabilization.

In the estimate to follow, we will nonetheless consider the alternative possibility, namely that $\mathrm{Kbl}_{\mathrm{g1,g2,g3,N155H}}$ lacks BioF function not because more changes are needed but because the twenty changes it carries have disrupted folding. Although this alternative will make functional conversion appear more feasible by the calculations to follow (which is why we consider it), its implications are actually more problematic than helpful. Specifically, if the ability of protein chains to fold into stable scaffold structures is strongly coupled to the identity of the side chains at active-site positions, then the challenge of functional conversion would be greatly compounded by this coupling. That is, it becomes much harder to reconfigure an active site to provide a new function if the kinds of changes needed for this conversion jeopardize the formation of the scaffold.

We now generalize the reasoning behind our inference that the shortest route to a Bio^{+} variant of Kbl will involve multiple changes, one of them being N155H, so that it may be applied to other mutants as well:

i. Based solely on the similarity of Kbl_2 to BioF_2, we hypothesize that the former may be made to perform the function of the latter with modest change.

ii. The Bio^{-} phenotype of the BioF →Kbl mutant in question shows that this mutant has at least one aspect of Kbl that is incompatible with the Bio^{+} phenotype.

iii. The Bio^{-} phenotype of the reciprocal Kbl →BioF mutant shows that this change, although increasing the sequence similarity to BioF, does not cause functional conversion.

iv. The facts that the changes made to Kbl in *iii* are modest in extent, drawn from the corresponding residues of BioF, and (except for N155H) localized to the active site argue against disrupted folding, implying that greater *structural* similarity has been achieved (not merely greater sequence similarity).

v. Since the expectation of feasible conversion is based on similarity (*i*), and the structural similarity of Kbl to BioF has been enhanced (*iv*) in one or more functionally critical respects (*ii*), we infer from the lack of conversion (*iii*) that successful conversion will require not only this enhancement of structural similarity, but at least one further enhancement.

On this basis we infer that each of the four Kbl mutants changed at the critical loci ($\mathrm{Kbl}_{\mathrm{N155H}}$, $\mathrm{Kbl}_{\mathrm{g1}}$, $\mathrm{Kbl}_{\mathrm{g2}}$, and $\mathrm{Kbl}_{\mathrm{g3}}$) have been altered in at least one necessary respect, though none of them have been altered sufficiently. Furthermore, having found that none of the individual amino-acid changes within group 3 cause inactivation of BioF, we deduce that at least two changes in group 3 must contribute to its effect. The above reasoning therefore leads us to infer that the shortest path to a Bio^{+} variant of Kbl must include: one or more of the amino-acid substitutions from group 1, one or more of the substitutions from group 2, two or more of the substitutions from group 3, and the N155H substitution. Combining these inferences, we conclude that successful conversion would require at least five amino-acid substitutions. If the changes we made to $\mathrm{Kbl}_{\mathrm{g1,g2,g3,N155H}}$ do not prevent proper folding (as argued), then this inferred minimum should actually be raised by two substitutions: one by applying the above reasoning to $\mathrm{BioF}_{\mathrm{g1,g2,g3,H152N}}$ and its reciprocal ($\mathrm{Kbl}_{\mathrm{g1,g2,g3,N155H}}$), and another in consideration of the fact that random mutagenesis did not produce a Bio^{+} variant of $\mathrm{Kbl}_{\mathrm{g1,g2,g3,N155H}}$.

Still, we will proceed with the lower requirement of five amino-acid substitutions in order to arrive at a lower-bound estimate of the time needed for functional conversion in a natural population. That number must be incremented slightly in order to estimate the minimum number of nucleotide substitutions for achieving conversion. Since an average of 1.5 nucleotide substitutions is needed per amino-acid substitution[11], seven is a reasonable lower-bound estimate of the specific nucleotide substitutions required for conversion. Although this is not exceptionally high compared to the extent of change used in other attempted conversions (see Introduction), it nonetheless places the Kbl→BioF conversion outside the bounds of what can be achieved by the Darwinian mechanism. Specifically, using a population model described previously [18] with the following assumptions:

A1– Duplications of *kbl* that are suitable starting points for paralogous innovation occur at a rate of 10^{-8} to 10^{-3} per cell,

A2– The metabolic cost of carrying one such duplicate allele (unconverted) decreases fitness by 0.01% to 16% relative to the wild-type, [12]

A3– Conversion to *bioF* function can be achieved with seven base changes, and

A4– The converted gene confers an overall growth advantage of 1%, averaged over the full range of environments encountered by the species,

we estimate that some 10^{30} or more generations would elapse before a *bioF*-like innovation that is paralogous to *kbl* could become established (Figure 11). This places the innovation well beyond what can be expected within the time that life has existed on earth, under favorable assumptions. In fact, even the unrealistically favorable assumption that *kbl* duplicates carry no fitness cost leaves the conversion just beyond the limits of feasibility (Figure 11).

Using appropriate caution, we conclude not that paralogous evolution absolutely cannot have accomplished a functional jump like the one examined here, but rather that there is now a scientific case for doubting this particular jump to be evolutionarily feasible. At first glance, this claim may seem so modest as to verge on insignificance. Indeed, it could be interpreted as nothing more than a curious exception if the standard dogma—that paralogous evolution readily explains most examples of small-scale innovation—were well supported by the evidence. But as we have discussed, the many attempts to confirm that dogma have left it in question.

It is worth reviewing what a convincing demonstration of the feasibility of paralogous innovation would look like. It would start with a pair of natural enzymes that use the same overall fold structure to perform functions that differ not merely in their substrates but in their reaction chemistries. Choosing one of these functions as the target function, it would proceed by demonstrating that the unaltered source enzyme cannot provide the target function *in vivo*. It would then identify a set of amino-acid substitutions that convert the source enzyme so that it does perform the target function *in vivo*, after which it would combine growth data with reasonable assumptions about natural growth conditions to estimate the overall effect that a duplicate gene with or without the converted function would have on the fitness of the organism carrying it. Based on these results, a population model would next be used to estimate the time needed for the same small-scale innovation to appear in a natural population. Finally, unless the converted enzyme performs the target function with a proficiency approaching that of the natural enzyme, it would be necessary to demonstrate the feasibility of evolutionary improvement to that level.

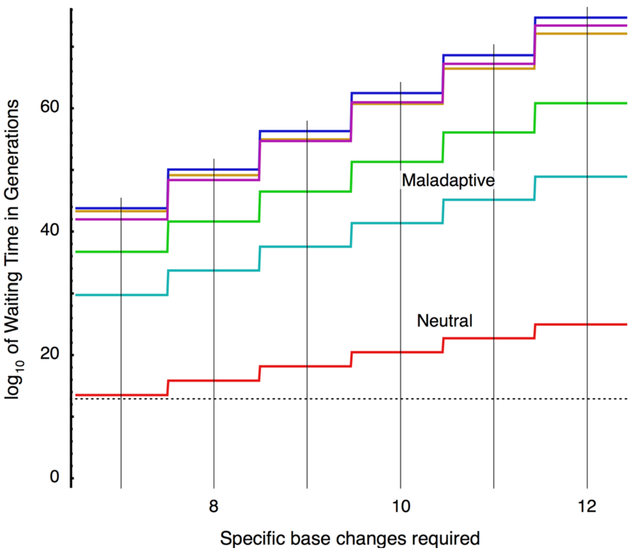


Figure 11: Expected waiting times for appearance and fixation of paralogous innovations requiring from seven to twelve specific base changes. The assumed starting point is a population lacking the required duplicate gene. Because gene duplicates that have not acquired new functions are known to carry a significant fitness cost [63], Equation 10 of reference 18 applies, with very long waiting times resulting. For comparison to Figure 4 of reference 18, three of the above staircase curves were calculated using an assumed *kbl* duplication rate of 10^{-8} per cell and a fitness cost of either 1% ($s^{-} = -0.01$; orange), 0.1% ($s^{-} = -0.001$; green), or 0.01% ($s^{-} = -0.0001$; cyan). The work of Reams et al. [63] provides direct evidence for higher duplication rates and higher fitness costs. Accordingly, we repeated the calculation using a duplication rate of 3×10^{-6} per cell with a fitness cost of 4% to represent the observed values for chromosomal gene *pyrD* (blue), and a duplication rate of 10^{-3} per cell with a fitness cost of 16% to represent the observed values for chromosomal gene *argH* (purple) [63]. Equation 20 of reference 18 would apply if there were no cost to carrying a duplicate (red). Other parameter values are as listed in Table 1 of reference 18. The dashed line marks the boundary between feasible waiting times (below) and waiting times that exceed the age of life on earth (above), assuming 10^3 generations per year. **doi:**10.5048/BIO-C.2011.1.f11

To the best of our knowledge, no study has yet met this description. If a future study does, it would provide the first example where small-scale innovation by paralogous evolution is demonstrably feasible. Even so, unless successful examples like that were to become much more numerous than the unsuccessful ones, there would be no basis for thinking that jumps to new functions are feasible as a *rule*. Indeed, as the evidence now stands, it seems more reasonable to doubt their feasibility than to presume it.

[11] Based on the actual amino-acid substitutions used.

[12] Specific combinations of duplication rate and fitness cost are explained in Figure 11 legend.

The implications of this come into full view when we begin to ask how much evolutionary significance can really be attached to structural similarity in the first place. Koonin and Wolf have recently exposed the fallacy of taking similarity as proof of homology [64], and yet in our judgment they commit another fallacy. It is abundantly clear that specific and extensive similarities, such as those shared by BioF$_2$ and Kbl$_2$, cannot be attributed to mere coincidence. This leads Koonin and Wolf to reject convergent evolution (extensive similarity appearing by evolution from dissimilar starting points) as implausible. But from this they conclude that homology, while not formally proven by similarity, is nonetheless overwhelmingly supported in cases where chance convergence is implausible. The problem with this is that *all* non-chance alternatives must be considered once chance is ruled out. Yet Koonin and Wolf consider only one of these alternatives—the standard Darwinian one.

We agree with their rejection of chance, but we argue here that the Darwinian explanation also appears to be inadequate. Its deficiencies become evident when the focus moves from similarities to *dissimilarities*, and in particular to functionally important dissimilarities—to innovations. The extent to which Darwinian evolution can explain enzymatic innovation seems, on careful inspection, to be very limited. Large-scale innovations that result in new protein folds appear to be well outside its range [5]. This paper argues that at least some small-scale innovations may also be beyond its reach. If studies of this kind continue to imply that this is typical rather than exceptional, then answers to the most interesting origins questions will probably remain elusive until the full range of explanatory alternatives is considered.

METHODS

Bacterial strains, bacteriophage, and plasmid vectors

EMG2 (a wild-type K12 strain) and bacteriophage P1*vir* were obtained from the *E. coli* Genetic Stock Center[13], as were the strains and plasmids necessary for the λ Red protocol [65]: BW25113 (*lacI*^{q} *rrnB*$_{T14}$ Δ*lacZ*$_{WJ16}$ *hsdR*514 Δ*araBAD*$_{AH33}$ Δ*rhaBAD*$_{LD78}$), BW25141 (*lacI*^{q} *rrnB*$_{T14}$ Δ*lacZ*$_{WJ16}$ Δ*phoBR580 hsdR514* Δ*araBAD*$_{AH33}$ Δ*rhaBAD*$_{LD78}$ *galU95 endA*$_{BT333}$ *uidA*(Δ*Mlu*I)::*pir*$^+$ *recA1*), pKD46, and pKD4. Chemically competent *E. coli* strain DH5α (F- φ80*lacZ*ΔM15 Δ(*lacZYA-argF*)*U169 recA1 endA1 hsdR17*(r$_k^-$, m$_k^+$) *phoA supE44 thi-1 gyrA96 relA1* λ-) was obtained from Invitrogen. Electrocompetent *E. coli* strain NEB 5-alpha (*fhuA2*Δ*(argF-lacZ)U169 phoA glnV44* Φ*80*Δ *(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17*) was obtained from New England Biolabs. Plasmid pKOV, a derivative of low copy plasmid pKO3 [66] carrying the *sacB* gene and having a temperature-sensitive origin of replication, was obtained from the laboratory of G. M. Church.

Construction of chromosomal deletion strains

The method of Link *et al.* [66] was used to make a precise in-frame deletion within the *bioF* gene of EMG2 (Figure 12A). The Δ*bioF* genotype of the resulting strain, designated 1D3, was confirmed by sequencing the region of the chromosome in the vicinity of the deletion.

To delete the *kbl-tdh* operon from the chromosome of strain BW25113, we used the λ Red protocol [65]. The presence of the desired Δ*kbl-tdh::kan*^{R} replacement was verified by PCR analysis, using inward-directed primers specific to sequences flanking the *kbl–tdh* operon and outward-directed primers that matched sequences internal to the *kan*^{R} replacement gene. The Δ*kbl-tdh::kan*^{R} replacement locus of this strain was then transferred to 1D3(Δ*bioF*) by P1*vir* transduction. The genotype of the resulting strain, designated AG2 (Δ*bioF* Δ*kbl-tdh::kan*^{R}), was verified by PCR analysis as above, and by DNA sequencing through the region of the replacement (see Figure 12B).

Plasmid construction

To generate a low copy vector that could be used for phenotype testing, the *bioF* gene and twenty bases of its upstream regulatory sequence was cloned into the pKOV vector between its unique NotI and HindIII sites, eliminating the *sacB* gene and intervening sequence. This plasmid, pKBF1, was further modified by inserting stop codons in all frames twenty bp upstream of the *bioF* insert, and by converting the NotI site to a BamHI site. All junctions and the entire *bioF* coding region in this construct, designated pKBF2 (Figure 12C, D), were verified by DNA sequencing.

Plasmid pKbl was constructed by inserting the PCR-amplified *kbl* gene (and twenty bases of its upstream regulatory sequence) from strain 1D3 into plasmid pKBF2 between the BamHI and HindIII sites, in place of *bioF*.

Mutations to *bioF* or *kbl* in these plasmids were introduced by the inverse PCR method. All plasmid constructs were initially propagated in strain DH5α in order for DNA methylation to occur without host restriction. Plasmid DNA prepared from DH5α was used for sequence confirmation and to transform experimental strains 1D3 or AG2 (where host restriction is active) for phenotype testing.

Testing for biotin autotrophy

Because bacterial cells require only trace quantities of biotin for growth, testing for biotin autotrophy (phenotype Bio$^+$) by colony growth in the absence of supplied biotin required careful measures to minimize exogenous biotin and controls to monitor batch-to-batch variations in medium. For each test, we used single batches of freshly prepared medium and plated both a positive control strain (1D3 or AG2 carrying pKBF2 for Bio$^+$ phenotype) and a negative control strain (1D3 or AG2 without plasmid for Bio$^-$ phenotype) in parallel with experimental strains. Strains were grown in 30-ml culture tubes for 48 hr at 30°C (250 rpm) in Minimal Davis (MD) medium with biotin (20 ng/ml) and, where appropriate, chloramphenicol (20 μg/ml) and washed four times in ice-cold phosphate-buffered saline solution (to remove all traces of biotin) before spreading at a density of about 500 cells per plate (100 mm diameter). Washed cells were plated on MD agar with and without bio-

[13] http://cgsc.biology.yale.edu/index.php

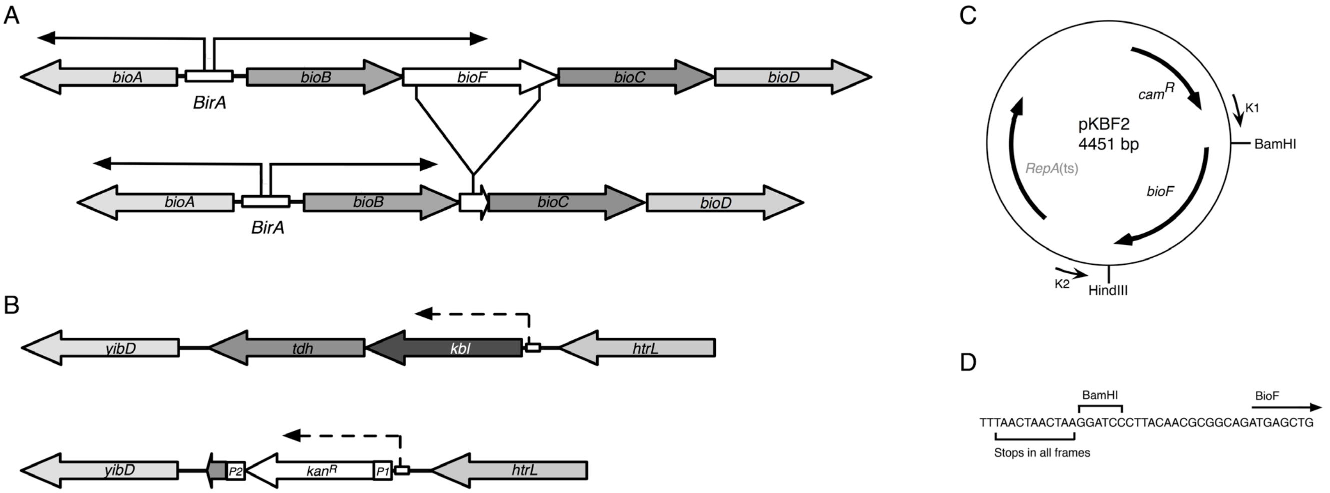


Figure 12: Structure of genetic constructs (not drawn to scale). A) Strain 1D3 (bottom) was constructed by removing 90% of *bioF* from the wild-type biotin operon (top), leaving a total of 117 bp from 5′ and 3′ ends. B) Strain AG2 carries both the above *bioF* deletion and the *kbl–tdh* replacement shown (mutant below wild-type), which leaves 21 bp from the 3′ end of *tdh*. C) Map of plasmid pKBF2. *RepA*(ts) is a temperature-sensitive origin of replication; *cam*^{R} encodes chloramphenicol acetyltransferase, an enzyme that confers resistance to chloramphenicol. D) Sequence in the vicinity of the upstream BamHI site of pKBF2. **doi:**10.5048/BIO-C.2011.1.f12

tin (20 ng/ml). Chloramphenicol (50 μg/ml) was added for all plasmid-containing strains. Plates were incubated at 30°C for 48 hours prior to inspection for colony growth. This procedure allows unambiguous assignment of phenotype (Figure 13).

At high plating densities some cells of *bioF*$^-$ genotype scavenge enough biotin from their neighbors to grow poorly on MD agar. When dense plating was called for (see below), we added streptavidin (a protein that forms a very tight complex with biotin) to the medium at a concentration of 100 ng/ml. This addition consistently prevented visible growth of *bioF*$^-$ strains even at high plating densities. As expected, strains that are genotypically *bioF*$^+$ are unaffected by streptavidin.

Preparation and screening of randomly mutagenized library

A plasmid library containing random mutations in the *kbl* gene variant of plasmid pKbl$_{g1,g2,g3,N155H}$ was prepared by PCR amplification using an error-prone DNA polymerase (Mutazyme II; Stratagene) with flanking primers K1 and K2 (Figure 12C). The amplification conditions used (20 cycles with 500 ng of initial plasmid template and 5 units of Mutazyme II) produce more than one base substitution per kilobase on average, though many genes will have none. Following amplification, the PCR product was digested with BamHI and HindIII, gel-purified, and ligated back into the complementary BamHI–HindIII fragment from unmutagenized plasmid. After dialysis and drying, the ligation product was resuspended in 5 μl H_2O, and 2 μl was added to 50 μl electrocompetent *E. coli* strain NEB 5-alpha for electroporation 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 1 ml SOC medium and incubated at 30°C, 250 rpm, for 90 min. Following incubation, cells were spread onto a 245x245 mm tray (Biodish XL, Becton Dickenson) containing LB agar with chloramphenicol (20 μg/ml). To estimate the number of transformed cells placed on the tray, a 1000-fold dilution of the same culture was spread on several plates containing the same medium. The tray and plates were incubated overnight at 30°C.

Based on the plate counts, approximately 0.9 million transformants were spread on the tray. Cells that grew on the tray were recovered by washing with 5 ml of Terrific Broth (TB). After thorough mixing, a portion of the wash was diluted in TB with chloramphenicol (20 μg/ml) and incubated at 30°C for 8 hours (250 rpm). Plasmid DNA was prepared from the resulting culture. After dialysis, approximately 5 ng of this plasmid mixture was used to transform (by electroporation) competent AG2 cells as described above.

Plate counts showed that roughly twelve million AG2 transformants were spread on the tray. The resulting cells were recovered by washing the tray with MD medium, and then diluted and grown in MD medium containing biotin (20 ng/ml) and chloramphenicol (20 μg/ml) for 48 hours at 30°C. Parallel cul-

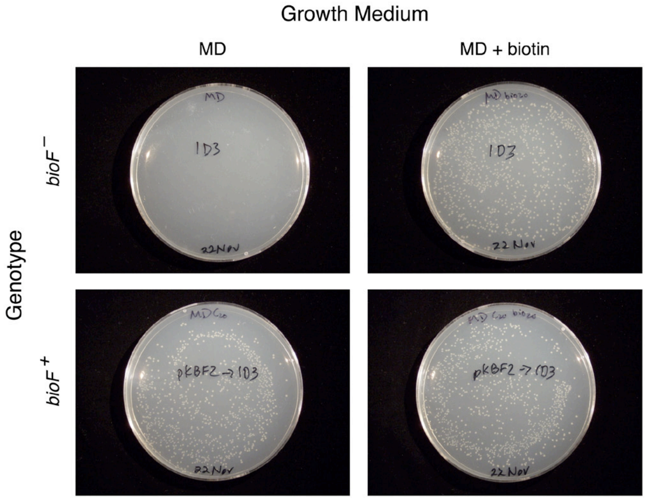


Figure 13: Correspondence between phenotype and genotype in biotin autotrophy test. **doi:**10.5048/BIO-C.2011.1.f13

tures of AG2 cells carrying plasmid pKBF2 (Bio$^+$ phenotype) and AG2 cells carrying unmutagenized pKbl$_{g1,g2,g3,N155H}$ (Bio$^-$ phenotype) were grown in the same medium. All three cultures were tested for biotin autotrophy as described, except that they were spread at high density (100-fold dilution of the day culture) onto both biotin-free MD agar with 100 ng/ml streptavidin and MD agar supplemented with biotin (20 ng/ml), and the biotin-free test of the experimental culture was done on a tray. Each culture was also diluted 10^5-fold and plated on LB agar with chloramphenicol for colony counts, which indicated that approximately 4 million plasmid-containing cells were spread on the biotin-free tray. The library size screened was therefore limited by the initial 0.9 million NEB 5-alpha transformants.

To sample the mutations present in the unselected mutagenized library, plasmid DNA was prepared from eight clonal lines isolated from the LB agar plates. One of these plasmids had two BamHI–HindIII inserts and was not sequenced. The other seven had a variety of *kbl* mutations, ranging in number from zero to five new substitutions per gene, with an average of two substitutions. All identified base substitutions were unique, with two of the seven plasmids carrying single substitutions.

Prolonged incubation of non-growing cells

Approximately 8 x 10^{10} AG2 cells (based on colony counts on rich medium) carrying plasmid pKbl$_{g1,g2,g3,N155H}$ were spread onto ten 245×245 mm trays, each containing 250 ml MD agar with 20 µg/ml chloramphenicol and 200 ng/ml streptavidin. Trays were wrapped in foil and placed in a humidified incubator at 25°C for 21 days with periodic inspection for colonies.

Acknowledgements

We thank the Church lab for generously providing the pKOV vector, the *E. coli* Genetic Stock Center for supplying phage and bacterial strains, Stan Roberts for technical assistance, and our colleagues and reviewers for helpful comments.

1. Wagner GP (2001) What is the promise of developmental evolution? Part II: A causal explanation of evolutionary innovations may be impossible. J Exp Zool 291: 305-309. doi:10.1002/jez.1130
2. Müller GB, Wagner GP (1991) Novelty in evolution: restructuring the concept. Annu Rev Ecol Syst 22: 229-256. doi:10.1146/annurev.es.22.110191.001305
3. Fontana W, Buss LW (1994) "The arrival of the fittest": toward a theory of biological organization. B Math Biol 56: 1-64.
4. Stadler BM, Stadler PF, Wagner GP, Fontana W (2001) The topology of the possible: formal spaces underlying patterns of evolutionary change. J Theor Biol 213: 241-274. doi:10.1006/jtbi.2001.2423
5. Axe (2010) The case against a Darwinian origin of protein folds. BIO-Complexity 2010(1): 1-12. doi:10.5048/BIO-C.2010.1
6. Ohno S (1970) Evolution by gene duplication. Springer-Verlag (New York).
7. Ohno S (1973) Ancient linkage groups and frozen accidents. Nature 244: 259-262. doi:10.1038/244259a0
8. Jensen RA (1976) Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30: 409-425. doi:10.1146/annurev.mi.30.100176.002205
9. Hughes AL (2002) Adaptive evolution after gene duplication. Trends Genet 18: 433-434. doi:10.1016/S0168-9525(02)02755-5
10. Chothia C, Gough G, Vogel C, Teichman SA (2003) Evolution of the protein repetoire. Nature 300: 1701-1703. doi:10.1126/science.1085371
11. O'Brien PJ, Herschlag D (1999) Catalytic promiscuity and the evolution of new enzymatic activities. Chemistry and Biology 6: R91-R105. doi:10.1016/S1074-5521(99)80033-7
12. 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
13. Gerlt JA, Babbitt PC (2009) Enzyme (re)design: lessons from natural evolution and computation. Curr Opin Chem Biol 13(1):10-18. doi:10.1016/j.cbpa.2009.01.014
14. 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
15. Wilson EM, Kornberg HL (1963) Properties of crystalline L-aspartate 4-carboxy-lyase from *Achromobacter* sp. Biochem J 88:578-587.
16. 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
17. 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
18. Axe D (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
19. Raillard S, Krebber A, Chen Y, Ness JE, Bermudez E, *et al.* (2001) Novel enzyme activities and functional plasticity revealed by recombining highly homologous enzymes. Chem Biol 8:891-898. doi:10.1016/S1074-5521(01)00061-8
20. 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
21. Wagner A (2005) Energy constraints on the evolution of gene expression. Mol Biol Evol 22: 1365-1374. doi:10.1093/molbev/msi126
22. Stoebel DM, Dean AM, Dykhuizen DF (2008) The cost of expression of *Escherichia coli* lac operon proteins is in the process, not in the products. Genetics 178: 1653-1660. doi:10.1534/genetics.107.085399
23. 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:10.5048/BIO-C.2010.2
24. 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

25. Patrick WM, Matsumura I (2008) A study in molecular contingency: glutamine phosphoribosylpyrophosphate amidotransferase is a promiscuous and evolvable phosphoribosylanthranilate isomerase. J Mol Biol 377:323-336. **doi:**10.1016/j.jmb.2008.01.043
26. Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Bio 10:866-876. **doi:**10.1038/nrm2805
27. 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
28. Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 60:2256-2268. **doi:**10.1107/S0907444904026460
29. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536-540. **doi:**10.1016/S0022-2836(05)80134-2
30. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406-425.
31. Capitani G, De Blase D, Aurizi C, Gut H, Bossa F et al. (2003) Crystal structure and functional analysis of *Escherichia coli* glutamate decarboxylase. EMBO J 22:4027-4037. **doi:**10.1093/emboj/cdg403
32. Tirupati B, Vey, JL, Drennan, CL, Bollinger, JM (2004) Kinetic and structural characterization of Slr0077/SufS, the essential cysteine desulfurase from *Synechocystis* sp. PPC 6803. Biochem 43:12210-12219. **doi:**10.1021/bi0491447
33. Outten FW, Wood MJ, Munoz FM, Storz G (2003) The SufE protein and the SufBCD complex enhance SufS cysteine desulfurase activity as part of a sulfur transfer pathway for Fe-S cluster assembly in *Escherichia coli.* J Biol Chem 278:45713-45719. **doi:**10.1074/jbc.M308004200
34. Zhang X, Roe SM, Hou Y, Bartlam M, Rao Z, Pearl LH, Danpure CJ (2003) Crystal structure of alanine: glyoxylate aminotransferase and the relationship between genotype and enzymatic phenotype in primary hyperoxaluria type 1. J Mol Biol 331:643-652. **doi:**10.1016/S0022-2836(03)00791-5
35. Rossi F, Lombardo F, Paglino A, Cassani C, Miglio G, et al. (2005) Identification and biochemical characterization of the *Anopheles gambiae* 3-hydroxykynurenine transaminase. FEBS 272:5653-5662. **doi:**10.1111/j.1742-4658.2005.04961.x
36. Han Q, Kim SR, Ding H, Li J (2006) Evolution of two alanine glyoxylate aminotransferases in mosquito. Biochem J 397:473-481. **doi:**10.1042/BJ20060469
37. Gelfand DH, Steinberg RA (1977) *Escherichia coli* mutants deficient in the aspartate and aromatic amino-acid aminotransferases. J Bacteriol 130:429-440.
38. Shen BW, Hennig M, Hohenester E, Jansonius JN, Schirmer T (1998) Crystal structure of human recombinant ornithine aminotransferase. J Mol Biol 277:81-102. **doi:**10.1006/jmbi.1997.1583
39. Miyazaki J, Kobashi N, Nishiyama M, Yamane H (2001) Functional and evolutionary relationship between arginine biosynthesis and prokaryotic lysine biosynthesis through α-aminoadipate. J Bacteriol 183:5067-5073. **doi:**10.1128/JB.183.17.5067-5073.2001
40. Aitken SM, Kim DH, Firsch JF (2003) *Escherichia coli* cystathionine γ-synthase does not obey ping-pong kinetics. Novel continuous assays for the elimination and substitution reactions. Biochem 42:11297-11306. **doi:**10.1021/bi035107o
41. Ono B-I, Tanaka K, Naito K, Heike C, Shinoda S, et al. (1992) Cloning and characterization of the *CYS3* (*CYI1*) gene of *Saccharomyces cerevisiae*. J Bacteriol 174:3339-3347.
42. 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
43. 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. EMBO J 24:3166-3177. **doi:**10.1038/sj.emboj.7600792
44. 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
45. Toney MD, Hohenester E, Cowan SW, Jansonius JN (1993) Dialkylglycine decarboxylase structure: bifunctional active site and alkali metal sites. Science 261:756-759. **doi:**10.1126/science.8342040
46. Liu W, Peterson PE, Carter RJ, Zhou X, Langston J et al. (2004) Crystal structures of unbound and aminooxyacetate-bound *Escherichia coli* γ-aminobutyrate aminotransferase. Biochem 43:10896-10905. **doi:**10.1021/bi049218e
47. Tanaka H, Esaki N, Soda K (1985) A versatile bacterial enzyme: L-methionine γ-lyase. Enzyme Microb Technol 7:530-537. **doi:**10.1016/0141-0229(85)90094-8
48. Webster SP, Alexeev D, Campopiano DJ, Watt RM, Alexeeva M et al (2000) Mechanism of 8-amino-7-oxononanoate synthase: Spectroscopic, kinetic, and crystallographic studies. Biochem 39:516-528. **doi:**10.1021/bi991620j
49. 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.
50. Del Campillo-Campbell A, Kayajanian G, Campbell A, Adhya, S (1967) Biotin-requiring mutants of *Escherichia coli* K-12. J Bacteriol 94:2065-2066.
51. Rolfe B, Eisenberg MA (1968) Genetic and biochemical analysis of the biotin loci of *Escherichia coli* K-12. J Bacteriol 96:515-524.
52. Streit WR, Entcheva P (2003) Biotin in microbes, the genes involved in its biosynthesis, its biochemical role and perspectives for biotechnological production. Appl Microbiol Biotechnol 61:21-31. **doi:**10.1007/s00253-002-1186-2
53. 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 Bioch 7:3033-3040. **doi:**10.1271/bbb.70438
54. Jarrett JT (2005) Biotin synthase: enzyme or reactant? Chemistry & Biology 12:409-415. **doi:**10.1016/j.chembiol.2005.04.003
55. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 22:4673-4680. **doi:**10.1093/nar/22.22.4673
56. Supplement to this paper. **doi:**10.5048/BIO-C.2011.1.s
57. Myers EW, Miller W (1988) Optimal alignments in linear space. Bioinformatics 4:11-17. **doi:**10.1093/bioinformatics/4.1.11
58. Axe DD, Foster NW, Fersht AR (1998) A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. Biochemistry 37:7157-7166. **doi:**10.1021/bi9804028
59. Mann S, Ploux O (2011) Pyridoxal-5′-phosphate-dependent enzymes involved in biotin synthesis: structure, reaction mechanism and inhibition. Biochim Biophys Acta (in press). **doi:**10.1016/j.bbapap.2010.12.004

60. Galhardo RS, Hastings PJ, Rosenberg SM (2007) Mutation as a stress response and the regulation of evolvability. Crit Rev Biochem Mol Biol 42:399–435. doi:10.1080/10409230701648502
61. Park HS, Nam SH, Lee JK, Yoon CN, Mannervik B, et al. (2006) Design and evolution of new catalytic activity with an existing protein scaffold. Science 311:535-538. doi:10.1126/science.1118953
62. Axe DD (2000) Extreme functional sensitivity to conservative amino-acid changes on enzyme exteriors. J Mol Biol 301:585-595. doi:10.1006/jmbi.2000.3997
63. 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
64. Koonin EV, Wolf YI (2010) The common ancestry of life. Biol Direct 5:64. doi:10.1186/1745-6150-5-64
65. Datsenko KA, Wanner BL (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc Natl Acad Sci USA 97:6640-6645. doi:10.1073/pnas.120163297
66. Link AJ, Phillips D, Church GM (1997) Methods for generating precise deletions and insertions in the genome of wild-type *Escherichia coli*: Application to open reading frame characterization. J Bacteriol 179: 6228-6237.